LINKING ENVIRONMENTAL EXPOSURES AND HEALTH EFFECTS USING EXISTING DATA TO EXPLORE THE RELATIONSHIPS BETWEEN ENVIRONMENT AND CHRONIC DISEASES

by

Xiaohui Xu

B. Med., Zhejiang University, 1999

M. Med., Fudan University, 2002

Submitted to the Graduate Faculty of Graduate School of Public Health in partial fulfillment of the requirements for the degree of Doctor of Philosophy

University of Pittsburgh

2007
UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Xiaohui Xu

It was defended on

June 6th, 2007

and approved by

Dissertation Advisor:
Evelyn O. Talbott Dr. P.H., Professor
Department of Epidemiology,
Graduate School of Public Health
Universith of Pittsburgh

Committee member:
Vincent C. Arena Ph.D., Associate Professor
Department of Biostatistics, Graduate School of Public Health,
Universith of Pittsburgh

Committee member:
Devra Lee Davis Ph.D., Visiting Professor,
Department of Epidemiology, Graduate School of Public Health,
Universith of Pittsburgh

Committee member:
Ravi K. Sharma Ph.D., Assistant Professor,
Department of Behavioral Community Health Sciences,
Graduate School of Public Health, Universith of Pittsburgh

Committee member:
Conrad D Volz Dr. P.H., Assistant Professor,
Department of Environmental and Occupational Health,
Graduate School of Public Health, Universith of Pittsburgh
The environment plays an important role in the health of communities. However, few health systems exist at the state and/or local levels to efficiently track the potential health effects associated with environmental exposure. The objectives of this dissertation are 1) to use secondary data for assessing the possible associations between health outcomes and environmental exposure and/or hazard; 2) to explore possible methods of data linkage and analyses which can be used by state and local environmental health tracking agencies and 3) to bring positive contributions to the development of national Environmental Public Health Tracking Network (EPHT). In this project, the Three Mile Island (TMI) cohort data (1979-1995) and Pennsylvania (PA) Cancer registry data were used to evaluate the associations between cigarette smoking and adult leukemia. A case-crossover analysis was performed with PA cardiopulmonary hospital admission data and local air pollution data to assess the health effects of air pollutants on cardiopulmonary disease before and after the elimination of a major point source of air pollution. A case-control study was also conducted to examine the associations between term low birth weight and particulate air pollution. The results showed that cigarette smoking could increase the risk of acute myeloid leukemia (AML). In addition, particulate air pollution is significantly associated with cardiovascular hospitalization and low birth weight in term infant.
In conclusion, the findings suggest that environmental hazards have adverse health effects on a number of health endpoints. Secondary data can be a great resource for environmental public health tracking, which is of public health relevance. The use of existing data is an effective way to assess the potential health effects associated with environmental exposures after an appropriate study design with a feasible data linkage and correct methods of data analyses was developed.

**Keyword:** Environmental exposure, Secondary data, Chronic disease
# TABLE OF CONTENTS

PREFACE

1.0 INTRODUCTION

1.1 ENVIRONMENTAL EXPOSURES AND CHRONIC DISEASES

1.2 NATIONAL ENVIRONMENTAL PUBLIC HEALTH TRACKING NETWORK

   1.2.1 THE CURRENT STATUS OF ENVIRONMENTAL HEALTH TRACKING SYSTEMS

   1.2.2 THE NATIONAL ENVIRONMENTAL PUBLIC HEALTH TRACKING PROGRAM

1.3 PROPOSED THREE DEMONSTRATIVE STUDIES AND OBJECTIVES

   1.3.1 CASE-CROSSOVER ANALYSIS OF AIR POLLUTION AND CARDIO-PULMONARY HOSPITAL ADMISSIONS

   1.3.2 PARTICULATE AIR POLLUTION AND TERM LOW BIRTH WEIGHT IN ALLEGHENY COUNTY

   1.3.3 CIGARETTE SMOKING AND ADULT LEUKEMIA

2.0 CASE-CROSSOVER ANALYSIS OF AIR POLLUTION AND CARDIOPULMONARY HOSPITALIZATIONS: USING ROUTINELY COLLECTED HEALTH AND ENVIRONMENTAL DATA FOR TRACKING

   2.1 ABSTRACT

   2.2 INTRODUCTION

   2.3 MATERIALS AND METHODS

      2.3.1 STUDY POPULATION

      2.3.2 AIR POLLUTION AND WEATHER DATA

      2.3.3 DATA ANALYSIS
5.3 ASSOCIATION OF RISK OF TERM LBW AND PARTICULATE AIR POLLUTION

5.4 ASSOCIATION OF RISK OF LEUKEMIA AND CIGARETTE SMOKING

5.5 STRENGTH AND LIMITATIONS OF SECONDARY DATA

5.6 OVERALL SUMMARY AND CONCLUSION

APPENDIX A - RISK OF ADVERSE PREGNANCY OUTCOMES AND AIR POLLUTION

APPENDIX B - USING SECONDARY DATA FOR ENVIRONMENTAL PUBLIC HEALTH TRACKING: SOURCES, METHODS, AND ISSUES

BIBLIOGRAPHY
LIST OF TABLES

Table 2-1 Distribution of specific primary diagnoses of hospital admissions among study sample in Hazelwood area, Pittsburgh, PA ................................................................. 21
Table 2-2 Percentiles and means of age, daily air pollutants and weather variables in Hazelwood, PA ................................................................. 21
Table 2-3 Distribution of daily air pollutants and weather factors in two-year period before and after the plant closure in Hazelwood, PA ................................................................. 22
Table 2-4 ORs ‡ (95% CIs) of cardiopulmonary hospitalizations for PM$_{10}$ before and after the closure of the plant .......................................................................................................................... 23
Table 2-5 ORs ‡ (95% CIs) of cardiopulmonary hospitalizations for SO$_2$ before and after the closure of plant .......................................................................................................................... 24
Table 2-6 ORs* (95% CIs) of cardiopulmonary hospitalizations per unit change in PM$_{10}$ which is fitted as a continuous variable in case-crossover analyses with different control sampling approaches .......................................................................................................................... 25
Table 3-1 Weights for estimating the accumulated exposure during pregnancy .................. 42
Table 3-2 Distribution of trimester-specific PM$_{10}$, Allegheny County PA, 1994-2000 .......... 42
Table 3-3 Distribution of characteristics of term singleton births by low birth weight status .... 43
Table 3-4 Risks for term low birth weight $^a$ according to trimester-specific exposure to PM$_{10}$ .......... 44
Table 4-1 Distribution of baseline characteristics among smoking groups .............................. 62
Table 4-2 Distribution of lymphatic and hematopoietic neoplasm incident cases occur during study period (1979-1995)................................................................. 63

Table 4-3 Distribution and risk of total leukemia among each status of cigarette smoking........ 64

Table 4-4 Distribution and risk of AML among each status of cigarette smoking.................. 65

Table A-1 Results of studies of ambient air pollutants and LBW and VLBW......................79
LIST OF FIGURES

Figure 3-1 Inverse distance interpolation.................................................................41
Figure 4-1 Association between risk of AML and the number years of smoking..........66
Figure B-1 Environmental agent and adverse health outcome.................................87
Figure B-2 General procedure of data linkage.........................................................103
PREFACE

I owe special thanks to all of my committee members for their great support and guidance during the dissertation process. Without their personal involvement, this work could not successfully have been done so well and so quickly. I would like to thank my advisor, Dr. Evelyn Talbott, for her patience, availability and advice. Without her great support and direction, I would not have been able to accomplish this work. I am grateful for the opportunities she has provided to me. I would also like to thank everyone in my research team, especially Dr. Jeanne V. Zborowski, whose professional suggestions and editorial support did me tremendous favor. I appreciate their support more than I can say. I am very thankful to my family and friends for their emotional and editorial support to my work.
1.0 INTRODUCTION

The environment plays an essential role in the health of community. An enormous number and variety of chemicals have been introduced into the environment with the process of industrialization and economic development. They have caused considerable concerns among the community. Exposures to environmental hazards including chemical, biological and physical agents have been linked to various adverse health outcomes such as cancer (Jurewicz and Hanke 2006; Navarro Silvera and Rohan 2007), cardiovascular diseases (Delfino et al. 2005; Monarca et al. 2006), respiratory diseases (Kunzli and Tager 2005), adverse pregnancy outcomes (Perera et al. 2005; Shi and Chia 2001; Sram et al. 2005) and others in the research literature. Efficiently tracking potential health effects associated with environmental exposure appears important, urgent and necessary.

1.1 ENVIRONMENTAL EXPOSURES AND CHRONIC DISEASES

Most chronic diseases are caused by the interaction between gene and environment, i.e. diseases result from an unfavorable combination of genetics and environment (Olden and Guthrie 2001). Population-based twin studies, which can distinguish between the contributions of genetics and environment, suggest that the environment plays a prominent role in disease development (Mucci et al. 2005; O'Brien 2000). A recent study by Lichtenstein et al found that
genetics only account for 21-42% of the risk for developing the 10 common cancers studied and shared environmental exposures could account for 58-82% of the risk (Lichtenstein et al. 2000). Environmental hazards are found not only in chemical agents used in agriculture and industrial emissions but also in household commercial products, food additives, medicine, research lab as well as natural events such as forest fire and volcano eruption. Scientific research has provided a better understanding of some environmental exposures that could cause a variety of adverse health outcomes such as arsenic and cancers in different systems (Ferreccio and Sancha 2006; Tapio and Grosche 2006), radon and lung cancer (Alavanja et al. 1994; Neuberger and Gesell 2002; Pershagen et al. 1992), lead and cognitive function (Banks et al. 1997; Rice 1996), and particulate matter and cardiopulmonary diseases (Peters and Pope 2002; Pope 2000; Pope et al. 2004). However, for many other environmental hazards, scientific evidence is less conclusive. Billions of pounds of toxic agents are released into the environment every year where we live, eat, drink, work and play. In the report of National Survey of Public Perceptions of Environmental Health Risks, 90% of Americans believed that environmental factors like pollution, waste and chemicals are important contributors to diseases (http://healthyamericans.org/reports/files/survey0620.pdf). To address the public concerns, it is very important to systematically and efficiently identify environmental hazards that cause adverse health conditions such as mortality or morbidity in order to remove or minimize the health impacts of the toxic agents in the environment.
1.2 NATIONAL ENVIRONMENTAL PUBLIC HEALTH TRACKING NETWORK

Health tracking systems can be classified into two categories: passive and active tracking system. Passive tracking systems are usually based on a registry where local health care providers routinely report every new-onset case of a disease. Active tracking system mostly relies on population-based survey methods which include conducting personal interviews and reviewing medical records to collect information about personal characteristics, health conditions and others. Tracking in Environmental health is a strategy for the identification of hazards of the environment that cause death, disease, or disability, in order to facilitate the goal of prompt removal or reduction of exposures to the offending agents (Hertz-Picciotto 1996).

1.2.1 THE CURRENT STATUS OF ENVIRONMENTAL HEALTH TRACKING SYSTEMS

The current tracking systems in environmental health are specific only on either environmental hazards or health outcomes monitoring. For example, federal and state environmental agencies have routinely collected much environmental hazard data on air pollution, water pollution and others. In addition, federal, state and local public health agencies have developed registries focusing on health outcomes like cancer, birth defects as well as asthma etc. Some familiar examples of health effects datasets include the Surveillance, Epidemiology, and End Results (SEER) program, the Behavioral Risk Factor Surveillance System (BRFSS), the Iowa Birth Defects Registry, National Health Interview Survey (NHIS), and vital statistics data (See Appendix B: using secondary data for environmental public health tracking). However, few current tracking systems in environmental health can effectively
communicate with other tracking systems, efficiently track environmental exposures, and link environmental contaminants with health endpoints.

In 2000, after 18 months of review, the Pew Environmental Health Commission stated that the current system does not have the capability to respond adequately to environmental threats and that the environmental public health system was fragmented, neglected and ineffective. The commission recommended establishing a national environmental public health tracking (EPHT) network to link information on environmentally related diseases, human exposures and environmental hazards (McGeehin et al. 2004).

1.2.2 THE NATIONAL ENVIRONMENTAL PUBLIC HEALTH TRACKING PROGRAM

The Centers for Disease Control and Prevention (CDC) is leading the initiative to build a national EPHT network which is the ongoing systematic collection, integration, analysis, interpretation, and dissemination of data about environmental hazards, exposure to environmental hazards, and health effects potentially related to exposure to environmental hazards (McGeehin et al. 2004). The establishment of this network is a vital step to address the public health needs of the United States, track chronic diseases as well as environmental exposures and eventually reduce the burden of disease on the nation’s population. Currently there are 17 states and local health departments and four schools of public health funded to conduct activities that will form the basis of the tracking network. As one of four academic partners funded by CDC, the University of Pittsburgh/Drexel Academic Consortium for Excellence in Public Health Tracking (UPACE-EPHT) is uniquely positioned to interact with health-related agencies in a local, state and regional initiative to facilitate environmental capacity
building, to evaluate existing surveillance methodologies and to develop innovative strategies and tools to link hazards, exposures and health effects databases and conduct demonstrative studies of using secondary data to assess the potential associations between environmental hazards and health outcomes. The purpose of the Academic Partners program is to provide expertise and support to the National Environmental Public Health Tracking Program (NEPHTP) in the development and utilization of the Tracking Network.

1.3 PROPOSED THREE DEMONSTRATIVE STUDIES AND OBJECTIVES

One of the important activities of the academic partners in the CDC EPHT program is to conduct pilot/demonstration epidemiological studies using existing hazard, exposure and health outcome data. The main aims of this work are to examine the potential associations between environmental hazards/exposures and health outcomes and to explore methods of data linkage and data analyses. The techniques employed in the demonstration studies will be applicable to state and/or local environmental health tracking agencies. In this project, three pilot studies have been developed using secondary data from state or local environmental public health tracking agencies.

1.3.1 CASE-CROSSOVER ANALYSIS OF AIR POLLUTION AND CARDIO-PULMONARY HOSPITAL ADMISSIONS

A case-crossover analysis is performed to assess the association between air pollution and cardiopulmonary hospital admissions in the Hazelwood neighborhood of Pittsburgh, PA. The
case-crossover design was first proposed by Maclure in 1991 (Maclure 1991). The design is an attractive approach to estimating the effects of triggers of acute health outcomes by environmental exposure. In the design, only cases are involved and the exposure of each case during an at-risk “hazard period” just before the event is compared with the level on one or more reference days when the event did not occur. This self-matching approach will control for all known or unknown time-invariant confounders by design. This method is an alternative to time-series analysis in air pollution studies. The secondary datasets used in the study included the hospital admission data from the Pennsylvania Health Care Cost Containment Council, the climatic data of Pittsburgh from the U.S. National Climatic Data Center database and the air pollution data (PM$_{10}$ and SO$_{2}$) in the Hazelwood monitor site from the Allegheny County Department of Health. The objectives of this study are to assess the health effects of air pollution on cardiopulmonary disease hospitalizations, to evaluate the effects of elimination of a major point source of air pollution, i.e. the closure of LTV coke plant, on these relationships and to examine how the different methods of control sampling in the case-crossover design influence the results (See SECTION 2).

1.3.2 PARTICULATE AIR POLLUTION AND TERM LOW BIRTH WEIGHT IN ALLEGHENY COUNTY

Low birth weight has been reported to influence the subsequent health status of individuals including morbidity in adulthood (Barker 1995; Hales 1997). Studies across the world have consistently provided evidence of the association between particulate matter and adult health including mortality and morbidity (Chen et al. 2007; Dockery et al. 1993; Kan et al. 2007; Kettunen et al. 2007; Pope 2000; Saldiva et al. 1995). A special concern of adverse health
effects of particulate air pollution on fetal health is generated. This study focuses on the associations between exposures to particulate air pollution (PM$_{10}$) and term low birth weight infant and further examines the effect period of particulate matter during pregnancy. The birth registration data from the Allegheny County Health Department and the air pollution data from RAND’s Center for Population Health and Health Disparities (CPHHD) are obtained in this study. (See SECTION 3)

1.3.3 CIGARETTE SMOKING AND ADULT LEUKEMIA

Benzene, a well-established carcinogen and leukemogen, has been shown to be present in cigarette smoke. Wallace observed that approximately 90% of personal exposure to benzene in the United States is due to smoking (Wallace 1996). The study was designed to assess the association between cigarette smoking and leukemia, as well as its subtype, acute myeloid leukemia (AML) and to further examine the dose-response relationship between health endpoints and smoking with the number of cigarettes per day, the number of years of smoking and pack years. The TMI cohort, assembled by the Pennsylvania Department of Health to evaluate the adverse health effects of exposure of low-level radiation emitted from the TMI nuclear power plant accident on 28 March 1979, and the PA cancer registry data are obtained in this study.

The TMI cohort and PA cancer registry data was linked by the key variable of social security number (SSN). The records from these two databases refer to the same individual if the SSN is identical. This is a simple example of deterministic data linkage strategy. Due to confidentiality, this part of the work was done by the PA Department of Health. Survival analysis is performed to estimate the associations after controlling for other important confounding factors.
2.0 CASE-CROSSOVER ANALYSIS OF AIR POLLUTION AND CARDIOPULMONARY HOSPITALIZATIONS: USING ROUTINELY COLLECTED HEALTH AND ENVIRONMENTAL DATA FOR TRACKING

Xiaohui Xu, Ph.D. ¹
Evelyn O. Talbott, Dr.P.H., M.P.H ¹
Jeanne V. Zborowski, Ph.D. ¹
Vincent C. Arena, Ph.D. ²
Judy Rager, M.P.H ¹

1. Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA 15261

2. Department of Biostatistics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA 15261

Manuscript in preparation
2.1 ABSTRACT

**Background:** From the early 1900s until its closure in February 1998, the LTV Steel coke oven in Pittsburgh (Hazelwood), Pennsylvania was a key source of air pollution. A case-crossover study was performed to assess the associations between daily air pollution and cardiorespiratory (ICD-9: 390-519) hospitalizations before and after plant closure and to evaluate how closure influenced these associations.

**Methods:** Air pollution data, climatic data and cardiorespiratory hospitalizations among Hazelwood-area residents 65 years and older were obtained for the period of 1996 through 2000. Data were analyzed using a case-crossover design and conditional logistical regression. Two distinct referent sampling approaches were compared.

**Results:** Significant associations were observed between the fourth quartile in PM$_{10}$ and cardiorespiratory hospitalizations (OR: 1.12; 95%CI: 1.02-1.23) as well as cardiovascular hospitalizations only (ICD-9: 390-459) (OR: 1.13; 95%CI: 1.01-1.26) before the plant closure. After closure of the plant, PM$_{10}$ was not significantly associated with cardiorespiratory or cardiovascular disease hospitalizations. Moreover, the referent sampling approaches did not greatly alter the estimations in the case-crossover analysis.

**Conclusion:** Existing secondary data is an economical source to assess the impact of point source polluters on the environmental landscape. The findings suggest that closure of the LTV coke plant was associated with a reduction risk of the cardiorespiratory and cardiovascular hospitalizations.

**Key word:** Case-Crossover Analysis Cardiopulmonary Air Pollution
2.2 INTRODUCTION

The environment plays an important role in health and human development. There are few surveillance systems that have the capability to adequately assess to environmental threats at the state or national levels in the United States. To address this gap, in 2002 the Centers for Disease Control and Prevention (CDC) began to establish the National Environmental Public Health Tracking (EPHT) Network, which is the ongoing collection, integration, analysis, interpretation, and dissemination of data on environmental hazards; exposure to those hazards; and related health effects (1). A hallmark of the Tracking Network is the use of exiting data in evaluating the relationships between exposures and health outcomes. In the present study, we demonstrated the use of data for purposes other than for which they were collected such as administrative purposes, applied available methods for linking and analyzing environmental and health outcome data, and made a positive contribution to closing gaps in methodology for conducting similar investigations within the context of the environmental public health tracking program.

Epidemiological studies consistently show that air pollutants are linked to cardiovascular and respiratory diseases (2-5). There are several potential biological mechanisms by which air pollutants cause adverse health effects. One of the potential mechanisms is related to inflammation and oxidative stress. High concentrations of oxidants and pro-oxidants contained in ambient air pollution, such as transition metals, reactive organic compounds and gases such as ozone (O₃) or nitrogen oxides (NO, NO₂) can generate reactive oxygen species(ROS)(6), alter function of mitochondria(7) or NADPH-oxidase (8), activate inflammatory cells capable of generation of ROS and reactive nitrogen species(9), and promote oxidative stress and respiratory inflammatory responses, which lead to DNA damage(10). Studies show that air pollution can increase the level of blood viscosity and modify the adhesive properties of red blood cells(11),
increase blood pressure(12) and C-reactive protein level(13), and increase heart rate(14), which lead to an increased risk of cardiovascular diseases. In addition, air pollutants, interacting with lung receptors, can activate pulmonary neural reflexes and change the autonomic function resulting in an decreased heart rate variability (HRV)(15) which partly explains the observed cardiovascular effects.

Coke plants produce a by-product of coal that burns more intensely and is used in the making of steel. These plants, as a major source of air pollution, emit gaseous air pollutants such as sulfur dioxide (SO$_2$), nitrogen oxide (NO$_2$), carbon monoxide (CO), airborne particulate matter which contains organic compounds like benzo[a]pyrene (B[a]P) and other polyaromatic hydrocarbons (PAHs) (16, 17) and inorganic compounds like lead and mercury as well as other hazardous pollutants. These emissions have been linked to cardiorespiratory diseases and threaten the natural environment and overall health of the human population(18, 19).

Hazelwood is a neighborhood located in Pittsburgh, Pennsylvania. The LTV Corporation coke plant, a major local source of air pollution, was operational in the area for more than half a century. Toxic emissions from the plant led to the deterioration of the surrounding environmental landscape in the Hazelwood, Greenfield, and Squirrel Hill neighborhoods. On February 28, 1998, after eighty years of operation, the LTV plant was closed. The plant's auxiliary facilities and much of the plant itself were demolished in early 1999. The closure of the plant provided a unique opportunity to conduct a “retrospective natural intervention” study to examine whether reductions of ambient pollution concentration related to plant emissions reduced the adverse cardiorespiratory health effects in the area.

The case-crossover design was first proposed by Maclure in 1991(20). The design is an attractive approach to estimating the effects of triggers of acute health outcomes by
environmental exposure and has subsequently been widely used in air pollution studies (21-25). In this design, only cases of events are involved. The exposure of each case during an at-risk “hazard period” just before the time of the event is compared with exposure during one or more periods when the event did not occur (control periods). This self-matching approach controls by design for all known or unknown time-invariant confounders.

In the present study, a case-crossover analysis was performed to assess the association between air pollution and cardiorespiratory hospital admissions before and after the closure of this plant among the population 65 years of age and older who lived in the Hazelwood area, and to evaluate how the closure of a major point source of air pollution influences the associations. We also compared two different strategies of selecting referent periods in the case crossover analysis and evaluated how the selection of referent periods influenced the results.

Existing data used in the study included hospital admission data from the Pennsylvania Health Care Cost Containment Council, climatic data of Pittsburgh from the U.S. National Climatic Data Center Database, and a pollution monitoring data from the Allegheny County Health Department.

2.3 MATERIALS AND METHODS

2.3.1 STUDY POPULATION:

The Pennsylvania Health Care Cost Containment Council (PHC4), Special Request Unit (Harrisburg, PA) provided the cardiorespiratory hospital admission data for subjects 65 years of age and older in Hazelwood and surrounding neighborhoods in the city of Pittsburgh,
Pennsylvania. The study location included all zip codes surrounding the LTV coke plant with the plant as the center, i.e. zip codes 15207, 15120, 15217 and 15218. A dataset that included all inpatient cardiovascular and respiratory (i.e. Cardiorespiratory) hospital admissions in the Hazelwood area between 1996 and 2000 was created. The records contained a primary discharge diagnosis of the circulatory system [International Classification of Disease, 9th Revision (ICD-9) codes of 390-459] or respiratory system (ICD-9 codes of 460-519]. The study was limited to the first event for those who were hospitalized for Cardiorespiratory diseases in a one month period.

2.3.2 AIR POLLUTION AND WEATHER DATA

Ambient air levels of specific criteria pollutants were obtained in electronic format for the Hazelwood monitoring site from December 1995 through January 2001. The data collected included 24-hour values for sulfur dioxide (SO$_2$) and PM$_{10}$. The daily minimum, maximum and mean values of SO$_2$ and PM$_{10}$ were computed for the present study. These daily measurements were assumed to represent the exposure level for the entire Hazelwood study area.

Daily meteorological data were obtained from the U.S. National Climatic Data Center database from the monitoring site at the Pittsburgh International Airport, Allegheny County (Coopid: 366993, Wbandid: 94,823, Latitude: 40°30', Longitude: -80°14'). The information for daily mean temperature and daily mean relative humidity was abstracted from the database for the time period of 1996-2000.
2.3.3 DATA ANALYSIS

Data were analyzed with the case-crossover technique which is an alternative to time-series analysis for assessing acute health effects of air pollution. In the design, cases serve as their own controls. A subject’s exposure at the time of a health event (case-period) is compared with exposures at previous or subsequent points of time when that subject was a non-case (control-period). In the present study, the case-period is defined as the date of hospitalization for any cardiorespiratory diseases. The control periods were chosen using two methods, either the bidirectional control sampling approach or the time-stratified sampling approach in order to control relevant time-varying confounders (26). In the bidirectional control sampling approach, the control periods were selected as 7 and 14 days before and after the date of hospitalization in order to control for day-of-week effect. In the time-stratified method, the stratum is defined as the month of event. All other same days of week as the case period in the month were selected as the control periods. For example, if a case occurs in March on a Wednesday, all other Wednesdays in March are the control periods.

In the bidirectional control sampling approach, the associations between air pollutants and hospitalization for cardiovascular diseases only or respiratory diseases only as well as both combined (cardiorespiratory) were estimated for the time period January 1, 1996 to February 28, 1998 (before closure of the plant) and for the period March 1, 1998 to December 31, 2000 (after closure of the plant), respectively. The SAS conditional logistic regression procedure (Proc Phreg) was applied to estimate the associations between hospital admissions and air pollution. Air pollutants were fitted into the model as continuous variables or categorical variables (quartiles). Odds ratios (ORs) were calculated for continuous variables and the quartiles of each
pollutant during each time period with the first quartile as a reference after adjusting for daily meteorological factors including current day’s mean temperature and relative humidity.

In the time-stratified control sampling approach, all analyses were performed with the Case-Crossover Analysis Tool (Beta V1.1 C-CAT), developed by Apex Epidemiology Research, LLC, in conjunction with the New York State Department of Health, to provide an easy-to-use interface to SAS software that implements time-stratified case-crossover analysis. Air pollutants were fitted into the model only as continuous variables. The results of this approach were compared with the previous approach. In addition, the impact of different referent period selections on the results was also examined.

2.4 RESULTS

Among the residents of Hazelwood area who were at least 65 years of age, there were a total of 13,437 cardiorespiratory disease admissions (9,935 cardiovascular disease admissions and 3,502 respiratory diseases admissions) during 1996-2000. Table 1 summarizes the number of admissions that had a specific diagnosis of cardiovascular disease and respiratory disease before and after the closure of the LTV coke plant. There were 4,491 and 5,444 cardiovascular disease hospitalizations before and after the closure of the plant, respectively. In addition, 1,596 and 1,906 respiratory disease hospitalizations were respectively observed before and after the closure of the plant in this study.

The distributions of air pollutants, age and meteorological variables before and after the closure of plant are described in Table 2-2. The distribution of age of hospitalizations is similar before and after the plant closure. The percentiles ($10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$ and $90^{th}$) of PM$_{10}$ and
SO_2_ were all higher before than after the closure of the plant. The daily averages of PM_{10} were 27.75 µg/m^3 before and 21.96 µg/m^3 after the plant closed, respectively. The average of SO_2 decreased to nearly half of the previous levels (from 16.4 ppb to 8.7 ppb) after the plant was closed. For considering the difference of length of time between before and after the plant closure, air pollutants and climatic factors are compared for two-year period before and after the plant closure, respectively. The results are described in Table 2-3. The change of air pollutants is the same regardless of the difference of length of time. The percentiles of daily temperature and daily relative humidity were similar in two-year period before and after the closure of the plant.

Table 2-4 shows the associations between quartiles of PM_{10} and cardiorespiratory hospitalizations after controlling for temperature and relative humidity. The results showed significant associations between the fourth quartile in PM_{10} and cardiorespiratory hospitalizations (OR: 1.12; 95%CI: 1.02-1.23) as well as cardiovascular disease hospitalization only (OR: 1.13; 95%CI: 1.01-1.26) before the closure of the plant. Moreover, significant trends for increasing risks on cardiorespiratory hospitalizations (p=0.008) and cardiovascular disease hospitalizations (p=0.02) with increasing concentration of PM_{10} were observed. After the closure of the plant, the associations with cardiorespiratory hospitalization were no longer statistically significant. In this study, no significant association between respiratory disease hospitalization and PM_{10} was found either before or after the plant closure.

The associations between SO_2 and cardiorespiratory hospitalizations were illustrated in Table 2-5. There were no significant associations between SO_2 and hospitalizations of cardiovascular diseases only or respiratory diseases only or both combined either before or after the closure of the plant.
Table 2-6 shows the associations between PM$_{10}$ and cardiorespiratory hospitalizations with different referent periods in the case-crossover analysis. The results from the bidirectional control sampling approach are similar to the results of the time-stratified control sampling method.

2.5 DISCUSSION

Few studies are currently available to assess how the addition or elimination of a single point major air pollution source affects the risk of adverse health effects in specific geographic area. This study, for the first time, used a case-crossover design to evaluate this relationship using existing secondary data. The case-crossover approach is a self-matched case control study, which has great advantages in comparison with a time series approach. This approach controls certain confounding factors by design rather than by modeling. With this design, personal characteristics such sex, age, race and time-invariant variables are controlled by the design. By using symmetric 7 and 14 day reference periods, the case-crossover design eliminates the confounding effects of long-term trends, seasonality and day of week(27). Therefore, the use of a case-crossover design avoids common concerns about the complex mathematical modeling and adequacy of seasonal control.

This study demonstrated that the levels of PM$_{10}$ and SO$_2$ decreased in Hazelwood area after the plant was closed. PM$_{10}$ declined an average of 6 µg/m$^3$ and SO$_2$ was reduced to half of the level when the plant was open. Significant associations between PM$_{10}$ and cardiovascular diseases as well as cardiorespiratory disease hospitalizations were observed before the closure of plant. Moreover, a trend of increased risk for cardiovascular disease as well as cardiorespiratory
disease hospitalizations with increasing particulate matter was observed before the plant closure in the study. Interestingly, after the closure, the associations were no longer statistically significant. Sulfur dioxide did not show a significant association with any outcomes in this study either before or after the closure of the plant. Because sulfur dioxide, a highly soluble gas, can be easily and predominantly stripped out of the upper airways(28), the role of sulfur dioxide in developing adverse health effects is less coherent. Several case-crossover studies reported that there are no significant associations between sulfur dioxide and respiratory disease(22, 29) as well as cardiovascular disease hospitalization(30, 31).

The reduced health effects of PM$_{10}$ after the plant closure could be due to two facts: the change in composition and/or concentration of particulate matter. After the elimination of the LTV coke plant, some toxic elements in the particulate matter associated with the plant emission have also been removed. The monitored data from the Allegheny County Health Department showed that the concentrations of benzo(α)pyrene [B(α)P], as a recognized carcinogen as well as respiratory, endocrine and immunological toxicant, was significantly lower after the plant closure. B(α)P can not however be used as a predictor in the study because it was measured every 7 days. Another potential fact is the reduced concentration of particulate matter in the area. At the low level of particulate matter, there might be no adverse health effect on cardiorespiratory hospitalizations or the association is too weak to be detected on this population sample size.

The findings of this study support that particulate air pollution could increase the risks of cardiorespiratory hospitalizations before the LTV coke plant closure and an elimination of a major air pollution source could lead to a reduced risk of adverse health effects, which is consistent with other epidemiological studies. Pope et al conducted a cross-sectional study to
assess how the operation of a steel mill influences the association between hospital admissions and PM10 in Utah Valley during the period April 1985-February 1988. The results showed that children's hospital admissions were two to three times higher during the winters when the mill was open compared to when it was closed (32). The findings in this study further strengthen the epidemiologic evidence of acute adverse health effects of airborne particle levels.

In this study, we also evaluated how the control sampling approaches influence the estimations in the case-crossover analysis. PM$_{10}$ was the only pollutant used to evaluate these approaches because sulfur dioxide had no associations with the outcomes in this study and Akaike's information criterion (AIC) was increased in both methods after SO$_2$ was fitted as a continuous variable, suggesting a poor model fit. The estimations from the bidirectional control sampling methods are similar to those from the time-stratified method. The results of the comparison suggested that the control sampling methods in case-crossover analysis did not alter the conclusion.

A potential limitation of this study is that only one air pollution monitoring site was available in the Hazelwood area. Individual exposure was estimated by using the measurement from this fixed outdoor monitoring station. The potential misclassification of exposure due to lack of personal measurement can not be avoided. However, the bias might underestimate the association and shift the results toward the null (33, 34). In addition, we did not assess the promoting role of other pollutants such as nitrogen dioxide or ozone because the data of these pollutants from the monitor site were not available or complete. However, particulate matter is likely a reasonable indicator of other pollutants due to the high collinearity between particulate matter and other pollutants (35). Another limitation is potential for selection bias because we were unable to include all cardiorespiratory disease cases in the study if unreported to the PHC4.
Finally, another limitation of this study is that the sample size of respiratory disease hospitalization might be not large enough to have sufficient power to detect a statistically significant difference.

In summary, this study demonstrates the utility of secondary datasets and the methodology of the case-crossover approach to evaluate the associations between environmental air pollutants and acute health outcomes. This investigation demonstrated important strategies for using existing data to address different health issues and developing the methodology of data linkage and data analysis in conducting similar studies in an environmental public tracking program. Moreover, this analysis provides, for the first time, evidence of a reduction in air pollution resulting in a decrease in risk of Cardiorespiratory disease hospitalization using a case-crossover design. The findings reinforce the deleterious impact of air pollution and provide new epidemiologic information that the elimination or addition of a major point source of pollution could change a local environmental landscape and influence the health of the population living in the area. Further studies are required to confirm these findings.
Table 2-1 Distribution of specific primary diagnoses of hospital admissions among study sample in Hazelwood area, Pittsburgh, PA

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>4,491</td>
<td>5,444</td>
<td>9,935</td>
</tr>
<tr>
<td>Respiratory</td>
<td>1,596</td>
<td>1,906</td>
<td>3,502</td>
</tr>
<tr>
<td>Combined</td>
<td>6,087</td>
<td>7,350</td>
<td>13,437</td>
</tr>
</tbody>
</table>

Table 2-2 Percentiles and means of age, daily air pollutants and weather variables in Hazelwood, PA

<table>
<thead>
<tr>
<th>Variables</th>
<th>Time periods</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>LTV Plant open</td>
<td>68</td>
<td>72</td>
<td>77</td>
<td>83</td>
<td>88</td>
<td>77.4</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>68</td>
<td>72</td>
<td>78</td>
<td>83</td>
<td>88</td>
<td>78.0</td>
</tr>
<tr>
<td>PM(_{10}), µg/m(^3)</td>
<td>LTV Plant open</td>
<td>13.1</td>
<td>17.8</td>
<td>24.0</td>
<td>35.0</td>
<td>46.8</td>
<td>27.8</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>10.4</td>
<td>13.5</td>
<td>19.1</td>
<td>27.7</td>
<td>37.2</td>
<td>22.0</td>
</tr>
<tr>
<td>SO(_2), ppb</td>
<td>LTV Plant open</td>
<td>4.7</td>
<td>8.2</td>
<td>14.1</td>
<td>22.4</td>
<td>30.1</td>
<td>16.4</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>3.4</td>
<td>5.0</td>
<td>7.7</td>
<td>11.0</td>
<td>15.3</td>
<td>8.7</td>
</tr>
<tr>
<td>Temperature (°F)</td>
<td>LTV Plant open</td>
<td>26.0</td>
<td>35.0</td>
<td>48.0</td>
<td>65.0</td>
<td>72.0</td>
<td>48.8</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>28.0</td>
<td>40.0</td>
<td>55.0</td>
<td>67.0</td>
<td>74.0</td>
<td>52.5</td>
</tr>
<tr>
<td>Humidity (%)</td>
<td>LTV Plant open</td>
<td>30.0</td>
<td>40.0</td>
<td>49.0</td>
<td>62.0</td>
<td>73.0</td>
<td>51.1</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>29.0</td>
<td>37.0</td>
<td>48.0</td>
<td>60.0</td>
<td>72.0</td>
<td>49.1</td>
</tr>
</tbody>
</table>
Table 2-3 Distribution of daily air pollutants and weather factors in two-year period before and after the plant closure in Hazelwood, PA

<table>
<thead>
<tr>
<th>Variables</th>
<th>Time periods</th>
<th>Mean</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{10}$, $\mu$g/m$^3$</td>
<td>LTV Plant open</td>
<td>28.2</td>
<td>13.5</td>
<td>18.2</td>
<td>24.6</td>
<td>35.4</td>
<td>47.4</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>21.3</td>
<td>10.2</td>
<td>13.1</td>
<td>18.5</td>
<td>26.9</td>
<td>36.3</td>
</tr>
<tr>
<td>SO$_2$, ppb</td>
<td>LTV Plant open</td>
<td>16.1</td>
<td>4.5</td>
<td>8.1</td>
<td>13.8</td>
<td>22.2</td>
<td>29.5</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>8.5</td>
<td>3.1</td>
<td>4.9</td>
<td>7.7</td>
<td>10.9</td>
<td>14.5</td>
</tr>
<tr>
<td>Temperature ($^\circ$F)</td>
<td>LTV Plant open</td>
<td>50.4</td>
<td>27</td>
<td>37</td>
<td>51</td>
<td>66</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>51.6</td>
<td>26</td>
<td>38</td>
<td>55</td>
<td>67</td>
<td>73</td>
</tr>
<tr>
<td>Humidity (%)</td>
<td>LTV Plant open</td>
<td>50.5</td>
<td>30</td>
<td>40</td>
<td>49</td>
<td>61</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>48.5</td>
<td>28</td>
<td>36</td>
<td>48</td>
<td>60</td>
<td>71</td>
</tr>
<tr>
<td>Hospitalizations</td>
<td>Exposure levels</td>
<td>PM$_{10}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-----------------</td>
<td>----------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular and Respiratory diseases combined</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;25th percentile *</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th to 50th percentile</td>
<td>0.99 (0.91-1.08)</td>
<td>1.03 (0.96-1.11)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50th to 75th percentile</td>
<td>1.08 (0.99-1.18)</td>
<td>1.02 (0.94-1.10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥75th percentile</td>
<td>1.12 (1.02-1.23) §</td>
<td>0.96 (0.88-1.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P for trend</td>
<td>0.008§</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular diseases only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;25th percentile *</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th to 50th percentile</td>
<td>1.01 (0.91-1.11)</td>
<td>1.09 (0.99-1.81)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50th to 75th percentile</td>
<td>1.10 (0.99-1.21)</td>
<td>1.04 (0.95-1.14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥75th percentile</td>
<td>1.13 (1.01-1.26) §</td>
<td>0.94 (0.85-1.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P for trend</td>
<td>0.02§</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory diseases only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;25th percentile *</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th to 50th percentile</td>
<td>0.97 (0.82-1.39)</td>
<td>0.89 (0.78-1.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50th to 75th percentile</td>
<td>1.05 (0.89-1.24)</td>
<td>0.97 (0.84-1.13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥75th percentile</td>
<td>1.11 (0.92-1.35)</td>
<td>1.03 (0.87-1.22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P for trend</td>
<td>0.22</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Reference group
† Odds ratios adjusted for daily weather variables (daily mean temperature and humidity)
§ P<0.05
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SO₂</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular and Respiratory</td>
<td>&lt;25&lt;sup&gt;th&lt;/sup&gt; percentile *</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>diseases combined</td>
<td>25&lt;sup&gt;th&lt;/sup&gt; to 50&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>1.00 (0.92-1.09)</td>
<td>1.03 (0.95-1.11)</td>
</tr>
<tr>
<td></td>
<td>50&lt;sup&gt;th&lt;/sup&gt; to 75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>1.06 (0.97-1.16)</td>
<td>0.99 (0.92-1.07)</td>
</tr>
<tr>
<td></td>
<td>≥75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>1.01 (0.93-1.11)</td>
<td>1.02 (0.94-1.11)</td>
</tr>
<tr>
<td></td>
<td>P for trend</td>
<td>0.47</td>
<td>0.68</td>
</tr>
<tr>
<td>Cardiovascular diseases only</td>
<td>&lt;25&lt;sup&gt;th&lt;/sup&gt; percentile *</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>25&lt;sup&gt;th&lt;/sup&gt; to 50&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>1.04 (0.95-1.15)</td>
<td>1.03 (0.96-1.12)</td>
</tr>
<tr>
<td></td>
<td>50&lt;sup&gt;th&lt;/sup&gt; to 75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>1.07 (0.96-1.19)</td>
<td>0.96 (0.88-1.05)</td>
</tr>
<tr>
<td></td>
<td>≥75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>1.05 (0.94-1.17)</td>
<td>0.99 (0.89-1.09)</td>
</tr>
<tr>
<td></td>
<td>P for trend</td>
<td>0.25</td>
<td>0.59</td>
</tr>
<tr>
<td>Respiratory diseases only</td>
<td>&lt;25&lt;sup&gt;th&lt;/sup&gt; percentile *</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>25&lt;sup&gt;th&lt;/sup&gt; to 50&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>0.89 (0.75-1.05)</td>
<td>1.02 (0.88-1.18)</td>
</tr>
<tr>
<td></td>
<td>50&lt;sup&gt;th&lt;/sup&gt; to 75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>1.05 (0.88-1.24)</td>
<td>1.09 (0.94-1.27)</td>
</tr>
<tr>
<td></td>
<td>≥75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>0.92 (0.77-1.10)</td>
<td>1.12 (0.95-1.32)</td>
</tr>
<tr>
<td></td>
<td>P for trend</td>
<td>0.59</td>
<td>0.09</td>
</tr>
</tbody>
</table>

* Reference group
† Odds ratios adjusted for daily weather variables (daily mean temperature and humidity)
<table>
<thead>
<tr>
<th>Hospitalizations</th>
<th>Time periods</th>
<th>Control sampling approach</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bidirectional control sampling</td>
<td>Time-stratified control sampling</td>
</tr>
<tr>
<td>Cardiovascular and Respiratory diseases combined</td>
<td>LTV Plant open</td>
<td>1.003 (1.001-1.005)</td>
<td>1.002 (1.00-1.005)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>0.998 (0.995-1.001)</td>
<td>0.997 (0.994-1.00)</td>
<td></td>
</tr>
<tr>
<td>Cardiovascular diseases only</td>
<td>LTV Plant open</td>
<td>1.003 (1.00-1.006)</td>
<td>1.003 (1.00-1.005)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>0.997 (0.994-1.00)</td>
<td>0.996 (0.993-1.00)</td>
<td></td>
</tr>
<tr>
<td>Respiratory diseases only</td>
<td>LTV Plant open</td>
<td>1.003 (0.998-1.008)</td>
<td>1.001 (0.996-1.006)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LTV Plant closed</td>
<td>1.00 (0.995-1.006)</td>
<td>0.998 (0.992-1.004)</td>
<td></td>
</tr>
</tbody>
</table>

* Odds ratios adjusted for daily weather variables (daily mean temperature and humidity)
2.6 REFERENCES


3.0 LOCAL VARIATION IN PARTICULATE MATTER DURING PREGNANCY AND TERM LOW BIRTH WEIGHT IN ALLEGHENY COUNTY, PA, USA

Xiaohui Xu, Ph.D. 1
Ravi K. Sharma Ph.D 2
Evelyn O. Talbott, Dr.P.H., M.P.H 1
Jeanne V. Zborowski, Ph.D. 1
Vincent C. Arena, Ph.D. 3
Devra Lee Davis Ph.D 1
Conrad D Volz Dr. P.H. 4
Judy Rager, M.P.H 1

1. Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh
2. Department of Behavioral Community Health Sciences, Graduate School of Public Health, University of Pittsburgh
3. Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh
4. Department of Environmental and Occupational Health, Graduate School of Public Health, University of Pittsburgh

Manuscript in preparation
3.1 ABSTRACT

Introduction: Low birth weight has been associated with increased risks of mortality and/or morbidity in childhood and adulthood. The aim of this study is to explore the association between particulate matter (PM$_{10}$) and term low birth weight (birth weight <2,500g).

Methods: Birth data from the Allegheny County Health Department, Pittsburgh PA and PM$_{10}$ air data generated with inverse-distance interpolation by RAND’s Center for Population Health and Health Disparities, were obtained. The study population consisted of all term singleton live births (gestational age $\geq$37 weeks) born between Jan 1$^{st}$, 1994 to Dec 31$^{st}$, 2000. Infants with birth weight <2,500g were classified as LBW. Logistic regression was performed to estimate the association per inter-quartile range increase in PM$_{10}$.

Results: The results showed that the odds ratios of term LBW per inter-quartile range increase in PM$_{10}$ were 1.13 (95%CI: 1.02-1.25) during the first trimester and 1.10 (95%CI: 1.00-1.22) during the second trimester after adjusted for other important covariates, respectively.

Conclusion: The findings of the study support that exposure to PM$_{10}$ could increase risks of term LBW. Further studies are warrant to corroborate these findings.

Keywords: LBW  PM10 air pollution  Inverse-distance interpolation
3.2 INTRODUCTION

A growing body of evidence suggests that maternal exposure to air pollution is associated with adverse pregnancy outcomes. Studies conducted worldwide have investigated the health effects of air pollution on preterm delivery (PTD) (1, 2), low birth weight (3-5), intrauterine growth restriction (IUGR) (6, 7) and birth defects (8, 9). The air pollutants of concern in these studies include carbon monoxide (CO), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ozone (O₃), and particulate matter (PM).

Low birth weight (birth weight less than 2500 g) is comprised of the two overlapping etiologies of PTD as well as IUGR. Term LBW is caused by fetal growth retardation instead of early delivery. LBW has been widely reported to influence the health status of individuals, including increased mortality and morbidity in childhood (10, 11) and an elevated risk of hypertension, coronary heart disease, Type II diabetes in adulthood, abnormalities of lipid metabolism and blood coagulation (12-15). The public health relevance on this important health issue is evident.

Studies across the world have consistently provided evidence that exposure to particulate matter increases risks of mortality and morbidity among adults (16-23). There is a great concern about the association between particulate matter and fetal health outcomes, especially low birth weight. However, the findings of particulate matter in fetal health research are inconsistent, especially regarding the effect period and the strength of association of particulate matter. In addition, many prior studies did not account for important confounders such as maternal smoking, gestational age and weight gain. Previous studies have also had limited spatial...
information on pollution sources and concentrations because the exposure mostly relied on the measurements at monitoring stations.

In the present study, we investigated the possible association between particulate matter (PM$_{10}$) and term LBW in Allegheny County, PA (USA). The specific purposes of this study are to use PM$_{10}$ data spatially and temporally being generated with inverse-distance interpolation, to estimate the levels of PM$_{10}$ corresponding to the first, second, and third trimesters of pregnancy, to evaluate the association between PM$_{10}$ and term low birth weight, and to explore the possible techniques of data management, data linkage and data analysis for environmental public health tracking to evaluate these relationships.

### 3.3 MATERIALS AND METHODS

#### 3.3.1 Live birth cohort

All live birth data in Allegheny County, PA was obtained from the birth registry data maintained by the Allegheny County Health Department, PA for the period January 1, 1994 through December 31, 2000. The information in the database includes date of birth, birth weight, gestational age, parity, birth order, maternal age, education, race and geographical location of birth (neighborhood in the city of Pittsburgh and municipality outside the city) and other reproductive data. Individual data of all singleton live births to the Allegheny County residents were abstracted from the database and used in the present study.
3.3.2 Air pollution data and geographical data

The quarterly measures of PM$_{10}$ in census tracts in Allegheny County were obtained from the RAND’s Center for Population Health and Health Disparities (CPHHD) Data Core. The geographic-specific data were derived from the publicly available Environmental Protection Agency (EPA) Air Quality System (AQS) Criteria pollutant data with the inverse-distance interpolation based on the 1990 based census tracts. The following are the basic processes used to construct the dataset.

The hourly data of PM$_{10}$ between 1994 and 2000 were obtained from the U.S EPA AQS. The data were aggregated from daily to quarterly levels for each monitor site. The quarterly PM$_{10}$ measures from the EPA monitor sites were used to estimate the levels at the centroids of census tracts, which represent the levels of the census tracts. All distances from the census tract centroid to the site locations were computed geospatially using ArcGIS (Version 9.1, ESRI Inc. Redlands, WA, USA). The measures from the monitor sites within 100 kilometers between the centroid and their locations are used to estimate the level of that census tract. The estimation involves various inverse distance-weighting schemes that are a function of the distance between the census tract centroid and the PM monitors (see Figure 1). Therefore, the sites further away will have less influence and thus less contribution overall to the estimation.

3.3.3 Exposure estimation

The quarterly PM$_{10}$ data were aggregated from census tract to neighborhood in the city or municipality outside the city in geographical level. The average of the quarterly PM$_{10}$ measures of all census tracts in each neighborhood or municipality was used to estimate the quarterly level
of that neighborhood or municipality. It assumed that all births in the same neighborhood or municipality had the same exposure at a single time.

Air pollution estimates for each individual birth were assigned using the neighborhood or municipality of the maternal residence at infant’s birth. For each birth, the accumulated first 3-month, 6-month and 9-month exposures was computed by weighting the quarterly PM10 measures based on the mother’s residence (neighborhood or municipality), the month, quarter and year of birth conception. The weights of estimating the accumulated exposures are shown in the table 1. The date of conception was computed based on the date of birth and gestational age. For example, for an infant who was conceived in February 1994, the accumulated first 3-month exposure is two times the first quarterly PM10 measure in 1994 plus one time the second quarterly PM10 measure in 1994. The monthly averages of trimester-specific exposures were calculated based on the accumulated exposures.

3.3.4 Statistical Analysis

In this study, our analyses were focused on the effects of PM$_{10}$ on birth weight mediated by reduced fetal growth as opposed to early delivery. Therefore, the study population was restricted to infants who were born at term with gestational age $\geq$37 weeks. The relationship between particles and term low birth weight was evaluated using logistic regression analyses. Adverse pregnancy outcomes, i.e. term LBW, defined as dichotomous categories, represent dependent variables in the analysis. A term LBW infant is defined as a live-birth infant weighing $<2,500$g and gestational age $\geq$ 37 complete weeks.

Several known risk factors for term LBW that could potentially confound the relationship between LBW and air pollution were also included in regression models: maternal age ($<$20, 20-
29, 30-34, 35-39, ≥40 years), maternal race (black, white, others), maternal education (<9, 9-11, 12, 13-15, ≥16 years), maternal tobacco use (yes/no), level of prenatal care (none, during first trimester, after first trimester), history of LBW or preterm infant (one or more vs. none), parity (first birth vs. second or subsequent birth), birth season, infant sex, gestational age (measured in weeks) and maternal weight gained.

Air pollution exposures were fitted into the logistic regression model as continuous variable. The odds ratios with per inter-quartile range increase in PM10, i.e. 7 \mu g/m^3 increase, were estimated for each trimester.

### 3.4 RESULTS

A total of 100,595 singleton birth records were available for Allegheny County between 1994 and 2000. Among 92,447 singleton term births (≥37 completed weeks gestation), 47,221 were male births and 45,226 female births. The sex ratio at birth is 104 per 100 females. A total of 2,058 (2.2%) low weight births occurred over the entire study period.

The mean trimester-specific exposures and 9-month period for PM10 during the study period were 28.7, 28.2, 28.1 and 28.1 \mu g/m^3 respectively. The inter-quartile range for each trimester is nearly 7 \mu g/m^3 (table 3-2).

Table 3-3 summarizes the distributions of characteristics of singleton births among term LBW. The results of crude estimations suggested that term LBW was significantly associated with maternal age, race, education, tobacco use, infant gender, previous LBW or PTD, parity, prenatal care, gestational age and weight gained. For term LBW, younger (<20) and older (≥40) maternal age, no previous birth, low levels of education and prenatal care, tobacco use and
previous LBW or PTD could increase the risks of term LBW. African-American women had a higher risk of term LBW than whites. Male infant sex was negatively associated with term LBW. Risk of term LBW decreased as the increase of gestational age and maternal weight gained (See Table 3-3).

Table 4-4 illustrates the risks of term LBW associated with per inter-quartile range increase in exposure to trimester-specific PM$_{10}$. Significant associations between term LBW and the first trimester exposure to PM$_{10}$ was observed in the study after controlling for other covariates including for maternal age, maternal race, maternal education, smoking, weight gain, gender of infant, gestation age, parity, previous LBW or preterm birth, level of prenatal care and birth season. The odds ratios of term LBW for per inter-quartile range increase in PM$_{10}$ were 1.13 (95%CI: 1.02-1.25) during the first trimester and 1.10 (95%CI: 1.00-1.22) during the second trimester, respectively. There is no significant association between term LBW and PM$_{10}$ exposure during the third trimester (OR=1.05, 95%CI: 0.96-1.16) and 9-month period (OR=1.07, 95%CI: 0.99-1.14).

3.5 DISCUSSION

The sex ratio at birth is one of the stable parameters of a population and it should lie in a narrow range of 100-108 males per 100 females. In this study, the sex ratio is 104 per 100 females, which is similar to the national level in the USA (24). We examined associations between term low birth weight and exposure to PM$_{10}$ at various stages of pregnancy. Increased risks of LBW were observed for mother’s exposure to PM$_{10}$ during the first trimester after adjusting for other important confounding factors including maternal age, race, education,
smoking, weight gain, infant gender, gestational age, parity, history of LBW or PTD, prenatal care and birth season. The findings of this study support that PM$_{10}$ exposure could increase risk of term LBW.

In adverse pregnancy outcome research, it is important to identify the critical time points during pregnancy when exposure to air pollutants might be most harmful (25). Our finding of an adverse effect of ambient particulates in the first trimester on pregnancy outcomes is consistent with some previous studies. Dejmek et al conducted a study in the Treptice District of the Czech Republic and revealed that exposure to PM$_{10}$ during the first month was associated with intrauterine growth retardation (IUGR) (6). The study from Seoul found that carbon monoxide, nitrogen dioxide, sulfur dioxide, and total suspended particle concentrations in the first trimester of pregnancy period are risk factors for low birth weight (26). The finding from a study in Brazil also showed that exposure from the first trimester was the most important for LBW (27). However, other studies found that exposures during the third trimester were importantly linked with birth weight. Chen et al found that exposure to PM$_{10}$ in the third trimester of pregnancy was negatively associated with birth weight in North Nevada, USA (28). Wang and colleagues reported that concentrations of TSP and SO$_2$ in the last trimester of pregnancy were associated with low birth weight in Beijing, China (29). A study in Poland also found that PM$_{2.5}$ exposure in the second trimester of pregnancy was negatively associated with birth weight (30). Therefore exposure during the earlier or later stage of pregnancy could present the possibility of interference with final infant weight gain.

Although the specific biological mechanisms that may account for the association between ambient air pollution and adverse pregnancy outcomes are not well known, the effect of particulate matter exposure during pregnancy on adverse pregnancy outcomes has a plausible
biological basis. It has been assumed that prenatal exposure to ambient air pollution might be similar to maternal smoking which can result in some adverse reproductive outcomes. The possible biological mechanisms of air pollution on birth weight might vary according the time of pregnancy, such as the implantation of the fetus and the formation of placenta during the first trimester as well as important weight gain during the third trimester. A study found that placental abnormalities due to exposure in the first trimester and complex vascular alterations in the second and third trimester could be the main causes of placental abnormalities and fetal growth retardation(31). PM10 contains many toxic elements including polycyclic aromatic hydrocarbons (PAHs), which could produce DNA adducts and cause DNA damage, resulting in activation of apoptotic pathways(32). Molecular epidemiological studies have been shown that the levels of DNA adducts are positively related to the risk of low birth weight(33-36). Another possible way is that toxic components and/or its metabolites may bind some receptors, which results in disruption of endocrine system and changes the placental function with decreasing exchange of oxygen and nutrients(37). Moreover, these toxic elements might also be able to provoke alveolar inflammation and release the mediators capable of increasing blood coagulability which increase blood viscosity (38). The change of viscosity can affect blood perfusion and have an adverse effect on placental functions, which may result in adverse birth outcomes. All these changes could result into consequent fetal growth retardation. Therefore, our finding of a significant effect during the first trimester is coherent with some potential mechanisms.

There are number of strengths to our study. First, the present study had a population based birth data with a large sample size to assess the effects of particulate air pollution on term LBW. By focusing on term LBW, the effect of PM10 on fetal growth independent from the effects of prematurity could be examined in this study. Second, the health effects of particulate air
pollutant on birth weight were estimated after controlling for potential confounding factors in this study. Important known risk factors for a low birth weight baby like maternal smoking and maternal weight gain were also considered in this study. Most previous studies of adverse pregnancy outcomes did not control for the information due to unavailability (3-5, 26, 27, 29).

Third, our study used the inverse-distance interpolation to predict the levels of particulate air pollutant in smaller geographical resolutions, which provided a more accurate exposure assessment for individual mothers. The inverse-distance interpolation, one of the popular methods to predict spatial distribution of air pollutants, uses real pollution measurements in the neighborhoods and estimates the weight average of neighboring values for un-sampled locations(39).

While the findings of this study are in accordance with other previous studies, some methodological aspects of this study should be elucidated. A certain degree of inaccuracy of maternal exposure might occur due to invalid assumption, difference between monthly period and gestational age, and difference between personal exposure and estimated quarterly measurement. For example, one assumption of this study is that mothers did not move and stayed most of time pregnancy in the place where they resided. However, we do not have information on maternal mobility during pregnancy to assess whether the assumption is true or not. In addition, although our analyses controlled for a number of important potential confounders, information on other factors such as second hand smoking, maternal occupational exposures and nutrition are not available in this study. Therefore, the potential confounding of these factors could not be controlled. However, we adjusted for maternal weight gain in our analyses, which could partially control for the effects of maternal nutrition. Another issue in this study is that we used quarterly levels of PM$_{10}$ to estimate maternal trimester-specific exposure.
Therefore, a certain degree of inaccuracy of maternal exposure might occur. However, such bias or inaccuracy would be non-differential and make the effects toward the null. Moreover, we applied the weight method to estimate the accumulated exposure, which could provide estimates very close to one based on monthly air data.

In conclusion, our findings suggested that exposure to PM10 during the first trimester of pregnancy is associated with an increased risk of term LBW among infants whose mothers resided in Allegheny County, PA, between 1994 and 2000. Further studies are required to elucidate this association and to corroborate the findings of potential public health significance.
Monitor sites
Points to be
Distance

\[
\hat{Z}_{0,IDp} = \sum_{i=1}^{N} Z(s_i) d_{0,i}^{-p} \quad \frac{\sum d_{0,i}^{-p}}{} 
\]

\( \hat{Z}_{0,IDp} \) estimated value
Z(s_i) observed value in “i” monitor site
N: number of monitor sites
d_{0,i} distance between estimated point and “i” site
p the power

Figure 3-1 Inverse-distance interpolation
### Table 3-1 Weights for estimating the accumulated exposure during pregnancy

<table>
<thead>
<tr>
<th>Month of conception</th>
<th>Accumulated Exposure first 3 months</th>
<th>Accumulated Exposure first 6 months</th>
<th>Accumulated Exposure first 9 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_c$</td>
<td>$Q_{a1}$</td>
<td>$Q_c$</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

$Q_c$: the level of PM10 during the quarter of birth conception  
$Q_{a1}$: the level of PM10 in the 1st quarter after the quarter of birth conception  
$Q_{a2}$: the level of PM10 in the 2nd quarter after the quarter of birth conception  
$Q_{a3}$: the level of PM10 in the 3rd quarter after the quarter of birth conception

### Table 3-2 Distribution of trimester-specific PM$_{10}$, Allegheny County PA, 1994-2000

<table>
<thead>
<tr>
<th>Trimester (monthly average)</th>
<th>Mean</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Trimester</td>
<td>28.7</td>
<td>25.1</td>
<td>28.3</td>
<td>32.2</td>
<td>37.4</td>
</tr>
<tr>
<td>2nd Trimester</td>
<td>28.2</td>
<td>24.9</td>
<td>27.6</td>
<td>31.3</td>
<td>36.9</td>
</tr>
<tr>
<td>3rd Trimester</td>
<td>28.1</td>
<td>24.9</td>
<td>27.3</td>
<td>31.0</td>
<td>36.6</td>
</tr>
<tr>
<td>9-month period</td>
<td>28.1</td>
<td>25.8</td>
<td>27.8</td>
<td>30.1</td>
<td>33.5</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Term LBW</td>
<td>OR (95%CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CASE (N=2,058)</td>
<td>Control (N=90,389)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20</td>
<td>307 (14.9)</td>
<td>7,620 (8.4)</td>
<td>1.69 (1.49-1.93)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-29*</td>
<td>949 (46.1)</td>
<td>39,907 (44.1)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-34</td>
<td>463 (22.5)</td>
<td>27,628 (30.6)</td>
<td>0.71 (0.63-0.79)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35-39</td>
<td>282 (13.7)</td>
<td>13,002 (14.4)</td>
<td>0.91 (0.80-1.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥40</td>
<td>57 (2.8)</td>
<td>2,232 (2.5)</td>
<td>1.07 (0.82-1.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White*</td>
<td>1,250 (60.9)</td>
<td>72,087 (79.8)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>731 (35.6)</td>
<td>15,808 (17.5)</td>
<td>2.67 (2.43-2.93)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>73 (3.5)</td>
<td>2,392 (2.7)</td>
<td>1.76 (1.38-2.24)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;9</td>
<td>43 (2.1)</td>
<td>641 (0.7)</td>
<td>4.51 (3.27-6.22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9-11</td>
<td>313 (15.2)</td>
<td>6,662 (7.4)</td>
<td>3.16 (2.74-3.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>761 (37.0)</td>
<td>27,512 (30.4)</td>
<td>1.86 (1.66-2.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13-15</td>
<td>425 (20.6)</td>
<td>20,884 (23.1)</td>
<td>1.37 (1.20-1.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥16 *</td>
<td>516 (25.1)</td>
<td>34,690 (38.4)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal tobacco use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>879 (43.1)</td>
<td>16,103 (17.9)</td>
<td>3.48 (3.18-3.80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1,160 (56.9)</td>
<td>73892 (82.1)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infant sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>814 (39.6)</td>
<td>46,407 (51.3)</td>
<td>0.62 (0.57-0.68)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female *</td>
<td>1,244 (60.45)</td>
<td>43,982 (48.7)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second or subsequent birth</td>
<td>1,005 (48.9)</td>
<td>52,856 (58.5)</td>
<td>0.68 (0.62-0.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First birth *</td>
<td>1,049 (51.1)</td>
<td>37,465 (41.5)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer (Jun-Aug)*</td>
<td>534 (25.9)</td>
<td>23,747 (26.3)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autumn (Sep-Nov)</td>
<td>541 (26.3)</td>
<td>22,155 (24.5)</td>
<td>1.09 (0.96-1.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter (Dec-Feb)</td>
<td>503 (24.4)</td>
<td>21,128 (23.4)</td>
<td>1.06 (0.94-1.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring (Mar-May)</td>
<td>480 (23.3)</td>
<td>23,359 (25.8)</td>
<td>0.91 (0.81-1.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous LBW or PTD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>53 (2.6)</td>
<td>364 (0.4)</td>
<td>6.56 (4.90-8.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1,998 (97.4)</td>
<td>89,931 (99.6)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prenatal care</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>45 (2.2)</td>
<td>517 (0.6)</td>
<td>3.95 (2.91-5.38)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st trimester *</td>
<td>1,945 (97.3)</td>
<td>88,347 (99.1)</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>After 1st trimester</td>
<td>10 (0.5)</td>
<td>235 (0.3)</td>
<td>1.93 (1.02-3.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gestational age</td>
<td>Mean±SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean±SD</td>
<td>38.2±1.1</td>
<td>39.4±1.2</td>
<td>0.40 (0.39-0.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal weight gain</td>
<td>Mean±SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean±SD</td>
<td>26.4±11.9</td>
<td>31.0±12.2</td>
<td>0.968 (0.96-0.97)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Reference group
** Number of missing for each variable: maternal race (106); maternal tobacco use (413); Parity (72); Previous LBW or PTD (101); Prenatal care (1348)
Table 3-4 Risks for term low birth weight * according to trimester-specific exposure to PM$_{10}$

<table>
<thead>
<tr>
<th>Exposure period</th>
<th>OR (95%CI) *</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
</tr>
<tr>
<td>1st Trimester</td>
<td>1.13$^b$</td>
</tr>
<tr>
<td>2nd Trimester</td>
<td>1.10</td>
</tr>
<tr>
<td>3rd Trimester</td>
<td>1.05</td>
</tr>
<tr>
<td>9-month period</td>
<td>1.07</td>
</tr>
</tbody>
</table>

* ORs were estimated by per inter-quartile range increase (per 7 µg/m$^3$ for trimester-specific exposure and 4.3 µg/m$^3$ for 9-month period) after adjustment for maternal age, maternal race, maternal education, smoking, weight gain, gender of infant, gestation age, parity, previous LBW or preterm birth, level of prenatal care and birth season

$^a$ All singleton births with gestational age $\geq$37 weeks

$^b$ P<0.05
3.6 REFERENCES


4.0 CIGARETTE SMOKING AND THE RISK OF ADULT LEUKEMIA: RESULTS FROM THE THREE MILE ISLAND COHORT STUDY

Xiaohui Xu, M.P.H.

Evelyn O. Talbott, Dr.P.H., M.P.H.

Jeanne V. Zborowski, M.S., Ph.D.

Judy Rager, M.P.H.

Department of Epidemiology
University of Pittsburgh Graduate School of Public Health

Manuscript in preparation
4.1 ABSTRACT

Smoking is an unconfirmed risk factor for the development of leukemia. The potential link was examined using data from the TMI cohort for the period of 1979 to 1995. A total of 24,539 subjects followed up over 16 years from the TMI cohort who were 14 years or older, were eligible for this analysis. All incident leukemia cases were identified through the Pennsylvania Department of Health Cancer Registry. The Cox proportional hazards model was employed to evaluate the relationships. A total of 42 incident leukemia cases, including 15 AML cases, were observed in the cohort. After controlling for other confounding factors, current smoking was associated with an increased risk of adult AML (RR=3.47; 95% CI: 1.002-11.99). A marginally significant linear trend of risk of AML associated with the number of years smoked was also observed (p=0.06). The results from this study suggested that cigarette smoking was associated with an increased risk of adult AML. Further investigation is required to confirm these findings.

Key words: Leukemia, AML, Smoking, Cohort Study.
4.2 INTRODUCTION

Benzene, a well-established carcinogen and leukemogen, has typically been associated with acute myeloid leukemia (AML) over the past century\(^1,2\) and has been shown to be present in cigarette smoke. Wallace observed that approximately 90\% of personal exposure to benzene in the United States is due to smoking. Smokers have an average benzene body burden approximately 6-10 times that of nonsmokers\(^3\). Other carcinogens including ionizing radiation (polonium and lead), nitrosamines, styrene, naphthalene and urethanes are also found in cigarette smoking\(^4,5\). Cigarette smoking has long been suspected as an etiologic factor in leukemogenesis\(^6\); an association between cigarette smoking and leukemia was reported as early as 1978\(^7\). Epidemiological studies to date have provided inconsistent results regarding the association between cigarette smoking and leukemia. Several case-control studies demonstrated an increased risk of leukemia among cigarette smokers\(^8-13\). However, other studies have observed no significant association\(^14,15\).

Only a few cohort studies have investigated the smoking-leukemia link because of the latency period associated with the development of hematological malignancies and the large sample size required. The conclusions from available prospective studies are inconsistent and some large cohort studies were unable to detect any excess risk of leukemia related to cigarette smoking. Adami et al reported that there was no significant association between smoking status, number of cigarettes smoked per day or duration of smoking and the risk of developing leukemias in 334,957 Swedish construction workers after following up 10 years\(^16\). Although the study is large, all participants are male construction workers. The healthy worker effect cannot be negligible in this study. Engeland et al also found no association between smoking and leukemia in 28 year follow-up of 26,000 Norwegian men and women\(^17\). However, this study
failed to elucidate subtype association and the dose-response relationship of cigarette smoking. Moreover, this study was unable to control for some confounding factors such occupational radiation exposures. In summary, few cohort studies have been carried out to assess the association between cigarette smoking and subtypes of leukemia such as AML, which is more related to the exposure of benzene.

While published evidence for leukemia and cigarette smoking is inconclusive, several case-control studies suggest a weak association between AML and tobacco use\textsuperscript{9,18,19}. The subtypes of leukemia vary in histological, molecular and clinical characteristics as well as in prognosis. The emphasis of etiologic studies on histological subtypes must be taken into consideration.

To further explore the role of cigarette smoking in the development of adult leukemia, we analyzed the risk of total leukemia and specifically acute myeloid leukemia (AML) as well as the dose-response relationship between cigarette smoking and adult leukemia in the large Three Mile Island (TMI) (Pennsylvania, U.S.A) cohort, which provided detailed information on tobacco use at baseline and the incidence of cancer with long-term follow-up.

### 4.3 METHODS

#### 4.3.1 Study Population

The TMI cohort was assembled by the Pennsylvania Department of Health (PADoH) to evaluate the adverse health effects of exposure of low-level radiation emitted from the TMI nuclear power plant accident on 28 March 1979. The cohort consisted of 32,135 individuals in the 1979 TMI census\textsuperscript{18} who were followed up from 1979 to 1995. The estimated average likely
and maximum gamma doses from the accident were 0.09 mSv (9 mrem) and 0.25 mSv (25 mrem) respectively. These exposures were therefore considered overall minimal (less than the dose of 2-3 chest X-rays). Previous mortality studies in these subjects also found no significant evidence of the association between the level of accident radiation exposure and leukemia mortality risk \(^{18,19}\). In this study, all analyses presented are based on cohort members aged 14 years or older (n= 24,539). Data collected included individual information at baseline on smoking status, demographic characteristics (age, gender and race etc), residential history, background radiation exposure, previous occupational or treatment radiation exposure, and estimated radiation exposure from the TMI nuclear power plant accident.

### 4.3.2 Smoking Information

Data on personal smoking history was collected in 1979 at baseline. Cigarette smoking in individuals was assessed as never smoked, former smoker and current smoker. Former smoker was defined as ever smoked at least 100 cigarettes. For those who ever or currently smoked, The number of cigarettes smoked per day as well as total number of years smoked was also obtained. Based on this information, a new variable of “pack-year” was created by multiplying total number of years smoked with number of cigarettes smoked per day then divided by 20.

### 4.3.3 Total Leukemia and AML Subtype Identification

The TMI cohort was followed up from March 1979 through December 1995 to determine vital status and cancer incidence. All cases of adult-onset leukemia (International Classification of Diseases for Oncology, Second Edition, and Code 980-994) and AML (ICD-O-2, code 9861,
9864, 9866, 9867 and 9891) diagnosed between March 1979 through December 1995 were identified through the Pennsylvania Cancer Registry (PCR), which is complete for those individuals who remain within Pennsylvania. Not Otherwise Specified (NOS) leukemia cases (ICD-O-2: 9800 and 9801) are further clarified as AML when subject died from AML (International Classification of Diseases, version 9: 205.0) in the National Death Index. The PADoH, along with the U.S. Postal Service, annually obtained the current addresses of persons to maintain follow-up of the cohort. The address confirmation was updated through December 1995. We verified that 92.1% of subjects remained in Pennsylvania during the 1979-1995 follow-up period.

4.3.4 Statistical analysis

Comparison of the distribution of demographic and other characteristics among each smoking status category was made using ANOVA for continuous variables and the Chi square test for categorical variables respectively. Total leukemia and AML were analyzed separately. The Cox proportional hazards model was employed to estimate Relative Risk (RR) with 95% Confidence Intervals (CI) for the association of adult leukemia and its subtype AML with cigarette smoking after controlling for potential confounding factors. Smoking status was assessed as a dummy variable in the model. The number of cigarettes smoked per day, number of years smoked and number of pack-years were also fit as ordinal categorical variables to evaluate the dose-response relationship between cigarette smoking and total leukemia as well as AML. All final models were adjusted for age, gender, race, education, background radiation exposure (<8 μR/hr vs. ≥8 μR/hr), occupational exposure, and estimated maxima gamma radiation and likely gamma radiation.
Subjects with missing values were excluded from relevant analyses. The statistical tests were two-sided with a 0.05 significant level. The SAS 8.02 software package (SAS Institute Inc., 100 SAS Campus Drive Cary, NC 27513) was used to analyze the collected data.

4.4 RESULTS

Within the study population, 33% (n=8,083) of the individuals were current smokers, 18% (n=4,330) were ever smokers and 49% (11,913) were non-smokers at baseline. More than 95 percent of subjects in the cohort are white. Table 4-1 shows the distribution of baseline characteristics among each group by smoking status. The results of the ANOVA and \( \chi^2 \) tests indicated that age, education level, race, occupational radiation exposure were significantly different among smoking groups. The percentages of female are 60% among never smokers, 36% among former smoker and 46% among current smokers respectively. There are 5.4% participants among never smokers, 7.4% among former smokers and 6.8% among current smokers who had experience of occupational radiation. The estimated Maximum and likely gamma exposure emitted from the TMI nuclear power plant accident was slightly higher among nonsmokers than smokers. However, the distributions of background radiation exposure were similar among each smoking group.

Table 4-2 describes the distributions of lymphatic and hematopoietic neoplasm incident cases occurred during study period. A total of 42 leukemia cases including 15 AML cases were observed within follow-up time period in the cohort.
The distributions of total leukemia among each status of cigarette smoking and risks of total leukemia associated with cigarette smoking are presented in table 4-3. We found higher proportion of leukemia cases among the current smoker group as compared with those who had never smoked (table 4-3). Of these, the cumulative incidences of total leukemia are 0.21% in current smokers, 0.14% in former smokers and 0.16% in non-smokers respectively.

To examine the association between dose of cigarette smoking and total leukemia, the age and sex adjusted model showed that a positive association between current cigarette smoking and total leukemia was suggestive but not statistically significant by using a dummy variable of smoking status (RR=1.49, 95% CI: 0.76-2.96 in current smoker). After controlling for other risk factors in the fully adjusted model, current cigarette smoking is still positively but not statistically significantly associated with total leukemia (RR=1.39, 95%CI: 0.69-2.82). Furthermore, a more powerful analysis using the number of cigarettes smoked per day, the number of years smoked and pack years of smoking was performed to provide evidence of an association between cigarette smoking and total leukemia. The results showed that there is no evidence of significant positive associations between cigarette smoking and total leukemia either in the age-sex adjusted model or in the fully adjusted model.

Table 4-4 reported the results of the distribution of the subtype of leukemia, i.e. AML among each cigarette smoking group and risk of AML associated with cigarette smoking. Only 15 AML cases were observed within follow-up time period in the cohort. The age-sex adjusted model indicated a significantly positive association between current cigarette smoking and AML using a dummy variable of smoke status (AML: RR=3.26, 95% CI: 1.05-10.17 in current smoker). After controlling for other risk factors in the fully adjusted model, the association between current cigarette smoking and AML is still significant (RR=3.47, 95% CI: 1.002-11.99).
An association between cigarette smoking and AML was further explored with a more powerful analysis using the number of cigarettes smoked per day, the number of years smoked and pack years of smoking. The age-sex adjusted models showed that those who smoked 11-20 pack years have a higher risk of AML as comparing with non-smokers (RR=4.25, 95%CI: 1.11-16.35) after controlling for other risk factors. In addition, there is a significant dose-response relationship between the number years of cigarette smoking and risk of AML (p=0.04) (see Fig 1). However, there is no evidence for a linear trend of risk of AML associated with the number of cigarettes per day as well as the pack years of smoking. We further analyzed the risk of AML associated with cigarette smoking after controlling for some other potential confounders, including age, gender, education level, race, occupational radiation, background radiation, estimated maximum and likely gamma exposures. In these analyses, we found a statistically increased risk of AML among those who smoked 1-10 cigarettes per day or 11-20 pack years (RR=5.02, 95% CI: 1.11-22.64; RR=5.07, 95% CI: 1.22-21.01). More interestingly, a dose-response relationship between the number year of smoking and risk of AML was marginally statistically significant (p=0.06). However, we found no evidence of any dose-response relationship between the number cigarettes per day or pack years of smoking and risk of AML.

4.5 DISCUSSION

In this large cohort study, 42 incident leukemia cases occurred during the study period. The Not Otherwise Specified (NOS) leukemia was further determined as its subtypes by scrutinizing PA Cancer Registry for cell type and death certificates for underlying cause of death (UCOD). Five NOS leukemia cases were determined as AML (UCOD: 205.0). A total of 15 AML cases
out of 42 leukemia cases were found in this study. The percentage of AML in leukemia is 35.7%, which is close to general population. The analyses of this study emphasized on AML because it had the largest number of cases in leukemia subtypes and epidemiological evidence suggested that AML is more likely linked to cigarette smoking.

In this study, the age-sex adjusted model and the fully adjusted model were applied to assess the associations between leukemia as well as AML and cigarette smoking. The estimations from these two models are very similar. It suggested that age and sex are the factors which confound the associations between smoking and leukemia as well as AML. In this study, neither the age-sex adjusted models nor the fully adjusted models could provide any evidence of a significantly positive association between cigarette smoking and total leukemia after adjusting for other confounders. Indeed, significant association between current cigarette smoking and AML is demonstrated after controlling for other risk factors. It also suggested that there is a significant dose-response relationship between the number year of cigarette smoking and risk of AML (trend \(p=0.04\)) after adjusted for age and sex. This linear relationship between the number years of cigarette smoking and risk of AML are still marginally significant (\(p=0.06\)) after controlling for other risk factors. However, there are no statistically significant linear relationships between AML and cigarettes smoked per day as well as pack years. Cigarettes smoked per day, number years of smoked, and pack years have their own limitations and strengths when they are used to estimate the dose of cigarette smoking. Cigarettes smoked per day have information on daily consumption but does not include information on duration of smoking. Number years of smoked emphasizes information on duration of smoking but neglects daily consumption. Pack years is a comprehensive indicator that combines information on daily consumption with duration of smoking. But inaccurate recalls in either cigarettes smoked per day or number years of smoked
could result into a misclassification of pack years. In this study, it could be possible that the recall of years smoked was more accurate than number of cigarettes smoked per day.

The findings of the present study are in agreement with the results from some other studies. The cohort study of 34,000 Seventh-Day Adventists demonstrated that a higher risk of myeloid leukemia was associated with cigarette smoking (RR=2.24, 95% CI 0.91-5.53)\(^{22}\). Several case-control studies reported an increased risk of AML among cigarette smokers\(^{6,8,9,13,15,23}\). Moreover, the childhood leukemia studies also provided significant evidence of an association between parental smoking and childhood leukemia. Chang et al found that parental preconception smoking was significantly associated with an increased risk of AML (OR=3.84, 95%:1.04-14.17)\(^{10}\). Ji et al reported that paternal preconception smoking was related to a significantly elevated risk of childhood cancers, particularly acute leukemia and lymphoma\(^ {24}\). These data are consistent with a possible mechanism linking cigarette smoking to an increased risk of leukemia in human subject.

On the other hand, a few studies reported that the risk of leukemia induced by cigarette smoking decreases or ceases after smoking cessation. Kane et al showed that the odds ratio was decreased as the number of years 'stopped smoking' increased, falling to 1.0 amongst those who had given up smoking for more than 10 years\(^ {25}\). Bjork et al found that the risk of AML declined in 5-10 years after smoking cessation among subjects with smoking history long enough to give a substantially elevated AML risk\(^ {1}\). These studies provided supporting evidence of an association between AML and cigarette smoking.

The number of cigarettes per day, the number of years smoked and pack years of cigarette smoking have been previously used to evaluate the dose-response relationship for leukemia and AML. Our study found no evidence of a trend in risk with number of cigarettes smoked per day.
and pack years, which was similar to the majority of studies\textsuperscript{9,14,26}. A number of studies reported that there were significant increasing trends in risk with increasing number of years smoked\textsuperscript{13,22}. These findings were consistent with the results from our study, which provided additional support for an association between cigarette smoking and leukemia.

The mechanism by which cigarette smoking increases risk of leukemia or AML are unknown. Cigarette smoke contained some known or suspected leukemogens, including benzene\textsuperscript{3}, ionizing radiation\textsuperscript{5} and other carcinogens (nitrosamines, styrene, naphthalene and urethanes)\textsuperscript{4,27}. It was reported that the level of trans, trans-muconic acid (t,t-MA), a urinary benzene metabolite, were about 3 times higher than those in nonsmokers\textsuperscript{28}. Other studies found higher level of polonium and lead in tissues from smokers\textsuperscript{29}. In addition, chromosomal defects in the peripheral blood were also observed to be increased in smokers\textsuperscript{30,31}. All these findings suggested a possible leukemogenic effect of cigarette smoking.

There are several strengths to the study. First, the study was considered noteworthy, as it included a large sample size of cohort and a relatively long period of follow-up, i.e. over 16 years complete follow-up. Second, the rate of loss of follow-up is low in this study. Ninety three percent of participants in the cohort were verified for leukemia outcome using the population-based Pennsylvania Cancer Registry system. Third, we assessed the risk of AML among cigarette smokers and further evaluated the dose-response relationship between cigarette smoking and outcomes, as this form of leukemia has more specifically been linked to benzene exposure in occupational and community setting. In addition, personal information about natural background radiation, TMI accidental exposure and previous occupational radiation was collected in the study and was able to be included in the final statistical models.
One of the limitations of our study lies in the fact that the baseline cigarette smoking information was only able to obtain in 1979. Any changes in smoking status during the follow-up (1979-1995) were not available. This could result into some misclassification bias in the analyses based on duration of smoking and pack-years of smoking. Second, an additional 12 years of follow up would be very beneficial and improve the power of the study if the cohort could be followed up until now. However, formal follow up of the cohort ceased in 1995 when the PA Department of Health ceased funding formal follow up through post offices of the cohort. Thus, there would no longer be complete ascertainment of the denominator of the population. Third, the analysis was based on leukemia cases identified through the Pennsylvania Cancer Registry (PCR), 1979-1995. Only individuals who remained in Pennsylvania were included in the case ascertainment. In this study, 1,960 cases (7.9%) left the state and lost to follow-up during the study period. The cigarette smoking status and mortality experience were assessed among these individuals of loss follow-up. The distributions of smoking status among these cases are 34% current smokers, 17% former smokers and 49% non-smokers, which is similar to the distributions of the cohort. Moreover, 179 deaths among 7.9% loss follow-up were identified and only one case that died from myeloid leukemia in current smoking group by scrutinizing death certificate for diagnostic code. This result suggested that the study could potentially underestimate the association between cigarette smoking and leukemia. Fourth, although there was a large cohort in this study, only a few new AML cases (n=15) were identified during the follow-up time period because of the low incidence of AML, i.e. 1 to 15 cases per 100,000 persons in the general population\textsuperscript{32}. It resulted into a wide range of 95% Confidence Interval of hazard ratios. However, we had statistical power to detect the risk difference among cigarette smoking groups. Finally, information regarding other potential risk factors of leukemia, such as
diet and medicine, was not obtained in the study. Thus, their potential confounding effects could not be ruled out.

In conclusion, the present study suggests that cigarette smoking is associated with AML. Moreover, the risk of AML was observed to be increased with the increasing number of years smoked. Further well-designed studies are needed to support the establishment of causal association between cigarette smoking and AML.
Table 4-1 Distribution of baseline characteristics among smoking groups

<table>
<thead>
<tr>
<th></th>
<th>Never smoked (n=11,913)</th>
<th>Former smoker (n=4,330)</th>
<th>Current smoker (n=8,083)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (X ±SD)</td>
<td>40.60 ± 20.73</td>
<td>46.55 ± 17.33</td>
<td>38.43 ± 15.52</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male (%)</td>
<td>4,746 (39.84)</td>
<td>2,770 (63.97)</td>
<td>4,382 (54.21)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Female (%)</td>
<td>7,167 (60.16)</td>
<td>1,560 (36.03)</td>
<td>3,701 (45.79)</td>
<td></td>
</tr>
<tr>
<td>Education level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤9 yr (%)</td>
<td>2,019 (17.39)</td>
<td>795 (18.51)</td>
<td>1,233 (15.42)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>10-12 yr (%)</td>
<td>7,037 (60.62)</td>
<td>2,587 (60.25)</td>
<td>5,615 (70.24)</td>
<td></td>
</tr>
<tr>
<td>&gt;12 yr (%)</td>
<td>2,553 (21.99)</td>
<td>912 (21.24)</td>
<td>1,146 (14.34)</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>11,621 (97.55)</td>
<td>4,227 (97.62)</td>
<td>7,836 (96.94)</td>
<td>0.016</td>
</tr>
<tr>
<td>Non-white</td>
<td>292 (2.45)</td>
<td>103 (2.38)</td>
<td>247 (3.06)</td>
<td></td>
</tr>
<tr>
<td>Occupational</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>radiation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>642 (5.39)</td>
<td>321 (7.41)</td>
<td>555 (6.87)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No</td>
<td>11,271 (94.61)</td>
<td>4,009 (92.59)</td>
<td>7,528 (93.13)</td>
<td></td>
</tr>
<tr>
<td>Background radiation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;8 mrem</td>
<td>9,039 (76.12)</td>
<td>3,331 (77.11)</td>
<td>6,191 (76.85)</td>
<td>0.30</td>
</tr>
<tr>
<td>≥8 mrem</td>
<td>2,836 (23.88)</td>
<td>989 (22.89)</td>
<td>1,865 (23.15)</td>
<td></td>
</tr>
<tr>
<td>Maximum gamma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exposure (X ±SD)</td>
<td>25.59±22.08</td>
<td>24.70±21.85</td>
<td>24.67±21.75</td>
<td>0.005</td>
</tr>
<tr>
<td>Likely gamma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exposure (X ±SD)</td>
<td>10.93±10.40</td>
<td>10.61±10.14</td>
<td>10.58±10.25</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Note: χ² test for the categorical variables and ANOVA test for the continuous variables
<table>
<thead>
<tr>
<th>Types of Cancer</th>
<th>Number</th>
<th>ICD-O-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>42</td>
<td>980-994</td>
</tr>
<tr>
<td>Acute myeloid leukemia (AML)*</td>
<td>15</td>
<td>9861, 9864, 9866, 9867 and 9891</td>
</tr>
<tr>
<td>Chronic lymphocytic leukemia (CLL)</td>
<td>11</td>
<td>9823</td>
</tr>
<tr>
<td>Chronic myeloid leukemia (CML)</td>
<td>7</td>
<td>9863</td>
</tr>
<tr>
<td>Other leukemia</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td>12</td>
<td>9732</td>
</tr>
<tr>
<td>Other Lymphatic Neoplasm</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
<td>C42.1</td>
</tr>
</tbody>
</table>

*Leukemia NOS (ICD-O-2: 9800, 9801) cases are clarified with death index (ICD-9:205.0)
<table>
<thead>
<tr>
<th>Smoking status</th>
<th>No.</th>
<th>Leukemia</th>
<th>Adjusted RR&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Adjusted RR&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
<td>(95% CI)</td>
<td>(95% CI)</td>
</tr>
<tr>
<td>Never smoked (%)</td>
<td>11,913</td>
<td>19</td>
<td>0.16</td>
<td>1.0</td>
</tr>
<tr>
<td>Former smoker (%)</td>
<td>4,330</td>
<td>6</td>
<td>0.14</td>
<td>0.60 (0.23-1.54)</td>
</tr>
<tr>
<td>Current smoker (%)</td>
<td>8,083</td>
<td>17</td>
<td>0.21</td>
<td>1.49 (0.76-2.96)</td>
</tr>
<tr>
<td>Number cigarettes per day</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>11,913</td>
<td>19</td>
<td>0.16</td>
<td>1.0</td>
</tr>
<tr>
<td>&lt;10</td>
<td>1,749</td>
<td>4</td>
<td>0.23</td>
<td>1.37 (0.46-4.05)</td>
</tr>
<tr>
<td>10-19</td>
<td>2,843</td>
<td>7</td>
<td>0.25</td>
<td>1.61 (0.67-3.86)</td>
</tr>
<tr>
<td>&gt;=20</td>
<td>7,821</td>
<td>12</td>
<td>0.15</td>
<td>0.84 (0.39-1.78)</td>
</tr>
<tr>
<td>P value for trend test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.79</td>
<td>0.57</td>
</tr>
<tr>
<td>Number of years smoked</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>11,913</td>
<td>19</td>
<td>0.16</td>
<td>1.0</td>
</tr>
<tr>
<td>1-19 yrs</td>
<td>7,608</td>
<td>5</td>
<td>0.07</td>
<td>0.55 (0.20-1.52)</td>
</tr>
<tr>
<td>20-29 yrs</td>
<td>2,015</td>
<td>8</td>
<td>0.40</td>
<td>1.76 (0.76-4.10)</td>
</tr>
<tr>
<td>&gt;=30 yrs</td>
<td>2,790</td>
<td>10</td>
<td>0.36</td>
<td>1.29 (0.58-2.88)</td>
</tr>
<tr>
<td>P value for trend test</td>
<td></td>
<td></td>
<td>0.33</td>
<td>0.53</td>
</tr>
<tr>
<td>Pack-Year *</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>11,913</td>
<td>19</td>
<td>0.16</td>
<td>1.0</td>
</tr>
<tr>
<td>≤10</td>
<td>5,897</td>
<td>7</td>
<td>0.12</td>
<td>0.998 (0.41-2.43)</td>
</tr>
<tr>
<td>11-20</td>
<td>2,358</td>
<td>7</td>
<td>0.30</td>
<td>1.74 (0.72-4.22)</td>
</tr>
<tr>
<td>&gt;20</td>
<td>4,158</td>
<td>9</td>
<td>0.22</td>
<td>0.88(0.39-2.0)</td>
</tr>
<tr>
<td>P value for trend test</td>
<td></td>
<td></td>
<td>0.98</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Note: *“Pack-year” is 20 cigarettes per day for one year.

1. “Adjusted RRs” were derived from Cox proportional hazards model after controlling for age and gender.
2. “Adjusted RRs” were derived from Cox proportional hazards model, adjusting for age, gender, education level, race, occupational radiation, background radiation, estimated maximum and likely gamma exposures.
Table 4-4 Distribution and risk of AML among each status of cigarette smoking

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>Total No.</th>
<th>AML No.</th>
<th>AML %</th>
<th>Adjusted RR (^1) (95% CI)</th>
<th>Adjusted RR (^2) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never smoked (%)</td>
<td>11,913</td>
<td>5</td>
<td>0.04</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Former smoker (%)</td>
<td>4,330</td>
<td>1</td>
<td>0.02</td>
<td>0.48 (0.05-4.20)</td>
<td>0.54 (0.06-4.98)</td>
</tr>
<tr>
<td>Current smoker (%)</td>
<td>8,083</td>
<td>9</td>
<td>0.11</td>
<td>3.26 (1.05-10.17)</td>
<td>3.47 (1.002-11.99)</td>
</tr>
</tbody>
</table>

Number of cigarettes per day

<table>
<thead>
<tr>
<th>Number cigarettes per day</th>
<th>Total No.</th>
<th>AML No.</th>
<th>AML %</th>
<th>Adjusted RR (^1) (95% CI)</th>
<th>Adjusted RR (^2) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11,913</td>
<td>5</td>
<td>0.04</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>&lt;10</td>
<td>1,749</td>
<td>3</td>
<td>0.17</td>
<td>4.13 (0.98-17.42)</td>
<td>5.02 (1.11-22.64)</td>
</tr>
<tr>
<td>10-19</td>
<td>2,843</td>
<td>3</td>
<td>0.11</td>
<td>2.77 (0.65-11.75)</td>
<td>3.31 (0.73-14.99)</td>
</tr>
<tr>
<td>&gt;=20</td>
<td>7,821</td>
<td>4</td>
<td>0.05</td>
<td>1.26 (0.32-4.90)</td>
<td>1.07 (0.23-5.01)</td>
</tr>
</tbody>
</table>

P value for trend test 0.67 0.82

Number years of smoking

<table>
<thead>
<tr>
<th>Number years of smoking</th>
<th>Total No.</th>
<th>AML No.</th>
<th>AML %</th>
<th>Adjusted RR (^1) (95% CI)</th>
<th>Adjusted RR (^2) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11,913</td>
<td>5</td>
<td>0.04</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1-19 yrs</td>
<td>7,608</td>
<td>2</td>
<td>0.03</td>
<td>0.79 (0.15-4.27)</td>
<td>0.91 (0.16-5.19)</td>
</tr>
<tr>
<td>20-29 yrs</td>
<td>2,015</td>
<td>3</td>
<td>0.15</td>
<td>3.03 (0.71-12.98)</td>
<td>3.79 (0.82-17.61)</td>
</tr>
<tr>
<td>&gt;=30 yrs</td>
<td>2,790</td>
<td>5</td>
<td>0.18</td>
<td>3.58 (0.95-13.50)</td>
<td>3.64 (0.80-16.49)</td>
</tr>
</tbody>
</table>

P value for trend test 0.04 0.06

Pack-Year *

<table>
<thead>
<tr>
<th>Pack-Year</th>
<th>Total No.</th>
<th>AML No.</th>
<th>AML %</th>
<th>Adjusted RR (^1) (95% CI)</th>
<th>Adjusted RR (^2) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11,913</td>
<td>5</td>
<td>0.04</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1-10 yrs</td>
<td>5,897</td>
<td>3</td>
<td>0.05</td>
<td>1.67 (0.38-7.27)</td>
<td>1.95 (0.42-9.03)</td>
</tr>
<tr>
<td>11-20 yrs</td>
<td>2,358</td>
<td>4</td>
<td>0.17</td>
<td>4.25 (1.11-16.35)</td>
<td>5.07 (1.22-21.01)</td>
</tr>
<tr>
<td>&gt;=20 yrs</td>
<td>4,158</td>
<td>3</td>
<td>0.07</td>
<td>1.44 (0.33-6.30)</td>
<td>1.11 (0.19-6.36)</td>
</tr>
</tbody>
</table>

P value for trend test 0.32 0.45

Note: * "Pack-year" is 20 cigarettes per day for one year.
1. "Adjusted RRs" were derived from Cox proportional hazards model after controlling for age and gender.
2. "Adjusted RRs" were derived from Cox proportional hazards model, adjusting for age, gender, education level, race, occupational radiation, background radiation, estimated maximum and likely gamma exposures.
Figure 4-1 the association between risk of AML and the number years of smoking.
4.6 REFERENCES


5.0 SUMMARY AND CONCLUSIONS

The project has been developed to explore the associations between environmental exposures and chronic diseases based on existing data. It has been segmented into three complementary topics as following:

1. Risk of cardiopulmonary diseases with exposure to air pollution
2. Risks of term LBW with exposure to particulate air pollution
3. Risk of leukemia and cigarette smoking

5.1 ASSOCIATION OF RISK OF CARDIOPULMONARY DISEASE AND AIR POLLUTION

A link between respiratory disease and air pollution has been well established in earlier air pollution studies (Holland et al. 1979; Logan 1953; Logan 1956). Recent short-term and long-term epidemiological studies across the world further showed that air pollution exposure is associated with increase risks of cardiopulmonary mortality and morbidity (Daggett et al. 2000; Dockery et al. 1993; Gehring et al. 2006; Katsouyanni et al. 1997; Krewski et al. 2005; Kunzli and Tager 2000; Nyberg and Pershagen 2000; Pope et al. 1995; Pope et al. 2002; Rosenlund et al. 2006; Wietlisbach et al. 1996). Pope et al found that each 10 mg/m^3 elevation in fine particulate air pollution was associated with approximately a 6% increased risk of
cardiopulmonary mortality (Pope et al. 2002). Cohen and his colleagues conducted a risk assessment of the global burden of disease attributable to urban ambient air pollution and showed that ambient air pollution, in terms of fine particulate air pollution, causes about 3% mortality from cardiopulmonary disease and about 1% mortality from acute respiratory infections in children under 5 years worldwide (Cohen et al. 2005).

In this study, we also found that exposure to particulate matter is associated with an increased risk of cardiopulmonary diseases before the LTV coke plant closure, which is consistent to the findings of previous studies.

Studies of cigarette smoking found that the cessation of smoking could result into reduction of health risks (Andrews and Tingen 2006; Bosetti et al. 2006; Krall et al. 2006). However, few air pollution studies have been conducted to assess how elimination or addition of a major point source of air pollution in an area affects the health among people who live there. In this study, we found that the elimination of a major point source of air pollution could reduce health risks in the area. This reduction of risk could possibly be explained by the declination of air pollution concentration and the change of components of particulate matter after the removal of the major point source of air pollution.

The studies across the world provided consistent and coherent toxicological and epidemiological evidence indicating that air pollution exposure is associated with cardiopulmonary disease. Further study might be focused on elucidating the biological mechanism, which is not yet clear.
5.2 ASSOCIATION OF RISK OF TERM LBW AND PARTICULATE AIR POLLUTION

Study of particulate matter on fetal health is an emerging area. Fetuses, a susceptible subgroup of population, are thought to be vulnerable to the effects of air pollution. Birth weight is one of the widely-used adverse pregnancy outcomes in fetal health air pollution studies. The findings of previous studies are not consistent and provide a mixed picture about the association between particulate matter and birth weight. Some studies reported a negative association over some exposure period (Ha et al. 2001; Lee et al. 2003; Mannes et al. 2005; Parker et al. 2005; Rogers and Dunlop 2006; Wang et al. 1997; Wilhelm and Ritz 2005; Yang et al. 2003) and others reported no evidence of an association between particulate matter and low birth weight (Chen et al. 2002; Dugandzic et al. 2006; Maisonet et al. 2001).

Birth weight is sensitive to many factors such as gestational age, maternal age, race, education, maternal weight gain during pregnancy, prenatal care level, smoking, infant sex, parity and others. Without adequate control for confounders, the association between low-level particulate matter and low birth weight might be difficult to quantify. In this study, we found that the odds ratios of term LBW per inter-quartile range increase in PM$_{10}$ were 1.13 (95%CI: 1.02-1.25) during the first trimester and 1.10 (95%CI: 1.00-1.22) during the second trimester after adjusting for other important covariates, respectively.

Further studies with more refined methodological designs such as high-quality exposure data, advanced methods of exposure assessment, and high-quality covariate information, are needed to clarify the adverse effect of particulate air pollution on fetal health.
Cigarette smoking is associated with many health outcomes such as chronic obstruction pulmonary disease (Davis and Novotny 1989; Kalucka 2006; Lee et al. 1990), cardiovascular disease (Cook et al. 1986; Lapidus et al. 1986; Milei et al. 1986), lung cancer (Benhamou et al. 1985; Kestner 1985; La Vecchia 1985), nasopharyngeal cancer (Mabuchi et al. 1985), bladder cancer (Marrett et al. 1985; Rebelakos et al. 1985), and adverse pregnancy outcomes (Weisberg 1985). Cigarette smoking has not classically been linked with leukemia until a few epidemiological studies reported an increased risk of leukemia among cigarette smokers (Brownson 1989; McLaughlin et al. 1989; Severson et al. 1990; Spitz et al. 1990). Several case-control studies suggested an association that is most pronounced for acute myeloid leukemia (AML) and provided evidence of a dose-response relationship between cigarette smoking and AML (Chelghoum et al. 2002; Kasim et al. 2005; Sandler et al. 1993).

In this study, we used a population based cohort to evaluate the association between leukemia as well as AML and cigarette smoking. The results showed that cigarette smoking is significantly associated with AML after controlling for other important risk factors. In addition, a dose-response relationship between cigarette smoking and AML, which represents additional evidence that cigarette smoking is an etiological factor of AML, was also observed. The finding in this cohort study is consistent with the results of previous case-control studies.

One of the strengths of a cohort study, as compared with a case-control study, is that the true risk of disease can be estimated in the study. A major limitation of cohort study for rare disease like AML is that the study sample size should be very large to detect significant associations for rare diseases like AML. Therefore, a few AML cases observed during the study period could
limit the value of this study. This might explain the wide range of risk estimation although there is not small population with 16 year follow-up in the study.

The epidemiological evidence concerning the possible etiologic role of cigarette smoking in leukemia as well as AML are few and controversial. A well designed cohort study could be more valuable than case-control study in establishing a relationship of causality between cigarette smoking and leukemia. Further studies could be improved by using less biased information of cigarette smoking, high quality covariate data, accurate classification of leukemia subtypes and a larger study sample size.

The findings of studies of cigarette-related leukemia could reinforce tobacco control including preventing initiation, promoting cessation, reducing exposure to cigarette smoking and finally reducing the health burden of cigarette-related diseases in the population.

5.4 STRENGTH AND LIMITATIONS OF SECONDARY DATA

Environmental public health tracking might cover a large population with a long time period in a region. Primary data gathering for environmental health tracking is unrealistic and inefficient. A variety of secondary data including environmental hazard, exposure and health outcomes is readily available in federal, state and local environmental and health agencies. It is a great source for environmental public health tracking. The review of three types of data including environmental hazard, environmental exposure and health outcomes is discussed in the section of Appendix B. Therefore, secondary data could provide an excellent and economical source for environmental public health tracking in comparison with collecting primary data.
Using secondary data might be an effective and quick way to identify health problems related to environmental exposure in a region.

However, there are a number of limitations when using secondary data for environmental public health tracking. The availability of information and data structure in secondary data heavily influence the study design and data analyses. There is often no way to validate the quality of secondary data because of lack of information about problems of sampling, method of data collection, response rates, data coding and others. We may lack information about the strengths and weaknesses of secondary data which are used in environmental public health tracking. The validity and reliability of results based on these data are difficult to assess.

Although secondary data is readily available, the accuracy and completeness may also vary. Therefore, some data issues including completeness of data, accuracy of information collected, registration period, data format, data accessibility and availability and possibilities of linkage with other data sources should also be aware of when secondary data are used in environmental public health tracking (See Appendix B).

Overall, the limitations of secondary data can not be avoided because it is usually collected for purposes other than for environmental public health tracking. However, secondary data is a valuable source and still attractive to environmental public health tracking agencies.

5.5 OVERALL SUMMARY AND CONCLUSION

The overall objectives of this project is to demonstrate how to using secondary data to link environmental exposures to chronic disease, to evaluate the possible associations between environmental exposure and health outcome and to explore the methodology of study design,
data linkage and data analyses. In this project, we succeeded in linking a variety of health outcome data including hospital admission data, birth registry data and research cohort data, with environmental exposures and examining the potential associations between environmental exposures and chronic diseases with appropriate methods of study design and data analyses. The techniques used in this project could be applicable to state or local environmental health tracking agencies for conducting similar studies.
APPENDIX A

RISK OF ADVERSE PREGNANCY OUTCOMES AND AIR POLLUTION

The study of adverse pregnancy outcomes is an emerging field of environmental epidemiology. Studies of air pollution on adverse pregnancy outcomes have been performed since the late 1990s. Low birth weight (birth weight less than 2500 g), as one of impaired birth outcomes, has been reported to influence the health status of individuals, including increased mortality and morbidity in childhood (1, 2) and an elevated risk of hypertension, coronary heart disease, type II diabetes in adulthood, abnormalities of lipid metabolism and blood coagulation (3, 4). Low birth weight is a heterogeneous outcome which could be caused by early delivery and fetal growth retardation and is associated with many factors. The early delivery due to premature rupture of the membranes and placenta abruption may result into a low birth weight infant. Some other maternal prenatal factors such as maternal diseases, maternal weight gain and maternal smoking could also be associated with a low birth weight baby. In addition, malformation may also increase the risk of low birth weight infant (5). There is still a limited literature regarding the effects of air pollution on birth weight. This appendix reviewed previous studies that have linked maternal exposure to ambient air pollution to birth weight.
We searched all publications included in the electronic databases PubMed and Ovid Medline with the combination of key words “air pollution” with any following: “birth weight”, “low birth weight”, “LBW”, “very low birth weight” or “VLBW”. Eighteen studies of assessing the relationship between birth weight and ambient air pollution were identified (6-23). These studies were summarized in Table 2. The air pollutant assessed in the studies included total suspended particles (TSP), PM$_{10}$, PM$_{2.5}$, SO$_2$, O$_3$, CO, NO$_2$ and NO$_x$. The estimations of an association between low birth weight and air pollutants reported in these studies ranged approximately from 1.01 to 1.5 (Table 2). The strengths that these studies shared include trimester-specific exposure estimations which could show the effect periods of air pollution on birth weight during pregnancy, standard definitions of adverse birth outcomes and adjustment for other important risk factors of low birth weight. However, the findings from these studies are inconsistent, especially regarding the effect period and the strength of association of particulate matter. In addition, many of these studies didn’t account for important confounders such as maternal smoking, gestational age and weight gain. Previous studies have also had limited spatial information on pollution sources and concentrations because the exposure mostly relied on the measurements at monitoring stations. This study, while not capable of clearing up all previous questions, could provide additional evidence regarding the health effects of particulate matter on LBW.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Design</th>
<th>Pollutants</th>
<th>Outcomes</th>
<th>Results</th>
<th>Covariates/stratified variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang X et al 1997</td>
<td>Cohort study (Beijing, China)</td>
<td>SO2(per 100µg/m³)</td>
<td>LBW</td>
<td>1.11 (1.06-1.16)</td>
<td>Gestational age, residence, year of birth, maternal age, and infant gender</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSP(per 100µg/m³)</td>
<td></td>
<td>1.10 (1.05-1.14)</td>
<td></td>
</tr>
<tr>
<td>Bobak M and Leon DA 1999</td>
<td>Ecological study (Czech Republic)</td>
<td>TSP(per 50µg/m³)</td>
<td>LBW</td>
<td>1.04 (0.96 -1.12)</td>
<td>Socioeconomic factors (mean income, car ownership, divorce rate, etc)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SO2(per 50µg/m³)</td>
<td></td>
<td>1.10(1.02- 1.17)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NOx(per 50µg/m³)</td>
<td></td>
<td>1.07(0.98- 1.16)</td>
<td></td>
</tr>
<tr>
<td>Bobak M 2000</td>
<td>Case-control study (Czech Republic)</td>
<td>TSP(50µg/m³)</td>
<td>LBW</td>
<td>LBW: 1.15[1.07-1.24] 1ST Trimester</td>
<td>Maternal sex, parity, age, education, marital status and month of birth.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SO2(50µg/m³)</td>
<td></td>
<td>LBW: 1.20[1.11-1.30] 1ST Trimester</td>
<td></td>
</tr>
<tr>
<td>Rogers JF 2000</td>
<td>Case-control study (Georgia, U.S.A)</td>
<td>Sum of TSP and SO2 (µg/m³)</td>
<td>VLBW (&lt;1,500g)</td>
<td>2.88[1.16-7.13] for &gt;95th percentile</td>
<td>Maternal race, age, education, smoking, medicine, weight gain, alcohol, stress and infant sex</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.27[0.68-2.37] for &gt;75th to 95th percentile</td>
<td></td>
</tr>
<tr>
<td>Ritz B and Yu F 1999</td>
<td>Cohort study (South California, U.S.A)</td>
<td>CO (&gt;5.5 ppm 3-month average) during the last trimester</td>
<td>LBW</td>
<td>1.22(1.03-1.44)</td>
<td>Commuting habits, infant sex, prenatal care, maternal age, ethnicity, and education</td>
</tr>
<tr>
<td>Reference</td>
<td>Design</td>
<td>Pollutants</td>
<td>Outcomes</td>
<td>Results</td>
<td>Covariates/stratified variables</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------</td>
<td>-----------------------------</td>
<td>-----------------------</td>
<td>---------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Bobak M 2001</td>
<td>Cross-sectional study</td>
<td>Pollution index</td>
<td>Birth weight</td>
<td></td>
<td>Gender, parental social class, maternal education, region, birth order, quality of house</td>
</tr>
<tr>
<td></td>
<td>(Britain)</td>
<td>1 (low)</td>
<td></td>
<td>3.9(24.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>-55.7(26.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td>-81.4(29.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4(high)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maisonet M 2001</td>
<td>Case-control study</td>
<td>CO (1ppm)</td>
<td>LBW</td>
<td>1.31[1.06-1.62]</td>
<td>Maternal age, education, race, smoking, alcohol, marital status, weight gain, previous termination, infant sex, gestational age, season of birth, firstborn, prenatal care.</td>
</tr>
<tr>
<td></td>
<td>(Six cities, U.S.A)</td>
<td>SO2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25th to 50th</td>
<td></td>
<td>1.18[1.12-1.25]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50th to 75th</td>
<td></td>
<td>1.12[1.07-1.17]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75th to 95th</td>
<td></td>
<td>1.13[1.05-1.22]</td>
<td></td>
</tr>
<tr>
<td>Ha EH 2001</td>
<td>Case-control study</td>
<td>CO</td>
<td>LBW</td>
<td>1.08[1.04-1.12]</td>
<td>Gestational age, maternal age, education, parity, infant’s birth order and sex.</td>
</tr>
<tr>
<td></td>
<td>(Seoul, South Korea)</td>
<td>NO2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SO2</td>
<td></td>
<td>1.07[1.03-1.11]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSP</td>
<td></td>
<td>1.06[1.02-1.10]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>O3</td>
<td></td>
<td>1.04[1.00-1.08]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.09[1.04-1.14]</td>
<td></td>
</tr>
<tr>
<td>Chen L 2002</td>
<td>Cross-sectional study</td>
<td>PM10(per 10µg/m³)</td>
<td>Reduction of</td>
<td>11 [2.3, 19.8]</td>
<td>Maternal age, race, education, smoking, drug, alcohol, prenatal visits, weight gain, residential city, infant sex, gestational age</td>
</tr>
<tr>
<td></td>
<td>(Northern Nevada, U.S.A)</td>
<td>Formaldehyde</td>
<td>Birth weight</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maroziene L 2002</td>
<td>Case-control study</td>
<td>Formaldehyde</td>
<td>LBW</td>
<td>ORmed. 1.86 [1.10-3.16]</td>
<td>Maternal age, parity, marital status, education, season of birth and smoking</td>
</tr>
<tr>
<td></td>
<td>(Kaunas, Lithuania)</td>
<td>NO2</td>
<td></td>
<td>ORhigh 1.84 [1.12-3.03]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ORhigh 1.64 [1.04-2.58]</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Design</td>
<td>Pollutants</td>
<td>Outcomes</td>
<td>Results</td>
<td>Covariates/stratified variables</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>---------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Lee BE 2003</td>
<td>Case-control study (Seoul, Korea)</td>
<td>PM10</td>
<td>LBW</td>
<td>1.03 [1.00-1.07]</td>
<td>Maternal age, education, infant sex, order, gestational age and date</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO</td>
<td></td>
<td>1.04 [1.01-1.07]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PM10</td>
<td></td>
<td>1.04 [1.00-1.08]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO</td>
<td></td>
<td>1.03 [1.00-1.06]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SO2</td>
<td></td>
<td>1.06 [1.02-1.11]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NO2</td>
<td></td>
<td>1.03 [1.01-1.06]</td>
<td></td>
</tr>
<tr>
<td>Yang CY 2003</td>
<td>Cross-sectional study (Kaohsiung, China)</td>
<td>PM10</td>
<td>Birth weight</td>
<td>β 0.52 [0.19-0.85]</td>
<td>Maternal age, education, marital status, season of delivery and infant sex</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SO2</td>
<td></td>
<td>β 0.52 [0.09-2.63]</td>
<td></td>
</tr>
<tr>
<td>Liu S 2003</td>
<td>Case-control study (Vancouver, Canada)</td>
<td>SO2 (5 ppb)</td>
<td>LBW</td>
<td>1.11 [1.01-1.22]</td>
<td>Maternal age, parity, infant sex, gestational age and season of birth</td>
</tr>
<tr>
<td>Gouveia N 2004</td>
<td>Cross-sectional study (Sao Paulo, Brazil)</td>
<td>PM10(per 10µg/m³)</td>
<td>Change of birth weight</td>
<td>-13.7[-27.0, -0.4]</td>
<td>Maternal age, education, parity, antenatal care, gestational age</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO (per 1ppm)</td>
<td></td>
<td>-23.1[-41.3, -4.9]</td>
<td></td>
</tr>
<tr>
<td>Parker JD 2005</td>
<td>Cross-sectional study (California, U.S.A)</td>
<td>PM2.5 (per 10µg/m³)</td>
<td>Birth weight</td>
<td>β -29.3g [-42.2, -16.4g]</td>
<td>Maternal race, age, education, parity, marital status, season of delivery and co-pollutant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO</td>
<td></td>
<td>β -38.2g [-54.9, -21.6g]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PM2.5(per 10µg/m³)</td>
<td></td>
<td>AOR 1.20[1.07-1.37]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO (per 0.5ppm)</td>
<td></td>
<td>AOR 0.89[0.77-1.03]</td>
<td></td>
</tr>
</tbody>
</table>
### Table A-1 continues

<table>
<thead>
<tr>
<th>Author</th>
<th>Study Type</th>
<th>Location</th>
<th>Pollutant</th>
<th>Outcome</th>
<th>RR  [95% CI]</th>
<th>Reference Range</th>
<th>Confounding Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mannes T 2005</td>
<td>Cohort study (Sydney, Australia)</td>
<td>PM10, PM2.5, CO, NO2, O3</td>
<td>PM10</td>
<td>Small</td>
<td>1.02 [.01-1.03]</td>
<td>1.03 [.01-1.05]</td>
<td>Infant sex, gestational age, maternal age, race, SES and season birth</td>
</tr>
<tr>
<td>Wilhelm M 2005</td>
<td>Case control study (Los Angeles, CA)</td>
<td>CO, PM10</td>
<td>LBW</td>
<td>LBW</td>
<td>1.36[1.04-1.76] 3rd trimester</td>
<td>LBW: 1.48[1.00-2.19] 3rd trimester</td>
<td>Infant sex, gestational age, maternal age, race, education, prior stillbirth, PTD, parity and birth season, other pollutants</td>
</tr>
<tr>
<td>Dugandzic R 2006</td>
<td>Retrospective cohort study (Nova Scotia, Canada)</td>
<td>PM10, SO2, O3</td>
<td>LBW</td>
<td>RR 1.33 [1.02-1.74] &gt;75th percentile, 1st trimester</td>
<td>RR 1.36 [1.04-1.78] &gt;75th percentile, 1st trimester</td>
<td>RR 1.16 [0.88-1.53] 25th to 50th percentile, 1st trimester</td>
<td>Gender of infant, gestation age, maternal age, parity. Smoking, weight change, prior neonatal deaths, stillbirth and LBW, and neighborhood family income</td>
</tr>
</tbody>
</table>
A.1 REFERENCE


The environment plays an important role in human health. In the 20th Century, the growing population and industrialization have adversely impacted the natural environment. Many studies have reported an association between chemical, physical and biological environmental exposures and adverse health effects. Adverse health effects related to chemical agents previously were the focus of studies primarily of the occupational environment. However, an enormous number and variety of chemicals have been introduced into the environment in the last decades and communities are now at the core of investigational efforts.

Chemical agents can enter human body through food, drinking water, air and skin contact. There is a considerable concern about the health effects of these agents on the general population. Certain specific environmental hazards have been linked with chronic diseases in the research literature. For instance, epidemiological studies have reported association between chlorinated water consumption and increased risk of various cancers, including cancers of the bladder, rectum, pancreas, kidney, stomach and lymphatic system (1-7). Air pollutant exposures have been associated with increased morbidity and/or mortality of cardiopulmonary disease (8-
11), asthma exacerbation (12-14) and adverse reproductive health outcomes (15-18). Benzene is well recognized as an etiological risk factor for leukemia(19, 20). Toxic effects related to metal or metal compounds have also been identified. For instance, exposure to arsenic has been linked with cancers of the skin(21-24), lung(25), urinary bladder(26), kidney(25) and liver(27), as well as non-cancer diseases such as peripheral vascular disease(28), cardiovascular and cerebrovascular disease(29), diabetes(30), and adverse reproductive outcomes(31, 32). Lead exposure has been associated with mental retardation(33-35). Physical environmental factors such as noise, extreme weather conditions, low-level radiation exposure and electromagnetic exposure, have also generated great concern in communities. Extremes of temperature has been reported to be associated with short-term increases in daily mortality(36, 37). Residential exposure to radon has been reported to cause lung cancer in human(38, 39). Several studies has linked exposure to electromagnetic fields with increased risk of cancer(40-42). There is little debate that agents in environment could cause adverse health effects in human. A systematic surveillance (e.g. tracking system) for environmental public health is clearly a necessary, rational and appropriate response to community, academic and governmental concerns about environmental exposures.

In 2000, the Pew Environmental Health Commission at the Johns Hopkins School of Hygiene and Public Health recognized that the environmental public health system was fragmented, neglected and ineffective in the United States (43). Under the recommendations of the Pew Environmental Health Commission, the Centers for Disease Control and Prevention (CDC) established the National Environmental Public Health Tracking (EPHT) Program, which is the ongoing collection, integration, analysis, interpretation, and dissemination of data on environmental hazards; exposure to those hazards; and related health effects (44). Its goal is to
provide information that can be used to plan, apply, and evaluate actions to prevent and control environmentally related diseases.

Thacker (1996) presented a useful framework of a “hazard-exposure-outcome” axis for conducting environmental public health tracking. It outlined the steps in the process of how an environmental agent produces adverse health outcomes (see figure B-1)(45). CDC applied the conceptual model proposed by Thacker to design the national EPHT network, which will meld data related to hazards, exposure and health outcomes into a network of standardized, distributed electronic data systems and will provide valid scientific information on environmental exposures and adverse health effects and the possible spatial and temporal relations between them (46).

![Figure B-1 Environmental agent and adverse health outcomes](image)

In this review, aspects related to the use of existing data for conducting environmental public health tracking will be discussed, including the availability of existing data, the strategy of data
linkage, statistical framework, issues related to the use of existing data for conducting environmental public health tracking.

B.1 OVERVIEW OF AVAILABLE EXISTING DATA SYSTEMS

The data on hazards, exposures and health outcomes are three types of data in the process by which an agent in the environment can produce adverse health effects in a host. The availability of these data is important for conducting environmental public health surveillance. Data on hazards, exposures, outcomes and spatial data have been collected at the national, regional or local levels as well as in scientific research studies. Much of the data has not been well utilized or never analyzed or interpreted in a way that could address other environmental issues or ever released to the public. A brief description of existing data on environmental hazards, exposures, health outcomes and geospatial field is presented in the following section.

B.1.1 HAZARD DATA

A hazard can be any chemical, physical or biological environmental agent which has a potential for adverse health effects. Information on the presence and quantity of contaminants in environmental media is systematically collected by various organizations. Most hazard data is collected by federal and state agencies and is mandated by legislation, such as Clean Air Act or Clean Water Act (1970). The U.S. Environmental Protection Agency (EPA) was established to consolidate in one agency a variety of federal research, monitoring, standard-setting and
enforcement activities to ensure environmental protection on December 2, 1970. Many hazard databases on air and water have been routinely established and maintained by EPA.

**Air**

The EPA Air Quality System (AQS) is a database containing hourly and daily measurements of criteria air pollutants and hazardous air pollutants from monitors across the United States. Detailed data from 1994 to the present can be downloaded from the EPA website (http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm). The EPA National Emissions Inventory (NEI) database includes estimates of annual pollutant emissions from point, non-point and mobile sources for 50 states, Washington DC, Puerto Rico and the Virgin Islands. Data drawn from the NEI and AQS databases consist of the EPA AirData/AIRS database. Theses data are available from 1971 to the present.

Toxic Release Inventory (TRI) is an EPA database containing estimated information about the amount of chemicals annually released to air, water, and land by the manufacturing industry. Any facility meeting established criteria (ref for criteria: http://epa.gov/tri/guide_docs/2001/brochure2000.pdf ) is required to report annual releases of nearly 600 chemical compounds. The data are available from 1987 to the current time.

**Water**

EPA maintains two data management systems containing water quality information for the nation's waters: the Legacy Data Center (LDC), and STORET. The LDC is a static, archived database and STORET is an operational system actively being populated with water quality data. The LDC contains historical water quality data dating back to the early part of the 20th century and collected through the end of 1998. STORET contains data collected beginning in 1999. Both databases contain raw biological, chemical, and physical data on surface and ground water.
collected by federal, state and local agencies, Indian Tribes, volunteer groups, academics, and others.

EPA also maintains several databases on drinking water. For example, the Safe Drinking Water Information System - Federal version (SDWIS/FED) database includes information about the nation's 160,000 public water systems and violations of drinking water regulations. The National Contaminant Occurrence Database (NCOD) contains occurrence data from both Public Water Systems (PWSs) and other sources (like the U.S. Geological Survey National Water Information System) on physical, chemical, microbial and radiological contaminants for both detections and non-detects. The Unregulated Contaminant Monitoring Rule (UCMR) database is collected to evaluate and prioritize contaminants on the Drinking Water Contaminant Candidate List, which could be possible new drinking water standards. The Federal Reporting Database System (FRDS) is a centralized database for information on public drinking water supplies. It is maintained by the EPA Office of Drinking Water. FRDS contains approximately 12 million records.

In addition, other organizations and institutions maintain nation wide hazard databases. For instance, the U.S. Geological Survey (USGS) created a National Water Information System (NWIS) database which provides access to water-resources data collected at approximately 1.5 million sites in all 50 states, plus border and territorial sites. There are nearly 70 million water-quality results from about 4 million water samples collected at hundreds of thousands of sites. The data are from as early as 1899 to the present.

The USGS implemented the National Water-Quality Assessment (NAWQA) Program in 1991 to develop long-term, consistent and comparable information on streams, rivers, ground
water, and aquatic systems. It includes 162 sample sites in 51 of the nation's river basins and aquifers.

**Climate**

The National Climate Data Center (NCDC) maintains the world’s largest active archive of weather data. The NCDC’s database provides a national resource for climate information from 1985 to present. The data include hourly temperature, wind, rain, pressure, clouds and snow in national or global scale.

**Agriculture**

The National Agricultural Statistics Service (NASS) database is maintained by the United States Department of Agriculture (USDA). The NASS conducts hundreds of surveys every year. It provides information about U.S. agricultural chemical use, agricultural-production and supplies of food and fiber, prices paid and received by farmers, farm labor and wages, farm finances, and changes in the demographics of producers. The available data are from 1867 to the present.

Moreover, some environmental hazard data were created in large national-wide survey studies. For example, the U.S EPA conducted a five-year study entitled National Pesticide Survey (NPS) in 1988. It provided information on specific contaminants, pesticides and nitrates in drinking water from groundwater supplies such as community water system wells and rural domestic wells.

The National Atrazine Occurrence Monitoring Program was conducted under the joint sponsorship of the Association Water Works Association’s (AWWA) Water Industry Technical Action Fund (WITAF) and the American Water Works Association Research Foundation
(AwwaRF). It is a study to characterize atrazine occurrence patterns across the United States. Forty-seven drinking water plants were included in this study.

**B.1.2 EXPOSURE DATA**

Exposure data include bio-monitoring for the presence and the quantity of a compound or its metabolites in human tissues or biologic samples such as fat, blood, hair and urine. Tracking actual exposure to hazards is frequently the missing link in evaluating the risk of environmental hazard. Exposure data are usually monitored at the individual level. However, little individual level exposure data currently exist in accessible datasets.

The National Human Adipose Tissue Survey database is a nation wide exposure data set. It provides information on 54 chemicals in human adipose tissue in three different age groups and in nine different geographic regions in the United States. The database contains results of analyses of human tissue samples collected between 1967 and 1990.

Biomonitoring measurements were also made in samples from participants in the National Health and Nutrition Examination Survey (NHANES) since 1999. More than 100 chemicals or their metabolites were measured in blood and urine samples from a subsample of participants from NHANES conducted by CDC’s National Center for Health Statistics.

The National Human Exposure Assessment Survey (NHEXAS), sponsored by the U.S EPA in 1994, is a pilot and long-term study of the ways that humans are exposed to potentially toxic chemicals in the environment. The environmental chemicals include lead and arsenic, benzene and related volatile organic compounds, and various pesticides. The field of study covers the counties in Minnesota, Wisconsin, Illinois, Indiana, Ohio, and Michigan(47).
Some exposure data are available from investigation studies. For example, several studies tested urinary dimethylthiophosphate (DMTP) to evaluate organophosphorus (OP) pesticide exposure. These studies have provided information of OP pesticide exposure in apple thinners (48), children of agricultural pesticide applicators (49), children living in a large metropolitan area (50), living in an agricultural community (51), and living in farmworker households (52).

Because exposure data are limited, levels of exposure are often estimated through sophisticated modeling. The Hazardous Air Pollutant Exposure Model, Version 4 (HAPEM4) is such a type of model and has been used by the U. S. EPA to estimate inhalation exposure of air toxics for specified population groups. The National-Scale Air Toxics Assessment (NATA) database were developed with this exposure modeling and provides information about exposure concentration estimates of the 33 air pollutants at the census-tract, as well as county/state level. The goal of the national-scale assessment is to identify those air toxics which are of greatest potential concern, in terms of contribution to population risk. Currently, however, only two years of data are available, i.e. 1996 and 1999.

B.1.3 HEALTH OUTCOME DATA

Health outcomes of interest can include any outcomes along the timeline for the development natural history of a disease from preclinical measurements to death. These outcomes of interest are in the effort to link an environmental hazard or exposure with them in human beings. Sources of existing health outcome data include national, state and local level disease registries, vital statistics data, nation-wide health surveys, administrative data systems and large epidemiological cohort studies.
a). Centers for Disease Control and Prevention (CDC) Databases

The Centers for Disease Control and Prevention (CDC), as the lead federal agency for protecting the health and safety of people, has housed health data from birth and death records, medical records, interview surveys, through direct physical exams and laboratory testing. These databases maintained by the CDC could serve as valuable sources of health outcomes for the Environmental Public Health Tracking.

The National Vital Statistics System in the CDC’s National Center of Health Statistics (NCHS) is one of the oldest and most successful databases. The database provides information about all vital events including births, deaths, marriages, divorces, and fetal deaths in the 50 States, 2 cities (Washington, DC, and New York City), and 5 United States territories (Puerto Rico, the Virgin Islands, Guam, American Samoa, and the Commonwealth of the Northern Mariana Islands).

The National Health and Nutrition Examination Survey (NHANES) is a major effort of the National Center for Health Statistics (NCHS) to assess the health and nutritional status of adults and children in the United States. The NHANES program began in the early 1960s and collects data on a nationally representative sample of about 5,000 persons each year. This survey consists of individual information on demographics, socioeconomics, dietary intake, and health. A physical examination conducts, which consists of medical and dental examination, physiological measurements, and laboratory tests administered by highly trained medical personnel.

The National Health Interview Survey (NHIS) is a multipurpose health survey conducted by the CDC’s NCHS. The NHIS has been an important source of information about health and health care in the United States since it was first conducted in 1957.
The National Health Care Survey (NHCS) of the CDC’s NCHS includes a series of health care provider surveys. These surveys provide information about the facilities that supply health care, the services rendered, and the characteristics of the patients served. This family of surveys includes National Ambulatory Medical Care Survey (NAMCS), National Hospital Ambulatory Medical Care Survey (NHAMCS), National Survey of Ambulatory Surgery (NSAS), National Hospital Discharge Survey (NHDS), National Nursing Home Survey (NNHS), National Home and Hospice Care Survey (NHHCS), National Employer Health Insurance Survey (NEHIS), and National Health Provider Inventory (NHPI). The detailed information of each survey refers to the CDC’s NCHS website (http://www.cdc.gov/nchs/nhcs.htm).


The Longitudinal Studies of Aging (LSOAs) is a collaborative project between CDC’s NCHS and the National Institute on Aging (NIA). The project consisted of four surveys: the 1984 Supplement on Aging, the 1984-1990 Longitudinal Study of Aging (LSOA), the 1994 Second Supplement on Aging (SOA II), and the 1994-2000 Second Longitudinal Study of Aging (LSOA II). The purpose of the project is to measure changes in the health, functional status,
living arrangements, and health services utilization among Americans of 70 years of age and over as they move into and through the oldest ages.

The CDC’s National Program of Cancer Registries (NPCR) has been funding 45 states, District of Columbia and 3 U.S. territories to collect population-based cancer incidence data since 1994. The database provides information about patient demographics, tumor characteristics, state at diagnosis and first course of treatment. NPCR has four products for data release: U.S. Cancer Statistics (USCS), U.S. County Cancer Incidence Dataset, USCS Expanded Dataset, and USCS Restricted Access Datasets.

b). Other agency and institute databases

In addition, there are many databases of health outcomes maintained by other agencies and organizations. The following review provides a brief description of the most frequently assessed databases available from other agencies excluding CDC.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is a comprehensive source of population-based information on cancer incidence and survival in the United States. The program began in 1973 and collected data from population-based cancer registries covering approximately 26 percent of the US population. It provides information about stage of cancer at the time of diagnosis and patient survival data.

The Agency for Toxic Substances and Disease Registry (ATSDR) has maintained an active, state-based Hazard Substances Emergency Event Surveillance (HSEES) since 1990. This system provides the information of the public health consequences associated with the release of hazardous substances. It includes 15 participating states(53).

The American Association of Poison Control Centers (AAPCC) maintains the Toxic Exposure Surveillance System (TESS), which is the comprehensive poisoning surveillance
database and covered the entire population of the 50 states, the District of Columbia, and Puerto Rico in the United States. TESS is a valuable resource for environmental public health tracking. It provides information on the patient, the caller, the exposure, the substance(s), clinical toxicity, treatment, and medical outcome. Developed in 1983, TESS contains detailed toxicological information on 36.2 million cases reported to U.S. poison centers through 2003. That includes more than 2 million reports to poison centers for 2000 alone, an estimated 96% of all poison exposures reported to poison centers in the U.S (54, 55).

The National Cancer Database (NCDB) was established by the American Cancer Society (ACS) and the American College of Surgeons (ACoS) in 1989. The database served as a comprehensive clinical surveillance resource for cancer care in the United States. It captures 75% of all newly diagnosed cancer cases in the United States annually and contains over 15 million cases of reported cancer diagnoses for the period 1985 through 2002. The database provides information about patient characteristics, tumor staging and histology characteristics, type of first course treatment administered, disease recurrence, and survival information.

c). Large research cohort study databases

Large cohort studies are other potential valuable sources of existing data for environmental public health tracking. Data collected in cohort studies are at the individual level. Detailed individual information such as contact information and address, lifestyle information like smoking and diet, and other individual factors are usually available in the data of cohort studies. A brief introduction of some well-documented cohort studies will be given in the following section.
The Framingham study, directed by the National Heart, Lung, and Blood Institute (NHLBI), is designed to identify the common factors or characteristics that contribute to CVD among 5,209 participants, their offspring and third generation. The study lasted over 50 years (56).

The Nurse’s Health Study, sponsored by the National Institutes of Health, is the largest prospective investigation of risk factors for major chronic diseases in women. The study included Nurses' Health Study and the Nurses' Health Study II. The Nurses' Health Study began in 1976 and approximately 122,000 nurses aged 30 to 55 from the 11 most populous states participated in the study (57). The Nurses' Health Study II began in 1989 and a total of 116,686 women between the ages of 25 and 42 years are involved in the study (58, 59).

The California Teachers Study, sponsored by the State of California, is a prospective study of 133,479 current and former public school teachers or administrators who participate in the California State Teachers Retirement System (STRS). The purpose of the study is to identify the risk factors associated with breast cancer research. The study began in the fall of 1995 (60).

The Agricultural Health Study was sponsored as a joint collaboration between the National Institutes of Health and the Environmental Protection Agency in 1994. Over 89,000 farmers as well as their families and commercial pesticide applicators from North Carolina and Iowa states participated in the project. The Study is designed to identify occupational, lifestyle, and genetic factors that may affect the rate of diseases in farming populations (61).

The American Cancer Society Cancer Prevention Study includes the Cancer Prevention Study I and the Cancer Prevention Study II. The Cancer Prevention Study I was a prospective mortality study and began in 1959. Over 1 millions participants were recruited into the study in twenty-five states and followed up through September 1972 (62). The Cancer Prevention Study II (CPS II) includes the CPS II Baseline Cohort, CPS II Nutrition Cohort and CPS II Biospecimen
Repository. The Cancer Prevention Study II is to examine the impact of environmental and lifestyle factors on cancer etiology. Approximately 1.2 million men and women in 50 states, the District of Columbia, and Puerto Rico participated in the CPS-II Baseline Cohort in 1982. The data have been examined extensively in relation to cancer mortality (63, 64). The CPS-II Nutrition Cohort was established in 1992 and 1993 as a subgroup of the larger CPS-II Cohort. A total of 184,194 men and women in middle-aged to elderly living in 21 states in the United States participated in the cohort. The study is designed to evaluate the associations between dietary as well as other risk factors and cancer incidence (65). The CPS-II Biospecimen Cohort was initiated to obtain blood samples from 40,000 surviving members of the CPS-II Nutrition Cohort in 1998.

The Women’s Health Initiative (WHI) Observational study was established by the National Institutes of Health (NIH) to address cardiovascular disease, cancer, and osteoporosis among women with age 50-79 years old. A total of 93,676 women were enrolled in this study (66).

The Cardiovascular Health Study (CHS) sponsored by the National Heart, Lung, and Blood Institute (NHLBI), is a population-based, longitudinal study of risk factors for the development and progression of CHD and stroke in the U.S. adults over the age of 65 years. The study stated in 1989 and will end in 2009. There are 5,201 participants in this study (67).

The Alameda County Study began in 1995 and is designed to examine the risk factors associated with health and mortality among 6928 adults living in Alameda County, California. The last update of this study occurred on 1997(68).
B.1.4 SPATIAL DATA

Geospatial data has been increasingly used in public health research. Many agencies or organizations provide geospatial databases. Some of these databases are now readily available free of charge on the internet. In the following section, sources of spatial data available that are potentially useful for environmental public health tracking are briefly described.

The U.S. Geological Survey (USGS) has long been a source of geospatial data. Most of these data are now available in digital formats. The Digital Cartographic Data products provided by USGS include Digital Orthophoto Quadrangle (DOQ), Digital Raster Graphics (DRG), Digital Elevation Models (DEM), Digital Line Graphs (DLG), and National Hydrography Dataset (NHD). The Digital Line Graphs data set provides information on transportation lines, hydrography, political boundaries, and elevation contours, vegetable surface cover, manmade features, non-vegetative features as well as township, range, and section lines. The National Hydrography Dataset (NHD) is a comprehensive set of digital spatial data that contains information about surface water features such as lakes, ponds, streams, rivers, springs, and wells.

The U.S. Census Bureau provides a number of spatial datasets. The Topologically integrated Geographic Encoding and Referencing (TIGER) system released by the U.S. Census Bureau contains information on street, railway, political boundaries of census, and major hydrographic features. It covers the 50 states, the District of Columbia, Puerto Rico, the Virgin Island, and the outlying area of the Pacific area. The Cartographic Boundary Files are selected generalized extracts from the Census Bureau's TIGER geographic database and are designed for use in a Geographic Information System (GIS) or similar mapping system. The files are available in three formats: ARC/INFO EXPORT (.e00) format, ArcView Shapefile (.shp) format, and ARC/INFO 100.
The Environmental Systems Research Institute (ESRI) specializes in geographic information systems (GIS) software. The spatial datasets provided by ESRI include the StreetMap data, Community Data, Basemaps and census data. The StreetMap Pro and StreetMap Premium enhanced street datasets provide routing, geocoding, and high-quality cartographic display for the entire United States. The ESRI Community Data encompasses a variety of datasets that help companies and organizations analyze markets, profile customers, evaluate competitors, and more. ESRI software users can access most of these datasets at on additional cost.

**B.2 STRATEGIES OF DATA LINKAGE**

Data linkage is the task of accurately identifying records corresponding to the same entity from one or more data sources and the methodology for joining together corresponding records from two or more databases. The entity in public health studies usually refers to individual, family or geographical region etc. Using existing data for evaluating possible associations between environmental hazard or exposure and health outcome, linkages of these types of data are the intermediate and important step.

**B.2.1 DATA LINKAGE**

Linking health outcome with data of environmental hazards or exposures depends on following aspects: purpose of study, study design and nature and structure of available data. The
general steps involved in computer-based linkage procedures for two or more databases are shown as the following scheme (Fig B-2).

The accuracy of data linkage greatly depends on the linkage strategy, methodology and quality of data. The first step of data linkage is to be familiar with the databases and to determine data fields or variables in each database to be studied. Therefore, some nuisance data fields can be eliminated from databases. This process makes the databases to be linked more compact. The next step is to identify the key variables across the different databases. These key variables are used as linkage variables, which should exist in each database to be linked. The selections of key variables are vital to successful data linkage. The quality of these key variables such as missing values and typographic errors can heavily influence the accuracy and quality of linked dataset. Standardization of key variables is another vital process for data linkage. Without standardization, many truly matched records might result in unmatched records. Standardization includes converting common words into standard spelling, for example “road” or “street” in an address is converted into “Rd” or “St” for each database, using the same system of unit, and applying the same coding system for same variables etc.
Determine the data variables or fields to be studied

Develop the strategy for match-merge

Restructure datasets and standardize the key variables

Make a program and start data linkage

Evaluate the linked dataset

Improve the program and run again

Create the linked dataset

Figure B-2 The general procedure of data linkage

B.2.2 METHODS OF DATA LINKAGE

The linkage of routinely collected data is a relatively cheap and readily available method of obtaining data for monitoring population health and for research purpose. With increasing capacity of data storage and development in data linkage techniques, there has been substantial
increase in linking existing data to obtain information on large population in research or health surveillance. There are two main types of data linkage methods: deterministic data linkage and probabilistic data linkage(69).

**a. Deterministic linkage**

Deterministic linkage, also called exact linkage, is the ability to link records from different databases that exactly match on error-free identifying fields or variables. Such identifying fields should be precise, robust, stable over time and highly available. Deterministic linkage can be used if a unique identifier is available in the record fields of all the dataset to be linked. However, in individual level datasets, the identifiers to be used in deterministic linkage usually are confidential information, such as Social Security Number (SSN), name and date of birth in health dataset. It is clear that such linkage of two or more datasets can infringe on individual privacy. For datasets aggregated in geographic regions or time dimensions, the issue of a breach in confidentiality or privacy is less problematic. The key variables for these type datasets might usually be the variables of geographic region such as zip code, county and state or time such as day, month, and year. For example, in a time series study of air pollution, the linkage of monitored air pollution data with the number daily hospital admissions (counts) data on a given day is based on the key variable of calendar day and not any individual level data element (70, 71).

Several studies have been conducted to evaluate the performance of deterministic linkage on health data with different combinations of key words. Herman et al used deterministic linkage methods to link birth registry data with infant mortality and morbidity surveillance data sets from the United States (Georgia, Missouri, Utah and Washington State), Israel, Norway, Scotland and
Western Australia in 1997. A unique identifier, national identification numbers, were available for these datasets. The study reported that the methodology was highly successful (72).

Li et al completed data linkage with the population registry, hospital discharge and Vital Statistics registry datasets using deterministic linkage with a combination of key variables surname, sex and date of birth. This study found that the combination of surname, sex and date of birth appears to be optimal using deterministic linkage(73).

Simon et al merged records of young women with invasive breast cancer identified through three population-based cancer registries, to state birth certificate records with deterministic linkage. The performance of data linkage with different combinations of key variables was compared. The study found that SSN appears to be fairly accurate for linkage for linking cancer registries to other data sources (74).

The deterministic data linkage method is simple and easier to interpret. Deterministic data linkage can be accomplished with standard statistical software packages such as SPSS and SAS. However, deterministic linkage can be very time-consuming and very practitioner-dependent. In addition, it can not handle partial agreements easily, which may lead to false negative matching(75).

b. Probabilistic linkage

Data linkage is inevitably probabilistic in nature. Deterministic linkage is an extreme example of probabilistic linkage in that it is 100% certain that records of different databases, exactly matched on key variables, indeed refer to the same entity. However it may not true due to
duplication of identifiers and errors in creating or transmitting records. Thus deterministic linkage does not adequately reflect the uncertainty that may exist for some potential links. Moreover, unique and reliable identifiers may not be available in the databases to be linked in many situations. It is desirable that different linkage strategy should be developed to address these issues. Probabilistic linkage is the alternative to deterministic linkage and is developed when the rules of classic probability theory have been applied in data linkage.

The method of probabilistic linkage is defined as record linkage of two or more files that utilizes the probabilities of agreement and disagreement between a range of matching variables (69). Newcomber et al first reported the possibility of probabilistically linking records and provided crucial insights that led to computerized approaches to record linkage in the absence of unique identifiers (76). Subsequently, Fellegi and Sunter introduced the mathematical and statistical foundation to formalize the theory of probabilistic data linkage, building on the concept provided by Newcomber (77). The basic principle of probabilistic linkage is to use the probability of a true match \( m(\gamma) \) and the probability of a true non-match \( u(\gamma) \) based on matching variable to determine its (dis)agreement weight. The (dis)agreement weights for all matching variables were determined by the same strategy. The total weight for a given record link is simply the sum of the (dis)agreement weights for each matching variables. If all or most matching variables agree, the total weight will be a large positive number; otherwise, it is a small positive number. By comparing the total weight with threshold values, a true match, possible match or true non-match is determined (69, 75).

Probabilistic data linkage is commonly used in health research studies. Several studies have evaluated the accuracy of probabilistic linkage in matching records of health databases. Newgard et al evaluated the accuracy of using probabilistic linkage for matching de-identified ambulance
records to a state trauma registry. It found that probabilistic linkage is a valid method for matching ambulance records to a trauma registry without the use of patient identifiers but the accuracy is related to the selection of common variables included in the analysis (78).

Nitsch et al assessed the validation of probabilistic record linkage to identify births to a cohort of women by linking the females in the cohort to birth records held by the Scottish Maternity Record System (SMR 2) based on surname, maiden name, initials, date of birth and postcode. It concluded that probabilistic record linkage to routine maternity records applied to population-based cohort can have high specificity, and as such may be reliably used in epidemiological research(79).

A few studies have been conducted to compare deterministic and probabilistic method on their performance of data linkage. Tromp M et al used the Dutch National Midwife Registry, the Dutch National Obstetrics Registry and the Dutch National Pediatrics Registry to compare probabilistic and deterministic record linkage techniques. The study reported that Probabilistic linkage identified 80% more links than a full deterministic linkage approach. External validation revealed an error rate of less than 1% (80).

Roos and Wajda compared deterministic and probabilistic methods and found that probabilistic linkage had great advantages in those situations where only a moderate amount of extra information was available(81).

Therefore, many studies have applied the deterministic and probabilistic linkage strategies to combine health data across the world. After a comprehensive review of these studies, it might provide some clues of data linkage in using secondary data for environmental public health tracking.
The relationship between environmental exposure and health outcomes is complex. For example, multiple relationships between environmental exposures and health outcomes often exist, as multiple environmental factors may contribute to a single disease. On the other hand, a single environmental factor may contribute to many diseases (82). The type of statistical model that is used to evaluate this possible association requires a careful consideration. The selection of a statistical model is much dependent on the purpose of study, study design and the availability of dataset.

Mather et al organized the statistical models for the analysis of environmental and health data as well as the environment and health relationship into three groups. The first group is composed of descriptive analyses which describes trends, generates baselines and compare temporal and spatial changes of health and environmental data. The second group is an ecologic analysis which uses aggregated health and environmental data to provide information on the relationship between hazards and health outcomes. The third group consists of etiologic research to test the relationships between environmental factors and health outcomes (83). Following the statistical framework proposed by Mather, statistical methods are organized into two categories: descriptive analysis and etiologic analysis. A variety of available statistical methods which are appropriate for the analysis of environment and health data has been reviewed based on these two categories.
B.3.1 DESCRIPTIVE ANALYSIS

Descriptive analysis often represents the first step in describing or investigating a new event or condition. It can be used to monitor the trends in hazards, exposures and health outcomes and to search for clues of cause of disease.

**Time trend:** Trend analysis of environmental hazards is important for environmental public health tracking to characterize the background and changes in environmental contaminants. In addition, Kyle et al suggested that for cases where associations between environmental hazards and health outcomes have been well studied and established, it may be more relevant to focus on tracking environmental contaminants (82). In such cases, trend analysis of environmental hazards can provide information for planning and action of prevention of diseases. Examining time trend of health outcomes is helpful to understand the temporal variation of disease and to identify disease time clustering. Statistical approaches for detection of clustering over time include scan statistic which is the maximum number of observed cases in an interval of pre-selected length, as the interval is allowed to scan, or slide along, the time frame of interest (84), and Ederer-Myers-Mantel test in which the test statistic, $m_1$ will be large when cases are clustered; otherwise, $m_1$ will be small when cases occur uniformly through time(85, 86).

**Disease cluster or clustering:** Disease cluster or clustering is useful to highlight areas at apparently high risk and to provide useful etiologic clues to disease origins. Many methods have been proposed to detect disease cluster or clustering. The methods include nearest neighbor test or Cuzick-and-Edwards’test (87) that compare the distance between cases to expected distance(88), spatial autocorrelation which summarizes the degree to which similar observations tend to occur near each other(89), and spatial scan statistic which identifies the collection of
diseases least consistent with the null hypothesis and provides a significance value representing the detected cluster’s “unusualness” (90).

**Visual analysis (mapping):** Mapping is a technique to provide insight into spatial and spatial-temporal variation in exposure and disease risk. In public health, disease mapping has a long history. The earliest example is Snow's famous cholera map in the mid-19th century. Disease mapping has been widely applied in the field of health studies and was often used in cancer studies (91-93).

For a fair comparison, the rates from map are required to be adjusted for potential confounders such as age, sex and race. Standardization methods include direct standardization and indirect standardization. With indirect standardization, disease maps typically show standardized mortality or morbidity ratios (SMRs or SIRs). With direct standardization, the adjusted rate was used to map.

The disease rate based on a large population may be a better estimator. However, the rates from small population are likely to be elevated artificially, which reflect lack of data rather than true elevated risk. Spatial smoothing is one method for reducing the noise in rates associated with geographic region. Common approaches to spatial smoothing include locally weighted average in which a smoothed value is obtained by averaging the values associated with neighboring regions, empirical Bayes smoothing which use a Bayesian approach to define the analytic form of the compromise estimator(47), and kernel smoothing in which a kernel function and a defined bandwidth are used to controls the amount of smoothing (94)

**Ecological analysis:** In ecological analysis, the purpose is to evaluate temporal and/or geographic variations in exposure to environmental factors across population groups in relation to health outcomes measured on time and/or geographic scale. The approach is easily adapted to
use data which are routinely collected. The major limitation of the approach is ecologic fallacy, which is that an association between exposure and health outcome observed at aggregated group level may not validate at the individual level. For this reason, ecological analysis can be useful to generate but not confirm a hypothesis for the causality of disease.

Temporal ecological analysis has been applied to evaluate data aggregated in time scale such as daily, weekly, monthly or yearly. It has been widely used in air pollution studies. Time-series analysis of air pollution is this type of study, which assess the association between daily variation of air pollutants and daily count of health outcomes after controlling for climatic factors, seasonality and secular trend. The major two statistical methods of time-series analysis are generalized linear model (GLM) with parametric natural cubic splines (95, 96) and generalized additive models (GAM)(97, 98).

Geographic ecological analysis to evaluate the correlation between aggregate exposure and aggregate health outcome in geographical unit is usually defined by administrative boundary such as city, county and states. Traditional statistics such as linear, Poisson and logistic regression may not be appropriate for spatial data analysis without incorporating the information on neighborhood relationships and spatially correlated error terms. Spatial autoregressive models are developed by treating observations of the outcome variable at other locations as additional covariates in the model with associated parameters defining spatial association(99). The generalized linear mixed model is another method for modeling regional counts and rates developed by introducing random effects and hierarchical modeling into models (100). Bayesian models are developed to fit very complicated hierarchical models including those with spatially correlated random effects. The basic idea of this model is that of using the data to define a probability distribution for each of the model parameters, then using these distributions to draw
inference. Several studies report that Bayesian models offer dramatic improvements in statistical accuracy over conventional statistical methods(101, 102).

B.3.2 ETIOLOGICAL ANALYSIS

Etiological analysis usually uses exposure/hazard and health outcome data at the individual level. The most common study designs to evaluate the possible association between health outcomes and environmental hazards/exposures consist of a cohort study and a case-control study including a case-crossover study.

The strength of association between environmental factors and health outcome is not usually high. In addition, the association between environmental factors and health outcomes is complicate. The ability to detect an association is influenced by many factors. Lack of validity of testing can be expected if there are unidentified confounding factors, if confounding factors are measured with errors or if controlling for confounding factors is inadequate(103). Therefore, the use of appropriate statistical models is important to detect an association between environmental factor and health outcome.

In cohort study, survival analysis such as Cox proportion hazard model was often applied to estimate the relative risk. In a case-control study, the statistical method is logistic regression, which is used to estimate odds ratios. Conditional logistic regression is required to estimate odds ratios in paired case control study or case-crossover study, which we will discuss later (see proposed demonstrative studies).
B.4 ISSUES OF USING SECONDARY DATA FOR ENVIRONMENTAL PUBLIC HEALTH TRACKING

B.4.1 DATA ISSUES

Using secondary data for environmental public health tracking is promising and feasible. Development of the EPHT program depends on the availability, quality, timeliness, compatibility and utility of existing data. However, secondary data collected for specific aims may not be sufficient for environmental public health tracking. A few publications discussed about the strengths and limitations of using secondary data for epidemiologic studies (104-106). Sorensen et al provided a comprehensive review of issues of using secondary data sources for epidemiologic research, which include completeness of data, accuracy of information collected, registration period, data format, data accessibility and availability and possibilities of linkage with other data sources (107).

Completeness of data is concerned with the function of registry system and data quality. Cases can be missed even in the best registration system. Improving the registry system may enhance completeness of data. On the other hand, completeness of data is also related to the completeness of information on subjects collected in the system. Significant missing or incomplete data can negate the value of a secondary data source. Completeness of data can also refer to time dimension. For example, many EPA monitors operate for only short periods of time or have intermittent periods with no data collection. It yields gaps in the data. This compromises the value of the data.

Accuracy of information collected reflects data quality as well. Incorrect information in secondary data may come from incorrect data entry or lack of entry of available information and
errors of measurement, i.e. information which may not reflect the true condition or characteristic of an individual or condition. It is very difficult or impossible to evaluate the accuracy of information in secondary data.

Data from different time periods have some potential issues. The issues are related to something changing over time in the data collection process. For example, in the NEI database, a different method for estimating emissions will be applied. Data produced at different periods of time might not be comparable. For a disease registration system, disease definition, assess to care, diagnostic technology, and disease reporting can change over time (108).

Data accessibility and availability and data format are the issues related to data sharing, data delivery and data application. If these issues are not well addressed, it might be very difficult or time and labor consuming to make use of secondary data.

**B.4.2 ISSUES OF ACCURATELY IN ASSESSING ENVIRONMENTAL EXPOSURE**

Accurately assessing environmental exposure is a major issue in environmental public health tracking. As mentioned before, available data systems lack adequate measures of human exposure. Many environmental agent exposures have not identified the specific and effective biologic indices. Assessment of human exposure will continue to be critical in environmental public health tracking. However, even if exposure data are available, these data usually provide information of intensity of exposure at single point of time, i.e. time of sample collection. To accurately assess environmental exposure, the data need to capture the duration, frequency and intensity of exposure over a specified time period. Moreover, timing of exposure is another import factor which could influence the correct estimation of exposure level. Rothman presented a useful structure to reveal the relation between timing exposure and the occurrence of disease. It
suggested that there would be a specific time period during which a given exposure is causally related to disease, i.e. time window of exposure. The exposure before or after the etiologically important time window might not be a good measurement of disease relevant exposure (109). Several publications reported the importance and methods for considering the time of period of exposure in addition to the dose of exposure in evaluating the effect of exposure on onset of diseases (110-112).

B.4.3 ISSUES OF CONFIDENTIALITY AND PRIVACY

Confidentiality and privacy are barriers to secondary data accessibility and availability. Health data usually include personal data information, which is important information for data linkage. Using these data might be an invasion of privacy and confidentiality. Moreover, it is difficult and potentially impossible to secure informed consent from individuals in identified in population-based administrative and registry data. Therefore, it is very important to clarify who has the right to use specific data and which authorities should approve the use of the data for environmental public health tracking. The strategies to address these issues should be developed in the future for balancing between the protection of privacy as well as confidentiality and data accessibility and availability.

In a word, all the above issues should be carefully considered when secondary data are used for environmental public health tracking.
B.5 REFERENCE:


44. CDC. Environmental Public Health Tracking Program. 2003.


119


78. Newgard CD. Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry. Acad Emerg Med 2006;13:69-75.


BIBLIOGRAPHY


Commission TPEh. 2000 America's Environmental Health Gap: Why the County Needs a Nationwide Health Tracking Network.


Drinker P. 1953. Deaths during the severe fog in London and environs, Dec. 5 to 9, 1952. *A M A Arch Ind Hyg Occup Med* 7:UNKNOWN.


Newgard CD. 2006. Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry. Acad Emerg Med 13:69-75.


