



**Michigan
Technological
University**

Michigan Technological University
Digital Commons @ Michigan Tech

Dissertations, Master's Theses and Master's Reports

2016

Data Management and Data Facilitation in Multi-Center Large Cohort Health Care Studies

Wei Wang

Michigan Technological University, wwang13@mtu.edu

Copyright 2016 Wei Wang

Recommended Citation

Wang, Wei, "Data Management and Data Facilitation in Multi-Center Large Cohort Health Care Studies", Open Access Master's Thesis, Michigan Technological University, 2016.
<https://doi.org/10.37099/mtu.dc.etr/303>

Follow this and additional works at: <https://digitalcommons.mtu.edu/etr>



Part of the [Environmental Public Health Commons](#), and the [Maternal and Child Health Commons](#)

**DATA MANAGEMENT AND DATA FACILITATION IN MULTI-CENTER
LARGE COHORT HEALTH CARE STUDIES**

By
Wei Wang

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Medical Informatics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2016

© 2016 Wei Wang

This thesis has been approved in partial fulfillment of the requirements for the Degree of
MASTER OF SCIENCE in Medical Informatics.

School of Technology

Thesis Advisor: *Jinshan Tang*

Committee Member: *Yu Cai*

Committee Member: *Yushin Ahn*

Committee Member: *Min Wang*

School Dean: *James Frendewey, Jr.*

Table of Contents

List of Figures	v
List of Tables	vii
Acknowledgment	viii
Abstract	x
Chapter 1. Introduction	1
1.1 Background and Significance	1
1.2 ECHO DMMC and ECHO study cohorts	3
1.2.1 The Fragile Families and Child Wellbeing Study (FF).....	7
1.2.2 The Early Life Exposures in Mexico to Environmental Toxicants (ELMENT)	8
1.2.3 The Center for Oral Health Research in Appalachia (COHRA).....	11
1.3 Data Management Plan	13
Chapter 2. Data Communication	17
2.1 Semi-automatic online data request process	17
2.2 Securely transfer data and privacy protection	21
2.3 Data follow-up communication.....	21
2.4 Data use permission and agreement.....	22
Chapter 3. Data Harmonization, Data preparation, and Data documentation...	25
3.1 Introduction to the Data Structure of ELEMENT Study	26
3.2 Our Plan for Data Harmonization and Data Documentation	30
3.3 A Case Study for Data Preparation Process: To investigate how the environmental exposures such as prenatal and childhood lead exposure affect adolescent's sexual maturation.....	34
3.3.1 Study goal	34
3.3.2 Fill in online request form.....	35
3.3.3 Data allocation in the database.....	36
3.3.4 Data harmonization and documentation	40
Chapter 4. Descriptive Statistics and Data Quality Control	43
4.1 Basic Statistical Tools in SAS for Descriptive Statistics and Detecting Potential Outliers.....	43

4.2	Example of Data Quality Control when Preparing Data for investigating how the environmental exposures such as prenatal and childhood lead exposure affect adolescent's sexual maturation.....	44
	Chapter 5. Summary.....	53
	References	54
	SAS code Appendix	62

List of Figures

Figure 1. Data collection procedure over time and the three waves of study cohorts of the ELEMENT study	9
Figure 2. The flow chart of ECHO studies, ECHO data, and their connections with our data management and modeling core	13
Figure 3. The flow chart of data collection, data entry, data management and data analysis in ELEMENT study	30
Figure 4. The new structure of our database with four modules and available sets of functional variables stored in each module for the ELEMENT study.....	32
Figure 5. Self-reported questionnaire for girls' sexual maturation in ELEMENT	37
Figure 6. Self-reported questionnaire for boys' sexual maturation in ELEMENT.	38
Figure 7. Boxplot of Child Blood Lead Over Time, with child's birth as the reference time.....	49
Figure 8. Boxplot of Mom Blood Lead Over Time, with child's birth as the reference time	50

Figure 9. Spaghetti plot of mom blood lead exposure trajectory over time, stratified by different measurements of sexual maturation stages	51
Figure 10. Spaghetti plot of child blood lead exposure trajectory over time, stratified by different measurements of sexual maturation stages	52

List of Tables

Table 1. ELEMENT Data Collected Across Children’s Life Stages for Cohorts 1, 2, and 3.	28
Table 2. ELEMENT Raw Database Collected over 2000 subjects.	31
Table 3. Descriptive Statistics of maternal characteristics and children’s characteristics stratified by girl’s self-reported stage of breast development.	45
Table 4. Descriptive Statistics of maternal characteristics and children’s characteristics stratified by girl’s self-reported stage of pubic hair development	46
Table 5. Descriptive Statistics of maternal characteristics and children’s characteristics stratified by boy’s self-reported stage of penis and scrotum development	47
Table 6. Descriptive Statistics of maternal characteristics and children’s characteristics stratified by boy’s self-reported stage of pubic hair development	48

Acknowledgment

First and foremost, I would like to thank my advisor, Dr. Jinshan Tang, for guiding me on my graduate education and Master thesis work. He has provided me a lot of valuable advice when I got problems. I am deeply grateful for his critical guidance and endless support. I am very honored to have Dr. Tang as my advisor. I have learnt a numerous amount from him over the past couple of years.

I would like to extend my gratitude to Dr. Peter Song, my Supervisor during my co-op at The University of Michigan, for supporting me with this interesting research topic. The experience of working with Dr. Song's team is really a great pleasure and valuable. I am truly thankful for their guidance and advice, which have been very helpful on my academic pathway.

I also own a special note of appreciation to my thesis committee members: Dr. Yu Cai, Dr. Yushin Ahn and Dr. Min Wang. I would like to take this opportunity to express my deep gratitude for their help and support. I would also like to thank many faculty and staff members in the School of Technology at MTU, as well as my friends and classmates here. It has been the first time I am away from my home and away from my country. But you guys make me feel so friendly here. It has been a wonderful journey.

Last but not the least, I would like to thank my family for always being there for me. I am truly grateful to my parents, my sister, my brother, and my little nephew

for their believes in me and their constant support, love, and encouragement. This thesis is dedicated to them!

Abstract

Data management and data facilitation is the foundation and a critical part for scientific investigations, especially with large cohort studies on health care, which may have multiple components collected from different institutes or different areas due to efficiency, cost, enrollment considerations. In this thesis, I consider how to design and to implement data management with multi-center health care research data. The research project that I joined focuses mainly on environmental health science, and the primary goal is to investigate the impact of nutritional and toxicant exposures in environment during fetal and early postnatal life on individual's health outcomes over the life course. More specifically, I will present my work in a newly established research project, ECHO (Environmental Influences on Child Health Outcomes), on designing the data management plan and program codes for the data management and modeling core, facilitating and accomplishing data communication, cleaning and unifying data labels and formats, preparing datasets to achieve desirable data structures for further statistical analysis, and controlling data quality to ensure the overall research quality.

Chapter 1. Introduction

1.1 Background and Significance

Basic and social sciences have independently advanced conceptual frameworks grounded in a developmental paradigm as a basis for understanding how early life exposures influence health over the life course. However, social and chemical exposures are not independent of each other. It is important to investigate not only the direct effects but also the interactions of social context and chemical exposures at key developmental periods as the health scientists try to understand how these confer to long-term susceptibility to obesity and adverse neurodevelopmental outcomes. The ‘developmental origins’ hypothesis postulates that nutritional and toxicant exposures during fetal and early postnatal life influence developmental plasticity through epigenetic mechanisms, altering susceptibility to chronic conditions in later life [1]. The Environmental Protection Agency (EPA) incorporated the concept of life stage exposures into its risk assessment and regulatory framework [2, 3], recognizing that exposures during vulnerable periods may affect health outcomes that do not manifest until many years later. In their influential article on critical windows of toxicant exposure, Selevan, et al.[4] noted that research on exposures *in utero* is more advanced than that considering adolescence, a biologically vulnerable period characterized by hormonal changes and maturation of many organ systems.

Perspectives from public health and social sciences underscore the relevance of multi-level influences during the perinatal and pubertal periods as well as other key developmental stages, including the transition from early to mid-childhood and from adolescence to young adulthood.

There has been a large body of literature working on environmental health, which has advanced the concept of the “exposome” that echoes developmental paradigms in basic and social sciences. At its essence, the exposome is the environmental equivalent of the genome and encompasses all lifetime exposures including chemical, lifestyle, dietary and other physical influences. Multiple definitions exist [5-8], all highlighting the broad range of exposures that may shape health and disease across the life course. Improving detection of internal biomarkers of exposure and response is an area of rapid technological development, but approximating the entire exposome may be decades away [6, 9]. Integrating data on biological responses to social and chemical exposures across multiple developmental time periods is essential to develop a more complete picture of the exposome.

Understanding how biologic responses to exposures may impact obesity, cognition and behavior requires, nevertheless, the selection of mediators that reflect change across developmental trajectories. Current evidence links social and toxicant exposures during discrete developmental windows to risk of obesity and/or poor neurocognitive health. Few birth cohorts with repeat toxicant exposures over

sufficient length of follow-up also collect nuanced measures of socioeconomic status (SES) and parenting. Thus, our ability to understand how interactions of social context with chemical exposures may affect neurodevelopment and risk of obesity and metabolic sequelae during multiple sensitive periods is limited. To address these challenges, a multi-center large cohort study has been proposed by leveraging the research infrastructures and data from environmental influences on Child Health Outcomes (ECHO). This study is a multi-center large cohort study, which includes three specific study cohorts: the Fragile Families and Child Wellbeing Study (FF), Early Life Exposures in Mexico to Environmental Toxicants study (ELEMENT), and Center for Oral Health Research in Appalachia study (COHRA). This dissertation work is mainly focusing on data management and data facilitation in this multi-center large cohort study. Below, I will introduce specifically more details of ECHO data management and each ECHO study cohort.

1.2 ECHO DMMC and ECHO study cohorts

ECHO (environmental influences on Child Health Outcomes) is a multi-center research investigation proposed to understand how the social and chemical environment is associated with the biophysiologic mediating mechanisms and how it may affect the obesity-related and cognitive outcomes from fetal life through the transition to young adulthood. This study involves four research institutes and collected data from three study cohorts with repeated measurements measured

longitudinally. The Fragile Families study has 925 participants; the Early Life Exposures in Mexico to Environmental Toxicants study has 800 participants; and the Center for Oral Health Research in Appalachia study has 1280 participants.

My work belongs to the data management and modeling core (DMMC) of ECHO. My primary goal is to provide high quality data management on these 3005 children to further enable scientific discovery, and to organize data ultimately useful in informing evidence-based interventions. In the research plan of ECHO, a sub-cohort of 200 participants from ELEMENT study and FF study will obtain richly characterized intergenerational exposure measures, including exposome and epigenome data. Each of the three studies has already collected large amounts of high quality, validated data on a broad set of areas. For example, all three studies already have measures of obesity, neurodevelopment, social variables, and so on. ECHO will collect new data, and further harmonize a variety of data across the three study cohorts: FF, ELEMENT, and COHRA.

The Fragile Families and Child Wellbeing Study (FF) is a landmark, nationally representative child development study of the effects of poverty, parenting and other social exposures on child health, obesity, cognitive and social development, school readiness, and academic achievement. The Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) is an award-winning cohort study, where the chemical exposures of ELEMENT participants have been characterized from the

prenatal period, and the health outcomes of interest include cognition, behavior, sexual maturation and an array of obesity-related outcomes, including metabolomics; the Center for Oral Health Research in Appalachia (COHRA) study is collecting a unique cohort of mother-infant pairs from northern Appalachia to evaluate the factors predisposing to disproportionately high rates of poor oral health.

I am mainly involved in the ECHO data Management and Modeling Core (DMMC). ECHO DMMC plays a central role of liaison in coordinating and giving updates on the progress of data collection, data QC, data transfers, and data analysis results, and publications among cohorts. Professional data managers and statisticians handle the data analysis part. In this dissertation, I designed a data management plan for ECHO DMMC. Since not all the data are available yet and some sub-studies are still collecting data, my major focus is on data processing from the ELEMENT study to illustrate how the data management and data organization process as proposed work. My main responsibilities in this work include data management, data integration, and data harmonization, which will handle parts of the existing databases from ELEMENT, COHRA and FF studies and new data to be collected to address ECHO's proposed scientific aims. My focus is to design a data management plan and provide some strategies to establish the centralized data coordination handling data inquiries pertaining to database from two or more cohorts, while individual cohort databases are stored in separate study sites. The data management strategies we are working on can maximally utilize the existing data management

infrastructure to support the ECHO proposed research activities. Specifically, I designed both online data require form and feedback form to facilitate data communications. I wrote the SAS [10-16] codes to implement data harmonization, data quality control (QC), and data transfer. SAS is one of the most popular statistical software. In this thesis work, I especially focused on SAS data process, SAS PROC IMPORT/EXPORT, SAS PROC SQL and some basic SAS STAT procedures. My work contributes to the goal of the part of ECHO DMMC, which is to manage all raw lab data including data storage, data QC and data transfer. Since the data are collected by different staffs at different research institutes, each institute may have multiple sites to enroll patients, each sub-study may include several waves of patients with data collected at different times, different ages, and repeated measures are collected over the time longitudinally, as one can imagine, one of the key functions of our DMMC is the coordination among the existing individual cohort data management teams, and integration of meta data and data elements needed for the proposed study aims. My work in this dissertation covers all of these aims, with a major focus on the multiple sub-cohorts in ELEMENT study. I will illustrate the data management process and some data analysis process through a research investigation on how prenatal and childhood lead exposure affect adolescent's sexual maturation.

1.2.1 The Fragile Families and Child Wellbeing Study (FF)

The Fragile Families and Child Wellbeing Study (FF) is a landmark, nationally representative child development study of the effects of poverty, parenting and other social exposures on child health, obesity, cognitive and social development, school readiness, and academic achievement. This birth cohort study enrolled approximately 4900 children born in large U.S. cities between 1998 and 2000. The study includes: (1) interviews with mothers and fathers at birth and again when children are 1, 3, 5, 9 and 15 years old; (2) medical records for a subset of mothers and children at birth; (3) in-home assessments of children and home environments when children are 3, 5, 9 and 15 years old; (4) interviews with children at age 9 and 15; (5) interviews with teachers when children are 3, 5, and 9 years old; and (6) saliva DNA samples from mothers and children when children are age 9 and 15, and neuroimaging data from a subset of children at age 15. When the corresponding sampling survey weights, the data are representative of all births in all US cities with populations of 200,000 or more. Additional weights make the data representative of births within each city [17]. Response rates for mothers at baseline, calculated as a percentage of eligible births, are 86% at year 1, 3, 5, 9, 15 interviews, and for children, calculated as a percentage of eligible respondents at each wave, are 90, 88, 87, 76, 91% respectively. Response rates for DNA at the year 9 interview are 85% for children and 79% for mothers; current age 15 response rates appear slightly higher. ECHO plans to collect a new wave of assessments, including blood and urine

collection on three of the twenty cities involved in FF (n=925): Detroit, Michigan and Oakland and San Jose, California. These three cities were chosen because they are in states that afford access to neonatal bloodspots. All participants were recently or are currently being interviewed at age 15. By the start of planned data collection the oldest would be 20 years old. Birth and early life outcomes will also be collected for the children of female FF participants. Of the approximately 430 females in the original sample for these three cities we conservatively expect approximately 175 births throughout the entire project timeline based on the race (10% non-Hisp white, 47% Black, 45% Hispanic, 3% other), SES (30% <100% the Federal Poverty line (FPL), 35% 100-200% FPL, 35% 200%+ FPL) and Age (Mom participants will be 19-24 during the data collection period).

1.2.2 The Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT)

The Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) study comprises three sequentially-enrolled epidemiologic birth cohort studies in Mexico city with the original aim of investigating the influence of lead exposure on various outcomes related to fetal and infant development. Pregnant women were recruited at maternity clinics of the Mexican institute of Social Security: Cohort 1 was recruited from 1994 to 1997; Cohort 2 from 1997-2001; and cohort 3 from 2001 to 2005. In 2006-2012, study continued follow-up of all 3 cohorts to assess the impact of

prenatal lead exposure on neurobehavioral outcomes with specific attention to genetic influences of cholesterol metabolism genes. Figure 1 below shows the data collection procedure and the timelines.

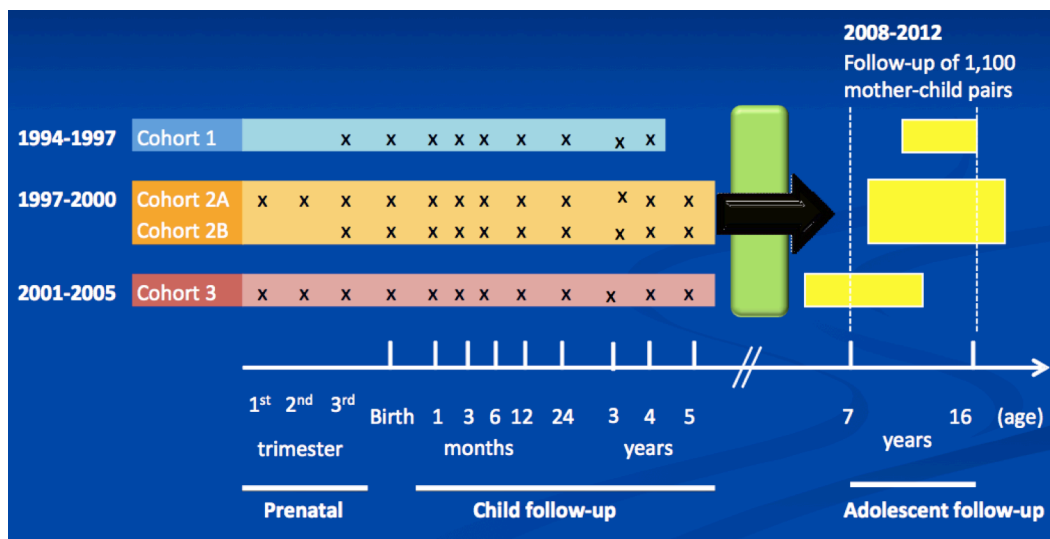


Figure 1. Data collection procedure over time and the three waves of study cohorts of the ELEMENT study.

All three ELEMENT cohorts have the same source of subjects (maternity clinics), eligibility criteria and methods when evaluating both fetal and postnatal toxicant exposures, covariates and outcomes. ELEMENT participants are generally low to middle-class and representative of the general population in Mexico City. All three cohorts of ELEMENT have archived biological samples that have been well-validated for chemical analysis and retrospective exposure assessment, a deliberate strategy to ensure the comparability of these subjects and associated data. Our work

will further organize the ELEMENT data in a better way and easily allow us to pool data across the ELEMENT population to address new scientific investigations. The ELEMENT study has been very successful, with its rigorous study design, outstanding follow-up rates and multidisciplinary expertise of their team, ELEMENT study group has generated over 60 international publications, including high impact research on cognitive outcomes related to prenatal exposures to lead [18-20] and gene-environment interactions [21]. Evidence from ELEMENT studies has informed U.S. lead exposure guidelines including many recent new policy statements [22-25]. The current outcomes collected in ELEMENT study include cognition and behavior in mid-childhood, obesity and metabolic outcomes including metabolomics in peripuberty as well as sexual maturation of physician-observed Tanner staging and hormone assays. The careful collection and storage of biological samples has allowed the study team to use archived samples to retrospectively assess a variety of exposures in the studies of the toxicity of manganese, bisphenol A, phthalates, and mercury and so on. Thus, the ongoing funding related to ELEMENT has enabled continued follow-up of participants, to assess the impact of EDC mixtures and their interaction with diet on growth, maturation and metabolic outcomes via epigenetic mechanisms. Potentially, a new wave of assessments, including structural brain MRIs at one time point will be collected on 800 ELEMENT participants currently aged 9-17 years. When combined with the other existing assessment collections of ELEMENT, these 800 participants will each have

at least two assessments during the transition late adolescence into young adulthood.

1.2.3 The Center for Oral Health Research in Appalachia (COHRA)

The Center for Oral Health Research in Appalachia (COHRA) study collected a unique cohort of mother-infant pairs from northern Appalachia to evaluate the factors predisposing to disproportionately high rates of poor oral health. COHRA applies a rigorous multidisciplinary strategy to understand the interrelated factors leading to poor oral health in Northern Appalachian children. To achieve this, a child- and family-centric study design is employed to combine assessments of individuals—their genomes and their personal “environments” (e.g., oral health status, oral microbiome, diet, behavior), with family- and community-level variables. The ultimate objective of COHRA is to determine targets for interventions that can alter the oral health trajectory begun in infancy, thus potentially improve oral health throughout life.

COHRA has recruited two samples from West Virginia and western Pennsylvania: COHRA1, a cross-sectional cohort of nuclear families (recruited from 2003 – 2009; [26]), and COHRA2, a longitudinal cohort of healthy mother-baby pairs recruited during pregnancy and followed until the children are at least age six (recruited from 2012-present). Trained and calibrated dental hygienists see the COHRA2 women and their children at least yearly, beginning before the children are born. In addition, the University of Pittsburgh Center for Social and Urban Research (UCSUR)

administers an extensive phone questionnaire to all women every six-months. New recruitment is ongoing, as are longitudinal follow-up protocols.

COHRA has produced many high quality peer-reviewed publications, which include the first GWAS of any dental disease—dental caries in the primary dentition [27]. COHRA data are part of the GLIDE consortium study of periodontal health and BMI [28], and also investigated the effect of sibship and caries status on the salivary metabolomes of sib pairs with different levels of dental decay [29], caries resistance as a function of age [30], demographic, SES, and behavioral factors [31], the nature of dental fear and/or anxiety [32, 33], sex differences in caries susceptibility [34, 35]; gene-environment interaction between enamel matrix genes and fluoride exposure [36], a gene-gender interaction [37], a gene-smoking interaction in periodontal disease [39], interaction between depression and rurality for several oral health indicators [39], and etc.

Continued follow-up are planned on 1280 current COHRA participants and additional assessments will be collected at 5 and 8 years to coincide with sensitive developmental periods for obesity and cognition and socioemotional behavior, as well as measures of these outcomes and to provide comparable data on exposures (e.g., with the ELEMENT study).

1.3 Data Management Plan

Since this is a multi-center large health care studies cohort, our data management and modeling center plays a central role in coordinating and providing updates on the progress of data collection, data QC, data transfers, and data analysis results among cohorts.

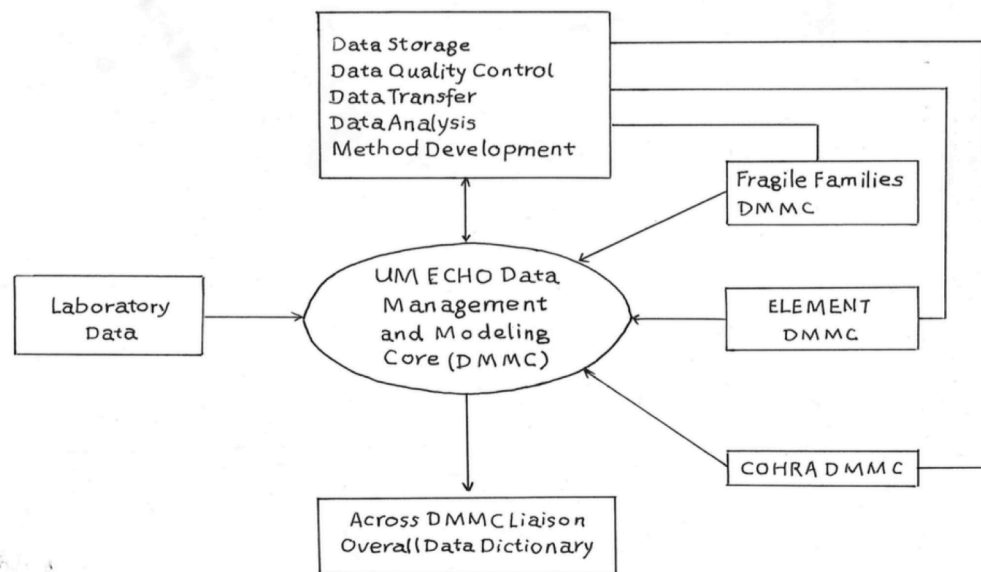


Figure 2. The flow chart of ECHO studies, ECHO data, and their connections with our data management and modeling core.

One of ECHO DMMC main role is to handle data management, integration and harmonization. FF, ELEMENT, and COHRA studies currently maintain databases that contain data elements beyond the scope ECHO-DMMC. Hence, the data

managed as part of this ECHO- DMMC includes parts of the existing databases from FF, ELEMENT, and COHRA and new data collected to address the proposed scientific investigations. The ECHO DMMC will lead centralized coordination to handle data inquiries pertaining to data from two or more cohorts, while individual cohort databases are stored in separate study sites. In this way, the existing data management infrastructure will be maximally utilized to support the newly proposed research activities, and not duplicated in the ECHO DMMC. The ECHO DMMC will manage all data required to address the specific study aims, including data storage, data QC and data transfer of newly collected data, and data harmonization of existing data elements. To maximize efficiency, each cohort-specific DMMC will be responsible for managing new questionnaire-based data from the existing cohort-specific participants, which will then be combined by the ECHO DMMC for analysis. The ECHO DMMC will manage all raw lab data including data storage, data QC and data transfer. Figure 2 present an organizational chart in which one of the key functions of the ECHO DMMC includes the coordination among the existing individual cohort data management teams, and integration of meta data and data elements needed for the proposed study aims.

Each individual cohort DMMC will include faculty statistician, a database manager and an analyst, covering basic data management needs, including integrating new data to be collected in this proposal to the cohort- specific databases, and basic data analysis of each cohort. These personnel already thoroughly understand the data

structure and contents of the individual cohorts and will also have key responsibilities related to data harmonization and integration for the ECHO projects. They will collaborate with the ECHO DMMC to create and maintain the overall data dictionaries and to extract and transfer any requested data used to fulfill analysis goals of the current proposal. As needed, each individual cohort DMMC will conduct data analysis of cohort- specific data and develop or apply statistical models needed to fulfill analytic needs related to this proposal. In terms of coordination, each individual cohort DMMT will be responsible to inform the ECHO DMMC of any updates and progress on data collection, data validation, data analysis results and publications for each project. In addition, the sharing of statistical models and methods used in cohort- specific analysis will be communicated with the statisticians serving the other cohorts.

Professional data managers and statisticians handle the data analysis part. For the part of data inquiry and communication, data harmonization and documentation, and data quality control defines my training spectrum and thesis. I ensure that dataset is adequately prepared with the appropriate variables in SAS format for further data analysis, and get the data easily understood by both research investigators (clinicians, medical professionals, professors, and researchers) and statisticians (those who perform data analysis). The major procedures that I used in SAS to deal with data management include SAS data process, SAS PROC IMPORT/EXPORT and SAS PROC SQL. I will also perform some basic analysis to ensure data quality

control and provide fundamental statistics using SAS stat procedures.

Chapter 2. Data Communication

As we have seen from the introduction in Chapter 1, this is a typical large cohort study with data collected at multiple locations, during multiple study waves, and at multiple longitudinal time points. To facilitate the communication between scientific investigators and the data management group, we design data request forms and data feedback forms for people who are interest using our existing meta database for further scientific investigations. At the same time when our data management group is preparing the desired data structure and dataset for further analysis, we need pay attention to the security when transferring data, and especially when connecting to different database servers at remote locations. After this, we dynamically modify and follow-up data communication online request form to allow more convenience and a more systematic data management procedure. Furthermore, we need to establish the appropriate notice and data documentation, which help explain the data usage right and permission.

2.1 Semi-automatic online data request process

According to what is potentially related to people's interest, and after several data process procedures handled manually face to face with the study team, I found it would be more efficient if we can develop some semi-automatic procedure for data request before face to face meetings. This will help both the scientific investigators

and the data manager to make things clarified, documented, and in the meanwhile save a lot of time from travelling to meetings.

The welcome page describes what this website is for, and you can find brief descriptions of the existing raw data structures and raw data contents. Along with this welcome page, I also provide some guidelines for the data file that will be generated by our data management group. For example, where are the variable names, how to code missing data, where to find the detailed data dictionary, and what the data dictionary contains, and some common coding of the demographics variables.

Welcome to our DMMC (Data Management and Modeling Core)!

You can request a dataset by providing the following information. We will prepare the dataset accordingly, along with our priority list. It may take a few days, but we will send it back to you as soon as the dataset is ready.

Please note some guidelines for data structure as you receive the requested dataset:

1. The first row contains a variable name.
2. Missing data are coded as NA.
3. Each data will come with a detailed data dictionary, including what each variable means, the units of continuous measurements, the value of each level for categorical variables, etc
4. Some common coding for demographics variables:

- 1) Gender: female=0, male=1
- 2) Race: 1=White 2=African American 3=Hispanic 4=Other
- 3) education level: 1=high school or less 2=two year college 3=four year college 4=post graduate degree

Then the investigator will complete the online questionnaire regarding the study they propose to investigate. Thus, we at the DMMC will know what variables from which cohort may potentially be of their study interest. We will merge and recreate new dataset for the investigators to use according to their need. Some sample survey questions are:

Which of the following databases do you need to use?

- Fragile Families (DMMC) include about 15 years data
- ELEMENT (DMMC) include about 20 years data which is related to chemical element
- COHRA (DMMC) include about 10 years data which is related to children's oral health
- Other database (please specify)

What is your project title?

Main Scientific Objective?

Principal Investigator:

Specific Scientific Questions (please enumerate):

Outcome (dependent variable):

Predictors (independent variables):

Potential Confounders:

This semi-auto data request form can be completed online at the DMMC website, and the completed form will be automatically sent to a specific data manager's mailbox. Once ECHO DMMC receives the completed data request form, ECHO DMMC manager will go over all existing raw data files and locate the appropriate study cohort(s) correspondingly, and then organize a new dataset or several datasets if needed according to years or other specified data request requirements. This semi-auto procedure can be done easily and quickly without too much interactions between personnel. Following this step, the investigator and the data manager can sit down and meet in person or over the phone with the semi-auto prepared dataset(s) of

their potential interest in front of them and see if they need to modify, further generalize or detailed tailoring of the data.

2.2 Securely transfer data and privacy protection

For complex computer world, hackers are more techs savvy and can infiltrate networks by exploiting vulnerabilities. Especially in the case when people's health care data are involved, it is very important to ensure data security and privacy protection. For ECHO DMMC, we consider the following methods that can help us achieve the aim of keep data security and confidential data privacy. First, before manipulating data for further structure changes, we make a copy of the raw data and safely save raw data with password protections, and even on private computer/server isolated from any Internet or other user use. Secondly, when carrying out the processing procedure, we always set safe password for connection servers. For example, with any remote connection to different locations when searching data, we always use SAS secure link/connections with password saved in a separate file directory. Another idea is setting encryption competence measure for ECHO DMMC data manager.

2.3 Data follow-up communication

The semi-auto process has been in use since I developed the online request form on their website this summer. Several studies have been proposed and got the dataset

generated by that way. We further developed a series of follow-up communications between data manager and scientific investigators to monitor the data creating process, data safety issues, and collect feedbacks from both sides on how to improve the process.

2.4 Data use permission and agreement

As many other study/protocol/grant, a data use permission and data use agreement (DUA) is necessary to ensure data safety and data privacy during the transfer of data, especially when the data is nonpublic or is otherwise subject to some restrictions on its use. Typically, this is a contractual document on the agreements that is required under the Privacy Rule and must be entered into before there is any use or disclosure of a limited dataset to an outside institution or party. Like in our study, any data created by our data management group is a necessary component of the research project and it does contain human subject data. So the institutes involved in ECHO want to ensure that DUA terms protect confidentiality when necessary, but permit appropriate publication and sharing of research results in accordance with University policies, applicable laws and regulations, and federal requirements. Therefore, the team established the corresponding data use permission and agreement to help restrict the use and disclosure of the data set, and, under these restrictions, our DMMC team prepare and unify the format of newly created datasets in the form appropriate for the transfer of data.

Principal Investigator (PI) creates and submits the data request in the secure intranet online system. To avoid any delays in processing, the PI/PI's department should ensure that the PI has completed his/her annual Conflict of Interest (COI) disclosure using the Faculty/Researcher form. Upon electronic receipt of a new submission, the project coordinator will begin initial submission review and will contact the PI or their administrative contact if any of the required documentation is missing or incomplete. Then data use agreement will be circulated for signatures. When all of the required documentation has been received completely, the data request will be assigned and forwarded to a data manager at DMMC officer for review and data preparation. The data use permission and agreement file may include the components similar as follows:

- Establish the permitted uses and disclosures of the limited data set;
- Identify who may use or receive the information;
- Prohibit the recipient from using or further disclosing the information, except as permitted by the agreement or as otherwise permitted by law;
- Require the recipient to use appropriate safeguards to prevent an unauthorized use or disclosure not contemplated by the agreement;
- Require the recipient to report to the covered entity any use or disclosure to which it becomes aware;

- Require the recipients to ensure that any agents (including any subcontractors) to whom it discloses the information will agree to the same restrictions as provided in the agreement; and
- Prohibit the recipient from identifying the information or contacting the individuals.

Chapter 3. Data Harmonization, Data preparation, and Data documentation

In this chapter, I will present details about the data process procedure that I have done every time I receive a data request. Following by the data request form, I will locate what parts of the whole data collection can be used to establish the specific scientific investigation according to their main scientific objectives and specific scientific questions. They mainly have listed their targeted outcome and potential predictors and confounds that they want to include. I will also go through the whole study and see if there is anything they may have missed and will add in, just for their potential interest. After I located every piece and where they are saved in the raw datasets, I will employ SAS DATA process, SAS PROC IMPORT/EXPORT, and SAS PROC SQL to perform Data Harmonization and Data preparation. I will also use both SAS and Word to prepare the corresponding Data archive files and Data documentations. In the following, I will mainly focus on the data management in ELEMENT with multiple waves and repeated measurements, as this part of the data is ok to present at this moment while all other part of the data are still confidential outside the study team. I will specifically illustrate the whole process through a specific project, which is mainly proposed to evaluate the effect of the environmental

exposures such as prenatal and childhood lead exposure on adolescent's sexual maturation.

3.1 Introduction to the Data Structure of ELEMENT Study

ELEMENT is comprised of three birth cohorts from Mexico City maternity hospitals that have been successfully followed for over two decades. The study cohorts (mother-infant pairs) were recruited from three maternity hospitals in Mexico City that serve a low- to moderate-income population (Mexican Social Security Institute, Manuel Gea Gonzalez Hospital, and National Institute of Perinatology). The initial cohort began with the support from the Harvard Superfund Basic Research Program and individual R01 grants and was an inter-institutional collaboration among Harvard University, the Center for Population Health Research of the national Institute of Public Health in Mexico, the American British Cowdray Medical Center and the National Institute of Perinatology of Mexico, with approval from each study location's Institutional Review Board (IRB). With the movement of multiple ELEMENT investigators from Harvard University to the University of Michigan between 2006 and 2008, and from the University of Michigan to the University of Toronto, the ELEMENT cohort has evolved into a collaboration among researchers at University of Michigan, Harvard, and the University of Toronto with continued interaction with all partner agencies in Mexico. At University of Michigan, School

of Public Health (UM-SPH) data were extracted or newly collected from the existing three birth cohorts.

The mother-child pairs of study participants were recruited over a series of years: Cohort 1, 1994-1997; Cohort 2, 1997-2000; Cohort 3, 2001-2005. Data was collected at various times throughout the prenatal period, early- to mid-childhood and adolescence in Cohorts 1, 2, and 3, with significant overlap in data collected during pregnancy, infancy and early childhood through age five (See Figure 1 on Page 9). A fourth cohort began in 2007 under a similar study design. Some of the information collected is common between cohorts and some differs slightly, as measures have evolved over time for improved accuracy and precision. Table 1 below is a general compilation of data collected across the children's life stages for all ELEMENT cohorts, so specific variables may not be available in all subjects. Since ELEMENT is an ongoing, longitudinal cohort, additional measures may be added in future studies. Table 2 reflects data that has been collected from 1994 through 2014, which provides a rich source of secondary data potentially proposed by the investigators.

All data collection for the ELEMENT studies occurs in Mexico, facilitated via collaboration with the Instituto Nacional de Salud Publica (INSP) that has existed since the longitudinal cohort was initiated in 1994. Study subjects include mother-child pairs who were originally recruited for ELEMENT participation before or duri-

Table 1. ELEMENT Data Collected Across Children’s Life Stages for Cohorts 1, 2, and 3

Child’s Lifestage	Study Data Collected	
	Maternal Measures	Child Measures
Prenatal	<ul style="list-style-type: none"> • Socio-economic status (2 PL • Urine • Venous blood (2 PL • Food frequency questionnaire (FFQ) • Socio-economic status questionnaire 	
Perinatal	<ul style="list-style-type: none"> • Socio-economic status • Patella and tibia bone lead • Venous blood • Urine • Anthropometry 	<ul style="list-style-type: none"> • Birth weight • Birth length • Umbilical cord blood • Bayley Scales of Infant Development II (Spanish)
Early Childhood (0-5 years old)	<ul style="list-style-type: none"> • Anthropometry • FFQ • Demographic information • Socio-economic status • Maternal IQ 	<ul style="list-style-type: none"> • Anthropometry • Bayley Scales of Infant Development II (Spanish) • Mental Development Index • McCarthy Scales • FFQ
Mid Childhood to Adolescence (7-16 years old)	<ul style="list-style-type: none"> • Screening questionnaire • Sociodemographic questionnaire • FFQ • Lead/Metals exposure questionnaire • GPS coordinates • HOME Inventory • Blood • Hair • Nails • Urine • Anthropometry (Height, weight, waist, chest, arm, and calf circumferences) • Blood pressure (simplified) • Skinfold thickness • Psychometric tests (BASC-2, CADS, BRIEF, Structured Interview) 	<ul style="list-style-type: none"> • Anthropometry (Height, weight, waist, chest, arm, and calf circumferences) • Blood pressure • Skinfold thickness • Bioelectrical impedance • Stage of sexual maturation • FFQ • Psychometric tests (CANTAB, WRAVMA, BASC, CADS, BRIEF, WASI, Connors’, Pre-Pulse Inhibition) • Blood • Hair • Nails • Urine • DNA (saliva) • Accelerometry data

ng pregnancy or shortly after giving birth, depending on the cohort. Current and future studies based on this longitudinal cohort re-recruit mother child pairs to explore related research study aims and continue to recruit new birth cohorts. Primary data is collected in multiple locations in Mexico, depending on the specific study requirements; these study locations include La Casita, the American British Cowdray Medical Center, and the Instituto Nacional de Perinatología (INPer).

Primary data for ELEMENT studies falls into three general categories: (1) data collected via filling out the Study Visit Data Collection Form or study questionnaires that is written on paper and directly entered into a database by the INSP Data Entry Technician or practitioners at one of the clinical study locations, (2) data collected from the visit that is directly entered by study practitioners into the computer via secure form and (3) biological specimens that require analysis prior to generating quantitative data that can be entered into a database. Examples of primary data that are recorded on paper during the study visit include: socio-economic status, demographic information, and responses to questionnaires. Data that is entered into a computer form during the study visit includes anthropometric data. Biological specimens may include urine, saliva, hair, nails, serum, and plasma. Multiple health outcomes can be assessed from each of these biological specimens and analyses vary by study.

3.2 Our Plan for Data Harmonization and Data Documentation

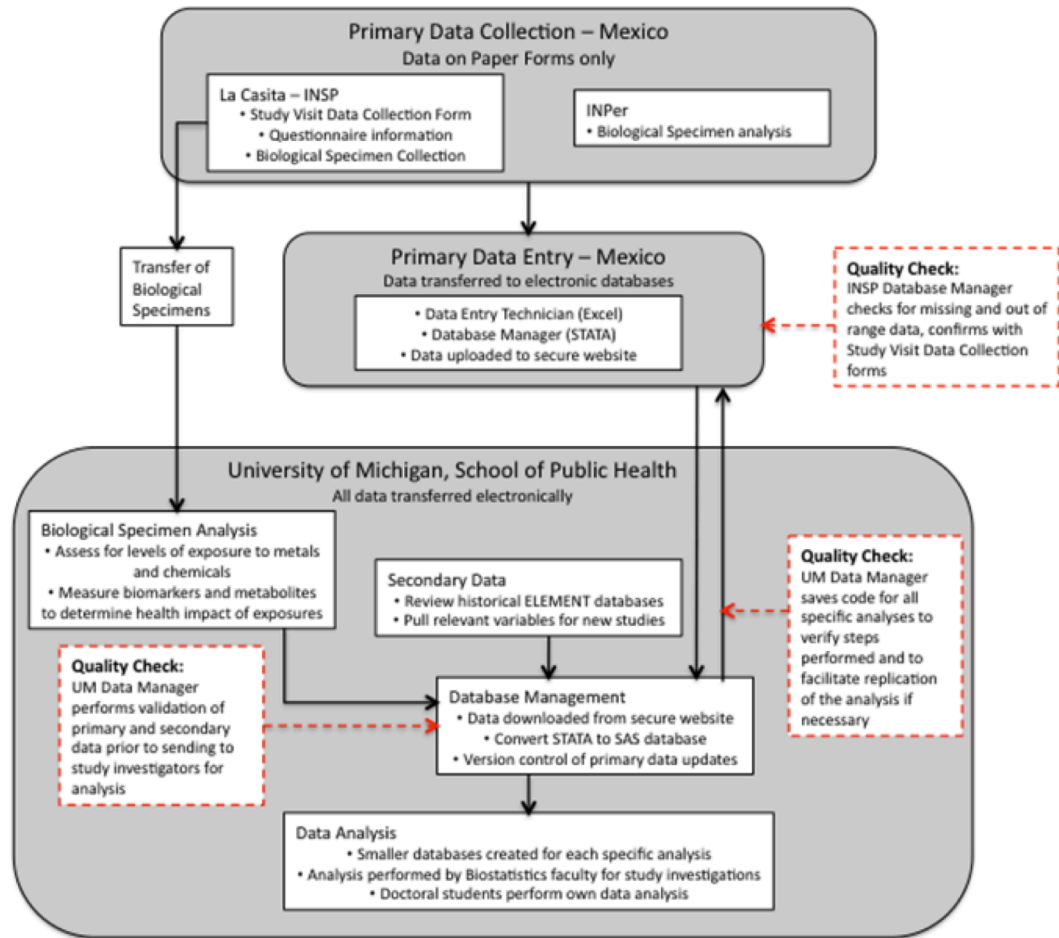


Figure 3. The flow chart of data collection, data entry, data management and data analysis in ELEMENT study.

Figure 3 presents the flow chart of data collection, data entry, data management and data analysis in the ELEMENT study. At the current stage, we have transferred all

the raw data in a secure, shared computer drive internally used by the research group members with authorization. The original folder was created for the multiple founded projects related to the ELEMENT cohort, and contains the most comprehensive historical data, including those from the earliest cohorts. A detailed description of all the cohorts can be found in the table below:

Table 2. ELEMENT Raw Database Collected over 2000 subjects.

Cohort 1 (1994-1997)	Cohort 2 (1997-2000)	Cohort 3 (2001-2005)
C1 – Cohort 1	PL – Plasma	SF – Super Fund
<ul style="list-style-type: none"> ▪ Followed 48 months ▪ Birth cohort ▪ Mother followed 12 months for breast feeding study ▪ Proyecto + filio 	<ul style="list-style-type: none"> ▪ Followed 60 months ▪ Pregnancy cohort 	<ul style="list-style-type: none"> ▪ Followed 60 months ▪ Pregnancy cohort
	BI – Biomarker <ul style="list-style-type: none"> ▪ Followed 60 months ▪ Birth cohort ▪ foliocc 	
Cholesterol (2006~2007)		
<ul style="list-style-type: none"> ▪ Subjects came from C1, PL, BI and SF 		
P20 (2010, n = 250)		
<ul style="list-style-type: none"> ▪ Subjects were subset of Cholesterol study (i.e., PL+SF+BI) 		
P01 (2015)		
<ul style="list-style-type: none"> ▪ New data collecting... 		

For security considerations, the raw data can only be accessed by the specific data manager assigned through the study team. All data are saved under a master Excel

sheet, where each separate Excel sheet is used to save data related to recruitment, surveys, and ultrasound data. Those files related to documentations, explanations, and presentations are also saved here and can be accessed in the same way. At DMMC, we have cleaned the data, unified the coding of common variables, created understandable labels for each variable in the dataset, converted them to SAS data files when needed, re-organized and restored the excel data files in the following way according to functional variables. Figure 4 shows the newly reconstructed data structure.

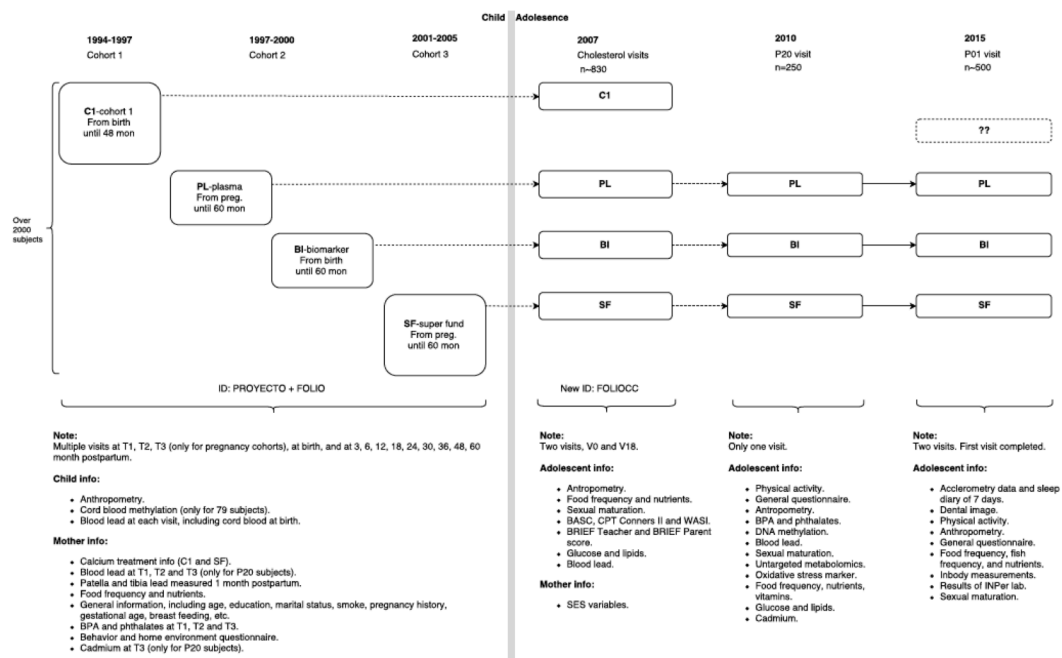


Figure 4. The new structure of our database with four modules and available sets of functional variables stored in each module for the ELEMENT study.

This new data structure has four modules: Historical (1994-2005), Cholesterol (2007), P20 (2010), and P01 (2015). Within each module, Figure 4 listed its sub-databases with specific functional meanings and documented the timing of measurements within that module. This new data structure is easier for data managers to locate the necessary variables of any newly proposed secondary data analysis or scientific investigation. In more details, my work in the group involved the following components:

- Create and maintain a common database for human studies related to the ELEMENT study cohorts
- Develop standard code for common outcomes, unify coding for common variables, create labels and unified format to harmonize the datasets
- Develop SAS code for data cleaning, data merging, and data reorganizations
- Develop standard code for data validation for data quality control
- Develop SAS codes to provide data statistics, and to perform simple data analysis.

Below I will illustrate my work through a concrete example, where the goal is to investigate how the prenatal and childhood environmental exposures such as lead exposure affect adolescent's sexual maturation.

3.3 A Case Study for Data Preparation Process: To investigate how the environmental exposures such as prenatal and childhood lead exposure affect adolescent's sexual maturation.

3.3.1 Study goal

ELEMENT uses self-reported and physician-assessed Tanner Staging for genital (boys), breast (girls), and pubic hair (both) development [40, 41] starting at approximately 8 years of age. FF collects Tanner Staging via self-report from the primary caregiver at age 9 and 15 years. COHRA children will have reached 8 years of age by the end of the proposed follow-up period; thus, at this point in time, the study plan to collect proxy-assisted Tanner Stage from the primary caregiver. By leveraging all these data from three studies in an appropriate way, one can evaluate how the prenatal and childhood environmental exposures such as lead exposure affect adolescent's sexual maturation. Obviously, since the data are collected by different study groups initially, and they have different formats and different codings. One major job of DMMC will include some time-consuming but necessary across cohort data harmonization and documentations. Since other data are confidential, we will only present the work and results done with ELEMENT data. Even within ELEMENT, as we described earlier, the data were collected during different waves on different cohorts, and at different locations, data harmonization is still a big challenge for further meta-analysis. The primary objective of data

harmonization in ELEMENT is to understand inter-cohort commonalities and heterogeneities, and to develop data harmonization and analytical approaches to address key data discrepancy issues.

3.3.2 Fill in online request form

By filling out the online data request form, the following major information has been provided by the study investigator, Dr. Peter Song, for the scientific investigation on how adolescent's sexual maturation can be potentially affected by the prenatal and childhood environmental exposures such as heavy metal exposures, lead exposures.

Project title: ELEMENT Sexual Maturation Study

Main Scientific Objective: How prenatal and childhood lead exposure affect adolescent's sexual maturation

Principal Investigator: Peter Song

Specific Scientific Questions:

1. Is there an association between Mom's prenatal lead exposure with kid's sexual maturation?
2. Is there a vulnerable window (like which trimester is more important)?
3. Is there a relationship between children's childhood lead exposure with kid's sexual maturation?

Outcome (dependent variable): sexual maturation outcomes (self reported, or physician diagnosed)

Predictors (independent variables): all kinds of Mom's lead exposures, kid's lead exposures, demographics.

Potential Confounders: kid's longitudinal measurements of health outcomes, age, gender, race...

3.3.3 Data allocation in the database

After I received the data request form, I went through all details filled in the form, and notice the main piont that this newly proposed secondary analysis is to investigate how prenatal and childhood lead exposure affects adolescent's sexual maturation. Then I dig into the massive database and try to locate the information that is related to express the levels each kid's sexual maturation. In ELEMENT, the sexual maturation inforamtion is collected for each kid through both self report questionnaires, and physician's examination and scoring. For physician scoring, they have a rigorous 4-page form to fill in, and all information is recorded in the data base from each question in that form. Specifically, the following variables are of our interest for the physician observed sexual maturation.

TN15_1: Physician observed pubic hair staging

TN15_2: Physician observed breast development (female only)

TN15_3: Physician observed genital development (male only)

TN15_4A: Physician observed testicle volume (male only), right

TN15_4B: Physician observed testicle volume (Male only), left

For girls, the self report questionnaire is as in Figure 5:

Questions on your physical maturation/development into adult.

We are going to ask you about your physical development. It is very important that you tell us about yourself honestly. This information will be kept confidential and only for study purposes.
There is no right or wrong answer.
Remember, you can choose to skip any question that you do not wish to answer.
For each question, mark your answer with a cross.

5. Have you started having periods?
₁ Yes
₂ No → skip question 6

6a. If Yes, at what age?
₁ 8 years ₅ 12 years
₂ 9 years ₆ 13 years
₃ 10 years ₇ 14 years
₄ 11 years ₈ 15 years or older

6b. AND at what month?
₁ January ₆ June ₁₁ November
₂ February ₇ July ₁₂ December
₃ March ₈ August ₁₃ Don't
₄ April ₉ September Remember
₅ May ₁₀ October

6c. What was the first day of your last menstrual period?
 Day ___ Month ___ Year ___

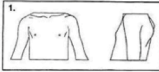
7. Do you shave or wax your pubic hair?
₁ Yes
₂ No → skip question 8

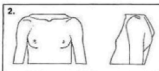
8. If Yes, in which area? (mark all areas where you shave or wax)
₁ On the inside of the thighs
₂ In the pubic area


Girls go through normal changes as they grow and develop into adults. Please look at the drawings and read the sentences near each of them.

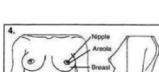
9. One of these changes is the breasts enlargement. Please choose the drawing closest to your stage of breast development by marking your answer with a cross.


10. Another change is the growth of pubic hair. Please choose the drawing closest to your stage of hair development by marking your answer with a cross.

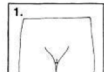
1.  The nipple is raised a little. The rest of the breast is still flat.


2.  The breast is a little larger and the nipple is raised a little more. The darker-colored area around the nipple (areola) is a little bit bigger than in Stage 1. Sometimes you can feel a small lump inside the nipples.


3.  The darker-colored area around the nipple (areola) and the breast are much larger than in Stage 2 and form a mound. The areola does not stick out away from the breast.

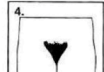
4.  The nipple and the areola stick out away from the skin and the rest of the breast.

5.  The nipple sticks out away from the skin and the rest of the breast. The areola is at the same level than the rest of the breast.

1.  There is no pubic hair.

2.  There is a little, long, lightly colored hair. This hair may be straight or a little curly.

3.  The hair is a little darker, coarser, and more curled. It has spread out and covers almost all the pubic area.

4.  The hair is now as dark, curly, and coarse as that of a grown woman. The hair covers all the pubic area but does not spread out to the thighs.

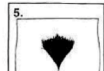
5.  The hair is like that of a grown woman (dark, curly, and coarse). The hair forms a triangle (▽) covering all the pubic area, and has spread out to the inner thighs.

Figure 5. Self-reported questionnaire for girls’ sexual maturation in ELEMENT.

The related variables that I locate in the database include the following:

- F5: Have you started having periods?
- F6: If yes, at what age?
- F6B_AA: First period (Year)
- F6B_MM: First period (month)
- F6C_AA: Last period (year)
- F6C_DD: Last period (day)
- F6C_MM: Last period (month)
- F9: Please choose the drawing closest to your stage of breast development by marking your answer with a cross

F10: Another changes is the growth of pubic hair. Please choose the drawing closest to your stage of hair development by marking your answer with a cross

The self-reported questionnaire for boys is as in Figure 6,

Questions on your physical maturation/development into adult.

We are going to ask you about your physical development. It is very important that you tell us about yourself honestly. This information will be kept confidential and only for study purposes.

There is no right or wrong answer. Remember, you can choose to skip any question that you do not wish to answer.

For each question, mark your answer with a cross.

5. Do you shave or wax your pubic hair?

₁ Yes

₂ No → skip question 6


6. If Yes, in which area?

₁ On the inside of the thighs

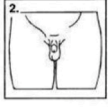
₂ In the pubic area

Boys go through normal changes as they get grow and develop into adults. Please look at the drawings and read the sentences near each of them.

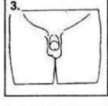
7. Please look at the penis and scrotum in the following drawings. Choose the drawing closest to your stage of penis and scrotum development by marking your answer with a cross.



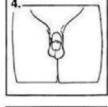
The scrotum and penis are the same size as when you were younger.



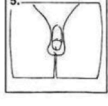
The scrotum has lowered a bit and the penis is a little larger.



The penis is much longer and the scrotum much larger.

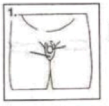


The penis is longer and wider. The scrotum is darker and larger than before.

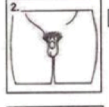


The penis and scrotum are the size and shape of an adult.

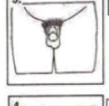
8. Another change is the growth of pubic hair. Please choose the drawing closest to your stage of development by marking your answer with a cross.
Please look only at pubic hair in the following drawings.



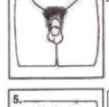
There is no pubic hair




There is little, long, lightly colored hair. Most of the hair is at the base of the penis. This hair might be straight or a little bit curly.



The hair is darker, coarser, and more curled. The hair has spread out and covers almost all the area above the penis.



The hair is as abundant, dark, curly, and coarse as that of a grown man. It covers all the area above the penis but has does not extend to the thighs.



The hair is as that of a grown man (dark, curly, and coarse) and has spread out to the inner thighs.

Figure 6. Self-reported questionnaire for boys’ sexual maturation in ELEMENT.

and the related variables that I locate in the database include the following:

M7: Your stage of penis and scrotum development

M8: Your stage of pubic hair development

38

Similar data information related to sexual maturation is saved within P20(2010) module separately in three different datasets (P20_Adolescent_Maturation_F.csv, P20_Adolescent_Maturation_M.csv, P20_Adolescent_Maturation_P.csv), as well as within P01(2015) module in one big dataset (P01_Adolescent_Maturation.csv with totally different variable names), and within Cholesterol(2007) module in one dataset but with simpler questionnaire designed at that time and thus less data information collected (Cholesterol_Adolescent_Maturation.csv with only similar questions to F9, F10, M7, M8 listed above).

Obviously, the main predictors for this project will need to include all longitudinal measurements of both Mom's prenatal lead exposures and kid's lead exposures in their early lives. These exposure measurements are all saved in the Historical(1994-2005) module. Specifically, we located two databases related to lead exposure records and extracted them into two data files: "Historical_Child_Blood_Lead.csv" and "Historical_Mom_Blood_Lead.csv", with much simpler data structure. One contains child blood lead level from month 1 to 60, and the other contains mom blood lead level from trimester 1 to time when child is 12 month of age. Besides, we also need to include other covariates in the database such as demographics(age, race, gender, weight, height, and BMI), and some other potential confounders, for instance, kid's longitudinal measurements of health outcomes, Mom's age of at delivery, smoking status of Mother, etc. Such information can be found in all four data modules: Historical (1994-2005), Cholesterol (2007), P20 (2010), and P01

(2015). The specific 4 datasets related to demographics and other health outcome records of either kid or mom are: “Cholesterol_Adolescent_Anthro.csv”, “Historical_Mom_General_Info.csv”, “P20_Adlescen_Anthro.csv” and “P01_Adlescen_Anthro.csv”.

3.3.4 Data harmonization and documentation

As we described in the previous session, the data saved in different modules may have different formats or different variable names, since they were collected during different time periods of the ELEMENT studies. Over the time as the design of the study gets improved, some new information is collected while it was not at an earlier stage or wave of cohorts. To achieve data harmonization, we (1) created overall data dictionaries of the existing data to transfer knowledge of the entire data availability to all investigators; (2) tried to understand commonalities and heterogeneities across different longitudinal study cohorts of ELEMENT, with respect to study design, data collection procedure, study population, measurements and cohesiveness of data sources across different cohort studies, and (3) identify any missing data, and examine the consistency, availability and variability of variables measured across different study cohorts, and develop data harmonization approaches to addressing key data discrepancy issues. After all these cleaning and harmonization work, I read in the corresponding files about the 11 datasets mentioned in Section 3.3.3 in SAS,

with unified data format, data structure, variable names, variable labels, and etc. According to the submitted data request, I selected 88 related variables to keep in the newly created datasets including kid's sexual maturation outcomes, Mom's lead exposures, kid's lead exposures, demographics of both mom and kid, and other potential confounders. Then I sort each dataset by ID and type of cohort. With this sorting, finally, I managed to merge those datasets according to ID and cohort type in one data file to deliver. The SAS code for data cleaning, data harmonization and data merging can be found in the appendix. Below is the output from SAS about the contents of our final data file.

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
7	AGE	Num	8
36	AGE_M	Num	8
63	BF_TOTAL_1STRECALL	Num	8
37	BIRTHDATE_M	Char	8
62	CALCIUM_TRT	Num	8
66	CD_AGE	Num	8
59	CIGSDAYPREG	Num	8
57	CIGSDAY_1MOPP	Num	8
49	DELIVERY_APGAR1MIN	Num	8
50	DELIVERY_APGAR5MIN	Num	8
45	DELIVERY_GESTAGE	Num	8
44	DELIVERY_HEADCIRC	Num	8
43	DELIVERY_HEIGHT	Num	8
46	DELIVERY_MULTIPLE	Num	8
51	DELIVERY_PBCORD	Num	8
48	DELIVERY_SEX	Num	8
47	DELIVERY_TYPE	Num	8
42	DELIVERY_WEIGHT	Num	8
3	ETAPACC	Char	8
75	F5	Num	8
76	F6	Num	8
11	F9	Num	8

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
55	SMHX	Num	8
56	SMNOW_1MOPP	Num	8
58	SMPREG	Num	8
72	SUB_SKIN	Num	8
71	SUP_SKIN	Num	8
53	TIBIA_QUALITYCHECK	Num	8
82	TN13	Num	8
83	TN14	Num	8
84	TN15_1	Num	8
85	TN15_2	Num	8
86	TN15_3	Num	8
87	TN15_4A	Num	8
88	TN15_4B	Num	8
70	TRI_SKIN	Num	8
69	WAIST	Num	8
67	WEIGHT	Num	8
5	WEIGHT_CH	Num	8
39	WRITE_READ	Num	8
8	X_AGEMONS	Num	8
9	X_CBMI	Num	8
10	X_ZBFA	Num	8

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
12	F10	Num	8
77	F6B_AA	Num	8
78	F6B_MM	Num	8
79	F6C_AA	Num	8
80	F6C_DD	Num	8
81	F6C_MM	Num	8
1	FAKEID	Num	8
64	FECHANAC	Char	8
68	HEIGHT	Num	8
6	HEIGHT_CH	Num	8
54	HRP_PARITY	Num	8
13	M7	Num	8
14	M8	Num	8
38	MARITAL_STATUS	Num	8
61	MONTHS_SINCE_QUIT_SMK	Num	8
29	PB_CORD	Num	8
15	PB_C_1	Num	8
16	PB_C_3	Num	8
17	PB_C_6	Num	8
18	PB_C_12	Num	8
19	PB_C_18	Num	8
20	PB_C_24	Num	8
21	PB_C_30	Num	8
22	PB_C_36	Num	8
23	PB_C_42	Num	8
24	PB_C_48	Num	8
25	PB_C_60	Num	8
26	PB_M_11	Num	8
27	PB_M_12	Num	8
28	PB_M_13	Num	8
30	PB_M_20	Num	8
31	PB_M_21	Num	8
32	PB_M_23	Num	8
33	PB_M_24	Num	8
34	PB_M_27	Num	8
35	PB_M_212	Num	8
60	PREPREGNANCYBMI	Num	8
2	PROYECTO	Char	8
73	PR_DIAS	Num	8
74	PR_SIS	Num	8
52	ROTULA_QUALITYCHECK	Num	8
41	SCHOOLP_T	Num	8
40	SCHOOL_T1	Num	8
65	SEXO_H	Num	8
4	SEX_CH	Num	8

Chapter 4. Descriptive Statistics and Data Quality Control

4.1 Basic Statistical Tools in SAS for Descriptive Statistics and Detecting Potential Outliers

Descriptive statistics provide simple summaries about the study cohort and the various measures collected in the dataset, which are typically employed by data manager to quickly take a snapshot of the data in a study, describe the basic features of the dataset prepared, and help detect some potential data errors, such as entry typos, missing values, outliers by mistakes. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. [42, 43]

For each univariate variable, descriptive analysis involves describing the distribution of a single variable (including the mean, median, range, variance and standard deviation). Characteristics of a variable's distribution may also be depicted in graphical or tabular format, including histograms, box-plot and stem-and-leaf display. With bivariate or multivariate analysis, descriptive statistics may be used to describe the relationship between pairs of variables. In this case, some descriptive statistics that can be used are: scatterplots, quantitative measures of dependence, spaghetti plot for longitudinal trajectory, and etc. Bivariate analysis can also be used

to describe the relationship between the dependent variable and the predictors. The slope in regression analysis, for example, can reflect the relationship between variables. Descriptive analysis can also help examine the skewness of data distribution and help justify whether certain transformation such as log transformation is necessary for each variable. Use of log transformation makes data distribution more symmetrical and look more similar to the normal distribution, making them easier to interpret intuitively. More complicated statistical methods can be found in the literature to validate the data integration procedure when performing formal statistical inferences for merged data. [44-47]

4.2 Example of Data Quality Control when Preparing Data for investigating how the environmental exposures such as prenatal and childhood lead exposure affect adolescent's sexual maturation.

We first examined every univariate to detect some potential problems, such as missing values, potential outliers due to entry mistakes, whether range, median makes sense, etc. Then we compared sub-cohorts between different sexual maturation level groups. The following tables summarize the maternal characteristics and children's characteristics stratified by self-reported sexual maturation stages and physician judged sexual maturation levels.

Table 3: Descriptive Statistics of maternal characteristics and children's characteristics stratified by girl's self-reported stage of breast development

Characteristics	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Maternal					
Age (median, IQR)	26 (8)	26 (5)	27 (9)	23.5 (6)	24 (5)
Pre-pregnancy BMI (median, IQR)	26.19 (8.41)	28.79 (4.95)	22.33 (1.97)	25.97 (3.22)	24.56 (.84)
Total months of breastfeeding (median, IQR)	8 (9)	6 (9)	7 (9)	10 (11.5)	7 (15)
History of smoking					
Yes	56 (47.86%)	56 (45.16%)	47 (45.19%)	20 (33.9%)	18 (47.37%)
No	61 (52.14%)	68 (54.84)	57 (54.81%)	39 (66.1%)	20 (52.63%)
Smoking during pregnancy					
Yes	4 (3.42%)	5 (4%)	8 (7.55%)	3 (5%)	2 (5.26%)
No	113 (96.58%)	120 (96%)	98 (92.45%)	57 (95%)	36 (94.74%)
Child					
Birth weight (median, IQR)	3 (0)	3 (0)	3 (0)	3 (0)	3 (0)
Birth length (median, IQR)	50 (2)	50 (3)	50 (2)	50 (4)	48.5 (3)
Head circumference (median, IQR)	34 (2)	34 (2)	34 (2)	34 (2)	33 (2)
Blood pressure (diastolic) (median, IQR)	63 (10)	68 (9)	71 (5)	68 (18)	68 (0)
Blood pressure (systolic) (median, IQR)	99 (15)	106 (19)	109 (12)	107 (29)	99 (0)
BMI (median, IQR)	16.99 (3.63)	18.98 (4.74)	21.41 (4.71)	22.7 (4.45)	21.60 (4.15)

Stage 1: The nipple is raised a little. The rest of the breast is still flat.

Stage 2: The breast is a little larger and the nipple is raised a little more. The darker-colored area around the nipple (areola) is a little bit bigger than is State 1. Sometimes you can feel a small lump inside the nipples.

Stage 3: The darker-colored area around the nipple (areola) and the breast are much larger than in Stage 2 and form a mound. The areola does not stick out away from the breast.

Stage 4: The nipple and the areola stick out away from the skin and the rest of the breast.

Stage 5: The nipple sticks out away from the skin and the rest of the breast. The areola is at the same level than the rest of the breast

Table 4: Descriptive Statistics of maternal characteristics and children's characteristics stratified by girl's self-reported stage of pubic hair development

Characteristics	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Maternal					
Age (median, IQR)	26 (7)	26 (5)	25 (7)	22 (7)	22 (2)
Pre-pregnancy BMI (median, IQR)	30.18 (9.49)	24.26 (0)	25.21 (2)	25.30 (.02)	24.56 (0)
Total months of breastfeeding (median, IQR)	7 (9)	7 (9)	6 (9)	7 (10)	5 (10)
History of smoking					
Yes	83 (47.98%)	42 (42.42%)	39 (37.5%)	29 (48.33%)	4 (66.67%)
No	90 (52.02%)	57 (57.28%)	65 (62.5%)	31 (51.67%)	2 (33.33%)
Smoking during pregnancy					
Yes	10 (5.78%)	1 (1.04%)	4 (3.7%)	7 (11.11%)	0 (0%)
No	163 (94.22%)	95 (98.96%)	104 (96.3%)	56 (88.89%)	6 (100%)
Child					
Birth weight (median, IQR)	3 (0)	3 (0)	3 (0)	3 (0)	3 (0)
Birth length (median, IQR)	50 (2)	49.5 (3)	50 (3)	50 (3)	48 (3)
Head circumference (median, IQR)	34 (2)	34 (2)	34 (2)	34 (1)	33 (3)
Blood pressure (diastolic) (median, IQR)	65 (9)	69 (9)	71 (4)	71 (1)	n/a
Blood pressure (systolic) (median, IQR)	100 (18)	99 (18)	110 (16)	111 (3)	n/a
BMI (median, IQR)	17.81 (4.28)	19.30 (6.17)	21.4 (4.36)	21.82 (4.01)	21.39 (5.07)

Stage 1: There is no pubic hair.

Stage 2: There is a little, long, lightly colored hair. This hair may be straight or a little curly.

Stage 3: The hair is a little darker, coarse, and more curled. It has spread out and covers almost all the pubic area.

Stage 4: The hair is now as dark, curly, and coarse as that of a grown woman. The hair covers all the pubic area but does not spread out to the thighs.

Stage 5: The hair is like that of a grown woman (dark, curly, and coarse). The hair forms a triangle covering all the pubic area, and has spread out to the inner thighs.

Table 5: Descriptive Statistics of maternal characteristics and children's characteristics stratified by boy's self-reported stage of penis and scrotum development

Characteristics	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Maternal					
Age (median, IQR)	25 (7)	23 (6)	26 (7)	26 (7)	24 (13)
Pre-pregnancy BMI (median, IQR)	21.71 (7.14)	23.95 (2.69)	33.26 (7.82)	26.94 (6.81)	23.80 (13.48)
Total months of breastfeeding (median, IQR)	6 (6)	7 (9)	6 (8)	7 (8)	11 (9)
History of smoking					
Yes	26 (45.61%)	72 (44.44%)	43 (35.83%)	46 (50%)	6 (35.29%)
No	31 (54.39%)	90 (55.56%)	77 (64.17%)	46 (50%)	11 (64.71%)
Smoking during pregnancy					
Yes	n/a	7 (4.22%)	43 (35.83%)	3 (3.13%)	n/a
No	57 (100%)	159 (95.78%)	77 (64.17%)	93 (96.88%)	17 (100%)
Child					
Birth weight (median, IQR)	3 (0)	3 (1)	3 (1)	3 (1)	3 (0)
Birth length (median, IQR)	50 (3)	50 (3)	50 (3)	51 (2)	50 (2)
Head circumference (median, IQR)	34 (1)	35 (1)	34 (2)	35 (2)	33 (2.5)
Blood pressure (diastolic) (median, IQR)	71 (13)	62 (9)	68.5 (9)	70 (4)	70 (9)
Blood pressure (systolic) (median, IQR)	99 (18)	102 (16)	107 (12.5)	108.5 (11)	111 (11)
BMI (median, IQR)	16.36 (5.27)	18.39 (4.92)	20.71 (6.17)	17.97 (6.94)	19.81 (4.86)

Stage 1: The scrotum and penis are the same size as when you were younger.

Stage 2: The scrotum has lowered a bit and the penis is a little larger.

Stage 3: The penis is much longer and the scrotum much larger.

Stage 4: The penis is longer and wider. The scrotum is darker and larger than before.

Stage 5: The penis and scrotum are the size and shape of an adult.

Table 6: Descriptive Statistics of maternal characteristics and children's characteristics stratified by boy's self-reported stage of pubic hair development

Characteristics	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Maternal					
Age (median, IQR)	25 (7)	26 (8)	26 (7)	25.5 (4)	26 (8)
Pre-pregnancy BMI (median, IQR)	25.05 (5.22)	22.47 (3)	23.8 (2.96)	26.06 (9.02)	n/a
Total months of breastfeeding (median, IQR)	7 (8)	7 (8)	7 (11)	6 (9)	5.5 (5)
History of smoking					
Yes	90 (41.86%)	42 (47.19%)	27 (43.55%)	26 (40.63%)	8 (44.44%)
No	125 (58.14%)	47 (52.81%)	35 (56.45%)	38 (59.38%)	10 (55.56%)
Smoking during pregnancy					
Yes	2 (.93%)	9 (9.47%)	1 (1.49%)	2 (3.03%)	n/a
No	213 (99.07%)	86 (90.53%)	66 (98.51%)	64 (96.97%)	18 (100%)
Child					
Birth weight (median, IQR)	3 (1)	3 (0)	3 (1)	3 (0)	3 (0)
Birth length (median, IQR)	50 (3)	50 (3)	51 (2)	50 (3)	51 (2)
Head circumference (median, IQR)	34 (2)	34.5 (1)	34 (2)	35 (2)	35 (2)
Blood pressure (diastolic) (median, IQR)	63 (9)	70 (7)	70 (1)	71 (0)	70 (0)
Blood pressure (systolic) (median, IQR)	100 (16)	111 (7)	110 (3)	109 (3)	119 (0)
BMI (median, IQR)	18.24 (4.93)	19.9 (6.1)	21.48 (6.92)	20.77 (7.96)	22.45 (4)

Stage 1: There is no pubic hair.

Stage 2: There is a little, long, lightly colored hair. This hair may be straight or a little curly.

Stage 3: The hair is a little darker, coarse, and more curled. It has spread out and covers almost all the pubic area.

Stage 4: The hair is now as dark, curly, and coarse as that of a grown woman. The hair covers all the pubic area but does not spread out to the thighs.

Stage 5: The hair is like that of a grown woman (dark, curly, and coarse). The hair forms a triangle covering all the pubic area, and has spread out to the inner thighs.

No obvious outlier has been detected from all sub-group descriptive analysis. However, we noticed that maternal BMI is missing for stage 5 group from Table 6, while all other variables still have values falling into reasonable range. Need to look at the data and see what is going on and whether it is a real missing value.

Notice we have repeated measurements of child blood lead exposure and mom's blood lead exposure. To capture the change over time, we plotted the boxplot trajectory of these two variables across different time in months (Figures 5 and 6).

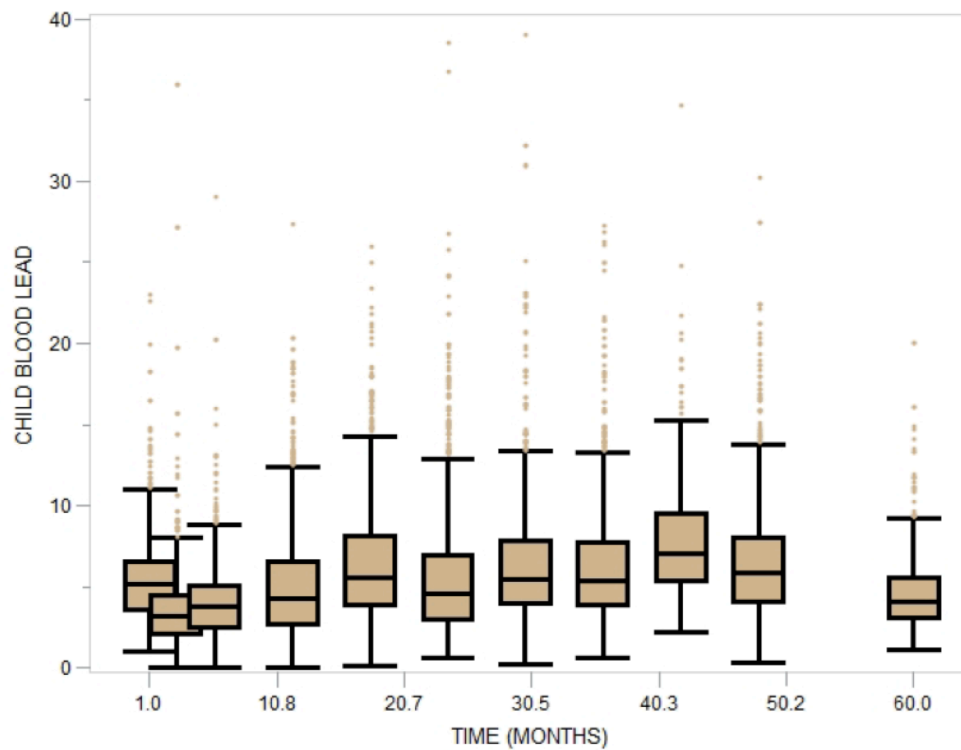


Figure 7. Boxplot of Child Blood Lead Over Time, with child's birth as the reference time.

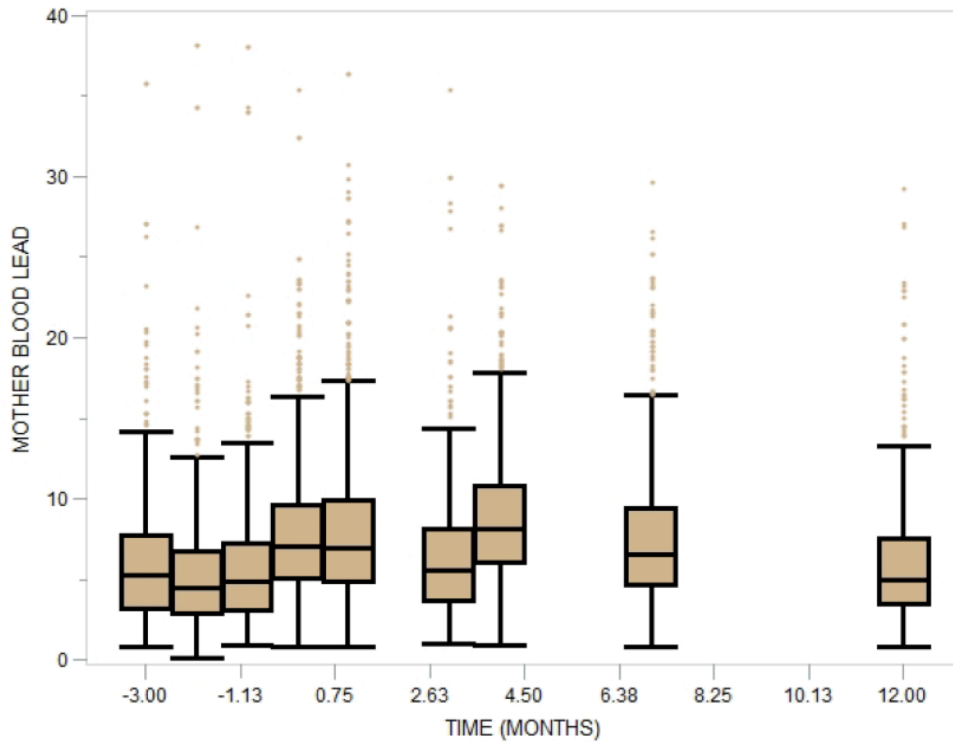


Figure 8. Boxplot of Mom Blood Lead Over Time, with child's birth as the reference time.

Both variables fluctuated over time, which indicated that both mom's prenatal lead exposure and children's early childhood lead exposure in our study vary as time goes on. This could be due to seasonal effect or food, diet, and nutrition effect. Over time, the variation of the measurement does not change much. To see how the lead exposure is related to kid's sexual maturation, we further examined the spaghetti plots of lead exposure over time stratified by sexual maturation stages (Figures 7 and 8). As one can see from the figures, across different sexual maturation groups, the

mother's blood lead exposure level do differ from one another, so as the child blood lead levels, which is relative higher for less sexually matured kids.

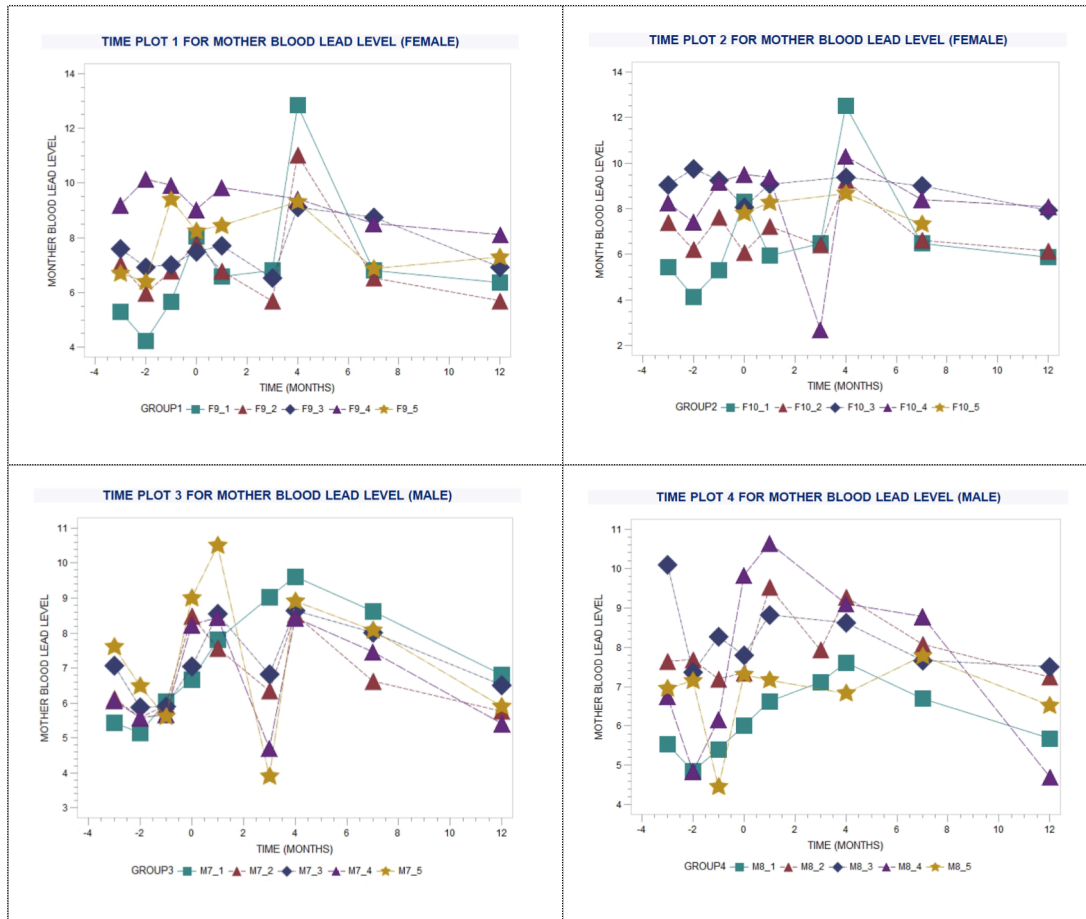


Figure 9. Spaghetti plot of mom blood lead exposure trajectory over time, stratified by different measurements of sexual maturation stages.

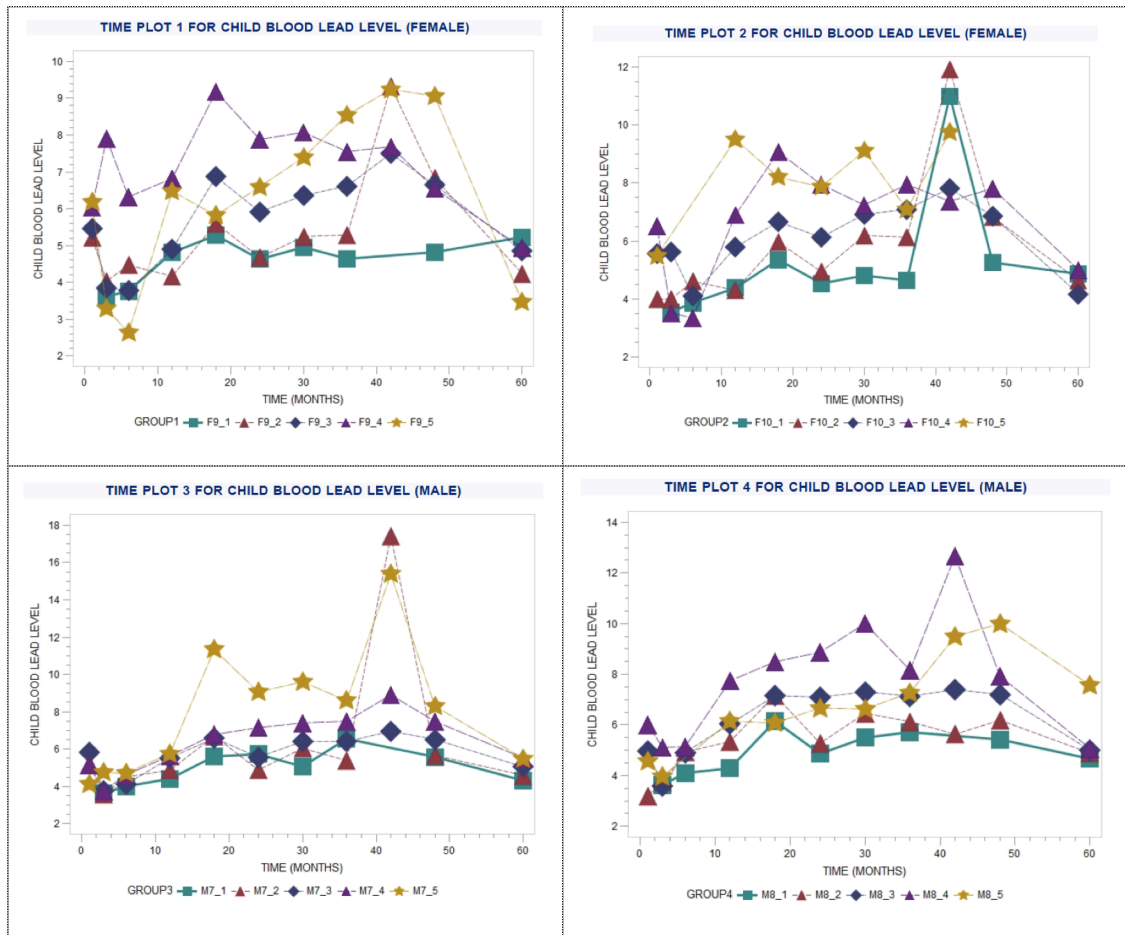


Figure 10. Spaghetti plot of child blood lead exposure trajectory over time, stratified by different measurements of sexual maturation stages.

Chapter 5. Summary

Data management and data facilitation is the foundation and a critical part for scientific investigations, especially with large cohort studies on health care, which may have multiple components collected from different institutes or different areas due to efficiency, cost, enrollment considerations. In this thesis, we have presented how we designed and implemented data management with a multi-center health care research project, ECHO (Environmental Influences on Child Health Outcomes). This research project is mainly focusing on environmental health sciences, and the primary goal is to investigate the impact of nutritional and toxicant exposures in environment during fetal and early postnatal life on individual's health outcomes over the life course. More specifically, I illustrated my work through an example project in ELEMENT study for investigating how the prenatal and childhood environmental exposures such as lead exposure affect adolescent's sexual maturation. I have presented my work on designing the data management plan and program codes for the data management and modeling core, facilitating and accomplishing data communication, cleaning and unifying data labels and formats, preparing datasets to achieve desirable data structures for further statistical analysis, and controlling data quality to ensure the overall research quality.

The study cohorts we considered also have another challenge, i.e. they all include repeated measures of social and/or chemical exposure through sensitive developmental periods. Approaches to disentangle whether there is variation in susceptibility due to timing of exposure are thus needed. In addition to using regression models as we presented where exposures measured during a given time period are fitted in separate models (one model for each time period of exposure), some novel approaches can be applied to consider the entire pattern of exposure through the life course as a predictor of the outcome [48, 49]. These methods have been used to identify windows of vulnerability to environmental toxicants [49], and have been extended to quantify the joint (interactive) effects of two factors (e.g., considering chemical exposures and social factors together). It is definitely of great future research interest, but is beyond the scope of this thesis.

References

1. Gluckman, P.D., et al., Effect of in utero and early-life conditions on adult health and disease. *N Engl J Med*, 2008. 359(1): p. 61-73.
2. U.S. Environmental Protection Agency, A framework for assessing health risks of environmental exposures to children. 2006. p. 1-145.

3. Brown, R.C., S. Barone, Jr., and C.A. Kimmel, Children's health risk assessment: incorporating a lifestage approach into the risk assessment process. *Birth Defects Res B Dev Reprod Toxicol*, 2008. 83(6): p. 511-21.
4. Selevan, S.G., C.A. Kimmel, and P. Mendola, Identifying critical windows of exposure for children's health. *Environ Health Perspect*, 2000. 108 Suppl 3: p. 451-5.
5. Wild, C.P., Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*, 2005. 14(8): p. 1847-50.
6. Rappaport, S.M., Biomarkers intersect with the exposome. *Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals*, 2012. 17(6): p. 483-489.
7. Buck Louis, G.M., et al., The exposome--exciting opportunities for discoveries in reproductive and perinatal epidemiology. *Paediatr Perinat Epidemiol*, 2013. 27(3): p. 229-36.
8. Miller, G.W. and D.P. Jones, The nature of nurture: refining the definition of the exposome. *Toxicol Sci*, 2014. 137(1): p. 1-2.

9. Go, Y.M., et al., Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicol Sci*, 2015. 148(2): p. 531-43.
10. Lora D. Delwiche; Susan J. Slaughter (2012). *The Little SAS Book: A Primer : a Programming Approach*. SAS Institute. p. 6. ISBN 978-1-61290-400-9.
11. Arthur Li (10 April 2013). *Handbook of SAS DATA Step Programming*. CRC Press. p. 149. ISBN 978-1-4665-5238-8.
12. Buck, Debbie. "A Hands-On Introduction to SAS DATA Step Programming" (PDF). SUGI 30: SAS Institute. Retrieved October 2, 2013.
13. Bass; K. Madhavi Lata & Kogent Solutions (1 September 2007). *Base SAS Programming Black Book*, 2007 Ed. Dreamtech Press. pp. 365–. ISBN 978-81-7722-769-7.
14. Der, G.; B. S. Everitt (March 10, 2009). "Basic Statistics using SAS Enterprise Guide". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 172 (2): 530. doi:10.1111/j.1467-985X.2009.00588_2.x.
15. John R. Schermerhorn (11 October 2011). *Exploring Management*. John Wiley & Sons. p. 3. ISBN 978-0-470-87821-7.

16. Spector, Phil. "An Introduction to the SAS System" (PDF). Berkeley. Archived from the original (PDF) on October 12, 2013.
17. Reichman, N.E. and E.M. Hade, Validation of birth certificate data. A study of women in New Jersey's HealthStart program. *Ann Epidemiol*, 2001. 11(3): p. 186-93.
18. Tellez-Rojo, M.M., et al., Longitudinal associations between blood lead concentrations lower than 10 microg/dL and neurobehavioral development in environmentally exposed children in Mexico City. *Pediatrics*, 2006. 118(2): p. e323-30.
19. Gomaa, A., et al., Maternal bone lead as an independent risk factor for fetal neurotoxicity: a prospective study. *Pediatrics*, 2002. 110(1 Pt 1): p. 110-8.
20. Hu, H., et al., Fetal lead exposure at each stage of pregnancy as a predictor of infant mental development. *Environ Health Perspect*, 2006. 114(11): p. 1730-5.
21. Wright, R.O., et al., Apolipoprotein E genotype predicts 24-month bayley scales infant development score. *Pediatr Res*, 2003. 54(6): p. 819-25.
22. Health, C.o.E., Lead Exposure in Children: Prevention, Detection, and Management. *Pediatrics*, 2005. 116.

23. U.S. Centers for Disease Control and Prevention, Preventing Lead Poisoning in Young Children. 2005, CDC: Atlanta, GA.
24. Advisory Committee on Childhood Lead Poisoning Prevention, Interpreting and Managing Blood Lead Levels <10 µg/dL in Children and Reducing Childhood Exposures to Lead, in MMWR. 2007, CDC. p. 1-14; 16.
25. Guideines for the Identification and Management of Lead Exposure in Pregnant and Lactating Women, A.S. Ettinger and A.G. Wengrovitz, Editors. 2010, U.S. Department of Health and Human Services: Atlanta, GA.
26. Polk, D.E., et al., Study protocol of the Center for Oral Health Research in Appalachia (COHRA) etiology study. BMC Oral Health, 2008. 8: p. 18.
27. Shaffer, J.R., et al., Genome-wide association scan for childhood caries implicates novel genes. J Dent Res, 2011. 90(12): p. 1457-62.
28. Shungin, D., et al., Using genetics to test the causal relationship of total adiposity and periodontitis: Mendelian randomization analyses in the Gene-Lifestyle Interactions and Dental Endpoints (GLIDE) Consortium. Int J Epidemiol, 2015. 44(2): p. 638-50.
29. Foxman, B., et al., Exploring the effect of dentition, dental decay and familiarity on oral health using metabolomics. Infect Genet Evol, 2014. 22: p. 201-7.

30. Wen, A., et al., Caries resistance as a function of age in an initially caries-free population. *J Dent Res*, 2012. 91(7): p. 671-5.
31. Shaffer, J.R., et al., Demographic, socioeconomic, and behavioral factors affecting patterns of tooth decay in the permanent dentition: principal components and factor analyses. *Community Dent Oral Epidemiol*, 2013. 41(4): p. 364-73.
32. Randall, C.L., et al., Gagging and its associations with dental care-related fear, fear of pain and beliefs about treatment. *J Am Dent Assoc*, 2014. 145(5): p. 452-8.
33. Younkin, S.G., et al., A genome-wide study of inherited deletions identified two regions associated with nonsyndromic isolated oral clefts. *Birth Defects Res A Clin Mol Teratol*, 2015. 103(4): p. 276-83.
34. Jindal, A., et al., Women are more susceptible to caries but individuals born with clefts are not. *Int J Dent*, 2011. 2011: p. 454532. 14
35. Shaffer, J.R., et al., Caries Experience Differs between Females and Males across Age Groups in Northern Appalachia. *Int J Dent*, 2015. 2015: p. 938213.
36. Shaffer, J.R., et al., Effects of enamel matrix genes on dental caries are moderated by fluoride exposures. *Hum Genet*, 2015. 134(2): p. 159-67.

37. Shaffer, J.R., et al., Genetic susceptibility to dental caries differs between the sexes: a family-based study. *Caries Res*, 2015. 49(2): p. 133-40.
38. Polk, D.E., et al., Effects of smoking and genotype on the PSR index of periodontal disease in adults aged 18-49. *Int J Environ Res Public Health*, 2012. 9(8): p. 2839-50.
39. Begum, F., et al., Regionally Smoothed Meta-Analysis Methods for GWAS Datasets. *Genet Epidemiol*, 2016. 40(2): p. 154-60.
40. Marshall, W.A. and J.M. Tanner, Variations in the pattern of pubertal changes in boys. *Arch Dis Child*, 1970. 45(239): p. 13-23.
41. Marshall, W.A. and J.M. Tanner, Variations in pattern of pubertal changes in girls. *Arch Dis Child*, 1969. 44(235): p. 291-303.
42. Mann, Prem S. (1995). *Introductory Statistics* (2nd ed.). Wiley. ISBN 0-471-31009-3.
43. van Belle, G., Fisher, L.D., Heagerty, P., and Lumley, T. (2004). *Biostatistics: A Methodology For the Health Sciences*, Edition 2. John Wiley & Sons
44. Wang, F., P.X. Song, and L. Wang, Merging multiple longitudinal studies with study-specific missing covariates: A joint estimating function approach. *Biometrics*, 2015. 71(4): p. 929-40.

45. Ding, W. and P.X.K. Song, EM algorithm in Gaussian coupla with missing data. *Computational Statistics and Data Analysis*, 2016. 101: p. 1-11.
46. Wang, F., L. Wang, and P.X. Song, Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements. *Biometrics*, 2016.
47. Tang, L. and P.X.K. Song, Fused lasso approach in regression coefficients cluster: Learning parameter heterogeneity in data integration, in *Journal of Machine Learning Resesarch*. 2016.
48. Ferguson, K.K., et al., Variability in urinary phthalate metabolite levels across pregnancy and sensitive windows of exposure for the risk of preterm birth. *Environ Int*, 2014. 70: p. 118-24.
49. Sanchez, B.N., et al., Statistical methods to study timing of vulnerability with sparsely sampled data on environmental toxicants. *Environ Health Perspect*, 2011. 119(3): p. 409-15.

SAS code Appendix:

```

/***** PART I read the seven csv files AND MERGE *****/
/*DATA 1.1*/
DATA DATA1_1;
    INFILE "C:\Users\Wei Wang\Google Drive\Wei\thesis\Thesis
Data\Cholesterol(2007)\Cholesterol_Adolescent_Anthro.csv" DSD FIRSTOBS=2;
    INPUT FAKEID PROYECTO $ ETAPACC $ SEX_CH WEIGHT_CH HEIGHT_CH AGE X_AGEMONS
X_CBMI X_ZBFA;
RUN;

PROC PRINT DATA=DATA1_1;
RUN;

/*DATA 1.2*/
DATA DATA1_2;
    INFILE "C:\Users\Wei Wang\Google Drive\Wei\thesis\Thesis
Data\Cholesterol(2007)\Cholesterol_Adolescent_Maturation.csv" DSD FIRSTOBS=2;
    INPUT FAKEID PROYECTO $ ETAPACC $ F9 F10 M7 M8;
RUN;

PROC PRINT DATA=DATA1_2;
RUN;

/*DATA 2.1*/
DATA DATA2_1;
    INFILE "C:\Users\Wei Wang\Google Drive\Wei\thesis\Thesis
Data\Historical(1994-2005)\Historical_Child_Blood_Lead.csv" DLM='2C0D'X DSD MISSEVER
LRECL=10000 FIRSTOBS=2;
    INPUT FAKEID PROYECTO $ PB_C_1 PB_C_3 PB_C_6 PB_C_12 PB_C_18 PB_C_24 PB_C_30
PB_C_36 PB_C_42 PB_C_48 PB_C_60;
RUN;

PROC PRINT DATA=DATA2_1;
RUN;

/*DATA 2.2*/
DATA DATA2_2;
    INFILE "C:\Users\Wei Wang\Google Drive\Wei\thesis\Thesis
Data\Historical(1994-2005)\Historical_Mom_Blood_Lead.csv" DLM='2C0D'X DSD MISSEVER
LRECL=10000 FIRSTOBS=2;
    INPUT FAKEID PROYECTO $ PB_M_11 PB_M_12 PB_M_13 PB_CORD PB_M_20 PB_M_21
PB_M_23 PB_M_24 PB_M_27 PB_M_212 ;
RUN;

PROC PRINT DATA=DATA2_2;
RUN;

/*DATA 2.3*/
DATA DATA2_3;
    INFILE "C:\Users\Wei Wang\Google Drive\Wei\thesis\Thesis
Data\Historical(1994-2005)\Historical_Mom_General_Info.csv" DLM='2C0D'X DSD MISSEVER

```

```

LRECL=10000 FIRSTOBS=2;
    INPUT FAKEID PROYECTO $ AGE_M BIRTHDATE_M $ MARITAL_STATUS WRITE_READ
SCHOOL_T1 SCHOOLP_T DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC
DELIVERY_GESTAGE DELIVERY_MULTIPLE
    DELIVERY_TYPE DELIVERY_SEX DELIVERY_APGAR1MIN DELIVERY_APGAR5MIN
DELIVERY_PBCORD ROTULA_QUALITYCHECK TIBIA_QUALITYCHECK HRP_PARITY SMHX SMNOW_1MOPP
CIGSDAY_1MOPP SMPREG CIGSDAYPREG PREPREGNANCYBMI
    MONTHS_SINCE_QUIT_SMK CALCIUM_TRT BF_TOTAL_1STRECALL FECHANAC $;
RUN;

PROC PRINT DATA=DATA2_3;
RUN;

/*DATA 4.1*/
DATA DATA4_1;
    INFILE "C:\Users\Wei Wang\Google Drive\Wei\thesis\Thesis
Data\P20(2010)\P20_Adolescent_Anthro.csv" DLM='2C0D'X DSD MISSOEVER LRECL=10000
FIRSTOBS=2;
    INPUT FAKEID PROYECTO $ SEXO_H CD_AGE WEIGHT HEIGHT WAIST TRI_SKIN SUP_SKIN
SUB_SKIN PR_DIAS PR_SIS X_AGEMONS X_CBMI X_ZBFA;
RUN;

PROC PRINT DATA=DATA4_1;
RUN;

/*DATA 4.2*/
DATA DATA4_2;
    INFILE "C:\Users\Wei Wang\Google Drive\Wei\thesis\Thesis
Data\P20(2010)\P20_Adolescent_Maturation.csv" DLM='2C0D'X DSD MISSOEVER LRECL=10000
FIRSTOBS=2;
    INPUT FAKEID PROYECTO $ SEXO_H F5 F6 F6B_AA F6B_MM F6C_AA F6C_DD F6C_MM F9
F10 M7 M8 TN13 TN14 TN15_1 TN15_2 TN15_3 TN15_4A TN15_4B;
RUN;

PROC PRINT DATA=DATA4_2;
RUN;

/*SORT AND MERGE THE DATA1*/
PROC SORT DATA=DATA1_1 OUT=DATA11;
    BY FAKEID PROYECTO;
RUN;
PROC SORT DATA=DATA1_2 OUT=DATA12;
    BY FAKEID PROYECTO;
RUN;

DATA DATA1;
    MERGE DATA11 DATA12;
    BY FAKEID PROYECTO;
RUN;

PROC PRINT DATA=DATA1;
RUN;

/*SORT AND MERGE THE DATA2*/

```

```

PROC SORT DATA=DATA2_1 OUT=DATA21;
    BY FAKEID PROYECTO;
RUN;
PROC SORT DATA=DATA2_2 OUT=DATA22;
    BY FAKEID PROYECTO;
RUN;
PROC SORT DATA=DATA2_3 OUT=DATA23;
    BY FAKEID PROYECTO;
RUN;

DATA DATA2;
    MERGE DATA21 DATA22 DATA23;
    BY FAKEID PROYECTO;
RUN;

PROC PRINT DATA=DATA2;
RUN;

/*SORT AND MERGE THE DATA4*/
PROC SORT DATA=DATA4_1 OUT=DATA41;
    BY FAKEID PROYECTO;
RUN;
PROC SORT DATA=DATA4_2 OUT=DATA42;
    BY FAKEID PROYECTO;
RUN;

DATA DATA4;
    MERGE DATA41 DATA42;
    BY FAKEID PROYECTO;
RUN;

PROC PRINT DATA=DATA4;
RUN;

/*SORT AND MERGE THE DATA*/
PROC SORT DATA=DATA1 OUT=DATANEW1;
    BY FAKEID PROYECTO;
RUN;
PROC SORT DATA=DATA2 OUT=DATANEW2;
    BY FAKEID PROYECTO;
RUN;
PROC SORT DATA=DATA4 OUT=DATANEW4;
    BY FAKEID PROYECTO;
RUN;

DATA DATA;
    MERGE DATANEW1 DATANEW2 DATANEW4;
    BY FAKEID PROYECTO;
RUN;

PROC PRINT DATA=DATA;
RUN;

```

```

/***** PART II TIME PLOT *****/
/*****/
/*DESCRIPTION OF DATA*/
PROC CONTENTS DATA=DATA;
RUN;

/* CATEGORY OF RESPONSE*/
PROC FREQ DATA=DATA;
TABLES F9 F10 M7 M8;
RUN;

/*CHILD BLOOD LEAD*/
DATA NEWDATA1;
    SET DATA;
    KEEP FAKEID SEX_CH F9 F10 M7 M8 PB_C_1 PB_C_3 PB_C_6 PB_C_12 PB_C_18 PB_C_24
PB_C_30 PB_C_36 PB_C_42 PB_C_48 PB_C_60 ;
RUN;

DATA NEWDATA2;
    SET NEWDATA1;
    IF F9=1 THEN GROUP1="F9_1";
    IF F9=2 THEN GROUP1="F9_2";
    IF F9=3 THEN GROUP1="F9_3";
    IF F9=4 THEN GROUP1="F9_4";
    IF F9=5 THEN GROUP1="F9_5";
    IF F10=1 THEN GROUP2="F10_1";
    IF F10=2 THEN GROUP2="F10_2";
    IF F10=3 THEN GROUP2="F10_3";
    IF F10=4 THEN GROUP2="F10_4";
    IF F10=5 THEN GROUP2="F10_5";
    IF M7=1 THEN GROUP3="M7_1";
    IF M7=2 THEN GROUP3="M7_2";
    IF M7=3 THEN GROUP3="M7_3";
    IF M7=4 THEN GROUP3="M7_4";
    IF M7=5 THEN GROUP3="M7_5";
    IF M8=1 THEN GROUP4="M8_1";
    IF M8=2 THEN GROUP4="M8_2";
    IF M8=3 THEN GROUP4="M8_3";
    IF M8=4 THEN GROUP4="M8_4";
    IF M8=5 THEN GROUP4="M8_5";
    Y=PB_C_1; TIME=1; OUTPUT;
    Y=PB_C_3; TIME=3; OUTPUT;
    Y=PB_C_6; TIME=6; OUTPUT;
    Y=PB_C_12; TIME=12; OUTPUT;
    Y=PB_C_18; TIME=18; OUTPUT;
    Y=PB_C_24; TIME=24; OUTPUT;
    Y=PB_C_30; TIME=30; OUTPUT;
    Y=PB_C_36; TIME=36; OUTPUT;
    Y=PB_C_42; TIME=42; OUTPUT;
    Y=PB_C_48; TIME=48; OUTPUT;
    Y=PB_C_60; TIME=60; OUTPUT;
    DROP F9 F10 M7 M8 PB_C_1 PB_C_3 PB_C_6 PB_C_12 PB_C_18 PB_C_24 PB_C_30
PB_C_36 PB_C_42 PB_C_48 PB_C_60;
RUN;

```

```

PROC PRINT DATA=NEWDATAC2;
RUN;

/*F9 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAC2G1;
    SET NEWDATAC2(WHERE=(Y NE .));
RUN;

DATA NEWDATAC2G1;
    SET NEWDATAC2G1(WHERE=(GROUP1 NE ""));
RUN;

PROC SORT DATA=NEWDATAC2G1;
    BY GROUP1 TIME;
RUN;

PROC MEANS DATA=NEWDATAC2G1;
    VAR Y;
    BY GROUP1 TIME;
    OUTPUT OUT=NEWDATAC2G1_MEAN1 MEAN=MEANY;
RUN;

SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;

PROC Gplot DATA=NEWDATAC2G1_MEAN1;
    AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
    AXIS2 LABEL=(ANGLE=90 "CHILD BLOOD LEAD LEVEL");
    PLOT MEANY*TIME=GROUP1/HAXIS=AXIS1 VAXIS=AXIS2;
    TITLE "TIME PLOT 1 FOR CHILD BLOOD LEAD LEVEL (FEMALE)";
RUN;

/*F10 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAC2G2;
    SET NEWDATAC2(WHERE=(Y NE .));
RUN;

DATA NEWDATAC2G2;
    SET NEWDATAC2G2(WHERE=(GROUP2 NE ""));
RUN;

PROC SORT DATA=NEWDATAC2G2;
    BY GROUP2 TIME;
RUN;

PROC MEANS DATA=NEWDATAC2G2;
    VAR Y;
    BY GROUP2 TIME;
    OUTPUT OUT=NEWDATAC2G2_MEAN1 MEAN=MEANY;
RUN;

```

```

SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;

```

```

PROC GPLOT DATA=NEWDATAC2G2_MEAN1;
  AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
  AXIS2 LABEL=(ANGLE=90 "CHILD BLOOD LEAD LEVEL");
  PLOT MEANY*TIME=GROUP2/HAXIS=AXIS1 VAXIS=AXIS2;
  TITLE "TIME PLOT 2 FOR CHILD BLOOD LEAD LEVEL (FEMALE)";
RUN;

```

```

/*M7 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAC2G3;
  SET NEWDATAC2(WHERE=(Y NE .));
RUN;

```

```

PROC PRINT DATA=NEWDATAC2G3;
RUN;

```

```

DATA NEWDATAC2G3;
  SET NEWDATAC2G3(WHERE=(GROUP3 NE ""));
RUN;

```

```

PROC SORT DATA=NEWDATAC2G3;
  BY GROUP3 TIME;
RUN;

```

```

PROC MEANS DATA=NEWDATAC2G3;
  VAR Y;
  BY GROUP3 TIME;
  OUTPUT OUT=NEWDATAC2G3_MEAN1 MEAN=MEANY;
RUN;

```

```

SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;

```

```

PROC GPLOT DATA=NEWDATAC2G3_MEAN1;
  AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
  AXIS2 LABEL=(ANGLE=90 "CHILD BLOOD LEAD LEVEL");
  PLOT MEANY*TIME=GROUP3/HAXIS=AXIS1 VAXIS=AXIS2;
  TITLE "TIME PLOT 3 FOR CHILD BLOOD LEAD LEVEL (MALE)";
RUN;

```

```

/*M8 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAC2G4;
  SET NEWDATAC2(WHERE=(Y NE .));
RUN;

```

```

PROC PRINT DATA=NEWDATAC2G4;
RUN;

```

```
DATA NEWDATAC2G4;
    SET NEWDATAC2G4(WHERE=(GROUP4 NE ""));
RUN;
```

```
PROC SORT DATA=NEWDATAC2G4;
    BY GROUP4 TIME;
RUN;
```

```
PROC MEANS DATA=NEWDATAC2G4;
    VAR Y;
    BY GROUP4 TIME;
    OUTPUT OUT=NEWDATAC2G4_MEAN1 MEAN=MEANY;
RUN;
```

```
SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;
```

```
PROC GPLOT DATA=NEWDATAC2G4_MEAN1;
    AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
    AXIS2 LABEL=(ANGLE=90 "CHILD BLOOD LEAD LEVEL");
    PLOT MEANY*TIME=GROUP4/HAXIS=AXIS1 VAXIS=AXIS2;
    TITLE "TIME PLOT 4 FOR CHILD BLOOD LEAD LEVEL (MALE)";
RUN;
```

/*MOTHER BLOOD LEAD*/

```
DATA NEWDATA1_1;
    SET DATA;
    KEEP FAKEID SEX_CH F9 F10 M7 M8 PB_M_11 PB_M_12 PB_M_13 PB_M_20 PB_M_21
PB_M_23 PB_M_24 PB_M_27 PB_M_212;
RUN;
```

```
DATA NEWDATAM2;
    SET NEWDATA1_1;
    IF F9=1 THEN GROUP1="F9_1";
    IF F9=2 THEN GROUP1="F9_2";
    IF F9=3 THEN GROUP1="F9_3";
    IF F9=4 THEN GROUP1="F9_4";
    IF F9=5 THEN GROUP1="F9_5";
    IF F10=1 THEN GROUP2="F10_1";
    IF F10=2 THEN GROUP2="F10_2";
    IF F10=3 THEN GROUP2="F10_3";
    IF F10=4 THEN GROUP2="F10_4";
    IF F10=5 THEN GROUP2="F10_5";
    IF M7=1 THEN GROUP3="M7_1";
    IF M7=2 THEN GROUP3="M7_2";
    IF M7=3 THEN GROUP3="M7_3";
    IF M7=4 THEN GROUP3="M7_4";
    IF M7=5 THEN GROUP3="M7_5";
    IF M8=1 THEN GROUP4="M8_1";
    IF M8=2 THEN GROUP4="M8_2";
```

```

IF M8=3 THEN GROUP4="M8_3";
IF M8=4 THEN GROUP4="M8_4";
IF M8=5 THEN GROUP4="M8_5";
Y=PB_M_11; TIME=-3; OUTPUT;
Y=PB_M_12; TIME=-2; OUTPUT;
Y=PB_M_13; TIME=-1; OUTPUT;
Y=PB_M_20; TIME=0; OUTPUT;
Y=PB_M_21; TIME=1; OUTPUT;
Y=PB_M_23; TIME=3; OUTPUT;
Y=PB_M_24; TIME=4; OUTPUT;
Y=PB_M_27; TIME=7; OUTPUT;
Y=PB_M_212; TIME=12; OUTPUT;
DROP F9 F10 M7 M8 PB_M_11 PB_M_12 PB_M_13 PB_M_20 PB_M_21 PB_M_23 PB_M_24
PB_M_27 PB_M_212;
RUN;

PROC PRINT DATA=NEWDATAM2;
RUN;

/*F9 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAM2G1;
    SET NEWDATAM2(WHERE=(Y NE .));
RUN;

DATA NEWDATAM2G1;
    SET NEWDATAM2G1(WHERE=(GROUP1 NE ""));
RUN;

PROC SORT DATA=NEWDATAM2G1;
    BY GROUP1 TIME;
RUN;

PROC MEANS DATA=NEWDATAM2G1;
    VAR Y;
    BY GROUP1 TIME;
    OUTPUT OUT=NEWDATAM2G1_MEAN1 MEAN=MEANY;
RUN;

SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;

PROC GPLOT DATA=NEWDATAM2G1_MEAN1;
    AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
    AXIS2 LABEL=(ANGLE=90 "MOTHER BLOOD LEAD LEVEL");
    PLOT MEANY*TIME=GROUP1/HAXIS=AXIS1 VAXIS=AXIS2;
    TITLE "TIME PLOT 1 FOR MOTHER BLOOD LEAD LEVEL (FEMALE)";
RUN;

/*F10 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAM2G2;
    SET NEWDATAM2(WHERE=(Y NE .));
RUN;

```



```

DATA NEWDATAM2G2;
    SET NEWDATAM2G2(WHERE=(GROUP2 NE ""));
RUN;

PROC SORT DATA=NEWDATAM2G2;
    BY GROUP2 TIME;
RUN;

PROC MEANS DATA=NEWDATAM2G2;
    VAR Y;
    BY GROUP2 TIME;
    OUTPUT OUT=NEWDATAM2G2_MEAN1 MEAN=MEANY;
RUN;

SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;

PROC Gplot DATA=NEWDATAM2G2_MEAN1;
    AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
    AXIS2 LABEL=(ANGLE=90 "MONTH BLOOD LEAD LEVEL");
    PLOT MEANY*TIME=GROUP2/HAXIS=AXIS1 VAXIS=AXIS2;
    TITLE "TIME PLOT 2 FOR MOTHER BLOOD LEAD LEVEL (FEMALE)";
RUN;

/*M7 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAM2G3;
    SET NEWDATAM2(WHERE=(Y NE .));
RUN;

PROC PRINT DATA=NEWDATAM2G3;
RUN;

DATA NEWDATAM2G3;
    SET NEWDATAM2G3(WHERE=(GROUP3 NE ""));
RUN;

PROC SORT DATA=NEWDATAM2G3;
    BY GROUP3 TIME;
RUN;

PROC MEANS DATA=NEWDATAM2G3;
    VAR Y;
    BY GROUP3 TIME;
    OUTPUT OUT=NEWDATAM2G3_MEAN1 MEAN=MEANY;
RUN;

SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;

```

```

PROC GPLOT DATA=NEWDATAM2G3_MEAN1;
  AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
  AXIS2 LABEL=(ANGLE=90 "MOTHER BLOOD LEAD LEVEL");
  PLOT MEANY*TIME=GROUP3/HAXIS=AXIS1 VAXIS=AXIS2;
  TITLE "TIME PLOT 3 FOR MOTHER BLOOD LEAD LEVEL (MALE)";

```

```
RUN;
```

```

/*M8 GROUP MEAN RESPONSE TIME PLOT*/
DATA NEWDATAM2G4;
  SET NEWDATAM2(WHERE=(Y NE .));

```

```
RUN;
```

```

PROC PRINT DATA=NEWDATAM2G4;

```

```
RUN;
```

```

DATA NEWDATAM2G4;
  SET NEWDATAM2G4(WHERE=(GROUP4 NE ""));

```

```
RUN;
```

```

PROC SORT DATA=NEWDATAM2G4;
  BY GROUP4 TIME;

```

```
RUN;
```

```

PROC MEANS DATA=NEWDATAM2G4;
  VAR Y;
  BY GROUP4 TIME;
  OUTPUT OUT=NEWDATAM2G4_MEAN1 MEAN=MEANY;

```

```
RUN;
```

```

SYMBOL1 INTERPOL=JOIN VALUE=SQUAREFILLED COLOR=VIBG HEIGHT=2;
SYMBOL2 INTERPOL=JOIN VALUE=TRIANGLEFILLED COLOR=DEPK HEIGHT=2;
SYMBOL3 INTERPOL=JOIN VALUE=DIAMONDFILLED COLOR=MOB HEIGHT=2;
SYMBOL4 INTERPOL=JOIN VALUE=HASHFILLED COLOR=VIP HEIGHT=2;
SYMBOL5 INTERPOL=JOIN VALUE=STARFILLED COLOR=VIOY HEIGHT=2;

```

```

PROC GPLOT DATA=NEWDATAM2G4_MEAN1;
  AXIS1 LABEL=(ANGLE=0 "TIME (MONTHS)");
  AXIS2 LABEL=(ANGLE=90 "MOTHER BLOOD LEAD LEVEL");
  PLOT MEANY*TIME=GROUP4/HAXIS=AXIS1 VAXIS=AXIS2;
  TITLE "TIME PLOT 4 FOR MOTHER BLOOD LEAD LEVEL (MALE)";

```

```
RUN;
```

```

/***** PART III BOXPLOT
*****/

```

```

/* BOXPLOT FOR CHILD BLOOD BY TIME*/
SYMBOL VALUE=DOT HEIGHT=.4 INTERPOL=BOXTF WIDTH=3 BWIDTH=5 COLOR=BLACK CV=TAN;

```

```

AXIS1 ORDER=(0 TO 45 BY 10) LABEL=(HEIGHT=1.25 ANGLE=90 'CHILD BLOOD LEAD')
MINOR=(NUMBER=1);

```

```

AXIS2 LABEL=(HEIGHT=1.25 'TIME (MONTHS)') OFFSET=(5,5);

```

```

PROC GPLOT DATA=NEWDATAC2 UNIFORM;

```

```

PLOT Y*TIME /HAXIS=AXIS2 VAXIS=AXIS1 HMINOR=0 SKIPMISS;
TITLE "BOXPLOT FOR CHILD BLOOD LEAD OVER TIME";
RUN;

/* BOXPLOT FOR MOTHER BLOOD BY TIME*/
SYMBOL VALUE=DOT HEIGHT=.4 INTERPOL=BOXTF WIDTH=3 BWIDTH=5 COLOR=BLACK CV=TAN;

AXIS1 ORDER=(0 TO 45 BY 10) LABEL=(HEIGHT=1.25 ANGLE=90 'MOTHER BLOOD LEAD')
MINOR=(NUMBER=1);

AXIS2 LABEL=(HEIGHT=1.25 'TIME (MONTHS)') OFFSET=(5,5);

PROC GPLOT DATA=NEWDATAM2 UNIFORM;
PLOT Y*TIME /HAXIS=AXIS2 VAXIS=AXIS1 HMINOR=0 SKIPMISS;
TITLE "BOXPLOT FOR MOTHER BLOOD LEAD OVER TIME";
RUN;

/***** PART IV
CHARACTERISTICS *****/
/*TABLE FOR F9*/
/*F91*/
DATA F91;
SET DATA(WHERE=(F9 EQ 1));
RUN;

PROC PRINT DATA=F91;
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=F91;
VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F91;
TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F91;
VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F92*/
DATA F92;
SET DATA(WHERE=(F9 EQ 2));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=F92;
VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F92;
TABLES SMHX SMPREG;

```

```

RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F92;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F93*/
DATA F93;
    SET DATA(WHERE=(F9 EQ 3));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=F93;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F93;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F93;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F94*/
DATA F94;
    SET DATA(WHERE=(F9 EQ 4));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=F94;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F94;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F94;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F95*/
DATA F95;
    SET DATA(WHERE=(F9 EQ 5));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=F95;

```

```

VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F95;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F95;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*TABLE FOR F10*/
/*F101*/
DATA F101;
    SET DATA(WHERE=(F10 EQ 1));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=F101;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F101;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F101;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F102*/
DATA F102;
    SET DATA(WHERE=(F10 EQ 2));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=F102;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F102;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F102;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F103*/

```

```

DATA F103;
    SET DATA(WHERE=(F10 EQ 3));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATA=DATADATA=F103;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F103;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F103;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F104*/
DATA F104;
    SET DATA(WHERE=(F10 EQ 4));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATA=F104;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F104;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F104;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*F105*/
DATA F105;
    SET DATA(WHERE=(F10 EQ 5));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATA=F105;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=F105;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=F105;

```

```
VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;  
RUN;
```

```
/*TABLE FOR M7*/  
/*FM71*/  
DATA M71;  
SET DATA(WHERE=(M7 EQ 1));  
RUN;
```

```
/*MATERNAL*/  
PROC UNIVARIATE DATADATA=M71;  
VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;  
RUN;
```

```
PROC FREQ DATA=M71;  
TABLES SMHX SMPREG;  
RUN;
```

```
/*CHILD*/  
PROC UNIVARIATE DATA=M71;  
VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;  
RUN;
```

```
/*M72*/  
DATA M72;  
SET DATA(WHERE=(M7 EQ 2));  
RUN;
```

```
/*MATERNAL*/  
PROC UNIVARIATE DATADATA=M72;  
VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;  
RUN;
```

```
PROC FREQ DATA=M72;  
TABLES SMHX SMPREG;  
RUN;
```

```
/*CHILD*/  
PROC UNIVARIATE DATA=M72;  
VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;  
RUN;
```

```
/*M73*/  
DATA M73;  
SET DATA(WHERE=(M7 EQ 3));  
RUN;
```

```
/*MATERNAL*/  
PROC UNIVARIATE DATADATA=M73;  
VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;  
RUN;
```

```

PROC FREQ DATA=M73;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=M73;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*M74*/
DATA M74;
    SET DATA(WHERE=(M7 EQ 4));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATA=M74;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=M74;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=M74;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*M75*/
DATA M75;
    SET DATA(WHERE=(M7 EQ 5));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATA=M75;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=M75;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=M75;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*TABLE FOR M8*/
/*FM81*/
DATA M81;
    SET DATA(WHERE=(M8 EQ 1));

```



```

RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=M81;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=M81;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=M81;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*M82*/
DATA M82;
    SET DATA(WHERE=(M8 EQ 2));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=M82;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=M82;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=M82;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

/*M83*/
DATA M83;
    SET DATA(WHERE=(M8 EQ 3));
RUN;

/*MATERNAL*/
PROC UNIVARIATE DATADATA=M83;
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;
RUN;

PROC FREQ DATA=M83;
    TABLES SMHX SMPREG;
RUN;

/*CHILD*/
PROC UNIVARIATE DATA=M83;
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;
RUN;

```

```
/*M84*/  
DATA M84;  
    SET DATA(WHERE=(M8 EQ 4));  
RUN;
```

```
/*MATERNAL*/  
PROC UNIVARIATE DATA=M84;  
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;  
RUN;
```

```
PROC FREQ DATA=M84;  
    TABLES SMHX SMPREG;  
RUN;
```

```
/*CHILD*/  
PROC UNIVARIATE DATA=M84;  
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;  
RUN;
```

```
/*M85*/  
DATA M85;  
    SET DATA(WHERE=(M8 EQ 5));  
RUN;
```

```
/*MATERNAL*/  
PROC UNIVARIATE DATA=M85;  
    VAR AGE_M PREPREGNANCYBMI BF_TOTAL_1STRECALL;  
RUN;
```

```
PROC FREQ DATA=M85;  
    TABLES SMHX SMPREG;  
RUN;
```

```
/*CHILD*/  
PROC UNIVARIATE DATA=M85;  
    VAR DELIVERY_WEIGHT DELIVERY_HEIGHT DELIVERY_HEADCIRC PR_DIAS PR_SIS X_CBMI;  
RUN;
```