2015

A Cross-Sectional Analysis of Health Impacts of Inorganic Arsenic in Chemical Mixtures

Paul Hargarten
Virginia Commonwealth University, hargartenp@vcu.edu

A Cross-Sectional Analysis of Health Impacts of Inorganic Arsenic in Chemical Mixtures

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science at Virginia Commonwealth University.

by
Paul Michael Hargarten
B.S. University of Richmond, Richmond, Virginia

Director: Yongyung Shin, Ph.D.
Assistant Professor,
Department of Biostatistics

Virginia Commonwealth University
Richmond, Virginia
April 2015

Acknowledgements

I would like to express my gratitude toward those who have helped me to achieve a master's degree in biostatistics. This thesis and all of the work would not have been possible without the assistance of my mentor, Dr. Chris Gennings. After she attended a national arsenic review board, I became interested in this topic. Starting as a summer project, it grew with Dr. Gennings' support and guidance to an interesting idea to incorporate the weighted sum regression approach into the study.

I would also like to thank the rest of the members of the committee: Dr. Yongyon Shin and Dr. Joseph Ritter for their time and assistance throughout the process. I would like to thank additional teachers in the Department of Biostatistics at VCU: Dr. Roy Sabo and Dr. Donna McClish for their assistance in development of the models.

Additionally, I would like to thank my fellow students in the department for the mutual encouragement and the shared sympathy (in the bad times) as we studied for exams and spent late nights completing challenging homework assignments. I would like to also thank the administrative staff in the department—Yvonne Hargrove and Cindy Sabo—for their continual faith and encouragement as I complete the thesis.

Lastly, I would like to thank my parents for their loving support and patience in my efforts in pursuing a masters degree.

Table of Contents

List of Tables

List of Figures

Abstract

A Cross-Sectional Analysis of Health Impacts of Inorganic Arsenic in Chemical Mixtures

By Paul Michael Hargarten,

A thesis submitted in partial fulfillment of the requirements for the M.S. degree, Department of Biostatistics at Virginia Commonwealth University.

Virginia Commonwealth University, 2015.
Director: Yongyun Shin, Assistant Professor, Department of Biostatistics.

Drinking groundwater is the primary way humans accumulate arsenic. Chronic exposure to inorganic arsenic (iAs) (over decades) has been shown to be associated with multiple health effects at low levels (5-10 ppb) including: cancer, elevated blood pressure and cardiovascular disease, skin lesions, renal failure, and peripheral neuropathy. Using hypertension (or high blood pressure) as a surrogate marker for cardiovascular disease, we examined the effect of iAs alone and in a mixture with other metals using a cross-sectional study of adults in United States (National Health and Examination Survey, NHANES, 2005-2010) adjusting for covariates: urinary creatinine level (mg/dL), poverty index ratio (PIR, measure of socioeconomic status, 1 to 5), age, smoking (yes/no), alcohol usage, gender, non-Hispanic Black, and overweight (BMI>=25).

A logistic regression model suggests that a one-unit increase in log of inorganic arsenic increases the odds of hypertension by a factor of 1.093 (95% Confidence Interval=0.935, 1.277) adjusted for these covariates , which indicates that there was not significant evidence to claim that inorganic arsenic is a risk factor for hypertension. Biomonitoring data provides evidence that humans are not only exposed to inorganic arsenic but also to mixtures of chemicals including inorganic arsenic, total mercury, cadmium, and lead. We tested for a mixture effect of these four environmental chemicals using weighted quantile sum (WQS) regression, which takes into account the correlation among the chemicals and with the

outcome. For one-unit increase in the weighted sum, the adjusted odds of developing hypertension increases by a factor of 1.027 (95% CI=0.882,1.196), which is also not significant after taking into account the same covariates. The insignificant finding may be due to the low inorganic arsenic concentration (8-620 μg /L) in US drinking water, compared to those in countries like Bangladesh where the concentrations are much higher. Literature provides conflicting evidence of the association of inorganic arsenic and hypertension in low/moderate regions; future studies, especially a large cohort study, are needed to confirm if inorganic arsenic alone or with other metals is associated with hypertension in the United States.

Chapter 1: Introduction

1. Hypertension and Arsenic

Chronic exposure to total arsenic has been shown to be linked to severe health risks in many literature studies. [1-11] Hypertension, or high blood pressure, has been shown in a study in Bangladesh to be associated with chronic arsenic exposure. [11] Jones et.al. 2011[12] did a study looking at total arsenic and other arsenic species on hypertension in the US; however, the authors did not examine the effect of inorganic arsenic (as defined in literature) [10] on hypertension as is proposed in this study.

Hypertension is selected as a surrogate as it increases the risk for diseases including: excessive bleeding, chronic kidney disease, heart attack, vision problems, and stroke. [13] In fact, the leading cause of death for New York residents was cardiovascular disease in 2007. [14] A stroke occurs when blood flow to the brain is cut off, causing the death of brain cells. High blood pressure increases the risk of stroke by 4 to 6 times. [15]

Additionally, hypertension data is easy to collect, and thus a large sample size is possible as only 3-4 blood pressure readings are required. However, the variance of blood pressure is high as blood pressure depends on an individual's current stress levels as well as a recent intake of caffeine, food, alcohol, and tobacco. The exclusion of these individuals did not affect the analyses though. [16-18]

2. Arsenic: Its Source

Found in minerals, arsenic (As), a natural element, is distributed via volcanic eruptions and mineral erosion, which consequently contaminates oceans, soils, and fresh water. [2,10]

Fish, seafood, and algae accumulate high levels of organic arsenic or arsenic meshed with lipids (arsenobetaine, arsenocholine, trimethylarsine oxide, and arsenosugars). The ingestion of these

organoarsenic compounds result in accumulations of only 1 to 2 microgram per liter (μg/L)* of organic arsenic. Studies have shown that organic arsenic is quickly excreted and thus has limited toxicity in humans. [2,10,19] By contrast, inorganic arsenic (iAs) is considerably more lethal to humans. Inorganic arsenic includes reduced and methylated metabolites, commonly consisting of: arsenite As(III), arsenate As(V), methylarsonite (MA(III)), methylarsonate MA, dimethylarsinite DMA(III), and dimethylarsonate DMA. [10] Although organic arsenic is less toxic than inorganic arsenic, many literature studies confuse "total arsenic" with "inorganic arsenic", which makes it challenging to interpret the results.

The main source of human exposure to arsenic is by drinking contaminated groundwater. About 50 million people in Bangladesh drink water at peak exposure with levels of arsenic higher than the World Health Organization's standard of 10 μg/L compared to only 30 million in the United States. [20] Not surprisingly, countries with active geothermal areas like Chile, Mexico, Taiwan, Geneva, Bangladesh, and Finland have been found to have higher concentrations of iAs both in groundwater and correspondingly in human blood. [6,10]

Another major source of arsenic exposure for humans is via industrial by-products, primarily through consumption of wood products that are contaminated by arsenic-rich soils. In a southern China study, for example, the burning of wood or coal contaminated by arsenic-rich soils and groundwater produced arsenic levels in kitchen dust as high as 3000 mg/kg. [6,10] Timber treated with chrome arsenate used to preserve wood is considered to be another major source of arsenic species found in humans. In fact, studies suggest that children who accidently ingest arsenic-contaminated wood contain high levels of arsenic in their bloodstream. [2] The burning of fossil fuels—which contains inorganic arsenic—can also lead to exposure. [2] Not surprisingly, the use of arsenic in industry is declining due to its toxicity. [10,19]

* (which is equivalent to 1-2 parts per billion, ppb).

Arsenic exposure through consumption of agricultural products is another area of concern. Frequently, food, such as rice, fish, milk and vegetables, contain arsenic if washed in arsenic-contaminated water or grown in arsenic soils. This becomes problematic since rice and beans are food staples in many countries. For instance, white rice was shown to have an average of 0.323 microgram of arsenic per gram in India, according to one study. [6,19] In the United States, the concentration is much less, the average concentration of most cooked rice products is 0.034 µg As/g in a sample of 1200 rice products. [21] Consequently, American children who reported consuming rice had total urinary arsenic levels 70-fold over those who did not eat rice; the median and interquartile range (IQR) were 8.9 µg/L (IQR: 5.3–15.6) and 5.5 µg/L (IQR: 3.1–8.4), respectively. [22] Compared to India and the developing world, the arsenic concentration in rice is much lower across European Union and North America, possibly due to the differences in overall sanitation practices and arsenic levels in soil. Pesticides used in agricultural production are also an issue. Two major methyl metabolites of inorganic arsenic—DMA and MMA--are used in some pesticides that might accidently be consumed in food products. [2,19,23] Tobacco is another agricultural product that contains arsenic. As a cigarette has 0.04 to 0.12 µg of arsenic, smoking tobacco increases arsenic levels; in fact, tobacco and As-contaminated drinking water synergistically increase the arsenic retention level, as much as 200 µg /L in blood. [2,6] Accidental consumption of groundwater, food, industrial products, and others are the major sources for arsenic. Most alarming is that a low concentration of inorganic arsenic (around 50-100 µg/L) [2,6,10] in drinking water can have adverse health effects in humans.

3. Arsenic: Its Effects

In short, inorganic "arsenic is one of the most notorious poisons of all time; it can be inhaled and digested, unnoticeable to sight, smell, or taste." [10] The earliest sign of chronic iAs disorder is skin lesions, including hyperpigmentation and keratosis, which is found in a quarter of these patients in Bangladesh. [10] A concise review is given by Simon, et.al. 2006. [6]

After arsenic-contaminated water is consumed, it is usually absorbed through the digestive system. Therefore, it affects almost every organ system since it disturbs essential cellular processes such as inhibiting numerous enzymes and substituting for phosphate (as arsenate). Thereby, the effects on cellular health are encompassing: it lowers ATP production, inhibits various signal transduction pathways, oxidizes fatty acids, and alters gene expression. The inorganic species, arsenite, is a neutral compound at physiological pH and substitutes glycerol which hinders glucose uptake. [10] The effects worsen at higher arsenic levels and/or with longer exposure. Since arsenic inhibits many cellular functions, there is an increase risk of cells eventually becoming cancerous or undergoing apoptosis.

The chronic intake (in increments of 5-10 years) of inorganic arsenic at levels as low as 50-100 µg/L may lead to failure of major organ systems, such as cardiac, renal, respiratory, and neurological. [2,10] A study in Bangladesh showed that symptoms may develop in a time period as short as six months. [5] Chronic exposure of inorganic arsenic is associated with cardiac failure (cardiac arrhythmia, hypertension), failure of immune system, [7,19] renal failure, and neurological disorders (peripheral neuropathy, constant pain, hypersensitivity to stimuli, muscle weakness, atrophy). [2,6] If lead is present, the neurological behavior are more severe. [6] Inorganic arsenic has also been shown to lead to respiratory lung disease, including chronic cough and chronic bronchitis, based on separate studies in Bangladesh and West India. [4,6,24]

Humans who are exposed to inorganic arsenic for decades experience catastrophic effects. [2] A variety of sleep disorders and further neural decay, including a possible link to Alzheimer's disease, are possible. [3] Arsenic also affects verbal IQ/memory. A study in Mexican children, found that those with >50 µg /L As had lower intelligence quotient (IQ) and lower long-term memory than those with <5 µg /L As. [6] Arsenic has also been shown to be a risk factor for diabetes although little is known about the mechanism. [6] Additionally, arsenic affects the reproduction system, leading to premature delivery and

low birth weights at very low concentrations (<10 µg /L). Preliminary evidence suggests that women who drink As-contaminated water (>50 µg /L) have a higher risk of premature delivery and fetal deaths.[6]

In several studies, inorganic arsenic has been indicated as a carcinogen for skin, lung, and urinary bladder cancers and may cause kidney and liver cancers. However, as the detailed mechanisms between arsenic toxicity and cancer are not well understood, additional research is needed.[10,19] Chronic exposure to levels typical of US drinking water (8-620 µg /L) has not been associated with cancer rates.

Interestingly, arsenic was used as a medicine beginning as early as 400 BC to treat various skin conditions, syphilis, epilepsy, and asthma until other drugs rendered arsenic obsolete in 1983. A form of arsenic, arsenic trioxide, was shown to be effective as an anti-cancer agent for acute promyelocytic leukemia in 2002. [10] Paradoxically, current research is examining whether arsenic can inhibit cancer. [10] Until the research is complete, other medicines may be used due to the toxic effects of arsenic.

Due to the fact that very small iAs levels contribute to adverse health effects, arsenic was recently found to be a risk to human health. Additionally, age has been found as a confounder for effects of arsenic. [6] The signs of inorganic arsenic toxicity are more severe at higher concentrations and/or longer durations of exposure. [6] The effects discussed have primarily been observed in world regions with higher levels of iAs in soils and groundwater. Although the studies in the United States are limited, the literature as a whole makes clear that adverse health risks are associated with chronic exposure to inorganic arsenic. A main purpose of this thesis is to examine the effect of inorganic arsenic on hypertension in the United States, which is a place of low/moderate iAs concentration in drinking water.

4. Current Arsenic Levels

As previously discussed, multiple organ failure can result from unusually small amounts of chronic arsenic exposure (on ppb scale). In places with high geothermic activity or arsenic-rich minerals like

Taiwan, India, Argentina, Chile, and Bangladesh, the concentration of arsenic is around 50 μg /L in drinking water. [6] A Bangladesh study reported that the concentration inorganic urinary arsenic in the highest quintile of pregnant women was: ~262-977 μg /L. [2,25] In the United States, the inorganic arsenic level in drinking water has been estimated between 8 and 620 μg /L, according to several studies. [6] The arsenic level varies widely by geographic location; for instance, parts of Michigan, northeastern Wisconsin, and New England have higher drinking water of arsenic than other areas. In one study, the 95th percentile for US population for urinary inorganic arsenic in 2009 was 18.9 μg /L and the median was 6 μg /L, [23] indicating that most Americans have 10-fold to 65-fold lower concentrations under those in Bangladesh.

The World Health Organization has set a safety limit of 10 μg /L=10 ppb for inorganic arsenic in drinking water. An ideal recommended limit would be 0.17 μg /L in drinking water, but it is unfeasible since the detection limit of high performance liquid chromatography/mass spectroscopy (HPLC-MS) to measure arsenic is within 3 μg /L. [6,26] The United States has followed similarly, lowering the drinking water limits from 50 to 10 μg /L of inorganic (or total)[†] arsenic in 2006; Canada has lowered its safe inorganic (or total) arsenic level to 5 μg/L. [6,9,27] The United States and those in high geothermic countries have urinary arsenic concentrations higher than the drinking water standards suggested by the WHO and FDA.

Using urinary arsenic to gauge blood concentrations of arsenic in humans has been established in the literature. The literature suggests that urinary arsenic levels are highly correlated with arsenic intakes from drinking water and dietary sources. Thus, this measurement is accepted as a valid biomarker from the WHO. [6] Urinary arsenic has been shown to be strongly correlated with arsenic levels in the bloodstream and in the groundwater supply. [10] Thus, urinary arsenic levels are used herein.

[†] the regulation is unclear since "inorganic" (Kapaj) and "total" (others) arsenic is cited in regulation. These limits may include both inorganic and organic arsenic, although organic arsenic is not considered toxic.

5. Outline

The goals of this thesis are two-fold. First, we investigate the effect of inorganic arsenic on hypertension in adults in the United States. As hypertension is defined as a binary variable, logistic regression is performed. As inorganic arsenic is frequently not the only toxic metal in the body at one time, other metals such as cadmium, total mercury, and lead, may also have an impact on hypertension. Due to the correlation among the metals, traditional logistic regression would lead to invalid results. A novel weighted quantile sum regression approach is used to determine if the collection of metals (including inorganic arsenic) has an effect on hypertension. Thus, we investigate the effect of inorganic arsenic alone and with other metals on hypertension using publically available data from the National Health and Nutrition Examinations Survey: NHANES 2005-2010.

Chapter 2 in this thesis is written as a standalone manuscript for peer review; introductiory information is repeated. Future work and conclusions are described in Chapter 3.

Chapter 2: A Cross-Sectional Analysis of Health Impacts of Inorganic Arsenic in Chemical Mixtures (Written as a Manuscript)

1. Introduction

Arsenic (As), a natural metalloid, is ubiquitous, primarily found in the Earth's crust and in groundwater. Groundwater becomes highly contaminated when minerals erode, which commonly occurs in regions with high volcanic activity, like Chile, Mexico, Taiwan, Geneva, Bangladesh, and Finland. [5,6,10] Rice, vegetables, and other foods become contaminated when rinsed in groundwater containing inorganic As, which is a concern since these items are a fundamental part of diets across the world. Consuming rice increases the urinary total arsenic concentration by as much as 70-fold in American children. [2] The consumption or burning of timber products, of fossil fuels, and pesticides (on fruit and vegetables) as well as smoking also contribute to inorganic arsenic (iAs) exposure. [2,6,10,19,23]

Organic arsenic compounds are mainly found in seafood, fish, and algae but have limited toxicity in humans since they are quickly excreted. [2,10,19] However, chronic exposure to inorganic arsenic compounds--arsenite As(III), arsenate As(V), methylarsonite (MA(III)), methylarsonate MA, dimethylarsinite DMA(III), and dimethylarsonate (DMA)-- have been shown to have adverse and systemic health effects [10]. These effects include skin lesions, stillbirths, cardiovascular diseases, diabetes, cancers (skin, lung, bladder, kidney, and liver), neurological disorders, respiratory diseases, memory loss, and diabetes. [1-3,5,6,10-12,25,28,29] In short, inorganic "arsenic is one of the most notorious poisons of all time as it can be inhaled and digested, unnoticeable to sight, smell, or taste". [10]

Most alarming is that these widespread toxic effects may occur at very low levels of inorganic arsenic arsenic (around 50-100 µg/L) [2,6,10] in drinking water. With high levels of arsenic in groundwater, the levels of urinary total arsenic in the highest quintile for Bangladesh pregnant woman was 262-977 µg /L, which increased the risk of lower respiratory infection by 69% compared to those with less than

39 µg/L. [2,25] By contrast, the ninety-fifth percentile of urinary inorganic arsenic in the United States was 18.9 µg/L in 2009. [2] The World Health Organization and other governments have been lowering the safe level of inorganic (or total)[‡] arsenic in drinking water to 10 µg/L, which may still be too high. [6,9,27] Instead, an ideal recommended limit would be 0.17 µg/L in drinking water, but it is unfeasible since the detection limit of high performance liquid chromatography/mass spectroscopy (HPLC-MS) to measure arsenic is within 3 µg /L. [6,26]. Using urine as a biomarker for iAs levels has been well established in literature, which will be used in this study.

Hypertension, or high blood pressure, is well known to be a surrogate for cardiovascular diseases, excessive bleeding, heart attack, stroke and chronic kidney diseases [13]. It has been estimated that around one billion people suffer from hypertension worldwide.

Several cross-sectional studies have shown that total arsenic levels are associated with developing hypertension. [5,9,11,30] However, most of these studies were conducted in countries like Bangladesh or Taiwan where the arsenic concentration in water is much higher (>100 µg /L) than in the United States. There was one study from Jones et.al. whose aim was to determine the association of inorganic arsenic exposure and hypertension in adults using the National Health and Nutrition Examination Survey (NHANES) 2003-2008 (Jones et.al 2011). [12] However, their study did not find evidence of an association with systolic or diastolic blood pressure and total arsenic or total arsenic minus one organic compound, but not the sum of inorganic arsenic species defined in other studies. [12] Few studies exist to show the effect of inorganic arsenic on hypertension in a population with low-to-moderate levels of inorganic arsenic, such as North America and Europe. Fewer still restrict the population to the United States.

[‡]. the regulation is unclear since "inorganic" (Kapaj) and "total" (others) arsenic is cited in regulation. These limits may include both inorganic and organic arsenic, although organic arsenic is not considered toxic.

The purpose of this study is to ascertain whether urinary inorganic arsenic, defined as the sum of the species, is associated with hypertension in adults in the United States. Additionally, as inorganic arsenic is not frequently the only toxic metal in the bloodstream, we interrogate whether a mixture of urinary inorganic arsenic with blood levels of cadmium, lead, and total mercury increase the risk of hypertension. Both use the NHANES data from 2005-2010.

2. Methods

2.1 The Dataset

The US National Center for Health Statistics (NCHS) through the Centers for Disease Control and Prevention conduct a National Health and Examination Survey (NHANES) on a continual basis to assess the health of American adults and children, using a complex sampling design that includes both laboratory tests and self-answered questionnaires. [31] The collection and publication of data occur every two years, which is called a cycle. NHANES recommends combining cycles to increase the sample size since the same individual never participates in subsequent cycles. [32] Our analysis has combined data from the D,E, and F cycles (2005-2010) giving a total of 31,034 observations. There were 17,132 adults (screening age $\geq$ 20) in the study; the age limit was due to the nature of the covariates. As urinary speciated arsenic was measured on one-third of the subsample, [26] the observations in this study reduce to 5,386. Lastly, we exclude "refused", "don't know", and missing values among all covariates, which leaves a final sample size of 4,386 subjects.

2.2 Covariates

To examine the association of inorganic chemicals and/or other chemicals with hypertension, some known risk factors for hypertension include: obesity, high alcohol consumption, age, gender, smoking, race (black vs. nonblack) and socioeconomic variables. [5,14,17] The covariates included in this study are: creatinine, PIR, age, alcohol usage, smoking status, gender, black, and overweight.

Metals and Arsenic Covariates

Total inorganic arsenic is the sum of urine concentration of species: arsenic acid (AS(V)O4), arsenite acid (AS(III)O3), DMA, and MMA. [2,10] We follow the convention used in NHANES to replace concentrations below the limits of detection (LOD) with LOD/$\sqrt{2}$. [26] The metal concentrations measured in blood are cadmium (Cd), total mercury (tHg), and lead (Pb). In order for all metals to have the same units, Cd is converted from nanomoles/liter to micromoles/liter. Since the concentration of urinary arsenic species is dependent on urine concentration (e.g. related to fluids, physical activity, and temperature) , the urine concentration is adjusted by including urinary creatinine, which was determined using Jaffe rate reaction measured with a CX3 analyzer, as a covariate in the model. [10]

Hypertension

Systolic (SP) and diastolic (DP) blood pressures were averaged across three and sometimes four measurements, which were collected on the same day and followed a standardized protocol. [26,33,34] Similar to other studies, [5,12,14] hypertension is defined for adults with average SP>=140, or average DP>=90, or currently taking anti-hypertensive medicine; conversely, normal blood pressure is having (average SP<140 and average DP<90) OR not taking current anti-hypertensive medicine. If individuals are missing SP, DP, or did not respond on the current use of anti-hypertension medicine, hypertension was considered missing.

Demographic Covariates

Covariates were coded to indicate increased risk of hypertension. Gender was recoded, so that males (male=1) would indicate an increased risk in hypertension compared to women (male=0). An individual who is overweight (BMI>=25) is more at risk of having high blood pressure compared to normal/underweight (0<BMI<25). The screening age was included in this study as recommended by NHANES protocol and was used as a continuous variable. [32]

Due to the different definitions of race across the three cycles, race is combined as suggested by NHANES. [35] Non-Hispanic African Americans (blacks) are known to be more likely to develop hypertension; in fact, they have a 50% higher prevalence of hypertension than whites although the reason is not well understood. [14] Consequently, race is dichotomized into "black" and non-black and is included as a covariate in this study.

Since socioeconomic status has been shown to be associated with hypertension,[14] a surrogate that measures the extent of poverty is the Poverty Index Ratio (PIR). [36] A PIR<=1 means the family is below the poverty line, while 1<=PIR<5 means that the family is above the poverty line. Index values exceeding 5 were truncated at 5. [35] Thus, PIR is centered at PIR=1 to highlight this interpretation: PIRdiff=PIR-1.

A surrogate for smoking was constructed using two NHANES questions--SMQ020 (smoke100, Smoked at least 100 cigarettes in life) and SMQ680 (smokefive, Used tobacco/nicotine last 5 days)--from self-based questionnaires from cycles D,E, and F: Smoking - Cigarette Use (SMQ)and Smoking – Recent Tobacco Use (SMQRTU). A smoker was defined as an individual who smoked at least 100 cigarettes in life or used tobacco and nicotine in the last 5 days. If the answers for only one question included a "refused", "don't know", or "missing", the other question was used to determine the smoking status. If responses for both questions were "don't know", "refused", or "missing", smoking status was also considered missing.

As alcoholics generally have lower blood pressure than non-alcoholics, questions ALQ120 and ALQ120U were converted to alcohol usage (alchl usg): "how often did you drink alcohol per week over the last 12 months?" Since those who answered "no", "refused", or "don't know" to the first question (ALQ101) are not asked any other questions including ALQ120, alcohol usage was assumed to be 0. For

the 2742 who responded to the question but gave missing units (ALQ120U=.), they were ignored in the calculations. [37]

In summary, the covariates—creatinine, PIR, age, alcohol usage, smoking status, gender, black, and overweight--included in this study are well known in literature to be risk factors for hypertension, allowing us to examine whether inorganic arsenic is also a risk factor.

2.3 Data Analysis:

(3a) Logistic Regression

Similar to other epidemiology studies, we are interested in investigating whether inorganic arsenic is a risk factor in developing hypertension. Since it is binary, logistic regression can be employed:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0^* + \beta_1 * \ln(iAs_i) + Z_i^{\boldsymbol{T}}\boldsymbol{\theta} \quad for\ i = 1 \ldots .4386 \qquad (1)$$

where π_i is the probability that the ith individual has hypertension with covariates in Z_i having effects $\boldsymbol{\theta}$. Creatinine, difference in PIR, age, and alcohol usage (# of drinks) are continuous. The assumptions of the logistic model include that each individual is independent and that each covariate is linearly associated with the log of odds of hypertension. The reference group for the model is an individual with the lowest risk: (non-Hispanic) black, non-smoking, and non-alcoholic (alchlusg=0) female with normal/underweight BMI (BMI<25) at twenty years of age with PIR=1.

Covariates considered biologically important were kept in the model regardless of significance. Quadratic effects of all continuous variables were considered using a likelihood ratio test with Type I error rate of 0.25. After selecting covariates, inorganic arsenic was subsequently added to the model to assess the risk of arsenic (alpha=0.05). If inorganic arsenic was significant, two-way interactions of inorganic arsenic with each of the covariates were considered (alpha=0.05).

(3b) Analysis of Metal Mixtures

In a logistic regression analysis, inorganic arsenic is the only chemical examined as a potential risk factor for hypertension (after adjusting for other covariates). However, there is often more than one metal detected in individuals at any given time: iron (which is needed as a cofactor in enzymes and in hemoglobin), copper, sodium (useful in action potentials), fluoride (in bones, teeth), iodine, copper, etc. Chemical concentrations—specifically, inorganic arsenic, cadmium, lead, and total mercury--allows us to answer if a mixture of these correlated metals increase the risk for hypertension and to identify "bad actors": i.e. specific metals that increase the risk of hypertension. [38-40] In addition to be known to be risk factors for hypertension, these metals are often correlated with each other.

We would like to determine which metals (and demographic covariates) are associated with hypertension; that is, we desire accurate variable selection (instead of low prediction error). Since there is high correlation among the metals, the collinearity needs to be addressed, but the standard techinques fail to apply. [41] Although ridge regression reduces the variance by adding a little bit of bias, it does not reduce the dimensionality of the data. The lasso and elastic net achieve data reduction and coefficient shrinkage but have low senestivity with correlated variables. Instead, we want to use a method that selects these metals based on their correlation with each other and the outcome. [41]

(3c) Weighted Quantile Sum (WQS) Regression

In short, the novel Weighted Quantile Sum (WQS) method addresses this problem of collinearity by empirically deriving a weighted score of exposure driven by both the correlation among the metals and the outcome so that variable selection may be achieved accurately and parsimoniously. Briefly, the weighted quintile sum method is described; for details, see Carrico, et.al. 2014. [41]

The dataset is split into a test and validation dataset, in this study: 40 percent was used in the test set leaving 60% in the validation dataset. Let c be the number of continuous chemicals be categorized into quartiles for j=1….c chemicals (here c=4).

In the test dataset, the weights are estimated in a non-linear model via bootstrapping:

$$g(\mu) = \beta_0 + \beta_1 \left(\sum_{j=1}^{c=4} \mathrm{w_j} x_{ijq} \right) + z_i^{\boldsymbol{T}} \boldsymbol{\theta}, \qquad i = 1, \ldots n \tag{2}$$

for some monotonic, differentiable link function g(μ) (as in generalized linear models), the intercept β_0, $x_{ijq} =$ the quartiles of the c metals for the ith subject ($x_{jq} = 0$ for the 1st quartile, $x_{jq} = 1$ for the 2nd quartile, $x_{jq} = 2$ for the 3rd quartile, and $x_q = 3$ for the 4th quartile), and covariate parameters θ. The weighted sum for the c chemicals of interest is $\sum_{j=1}^{c} \mathrm{w_j} x_{jq}$. While the weights can be interepereted as proportion of the effect of each chemical on hypertension, β_1 describes the aggregrated effect of the chemical sum on the outcome. The model is subject to the constraints for each bootstrap:

$$\Sigma w_j = 1 \ , \ 0 \leq w_j \leq 1 \text{ for all j, and } \beta_1 > 0.$$

The constraint on β_1 was chosen so that the selected weights would increase the risk of hypertension. Using the trust region algorithm, the adjusted parameter estimates for weights are averaged across all B bootstrap samples:[§]

$$\overline{w}_j = \frac{\sum_{b=1}^{B} w_{jb}}{B} \tag{3}$$

Using observations in the validation dataset, the weighted quantile sum for subject i is the sum across the four quantiles with the weights being the mean proportion of the chemical's effects on hypertension:

$$WQS_i = \sum_{j=1}^{c=4} \overline{w}_j x_{ijq} \tag{4}$$

and the relationship of WQS with the outcome can be assessed by ordinary least squares regression in the validation dataset:

[§] *(The signal function $f(\hat{\beta}_{1b})$ mentioned in Carrico et.al is 1 in this analysis.)*

$$g(\mu_i) = \beta_0 + \beta_1 WQS_i + z_i^T \theta \quad (5)$$

The $\beta\ parameter$ estimates the mixture effect of the metals on the outcome, where each weight identifies how much risk a particular chemical imposes on the outcome. However, the interpretation of WQS may change depending on the relative correlation between the metals and outcome versus correlation among the metals. On the one hand, if the correlation with the outcome is higher, WQS performs well as the association with the outcome drives the weights into proper association. On the other, both the weights individually and the sum are less interpretable if the outcome has a low correlation (about 0.1) relative to high correlation of metals. [41]

In this study, examining the metals (c=4)—inorganic arsenic, cadmium, lead, and total mercury, observations were gathered across NHANES cycles so that the sample size was considered large enough to split into the test and validation datasets (proportion=0.4). Since the outcome variable –having hypertension or not—is binary, the link function is the logit: $g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ where π_i denotes the probability of having hypertension.

The number of bootstrap samples was 100 and the number of observations per bootstrap sample was the length of test dataset*0.4=1743. The random seed used in data splitting and in bootstrap estimation of the weights was 506079. The initial value for β_1 was chosen to be a small positive number (0.005) since the trust region algorithm failed if it was at the boundary. The initial values of the weights were equal among the four metals (i.e. 0.25) and the initial values of the covariates were estimated from the final logistic model (containing arsenic).

(3d) Computation

All analyses were done using SAS 9.4. A non-black, non-smoking and non-alcoholic 20-year old female with normal/underweight BMI (BMI<25) and a PIR of 1 served as the reference for the analyses; such a person would be at the lowest risk for hypertension.

3. Results

3.1 Preliminary Findings

Out of the total of 4386 individuals examined, 1537 (35.04 %) had hypertension. Overall, there were 855 (19.49%) non-Hispanic African Americans, 2168 (49.43%) males, 3106 (70.82%) overweight, 2222 (50.66%) smokers, and 1048 (23.89%) over the age of 65.

A comparison of the covariates and arsenic between hypertension and non-hypertension was examined (**Table 1a**). Descriptive statistics indicate that inorganic arsenic, PIR, and alcohol usage were similar in the two groups, which may indicate that they are not risk factors. Age and categorical univariate results, perhaps except gender, follow what is expected from literature (**Table 1b**). There appears to be no association with gender and hypertension. (In this case, males have a slight risk so this could be expected with a small sample size).

Variable	Description	Normal (n=2849) mean(std)	Hypertension (n=1537) mean(std)
ln(iAs)	log of inorganic arsenic	1.91 (0.56)	1.89 (0.56)
creatinine	Creatinine, urine (mg/dL)	128.77 (79.03)	112.84 (70.41)
Difference in PIR	poverty index ratio (PIR) centered at poverty line (PIR=1).	1.57 (1.62)	1.54 (1.55)
Age	Screen Age in Years	42.35 (15.97)	62.31 (13.68)
Alcohol Usage	# of drinks per week in the past year	1.06(1.75)	1.10 (2.03)
		Number (%)	**Number (%)**
black: yes	non-Hispanic black	467(16.39)	388(25.24)
no		*2382 (83.61)*	*1149 (74.76)*
gender: male		1384 (48.58)	784(51.01)
female		*1465 (51.42)*	*753 (48.99)*
smoker	individual who smoked at least 100 cigarettes in life or used tobacco and nicotine in the last 5 days	1407 (49.39)	815(53.03)
non-smoker		*1442 (50.61)*	*722 (46.97)*
overweight: yes	$BMI \geq 25$	1890(66.34)	1216(79.12)
no		*959 (33.66)*	*321 (20.88)*

Table 1a : Descriptive Statistics. Out of 4386 total individuals, here are 1537 individuals (35.04%) with hypertension and 2849 without. The percent specified is the percentage of risk factor given that do/don't have hypertension.

Effect	Estimate of Odds Ratio (95% CI)
is_smoke yes vs no	1.157 (1.022,1.310)
gender male vs. _female	1.102 (0.974,1.248)
black yes vs no	1.722 (1.480,2.005)
overweight yes vs no	1.922 (1.662,2.223)

Table 1b: The Individual Odd Ratios of Obtaining Hypertension, Categorical variables. An univariate logistic regression of each categorical covariate on hypertension yielded odds ratio. For instance, smokers are 1.157 times more likely to develop hypertension than non-smokers. These univariate results, perhaps except gender, follow what is expected from literature.

Since urinary inorganic arsenic has an unbalanced distribution (min=2.69 $\mu g/L$, median=6.14, mean=8.41, 90th percentile =13.80 , 95th =18.86, 99th Percentile=38.39, max=628.2 $\mu g/L$, standard

deviation=14.03), the natural logathrim transformation allowed the data to be symmetric, removing the effect of influential observations. As the range is similar for the other metals: Pb, iHg, and Cd, the natural logarithm transformation was also taken. These transformations impact the results of logistic regression analysis but do not impact WQS as it uses the quartiles of the metals. The effect of arsenic on hypertension, adjusted for creatinine was also assessed; from the graph, there appears to be no relationship between hypertension and lniAs (**Figure 1**).

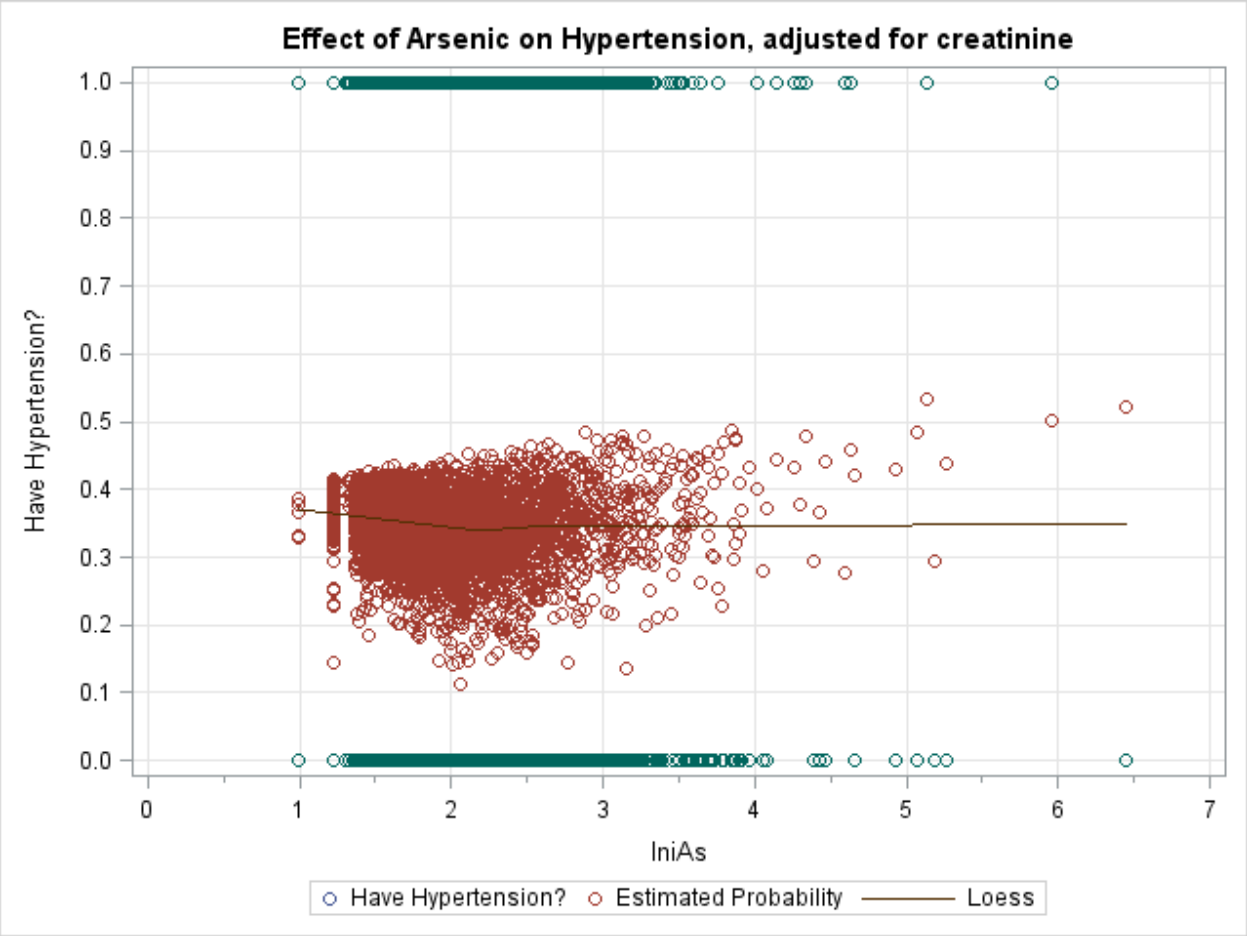


Figure 1: A LOESS Curve of Log of Inorganic Arsenic versus Hypertension: Since iAs has an unbalanced distribution (min=2.69 $\mu g/L$, median=6.14, mean=8.41, 75th percentile=9.02, 90th =13.80 , 95th =18.86, 99th Percentile=38.39, max=628.2 $\mu g/L$, standard deviation=14.03), a logarithmic transformation removed the effect of influential observations. To get a sense of relationship with log inorganic arsenic (lniAs), a loess curve was constructed versus the binary response hypertension using automatic smoothing procedure from the data (in sgplot), compared against the predicted values from a logisitic regression containing lniAs and creatinine as the only two covariates. There is little data beyond lniAs=5 so the relationship should be ignored.

3.2 Logistic Regression

For the covariate model, the quadratic terms of continuous covariates were considered **(Table 2).** The likelihood ratio tests using a significance of 0.25 showed that only adding polynomic effects of age should be included; all other quadratic effects do not significantly improve the likelihood. A LOESS plot of age with hypertension and the likelihood ratio test agree that the relationship between age and hypertension is quadratic (**Figure 2**). Thus, age^2 is kept in the final model, as indicated by the vector θ in Equation *(1)*.

Obs		Added Parameter Estimates	-2LL	Difference in -2LL	Decision
1	Cov	---	-2085.9	---	---
2	$Cov + alclusg^2$	0.00225	-2085.88	0.0488	.
3	$Cov + PIR^2$	0.0159	-2085.56	0.6727	.
4	Cov+age^2	-0.00110	-2061.98	47.8311	include age^2

Table 2: Likelihood Ratio Results from the Inclusion of Quadratic Effects of Continuous Variables. The base covariate model (not considering any quadratic effects: Cov) serves as the null model. Individual models that add quadratic effects for alcohol usage (#2), poverty index ratio (PIR, #3), and age (#4) The -2LL for each model, their difference, and the decision to keep the model is given (where a dot "."signifies the added variable is dropped). Compared with a significance level of 0.25, only the quadratic effect of age should be included.

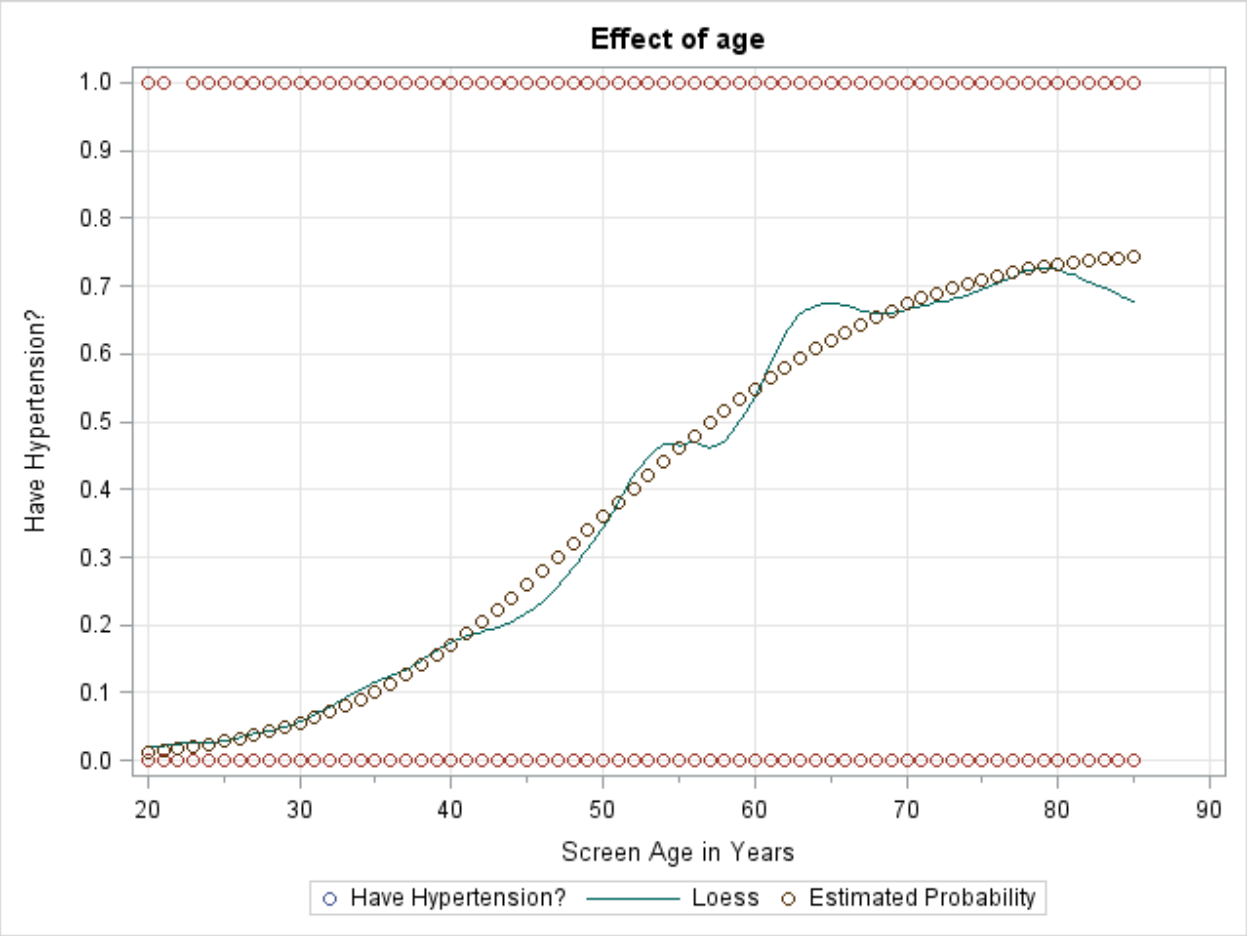


Figure 2: Comparing Age versus Hypertension: The predicted probabilities from a logisitic model containing age and age^2 as the covariates matches the loess curve from orginal data (smoothing parameter automatically selected). This suggests that hypertension and age have a quadratic relationship, due to the curvature seen from age 45 to 60, which is consistent with the results from the likelihood ratio test.

After adding inorganic arsenic, the parameter estimates and odds ratio are given (**Table 3)**. Although some covariates like alcohol usage, smoking, and gender do not increase the odds of obtaining hypertension, they are included as they are important biological variables so they remained in the model. The adjusted odds of obtaining hypertension for 1-unit increase in log of inorganic arsenic is 1.093, which is not significant (P-value=0.2659). A LRT suggests that adding inorganic arsenic term does not improve the model, supporting the results from **Table 3**. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) support the decision of adding the quadratic effect. The AIC and BIC increase when adding log inorganic arsenic, and the log likelihood decreases slightly. All of the criteria

agree that inorganic arsenic, adjusted for common covariates, is actually a worse model than one that consists of covariates + age^2 (**Table 4**).

Parameter	Estimate	Standard Error	Odds Ratio (95% CI)	P-value Wald's Test
Intercept	-8.2707	0.5146	----------	---------
lniAs	0.0886	0.0797	1.093 (0.935 , 1.277)	0.2659
creatinine	-0.00162	0.000675	0.998 (0.997 , 1.000)	0.0168
Difference in PIR	-0.0725	0.0247	0.930 (0.886 , 0.976) *	0.0033
age	0.1979	0.0186	1.218 (1.174 , 1.262) *	<.0001
age^2	-0.00109	0.000165		
Alcohol Usage: # of drinks per week in the past year	0.00610	0.0204	1.006 (0.967 , 1.047)	0.7654
smoker: yes	-0.0594	0.0812	0.942 (0.804 , 1.105)	0.4647
gender: male	0.0310	0.0825	1.031 (0.877 , 1.213)	0.7075
black: yes	0.8447	0.0991	2.327 (1.916 , 2.826) *	<.0001
overweight: yes	0.6194	0.0910	1.858 (1.554 , 2.221) *	<.0001

Table 3: Parameter Estimates and Odds Ratio for Final Logistic Model: The main variable of interest along with its odds ratio is highlighted. Significance at the 0.05 level is indicated by an asterisk.

Criterion	Intercept Only	Covariate Model	Covariates+ Age^2	COVARIATE+AGE^2+lniAs
AIC	5683.751	4189.793	4143.969	4144.731
BIC	5690.137	4247.269	4207.831	4214.979
-2 Log L	5681.751	4171.793	4123.969	4122.731

The difference in from final covariate model(#3) to lniAs(#4) is as follows:

-2LL (Δ) : 1.22905 **($\Delta \sim X_1^2$** ($Crit\ value = 3.84$) $P-value$: **0.1368**

Table 4: Model Selection Criterion across Various Models: The Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and -2 Log Likelihood. The base covariate model is considered along with adding the quadratic effect of age and log inorganic arsenic. Various information

criteria including Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) suggest that the model with lniAs is slightly worse.

As the data suggests that risk factor of inorganic arsenic adjusted for the covariates is not significant, the interaction terms were not considered. The question becomes: after controlling for these known confounders (smoking, age, alcohol usage, gender, etc.), does inorganic arsenic along with other metals—cadmium (Cd), total mercury (tHg), and lead (Pb)—significantly increase the risk of hypertension?

3.3 Weighted Quantile Sum Regression Analysis

The Weighted Quantile Sum approach takes advantage of the correlation among the metals and between the metals and outcome. The Pearson correlations between log of the metals range from -0.0426 to 0.325, which is significant (P-values not shown) but relatively low (**Table 5a**). Lead and cadmium as well as total mercury and inorganic arsenic appear to be fairly correlated indicating that these correlations must be taken into account in subsequent regression analyses. Lead seems to be fairly correlated with avgSP, which may indicate that the lead is the primary contributor in the sum. lnPb and all the other metals have P-value<0.0001. The correlations between log inorganic arsenic and average SBP was -0.023 and -0.003, respectively. Inorganic Arsenic, Cadimum, and total mercury have low correlations with systolic or diastolic blood pressure.

To get a sense how the metals are related to the outcomes, odds ratios for the quartiles of each metal are performed (adjusted for covariates θ) (**Table 5b**). Looking at quartiles enables to see the effect at higher doses of each metal. The table suggests that there is no evidence of a relationship of each of the metals with hypertension. For inorganic arsenic, the risk increases. For lead, there appears to be no effect as the concentration increases. Moderate levels of total mercury or cadmium appear to decrease the risk, but low or high doses increase it. Since all the confidence intervals for the odds ratio contain 1, the table suggests that each of the metals alone is not associated with hypertension. Due to

the correlated nature of these metals, the weighted quantile sum method is justified in identifying the bad actors for hypertension.

		avgSP	avgDP	lniAs	lnCd	lnPb	lntHg
	N	4307	4307	4386	4386	4386	4386
avgSP	4307	1.000					
avgDP	4307	0.337	1.000				
lniAs	4386	-0.027	-0.003	1.000			
lnCd	4386	0.083	-0.004	-0.035	1.000		
lnPb	4386	**0.271**	0.062	0.076	0.325	1.000	
lntHg	4386	0.005	0.060	0.276	-0.043	0.098	1.000

Table 5a: Pearson Correlation Coefficients for Log of Metals and Blood Pressure: The logs are taken to prevent the impact of influential observations as these distributions are skewed. Correlations of the metals (N=4386) with the average systolic blood pressure (avgSP) and diastolic (avgDP) (N=4307) uses the smaller sample size.

Obs	Effect		Adjusted Odds Ratio Estimate	Lower 95% CL	Upper 95% CL
1	iAsq	1 vs 0	1.055	0.846	1.316
2	iAsq	2 vs 0	1.093	0.861	1.387
3	iAsq	3 vs 0	1.110	0.862	1.431
4	Cdq	1 vs 0	1.085	0.862	1.367
5	Cdq	2 vs 0	0.965	0.764	1.220
6	Cdq	3 vs 0	1.064	0.827	1.368
7	Pbq	1 vs 0	1.004	0.782	1.290
8	Pbq	2 vs 0	1.071	0.829	1.384
9	Pbq	3 vs 0	0.955	0.728	1.251
10	tHgq	1 vs 0	1.067	0.857	1.328
11	tHgq	2 vs 0	0.930	0.746	1.160
12	tHgq	3 vs 0	1.017	0.813	1.271

Table 5b: Adjusted Odd Ratios for Quartiles of Metals in Mixture using Logistic Regression Individually. The odds ratio and its 95% Wald CI's are adjusted for the covariates θ in the model.

The data were randomly split between the test and validation test sets with 40% of the data going to the test set. The covariates and metals are similar in their summary statistics in each dataset **(Table 6).**

Variable	Dataset		
	Test Mean(std)	**Validation Mean (std)**	**Total**
Total # (%)	*1743 (39.74)*	*2643(60.26)*	*4386*
Creatinine, urine (mg/dL)	122.01 (76.46)	123.97 (76.51)	-------
Difference in PIR	1.59 (1.58)	1.54 (1.61)	-------
Screen Age in Years	49.55 (18.08)	49.21 (17.86)	-------
Square of Age	2781.42 (1856.28)	2740.27 (1834.54)	-------
Alcohol Usage: # of drinks per week in the past year	1.11 (1.90)	1.06 (1.83)	-------
Total Urinary Inorganic Arsenic (μg/L)	8.27 (9.90)	8.50 (16.19)	-------
Cadmium (μmol/L)	0.00 (0.00)	0.00 (0.01)	-------
Lead (μmol/L)	0.09 (0.07)	0.09 (0.09)	-------
Mercury, total (μmol/L)	8.17 (12.55)	7.58 (10.01)	-------
Quartiles of iAs	1.48 (1.13)	1.51 (1.11)	-------
Quartiles of Cd	1.52 (1.11)	1.51 (1.10)	-------
Quartiles of Pb	1.51 (1.13)	1.50 (1.11)	-------
Quartiles of tHg	1.52 (1.12)	1.49 (1.12)	-------
Variable	**Test #(%)**	**Validation #(%)**	
hypertension (% Disease)	609 (34.94)	928 (35.11)	1537
is_smoke (% yes)	837 (48.02)	1327 (50.21)	2164
Gender (% male)	861 (49.40)	1307 (49.45)	2168
black	337 (19.33)	518 (19.60)	855
overweight	1211 (69.48)	1895 (71.70)	3106

Table 6: Comparison of Covariates between Test and Validation Datasets: There is no difference between datasets as the means and percentages are very close between subjects. The total number indicates that all observations are included in the analysis (Table 1).

Estimation in Test Dataset

To estimate the covariates, one hundred bootstraps were performed on the nonlinear regression (of equation (*2*) in maximizing a binomial likelihood in the test dataset. For all replicates, the

weight estimates and beta are given. (Appendix I). A summary of the weights and the beta term adjusted for the covariates show that lead contains most of the weight and thus is the primary contributor in increasing the risk of hypertension (**Table 7**) (**Figure 3**) **.**

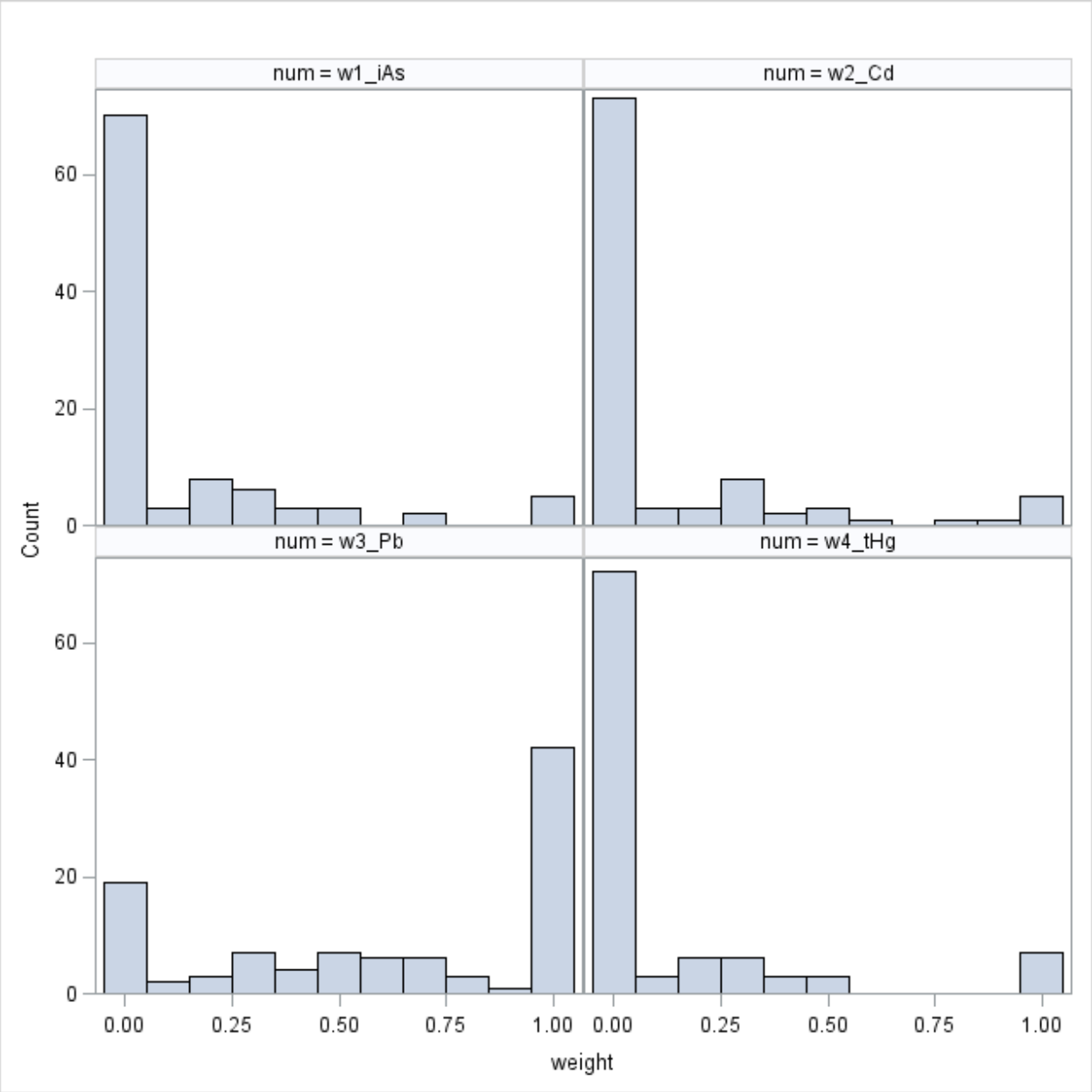


Figure 3: Histograms of weights across bootstraps after processing. After removal of small negative weights (to 0), the histograms reveal the spread of the data due to the complex correlations of the metals. While iAs, Cd, and tHg are mainly 0, there are some nonzero weights. The mean is like "average effect" of each metal on hypertension. Lead is the primary contributor to hypertension.

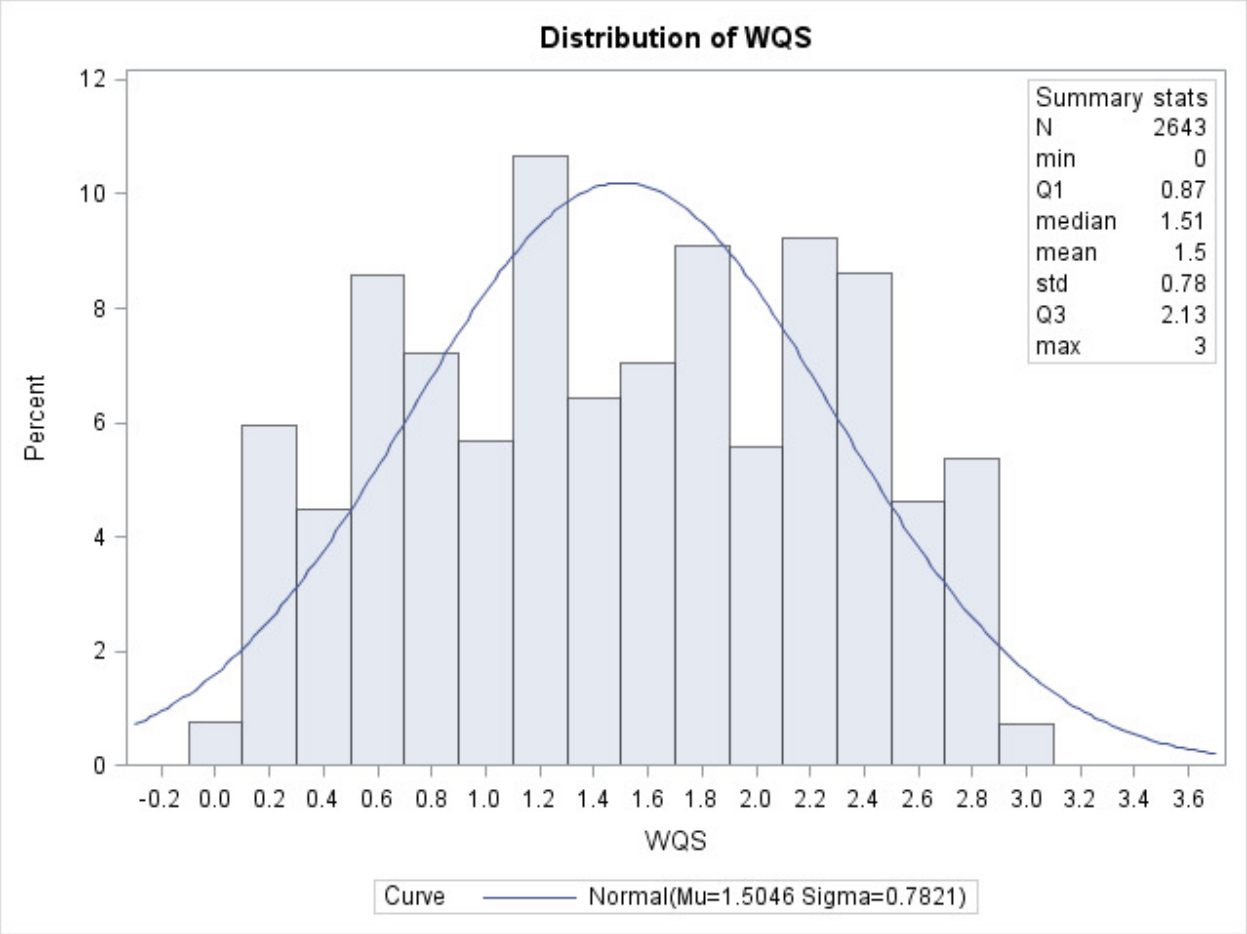
Variable	beta	w1	w2	w3	w4
		weights for iAs	weights for Cd	weights for Pb	weights for tHg
N	100	100	100	100	100
min	0	0	0	0	0
Q1	0	0	0	0.2589	0
median	0.0211	0	0	0.7296	0
mean	0.0378	0.1273	0.1269	0.6154	0.1305
std	0.0468	0.2535	0.2656	0.3962	0.2719
Q3	0.0616	0.1809	0.1276	1	0.1436
max	0.2058	1	1	1	1

Table 7: Statistical Summary of Bootstrap Weights and Beta Term: After removing the small negative value as 0, statistics over the 100 bootstraps are shown. Note that all constraints are met. The mean weight of lead consist the majority of contribution to hypertension, but the standard deviation is quite large from sample to sample, indicating the complex correlation among these metals.

3.4 Validation Dataset—Testing WQS

To get a sense of what the distribution of WQS looks like, a histogram describes the total effect of metals on hypertension **(Figure 4).** The sum ranges from 0 (no concentration of metals) to 3 (the hypothesized most impact on hypertension).

From ordinary logistic regression, an individual was 1.027 (0.882, 1.196) times as likely to develop hypertension from a one-unit increase in the weighted quantile sum, adjusted for other covariates **(Table 8)**. Or in other words, the aggregated sum of these four metals—inorganic arsenic, cadmium, lead, and total mercury—did not significantly increase the risk of hypertension. A likelihood ratio test of WQS+covariate against the covariate model agrees that WQS is not significantly associated ($\Delta 2LL=-0.101$, P-value=0.7404). The largest contributor to hypertension was lead, which constitutes 62% of the weight.



Variable	w1	w2	w3	w4
Label	weights for iAs	weights for Cd	weights for Pb	weights for tHg
Mean	0.13	0.13	0.62	0.13

Figure 4: Distribution of WQS in Validation Dataset: Using the average weights, the weighted quantile score was conducted. The means are the "weights" in the sum, and range from 0 (no effect on hypertension) to 3 (total effect on hypertension) depending on the levels for each chemical. Such a distribution shows the total mixture sum, which is usually between 1 and 2 in this dataset.

Parameter	Estimate	Standard Error	Odds Ratio (95% CI)	P-value (Wald's)
Intercept	-7.6053	0.6452	----------	<.0001
WQS	0.0269	0.0777	1.027 (0.882 , 1.196)	0.7292
creatinine	-0.00145	0.000752	0.999 (0.997 , 1.000)	0.0541
Difference in PIR	-0.0846	0.0319	0.919 (0.863 , 0.978)	0.0080
age	0.1792	0.0241	1.1952 (1.141, 1.252)	<.0001
age^2	-0.00091	0.000212		
alcohol usage	0.0107	0.0271	1.011 (0.958 , 1.066)	0.6925
smoker: yes	-0.1743	0.1069	0.840 (0.681 , 1.036)	0.1032
gender: male	0.0376	0.1084	1.038 (0.840, 1.284)	0.7290
black: yes	0.9132	0.1280	2.492 (1.939, 3.203)	<.0001
overweight: yes	0.5590	0.1188	1.749 (1.386 , 2.208)	<.0001

Table 8: Significance of WQS in Validation Dataset: A logistic regression was completed using the weighted quantile sum as a predictor in the model controlling for the same set of covariates as before. Parameter Estimates, standard errors, p-values, and odds ratios are shown. The adjusted odds ratio for WQS was 1.027, indicating that the overall effect of metals is not a risk factor for hypertension.

Individuals who immediately smoked, drank coffee, ate, or drank alcohol 30 minutes before, which would affect the measaurements on blood pressure, were included in the study. [16,18]If these individuals are excluded, the sample size is further reduced to 3183. The odds ratio (1.084) and 95% CI for lniAs are approximately the same as in logisitic regression case (95% CI=0.906 to 1.297) (P-value: 0.3795). For weighted quantile regression, the primary contributor for hypertension became Cadmium, instead of lead. However, this could be due to instability in the bootstrapping algorithm. The point

estimate of the adjusted odds ratio for WQS changed to 0.942, but it is still within the original 95% confidence interval, and therefore remains unsignificant. There was decrease in sample size, roughly 1,200 individuals were removed, so one would expect a lost of power.

4. Discussion

In conclusion, the data argue that there was not significant evidence to claim inorganic arsenic alone and with lead, cadmium and total mercury are associated with hypertension. Using logistic regression after adjusting for confounders of hypertension in literature, an individual is 1.093 (95% CI=0.935, 1.277 =95%) times more likely to have hypertension for a one-unit increase in log arsenic. This result could be misleading as there are often other toxins in the bloodstream, such as lead, inorganic mercury, and cadmium. Since these metals correlate with each other and are known to be risk factors for hypertension, [38-40] the weighted quantile sum method was employed, capitalizing on these correlations. The "bad actor" was found to be lead, on average consisting of 63% of the weight and poising the most risk for hypertension. However, the effect of the aggregated weighted sum on hypertension was also not significant, only increasing the odds of obtaining hypertension by 1.027 (95% CI=0.882 , 1.196) for a unit increase in WQS, after adjusting for the same covariates.

Due to the national complex sampling of NHANES, geographic locations in the United States could not be assessed [12] but is known to affect arsenic levels in drinking water. [6,19] In other words, groundwater in some states would contain more inorganic arsenic and consequently higher levels than other states. Likewise, the diets across the country are different, so that areas where people eat more protein, which is known to quickly eliminate arsenic, [6] would affect urinary arsenic levels. By ignoring these likely significant variables, the parameter estimates for arsenic may be biased.

Despite having a large sample size of roughly 4,400, the study may have been underpowered. The amount of urinary inorganic arsenic is low in United States compared to the rest of the world (like

Bangladesh); specifically, the total urinary arsenic concentration in United States is roughly 13 to 50-fold under Bangladesh. The clinical difference is smaller still since only low concentrations of urinary arsenic (5 to 10 μg/L) are needed for the effect. Thus, additional subjects may be needed to see the effect of inorganic arsenic on hypertension in the United States.

An advantage of a cross-sectional analysis is to explore many different risk factors for a disease, like hypertension in this study. Poverty index ratio (PIR), age, non-Hispanic black, and overweight were discovered to be risk factors for hypertension. Although biologically important in literature, there was not sufficient evidence to suggest that smoking, alcohol usage, and gender were associated with hypertension in this study. The other covariates may explain the variation of these three covariates and/or the sample size was too small to detect a difference. However, some limitations of a cross-sectional study include the lack of a controlled/low-exposed arsenic group, inability to assess incidence rates of hypertension, and the tendency toward bias. A cohort study can overcome these limitations since it has more power and is able to identify whether arsenic is a risk factor for hypertension.

Still, few studies exist that explore the relationship between urinary inorganic arsenic and hypertension in the United States. Additionally, this study presents a novel approach considering the joint effect of inorganic arsenic and other metals commonly found in humans on hypertension. Interestingly, the results from this study were similar to those found in Jones et.al. 2011, which did not find that total urinary arsenic was associated with hypertension. Some of the covariates were defined differently, such as using poverty index ratio instead of education to measure socioeconomic status or using a binary smoking variable instead of two or more categories to define smoking. In contradiction to the claims made here, a cross-sectional study that compared two Romanian towns concluded that low-to-moderate iAs exposure increases hypertension. The water levels in Romania of low/moderate levels (<0.1–240 μg/L) of inorganic arsenic in drinking water are comparable to the United States. [8] Thus, the

effect of low/moderate inorganic arsenic exposure on hypertension is still unclear. (Although what is considered "low-to-moderate" varies across studies, many fall within the range of 10-140 μg/L; [8,9,23,24,30] What is considered "high" is above 150 μg/L. [23])

Further studies, ideally a large cohort is needed to answer if inorganic arsenic—the sum of the species (arsenic acid (AS(V)O4), arsenite acid (AS(III)O3), DMA, and MMA [2,10])--increases the risk of hypertension after taking into account the covariates: creatinine, poverty index ratio, age, alcohol usage, non-Hispanic black, gender, smoker, and overweight as well as geographic location.

Chapter 3: Future Work and Conclusions

1. Conclusions

The goal of this thesis was twofold: to determine if inorganic arsenic alone and with other chemicals (cadmium, lead, and total mercury) had an effect on hypertension. Although inorganic arsenic in review articles has been indicated to affect nearly every major organ system after consuming groundwater, total arsenic has typically been used in several literature studies to be a risk factor for hypertension, which serves as a surrogate for cardiovascular disease and stroke. Most of these studies used data from countries where the arsenic levels in the water were higher than in the United States. [1,5,11,24] Additionally, as inorganic arsenic is not the only metal found in humans, other metals such as lead, cadmium, and total mercury are known to interact and be associated with hypertension. [38-40] There was no study to the best of our knowledge in determining if inorganic arsenic (and with other metals) as defined in literature affects hypertension in the United States.

The data from this study gave evidence that inorganic arsenic alone and with lead, cadmium, and total mercury are not significantly associated with hypertension. Using logistic regression after adjusting for covariates—which were found to influence hypertension in literature—an individual is 1.093 (95% CI=(0.935, 1.277)) more likely to develop hypertension for a one-unit increase in log arsenic. Looking at inorganic arsenic with the other metals, the individual is 1.026 (95% CI=(0.881, 1.194)) times as likely to develop hypertension for a one-unit increase in weighted quantile sum, adjusted for these same covariates.

This result fits into the ambiguity found in the literature, which contradictorily suggests that inorganic arsenic is a risk factor for hypertension in low-to-medium exposures. The non-significance found here can be explained as this study is underpowered; the range of inorganic arsenic in drinking

water in the United States is 70-fold lower under those in foreign countries, like Bangladesh. Some, like Jones, et.al., agreed with this study and did not find evidence that total arsenic is associated with hypertension. Conflicting with this study, the conclusion from a Romanian study (which has comparable arsenic exposure as the United States) was that inorganic arsenic is associated with hypertension.[8] Hoping to clarify the relationship, a large cohort study that includes these variables in this study as well as geographic location would realistically assess if inorganic arsenic (alone or with other metals) plays a role in the development of cardiovascular disease and stroke for low-to-moderate concentrations in drinking water (10-140 μg/L).

2. Future Work & Limitations

There were some limitations in the study. Against standard medical practices, NHANES defined hypertension as using three and sometimes four blood pressure readings on the same day. Since blood pressure varies tremendously day-to-day, the medical field determines hypertension with readings over weeks, not days. Although those who ate, drank caffeine, and/or smoked thirty minutes before being read affect blood pressure and were included in the study, [16-18] the exclusion of these individuals did not affect the analyses (results not shown).

In the near future, we would like to expand the methodology used in this paper to determine if there are other methods that could better explain the harmful effect of arsenic alone and with cadmium, total mercury, and lead in the human body. Looking at Figure 1, the logistic model explains some variability that the data do not indicate; could this be a misspecification in the model or too flat of a LOESS curve? If we use the predicted probabilities from the full model, does it explain the data better? If we do a similar plot with the other metals, are the relationships similar? Other exploratory analyses such as factor analysis or cluster analysis using the correlation matrix between the metals as a measure of similarity might give some insight in explaining the results.

Due to the modest correlation among the metals (from -0.04 to 0.325), it would be interesting to compare principal components analysis and regular linear regression analysis with WQS. Principle component analysis (PCA), which creates a series of variables that are linear combinations of correlated variables which are regressed with the outcome, may be a meaningful comparison to weighted quantile sum regression. Additionally, a regular multiple regression model with the different metals (lniAs, lnCd, lnPb, lntHg) using (a) hypertension and (b) SBP and DBP jointly may give different results. Would the results change if we do a stepwise or criterion-based model selection on these models?

The hypertension outcome in this study results in a loss of information as it dichotomizes blood pressure. Instead, we might jointly model systolic and diastolic blood pressure with the use of antihypertensive drug as a covariate. Although hypertension is a precursor for cardiovascular disease, it may fail to be a surrogate for general human health. High levels of an inflammatory protein CRP signal a wide range of toxic effects, including cardiovascular diseases, autoimmune diseases, cancer, and pneumonia as it is a general indicator of amount of inflammation in the body.[43] If we use cholesterol or CRP instead as an outcome, would inorganic arsenic have an effect? All of these variables are collected in the NHANES datasets. Different models of hypertension would be aided by a literature survey first.

In Weighted Quantile Sum Regression, we used quartiles of the four metals because we were attempting to control for influential observations. However, again, the quartiles lead to a loss of information. What if we perform a WQS using quintiles or deciles, instead of quartiles? Even better still, what if we perform the weighted bootstrap on the continuous metals (Pb, tHg, Cd, and As) instead of quartizing them? Second, in order to boost the power in the validation set, we may do the regular WQS regression on the whole dataset using quartiles, quintiles, and deciles on hypertension (Equation (5)). Third, as Carrico et.al. [41] suggest that the choice of the signal function has little effect on the significance of WQS, is this true in this study? The signal function of one was used in this study; instead, if the signal

function selected the weights from significant nonzero betas, would the results change and where would they fall on the distribution of the weights?

Lastly, literature suggests that metabolites of lniAs--methylarsonite (MA) and dimethylarsonite (DMA)—may be relatively harmless (or even beneficial) compared to arsenite and arsenate. Thus, if true, the combination of these species as done in this study may have null effect on hypertension. Then, it might be of interest to model arsenite, arsenate, MA species, and DA species on hypertension. This model may change depending on the literature search regarding the verification of this claim. It would be interesting to see if arsenic and other compounds (such as cholesterol, iodine, urinary mercury, copper, phthalates, barium, etc.) in the bloodstream would affect hypertension.

Does the model selection procedure affect the result? In this analysis, all covariates were considered biologically important, and the quadratic terms were considered using likelihood ratio test. One alternative approach, the Factor-Litvak Procedure, would remove the covariate from the full model if its deletion results in insignificant change in regression coefficient of exposure (β_1). Or, using a manual forward algorithm, the pool of covariates is added, one at a time, and is kept until the AIC is improved. The final model is one with the lowest AIC. Using these two models--the Factor-Litvak Procedure and forward approach—would the results change?

Subsequently, if covariates are defined differently, the significance of these effects might change. If we defined hypertension only by SP and DP (i.e. exclude use of the antihypertensive medications) and define alcohol usage from the NHANES questions (at least twelve drinks per year (ALQ101) and per lifetime (ALQ110), the effect of log inorganic arsenic on hypertension is quasi-significant (P-value=0.081). Furthermore, if we also define smoking only using yes/no from at least 100 cigarettes in a lifetime, preliminary results indicate a significance in WQS (P-value=0.028). Lastly, would the effect change if the surrogate for socioeconomic status changed from PIR to Family monthly poverty level

category (INDFMMPC), which could be more interpretable than PIR and has slightly more data? These preliminary results indicate the importance of defining covariates and the lack of power that exists in this study. Nonetheless, additional work is needed to confirm these results.

After establishing that inorganic arsenic is a risk factor for hypertension or other numerous diseases cited in literature, the next aim would be to find lifestyle variables that would reduce the risk of inorganic arsenic. A previous study, however, identified that vitamin A, vitamin C, methionine, low-carbohydrate, low-protein, and low-fat diets decrease arsenic levels. [19] As the results are questionable due to the inconsistencies found in literature, these dietary factors, the percentage of high body fat, percentage of calories that come from fat, or the amount of salt could be examined to confirm that toxic effects of arsenic can be reduced.

3. Summary

This study presents a novel approach considering the joint effect of inorganic arsenic and other metals commonly found in humans on hypertension. This study did not find evidence that inorganic arsenic alone as well as with three other metals—lead, cadmium, and total mercury—are risk factor(s) for hypertension in the United States, where the arsenic concentration in drinking water is low-to-moderate (10-140 μg/L) [8,9,23,24,30]. However, this study gives first glimpses in how inorganic arsenic and the joint collection of metals in a low/moderate enviornment generally affect human health; larger studies, ideally a cohort one, are needed to confirm these results.

References

1. Ahsan H, Perrin M, Rahman A, et al. Associations between drinking water and urinary arsenic levels and skin lesions in bangladesh. *J Occup Environ Med*. 2000;42(12):1195-1201.

2. CDC. Biomonitoring Summary Arsenic CAS no. 7440-38-2. http://www.cdc.gov/biomonitoring/Arsenic_BiomonitoringSummary.html. Updated Oct 2012.

3. Gharibzadeh S, Hoseini SS. Arsenic exposure may be a risk factor for alzheimer's disease. *J Neuropsychiatry Clin Neurosc*i. 2008;20(4):501-501.

4. Guha Mazumder DN. Arsenic and non-malignant lung disease. *J Environ Sci Health A Tox Hazard Subst Environ En*g. 2007;42(12):1859-1867.

5. Islam MR, Khan I, Attia J, et al. Association between hypertension and chronic arsenic exposure in drinking water: A cross-sectional study in bangladesh. *Int J Environ Res Public Healt*h. 2012;9(12):4522-4536.

6. KAPAJ S, PETERSON H, LIBER K, BHATTACHARYA P. Human health effects from chronic arsenic Poisoning–A review. *Journal of Environmental Science and Health, Part* A. 2006;41(10):2399-2428. http://www.tandfonline.com/doi/abs/10.1080/10934520600873571. doi: 10.1080/10934520600873571.

7. Kozul CD, Ely KH, Enelow RI, Hamilton JW. Low-dose arsenic compromises the immune response to influenza A infection in vivo. *Environ Health Perspec*t. 2009;117(9):1441-1447.

8. Kunrath J, Gurzau E, Gurzau A, et al. Blood pressure hyperreactivity: An early cardiovascular risk in normotensive men exposed to low-to-moderate inorganic arsenic in drinking water. *J Hypertens*. 2013;31(2):361-369.

9. Li X, Li B, Xi S, Zheng Q, Wang D, Sun G. Association of urinary monomethylated arsenic concentration and risk of hypertension: A cross-sectional study from arsenic contaminated areas in northwestern china. *Environ Healt*h. 2013;12:37-069X-12-37.

10. Lindberg A, Vahter M. Chapter 5: Health effects of inorganic arsenic . *Arsenic in Groundwater: A World Proble*m. 2008:64-72.

11. Rahman M, Tondel M, Ahmad S, et al. Hypertension and arsenic exposure in bangladesh. . *Hypertensio*n. 1999;33(1):74-78.

12. Jones MR, Tellez-Plaza M, Sharrett AR, Guallar E, Navas-Acien A. Urine arsenic and hypertension in US adults: The 2003-2008 national health and nutrition examination survey. *Epidemiolog*y [similar nHANES study]. Mar 2011;22(2):153-161.

13. National Center for Health Statistics. Health, united states, 2010: With special feature on death and dying-hypertension prevelance. . 2011.;76-641496:24-67, 250.

14. Angell SY, Garg RK, Gwynn RC, Bash L, Thorpe LE, Frieden TR. Prevalence, awareness, treatment, and predictors of control of hypertension in new york city. *Circ Cardiovasc Qual Outcome*s. 2008;1(1):46-53.

15. Beckerman J, MD. **High blood pressure and stroke**. http://www.webmd.com/hypertension-high-blood-pressure/guide/hypertension-high-blood-pressure-stroke. Published October 31, 2013. Accessed 01/07, 2015.

16. Vorvick LJM. Blood pressure measurement. http://www.nlm.nih.gov/medlineplus/ency/article/007490.htm. Accessed 1/17, 2015.

17. National Heart Lung and Blood Institute (NHLBI). Your guide to lower blood pressure. http://www.nhlbi.nih.gov/hbp/index.html. Accessed 06/18, 2013.

18. Handler J. The importance of accurate blood pressure measurement. *Perm* J. 2009;13(3):51-54.

19. Otles S, Cagindi O. Health importance of arsenic in drinking water and food. *Environ Geochem Healt*h. 2010;32(4):367-371.

20. Ravenscroft P. Predicting the global distribution of arsenic pollution in groundwater. *Royal Geographical Society Annual International Conferenc*e. 2007.

21. U.S. Food and Drug Administration. FDA looks for answers on arsenic in rice. Food and Drug Adminstration (FDA) Web site. http://www.fda.gov/forconsumers/consumerupdates/ucm319827.htm. Published September 19, 2012. Updated 2013. Accessed Jan 2, 2015.

22. Davis MA, Mackenzie TA, Cottingham KL, Gilbert-Diamond D, Punshon T, Karagas MR. Rice consumption and urinary arsenic concentrations in U.S. children. *Environ Health Perspec*t. 2012;120(10):1418-1424.

23. National Toxicology Program--US Department of Health and Human Services. Arsenic. .

24. Mazumder DNG, Haque R, Ghosh N, et al. Arsenic in drinking water and the prevalence of respiratory effects in west bengal, india. *International Journal of Epidemiolog*y. 2000;29(6):1047-1052.

25. Spivey A. Arsenic and infectious disease: A potential factor in morbidity among bangladeshi children. *Environ Health Perspec*t. 2011;119(5):A218-A218.

26. National Center for Health Statistics. NHANES 2009-2010: Urinary total arsenic and speciated arsenics website. http://wwwn.cdc.gov/nchs/nhanes/2009-2010/UAS_F.htm. Updated 2011.

27. A. Gomez-Caminero, P. Howe, M. Hughes, et al, eds. *Arsenic and arsenic compounds.* Second ed. Geneva: United Nations Environment Programme, the International Labour Organization, and the World Health Organization; 2001. Dr J. Ng , ed.

28. Selene Chou PD, Carolyn Harper PD, ATSDR, Division of Toxicology and Environmental Medicine, Atlanta, GA, et al. Toxicological profile for arsenic. [CDC Arsenic Review]. August 2007.

29. Gruber J, Karagas M, Gilbert-Diamond D, et al. Associations between toenail arsenic concentration and dietary factors in a new hampshire population. *Nutrition Journa*l. 2012;11(1):45.

30. Li X, Li B, Xi S, Zheng Q, Lv X, Sun G. Prolonged environmental exposure of arsenic through drinking water on the risk of hypertension and type 2 diabetes. *Environ Sci Pollut Res In*t. 2013;20(11):8151-8161.

31. US Department of Health and Human Services: Center for Disease Control and Prevention (CDC). National health and nutrition examination survey, 2013–2014 overview . .

32. National Center for Health Statistics. National health and nutrition examination survey: Analytic guidelines, 1999–2010 data evaluation and methods research. *Vital and Health Statistic*s. 2013;2(161).

33. National Center for Health Statistics. NHANES 2007-2008:Urinary total arsenic and speciated arsenics website. http://wwwn.cdc.gov/nchs/nhanes/2007-2008/UAS_E.htm. Updated 2011.

34. National Center for Health Statistics. NHANES 2005-2006:Urinary total arsenic and speciated arsenics website. http://wwwn.cdc.gov/nchs/nhanes/2005-2006/UAS_D.htm. Updated 2011.

35. National Center for Health Statistics. NHANES 2009-2010: Demographic variables and sample weights (DEMO_F). http://wwwn.cdc.gov/nchs/nhanes/2009-2010/DEMO_F.htm. Updated 2011.

36. Wang Y, Zhang Q. Reply to M bishop: Poverty threshold as an indicator of the association between childhood overweight and socioeconomic status over time. *The American Journal of Clinical Nutritio*n. 2007;85(5):1437-1438.

37. National Center for Health Statistics. NHANES 2005-2010: Alcohol use (ages 20+) Websites. http://wwwn.cdc.gov/nchs/nhanes/2007-2008/ALQ_E.htm. Updated 2009.

38. Kopp SJ, Barron JT, Tow JP. Cardiovascular actions of lead and relationship to hypertension: A review. *Environ Health Perspec*t. 1988;78:91-99.

39. Tellez-Plaza M, Navas-Acien A, Crainiceanu CM, Guallar E. Cadmium exposure and hypertension in the 1999â€“2004 national health and nutrition examination survey (NHANES). *Environ Health Perspec*t. 2008;116(1):51-56.

40. Houston MC. Role of mercury toxicity in hypertension, cardiovascular disease, and stroke. *The Journal of Clinical Hypertensio*n. 2011;13(8):621-627.

41. Carrico C, Gennings C, Wheeler D, Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *Journal of Agricultural, Biological, and Environmental Statistic*s. 2014:1-21.

42. Nwankwo T, Yoon SS, Burt V, Gu Q. Hypertension among adults in the united states: National health and nutrition examination survey, 2011–2012. *NCHS data brief, no 133*. 2013.

43. Institute of Medicine (US) Committee on Qualification of Biomarkers and Surrogate Endpoints in Chronic Disease. Chapter 4: Case studies. . In: Micheel CM BJ, ed. *Evaluation of biomarkers and surrogate endpoints in chronic disease.* Vol 4. Washington (DC): National Academies Press (US); 2010. http://www.ncbi.nlm.nih.gov/books/NBK220298/. Accessed 3/11/2015.

Appendix I: **Estimates for Beta and Weights in Bootstrap for Test Dataset**

To estimate the covariates, one hundred bootstraps were performed on the nonlinear regression in maximizing a binomial likelihood in the test dataset. For all 100 replicates, the weight estimates and beta are given, adjusted for the other covariates. The constraint on beta was 1E-6, which occurred 43 times. For these bootstraps, the chemicals are not associated with hypertension. There are also some violation (on order of 10-17) to the weights constrained to be between 0 and 1. Rounding these small negative weight values did not change the mean weight in the WQS analysis.

Replicate	beta	w1	w2	w3	w4
1	1E-6	56E-18	0	1	0
2	0.0346	0	0	0.723	0.277
3	0.0122	0.1801	0.1124	0.7074	0
4	1E-6	0.3191	0.2798	0.3378	0.0633
5	1E-6	0.0334	0.9666	-7E-18	0
6	0.0398	12E-17	0	0	1
7	1E-6	0.5068	0	0.4932	0
8	1E-6	0	0.4538	0.5462	0
9	0.1428	0.0106	0	0.9894	0
10	0.1522	0	0.4677	0.5323	0
11	1E-6	0	11E-17	1	0
12	1E-6	69E-18	0	1	28E-18
13	0.0195	1	0	28E-18	0
14	0.0864	0.2213	0	0.4918	0.2869
15	0.0644	0.1816	0	0.8184	0
16	0.1186	0	0.2719	0.7281	0
17	0.108	0	0	1	19E-17
18	0.0814	0	69E-18	1	0
19	0.1078	0	0.802	0	0.198
20	1E-6	0.6678	0	0.1405	0.1917
21	0.0462	0	0	1	0
22	0.0693	0	1	0	0
23	0.0237	0	28E-18	0.9514	0.0486

Replicate	beta	w1	w2	w3	w4
24	0.0003	0	17E-17	1	0
25	0.0323	0	0	1	-3E-17
26	0.0372	0	56E-18	1	0
27	1E-6	0.2513	0.2011	0.5476	0
28	1E-6	83E-18	0	1	0
29	1E-6	0.3002	0	0.2451	0.4546
30	0.0126	0	42E-18	1	0
31	0.0241	0	28E-18	1	0
32	0.0535	0	1	49E-18	0
33	0.0628	0	0	1	0
34	0.0049	0	11E-17	1	0
35	0.0557	0	0.1003	0.8997	0
36	0.0489	0	0	1	0
37	0.0263	0	0	11E-17	1
38	1E-6	0	35E-19	11E-17	1
39	0.1232	2E-16	69E-19	1	0
40	1E-6	0.2672	0.285	0.4477	0
41	1E-6	0.5034	0	0.4966	0
42	0.2058	0.1168	0.0348	0.731	0.1173
43	1E-6	0.1865	0.1697	0.3453	0.2984
44	0.0404	1	0	93E-18	0
45	1E-6	0	0	69E-18	1
46	1E-6	0	0	0	1
47	0.066	1	0	0	0
48	0.028	0	28E-18	1	0
49	1E-6	0	0.1575	0.3129	0.5295
50	1E-6	0	1	0	0
51	1E-6	0.2353	0.2532	0.2643	0.2473

Replicate	beta	w1	w2	w3	w4
52	1E-6	0.3133	0	0.6867	0
53	1E-6	0.1869	0.3128	0.291	0.2093
54	0.0122	0	11E-17	1	0
55	0.002	0	-3E-17	28E-17	1
56	0.109	0	0.2675	0.7325	0
57	1E-6	0	11E-17	1	0
58	1E-6	0	0	1	0
59	1E-6	0	0.3921	0.6079	0
60	1E-6	1	0	28E-18	0
61	0.0462	0	0	0.6456	0.3544
62	1E-6	0	56E-18	1	-3E-17
63	0.0442	0	11E-17	1	0
64	0.0604	28E-18	0	28E-18	1
65	0.0294	0	0	1	0
66	0.1589	28E-18	0	1	12E-17
67	0.0325	0	15E-17	1	0
68	1E-6	0.6744	0	0.3256	0
69	1E-6	0	22E-17	1	0
70	0.0585	83E-18	0	1	0
71	0.1434	0.3589	0	0.5269	0.1142
72	0.105	0.1685	0	0.8315	0
73	1E-6	17E-19	0	0.8302	0.1698
74	1E-6	0.0558	0	0.6038	0.3404
75	1E-6	0.2493	0.2756	0.2536	0.2215
76	0.0712	0	0	1	0
77	1E-6	0	0	0.59	0.41
78	0.0851	31E-18	0	1	0
79	0.0253	0.4078	0	0.5922	0

Replicate	beta	w1	w2	w3	w4
80	1E-6	0.3147	0.405	0.2369	0.0435
81	0.1331	0	0.2927	0.3959	0.3114
82	1E-6	0	28E-18	1	-3E-17
83	0.0368	0	22E-17	1	0
84	1E-6	0.4687	0	0	0.5313
85	0.0143	0	56E-18	1	-1E-17
86	1E-6	0	0	1	17E-17
87	0.0834	1	0	0	0
88	0.0834	0	0.6054	0.3946	0
89	1E-6	0	0.9444	0.0556	0
90	0.0924	0	17E-10	1	0
91	1E-6	0	0.1427	0.5876	0.2697
92	0.03	0	17E-17	1	28E-18
93	0.0694	56E-18	0	1	0
94	1E-6	0.1293	0.4978	0.3729	35E-19
95	0.0521	0.407	0	0.2343	0.3587
96	0.0601	0.0094	0	0.9906	0
97	1E-6	0	-3E-17	1	97E-18
98	0.0435	62E-18	0	1	0
99	0.1312	12E-17	0	1	35E-19
100	0.0354	0	1	0	-3E-17

Appendix II: SAS Code

A2.1 Formats and Macros

```
*(1) General Macros;
      /*InputnHanes allows the user to create SAS datasets from .xport files
from nHanes database (change final extension where they are stored)
      filename=name of file needed to input
      outlib=the output SAS library, the default is the Work library
      */

      %Macro InputnHanes (filename= ,outlib=work) /
                  des='Create SAS datasets from .xport files from nHANES
database';
            libname into xport "C:\Users\pablo\Desktop\research\Arsenic
Project\Data\NHANES DATA\&filename";
            proc copy in=into out=&outlib; run;
            *proc contents VARNUM; run;
      %Mend InputnHanes;

      /*Copied from Categorizing a continuous outcome;
      Categorize-Macro that categorizes a continuous variable into
quartiles*/
      *The macro takes three arguments:
            *data--name of the dataset, default is to take the last ;
            *var--name of continuous variable;
            *label--label of that continuous variable;
            *quantiles--how the data should be divided. the default is 4,
which is quartiles;
            *the new variable is varq. ;

      %Macro Categorize (data=&SYSLAST , var= , label=&var, quantiles=4)/
            des='Categorizes a continuous variable into quartiles';     ;

                        proc rank data=&data  out=&data groups=4;
                              var &var;
                              ranks &var.q;
                        run;
                        data &data;
                        set &data;
                              attrib &var.q label="Quartiles of &label";
                        run;
    %Mend Categorize;


      /*From: Analyzing Data Libraries;
      The Macro DataLook creates a nice summary of the data using proc
univariate*/
      %Macro DataLook (data=&SYSLAST, var= ,format=best4.);

                  *ods rtf select Histogram;
                  *ods listing select  Histogram;
                  proc univariate  data=&data noprint;
```

```
                        var &var;
                        histogram /nmidpoints=20 normal(color=red);     /*The
number of “bins” */
                        inset
                              /*INSET Statement Adds a text box inside of the
axes of the plot. This provides a summary statistics.
                              SEE: *SAS options=“Labels of variables." */

                              n="N" min="min" q1="Q1" median="median"
mean="mean" std="std"               /*standard deviation*/
                              q3="Q3" max="max"
                        /
                                    position= ne
      /*Takes the plot in top right (eg. NE*/
                                    header= "Summary stats"
      /*Title of summary*/
                                    format= &format
                              /*The Slash indicates options within options.
Inset is an option of proc univariate.*/
                              ;
                        title3 "Summary Statistics for &var";
                  run;
                  ods select all;
      %Mend DataLook;


******#2: Logistic Regression;
      ********#1C: Assessing Linearity**************;
      *(C) Can do it by treating each variable as categorical;
            %Macro lineartest(data=  , y=  , var=);
                  title2="&var";
                  ods graphics on;
                  proc logistic data=&data noprint;
                        class &var;
                        model &y=&var/outest=estimate;
                  run;

                  data estimate;
                  set estimate;
                  level=_n_;
                  run;

                  *not sgdesign. Great if i could fit a regression line and
get R^2 value;
                  symbol1 v=star i=join;
                  proc gplot data=estimate;
                        plot estimate*level;
                  run;quit;
            %Mend lineartest;

      *Univariate Models*************;
                  ******#2: Univariate Models*************;

                  *The Constant Logistic Model;
                  %Macro Constant();
                        ods rtf select  NObs ResponseProfile FitStatistics
GlobalTests ParameterEstimates OddsRatios  LackFitChiSq;
```

```
                        ods listing select  NObs ResponseProfile
FitStatistics GlobalTests ParameterEstimates OddsRatios  LackFitChiSq;
                        proc logistic data=&data;
                              model &y=;
                              title2 "One-way analysis for Constant";
                        run;
                  %Mend Constant ();

                  %Macro BackwardElim ( cat=  ,  x=        , n=1) ;

                        option spool;

                        proc logistic data=&data outest=parm&n;
                                    class &cat /param=ref ref=first;
      *Categorical variables;
                                    model &y=&x &cat/Rsq lackfit;
                              run;

                        ods select all;
                  %Mend BackwardElim;

                  %Macro MultivariateLog  (cat=&cat2  ,  x=&x2 ,
interaction=   , n=1);
                        option spool;
                        title 'Multivariate interactions';
                        ods selecT NObs ResponseProfile ClassLevelInfo
FitStatistics Type3  GlobalTests   LackFitChiSq ParameterEstimates;
                              proc logistic data=&data outest=parm&n;
                                    class &cat/param=ref ref=first;
      *Categorical variables;
                                    model &y=&x &cat &interaction;
                                    title2 "for &interaction";
                              run;
                        ods select all;
                  %Mend MultivariateLog;



*-------------------------------------------------------------------------
----------------;

*(2) Formats
*Standard labelling in nhanes;
      proc format; *library=nhanes;
      value standardf         -1="too young"    /*usually children(age<=8)
but varies...data not collected*/
                                    1="yes"
                                    1.5="somewhat"
                              2="no"
                                    3="unable to obtain"
                                    7="refused"
                                    9="don't know"
                                    ;
      value standard2f  0="no"  1="yes";

      value sessionf    0="morning"  1="afternoon" 2="evening";
```

```
      *Refused/Don't know formats for continuous variables: Why are there 3
different formats?;
      *The number refers to the repetition of the "7" for each format;
      value refused1000f      77777="refused"
                                    99999="don't know"
                                    ;
      value refused100f 777    ="   Refused     "
                                    999    ="   Don't know  "
                                    ;
      value refused800f   888  = "could not obtain"         /*Unsure if i can
combine with refused 100f...*/
                                    ;
      value refused10f  77="Refused"
                                    99="Don't know"
                                    ;
      value responsef     1="Have response"
                                    7="Refused"
                                    9="Don't know"
                                    ;
      *-------------------------------------------------------------------
----------------------;

*General Demographics Characteristics--Hopefully more general than just
NHanes;

*Respondent Info;
value statusf
            1      ="   Interviewed Only  "
            2      ="   Both Interviewed and MEC examined   "
            ;
value examperiodf 1="November 1 through April 30"
                        2="May 1 through October 31"
                       ;
*NHANES DEFINITION OF Race, Gender, pregnancy;
value racef
                  1="Mexican American"
                  2="other Hispanic"
                  3="Non-Hispanic White"
                  4="Non-Hispanic Black"
                  5="Other Race/MultiRacial"
                  ;
            value race2f  1="_white_      "           /*Reference: Least risk
for hypertension. Can be used to compare against minority effect.*/
                              2="y-hispanic"
                              3="z-black "                /*Most risk for
hypertension*/
                              4="other/multiracial"
                             ;

            value malef 0="_female_"
                              1="male"            /*Slightly higher
risk for hypertension*/

                  ;                 /*RIAGENDR: NHanes--different order
1="Male", 2="Female"*/
*Pregnacy;
            value pregnagef   -1="Children(0<=age<8)"
```

```
                              0="Teens (8<=age<20)"
                              1="Adults (20<=age<=44)"
                              2="Menopause Years(44<=age<55)"
                              3="Old Women(age>=59)"
                    ;

value pregnantf 1="yes (postive lab test or self-reported)"
                    2="no (at time of exam)"
                    3="cannot ascertain pregnancy"
                    ;

      value gendpregnf
                    0="Male"    /*Reference for logistic regression*/
                    1="Pregnant Female"  /*(postive lab test or self-
reported)*/
                    2="Non-Pregnant Female" /* (at time of exam)*/
                    3="Unable to tell if Pregnant" /*cannot ascertain
pregnancy*/
                    4="Uncertain if pregnant"  /*"missing female pregnant"*/
                    ;
*Immigration;
 value bornf 1     = "born in USA"
                    2       ="   born in Mexico    "
                    3       =" born elsewhere"
                    4       ="   born in other Spanish speaking country    "
                    5       ="   born in other Non-Spanish speaking country
      "
                    7       ="   refused     "
                    9       ="   don't know  "
                    ;
value citizenf 1="citizen by birth or naturalization"
                    2="not a citizen of US"
                    7="refused"
                    9="don't know"
                    ;
value  USTimef
                    1       ="   <1 yr"
                    2       ="   1 to<5 yr"
                    3       ="   5 to <10 yr "
                    4       ="   10 to <15 yr"
                    5       ="   15 to <20 yr"
                    6       ="   20 to <30 yr"
                    7       ="   30 to <40 yr"
                    8       ="   40 to <50 yr"
                    9       ="   >=50 yr     "
                    60      ="life"          /*US citizens are in US for life.
Of course...some people may have different ages*/
                    77      ="   refused     "
                    99      ="   don't know  "
                    ;

value SPLangf 1="English"
                     2="Spanish"
                     ;

*Education;
      *Education Levels and Ages;
```

```
      value agecatf 1="Infants (age<=6)"
                    2="Children (6<age<20)"
                    3="Adults (age>=20)"
                    ;

      value inschoolf
                   -1="Too Young (Age <6)"
                   1="in school"
                   2="on vacation from school (between grades)"
                   3="not in school"
                   7="refused"
                   9="don't know"
                   11="Children Unsure if in school"
                   13="Done with school (adults)"            /*should be
adults, age>19*/
                   /*.="Age missing"*/
                   ;
      *Children Education LEvel;
      *The highest grade last completed;
      value ednlvlf
             -1    ="Too Young (Age <6)"
                  0      ="   Never Attended / Kindergarten Only  "
                  1      ="   1st Grade   "
                  2      ="   2nd Grade   "
                  3      ="   3rd Grade   "
                  4      ="   4th Grade   "
                  5      ="   5th Grade   "
                  6      ="   6th Grade   "
                  7      ="   7th Grade   "
                  8      ="   8th Grade   "
                  9      ="   9th Grade   "
                  10     ="   10th Grade  "
                  11     ="   11th Grade  "
                  12     ="   12th Grade, No Diploma  "
                  13     ="   High School Graduate    "
                  14     ="   GED or Equivalent "
                  15     ="   More than high school   "
                  55     ="   Less Than 5th Grade     "
                  66     ="   Less Than 9th Grade     "
                  77     ="   Refused     "
                  99     ="   Don't know  "
                  113  =" Too Old (Adults)";
                  ;
      *Adult education level;
      value adultednf
             -1    ="Too Young (Age <20)"
                  1="Less Than 9th Grade"                  /*Change to 8.5*/
                  2="Dropped out of HS"                          /*12.5*/
                  3="High School Grad/GED or Equivalent"   /*13.5*/
                  4="Some College or AA degree"            /*16*/
                  5="More than college"                          /*17*/
                  7="Refused"                                          /*77
(same as child)*/
                  9="don't know"                                       /*99
(same as child) */
                  111="Adults with unknown education levels (not said,
missing)" /*111 (same as child)*/
```

```
                ;
      *My variable;
      value eduf
             -1="Too Young (Age <6)"
                0="Not Educated (Dropped out/not attend school)"
                /*12.5*/
                1="Currently in School" /*13.5*/
                2="High School Degree or less"                  /*16*/
                3="some college"
                4="More than college"                           /*17*/
                7="Refused/don't know"
      /*77 (same as child)*/
                13="Too Old"
      ;
      *Is educated?;
      value edu2f
                -1="too young"          /*NOTE: THIS TOTALLY CONFOUNDS WITH
AGE (THAT IS IF AGE<=6 is considered in model). NEED TO COMBINE WITH AGE*/
                0="not educated"
                1="educated"
                7="Refused/don't know"
                ;

*Married;
      value marriedf 1="married"
              2="widowed"
              3="divorced"
              4="seperated"
              5="single"
              6="living with partner"
                77="refused"
                99="dont know"
            ;
      *My Variable;
      value married2f 1="married"
                           2="once married"
                           3="single"
                           7="don't know or refused"
                           /*999="issue: investigate"*/
                ;
*Household Size;
value sizef 7=">=7 indv";

*Socioeconomic Status
*PIR: no format;
*HS and Family Income;

value incf  1      ="   $ 0 to $ 4,999    "
                2      ="   $ 5,000 to $ 9,999      "
                3      ="   $10,000 to $14,999      "
                4      ="   $15,000 to $19,999      "
                5      ="   $20,000 to $24,999      "
                6      ="   $25,000 to $34,999      "
                7      ="   $35,000 to $44,999      "
                8      ="   $45,000 to $54,999      "
                9      ="   $55,000 to $64,999      "
                10     ="   $65,000 to $74,999      "
```

```
                  11    =" $75,000 and Over"      /*Only Done in 1999-2006*/
                  12      ="   Over $20,000       "
                  13      ="   Under $20,000      "
                  14      =" $75,000 to $99,999   "      /*New Variables
starting in 2007*/
                  15      ="   $100,000 and Over  "
                  77      ="   Refused      "
                  99      ="   Don't know   "
                  ;

      *Monthly family Income;
      value mincf
      1       ="   $0 - $399    "
      2       ="   $400 - $799  "
      3       ="   $800 - $1249       "
      4       ="   $1250 - $1649      "
      5       ="   $1650 - $2099      "
      6       ="   $2100 - $2899      "
      7       ="   $2900 - $3749      "
      8       ="   $3750 - $4599      "
      9       ="   $4600 - $5399      "
      10      ="   $5400 - $6249      "
      11      ="   $6250 - $8399      "
      12      ="   $8400 and over     "
      77      ="   Refused      "
      99      ="   Don't know   "
      ;
      *Monthly family pir categories;
      value mpirf
      1       ="   Monthly poverty level index <= 1.30 "
      2       ="   1.30 < Monthly poverty level index <= 1.85       "
      3       ="   Monthly poverty level index > 1.85   "
      7       ="   Refused      "
      9       ="   Don't Know   "
      ;

      *Total Savings      ="          ";
      value savtotf
             1      ="   Less than $500     "
            2       ="   $501- $1000 "
            3       ="   $1001-$2000 "
            4       ="   $2001-$3000 "
            5       ="   $3001-$4000 "
            6       ="   $4001-$5000 "
            77      ="   Refused      "
            99      ="   Don't know   "
      ;


*-------------------------------------------------------------------------------
---------------;
*Format for My Variables (OTHER);
      *Body Measurements;
                  value weightcommentf 1="Could not obtain"
                                       2="Exceeds capacity"
                                       3="Clothing"
                                       4="Medical Appliance"
```

```
                                        .="Missing" /*Nothing Wrong*/
                                        ;
                value heightcommentf 1="Could not obtain"
                                        2="Exceeds capacity"
                                        3="Not Straight"
                                        .="Missing" /*Nothing WRong*/
                                        ;
                */////////;
                value BMIcatf 1="underweight"
                              1.5="_underweight/normal_"
                               2="normal"
                               3="y-overweight" /*at risk*/
                               4="z-obese"
                               ;
      *smoking;
            */////////;  value  smokef  -1="too young"  /*age<12,  removed in
analysis*/
                           1="smoker"     /*Defined (08/09/13): Smoked
at least 100 cigarettes in life OR  Used tabacco/nicotine last 5 days */
                           0="_non-smoker_"
                           7="don't know/refused" /*removed in
analysis*/
                           ;

                        value sechndf     1="second-hand smoker"
                                      1.5="somewhat second-hand"
                                      0= "not 2nd hand smoker"
                              ;

      *alcoholic;
                 value alcoholicf  1="alcoholic"
                                   2="somewhat alcoholic"
                                   3="borderline alcoholic"
                                   4="non-alcoholic"
                              ;
            */////////;
                  value isalcholicf  -1="too young (age<12)"
                                        -0.5="classified (12<=age<20)"
                                        1="alcoholic"   /*Defined
(08/09/13: Had at least 12 alcoholic drinks last year OR in lifetime*/
                                    0="_non-alcoholic_"
                                        ;
             *Alcoholic age category;
      value aagecatf          1="inegilble (age<12)"
                              2="classified (12<=age<20)"
                              3="eligible (age>20)"
                              4="age missing yet values ???"
                              ;

*************************************************************************
****************************************;
/*Non-Demographic Variables Formats*/

/*Toxic Metals*/
      value limitf      0="above the limit of detection"
                        1="below the limit of detection"
                        2="Corresponding missing value"   /*for metal XXX*/
```

```
                        ;

/*Blood Pressure and pulse*/
      value missingf                              /*Called Blood Pressure
Status*/
            1      ="   Complete     "
            2      ="   Partial      "
            3      ="   Not done     "
            /*     .     =" Missing  "*/
            ;
            value bpcommentf                    /*Blood Pressure Comment--
useful to address where missing data/partial data is*/
                  0    = "Missing"
                  1      ="   Safety exclusion   "
                  2      ="   SP refusal   "
                  3      ="   No time      "
                  4      ="   Physical limitation      "
                  5      ="   Communication problem    "
                  6      ="   Equipment failure  "
                  7      ="   SP ill/emergency   "
                  14     ="   Interrupted        "
                  56     ="   Came late/left early     "
                  84     ="   SP with child      "
                  99     ="   Other, specify     "
                  122    ="   Language barrier   "
                  ;
            value armslf
                  1      ="   Right  "
                  2      ="   Left   "
                  8      ="   could not obtain   "
                  ;
            value cuffsizef
                  1      ="   *1=Infant (6X12)   "
                  2      ="   2=Child (9X17)     "
                  3      ="   3=Adult (12X22)    "
                  4      ="   4=Large (15X32)    "
                  5      ="   5=Thigh (18X35);   "
                  ;
            value enhancef                /*Enchament-1*/
                  1="yes"
                  2="no"
                  8="could not obtain"
            ;

            *Pulse covariates;
            value pulsetypef
                  1     =     "     Radial      "
                  2     =     "     Brachial    "
                  8     =     "     Could not obtain  "
                  ;
            value pulseregularf
                  1     =     "     Regular     "
                  2     =     "     Irregular   "
                  /*.   =     "     Missing     "*/
                  ;
      *High Blood Pressure and Hypertension;
             value highBPf  1="normal"
```

```
                              2="pre-hypertension"
                                      3="Stage I hypertension"
                                      4="Stage II hypertension"
                                     -1="too young"
                                      ;
                value diseasef  0="normal"
                                      1="disease"
                                  -1="too young"
                                   ;
run;
```

A2.2 Obtaining the dataset

```
/*A complete set of code that includes all variables in one dataset
for inorganic Arsenic

libname data "C:\VCU Biostatistics\research\Arsenic Project\Data\Arsenic
Mixture Project Data (Final_Fall 2013)";              /*The output dataset*/
libname demog "C:\VCU Biostatistics\research\Arsenic Project\Data\Demographic
Datasets";  /*short for demographics*/
*libname format "C:\SAS and R Help\FORMAT LIB"; /*Save Common Demographic
Formats*/;
      OPTIONS FMTSEARCH=(data.nHANES);    /*Find format*/


*************************************************************************
**;
*************************************************************************
**;
```

*Summary of data;

```
            %let y=hypertension;
            %let bio=lniAs highfat highfat2;
            %let x=     iAs Cd      Pb    iHg   tHg   ;
                  %let lnx=   lnCd  lnPb  lniHg lntHg;
                  %let x_comm=      Cd_comm     Pb_comm     iHg_comm
      tHg_comm;
            %let theta=creatinine pirdiff agediff alchlusg;       /*list main
effects here*/
            %let cat2=is_smoke male black overweight;
      /*list any categorical variables here*/
            %let x2extra=     actscore    Ca    Mg    K     seafood;

            %let catextra=    veteran     SIALANG     citizen     highsalt
      married     hs_sz groundH20   edu   is_edu;
      *Categorizations of Continuous Variables;
                  %let xq=    iAsq Cdq    Pbq   iHgq  tHgq  ;
                  %let thetaq=creatinineq pirq ageq alchlusgq;

*************************************************************************
*********************************;
```

Demographics

```
*Covariates: Gender, Race, Age, Old;
      **I) create the demographic dataset;
            %InputnHanes (filename=DEMO_F.xpt, outlib=work);
            %InputnHanes (filename=DEMO_E.xpt, outlib=work);
      /*Demographics-Part E*/
            %InputnHanes (filename=DEMO_D.xpt, outlib=work);

            data demographics3;
            set demo_d demo_e demo_f;
            *Modify Where you are born;
                  if (sddsrvyr=5) then do;
                  if dmdborn2 in (4,5) then dmdborn = 3;
                  else dmdborn = dmdborn2;

                  if DMDHRBR2 in (4,5) then DMDHRBRN = 3;
```

```
                  else DMDHRBR2 = DMDHRBRN;

                        /*DMDBORN2: Country of birth was recoded into five
categories: 1) Born in 50 US States or Washington, DC; 2) Born in Mexico; 3)
Born Elsewhere 4) Born in Other Spanish Speaking Country (Not Mexico); and 5)
Born in Other Non-Spanish Speaking Country. [Note: category 3 is a place
holder to combine across other survey years- see analytic notes]
                        DMDBORN: Country of birth was recoded into three
categories: 1) born in one of the 50 U.S. states or Washington, D.C.; 2) born
in Mexico; and 3) born in any other location or foreign country.
                                          */

                  *Household Income;
                        if indhhin2 in (14,15) then indhhinc = 11;
                Else indhhinc = indhhin2;

                  end;
                  *drop old variables--okay to do this here???. I'm getting
rid of data--is that okay?;
                  drop dmdborn2 dmdhrbr2 indhhin2 indfmin2;
            run;

            *Income File;
                  *In NHANES DOCUMENTATION, the files are same;
            %InputnHanes (filename=INQ_F.xpt, outlib=work);
            %InputnHanes (filename=INQ_E.xpt, outlib=work);
*           %InputnHanes (filename=INQ_D.xpt, outlib=work);
                        *Somehow, although mentioned in demogrpahic file is
missing in questionaires. Maybe started collected in 2007?;
            data income;
            retain SEQN
                  /*monthly family income*/ IND235    INDFMMPI    INDFMMPC
                  /*source of income*/    INQ012      INQ020      INQ030
      INQ060      INQ080      INQ090      INQ132      INQ140      INQ150
                  /*savings*/ INQ244 IND247
                  ;
            set inq_e inq_f;
            run;
            *One dataset: For income, only kept family income #'s. Source of
income was dropped;
            data demographics2;
            merge demographics3 income(keep=SEQN IND235     INDFMMPI
      INDFMMPC);
            by SEQN;
            run;

            ***II) Split up the dataset into 4 smaller datasets--that make
more sense;
                  *Sample Weights for analysis;
                  data weights;
                  set demographics2;
                  keep SEQN SDDSRVYR
                        /*Weights*/ WTINT2YR WTMEC2YR SDMVPSU     SDMVSTRA;
                  run;

            *Respondent info;
                  data respondinfo;
```

```
                retain SEQN SDDSRVYR
                      /*Exam Period*/         RIDSTATR RIDEXMON
            /*unsure if i need RIDSTATR or RIDEXMON in this dataset*/
                      /*Respondent Info*/ DMDHRGND  DMDHRAGE    DMDHRBRN
      DMDHREDU    DMDHRMAR    DMDHSEDU    SIALANG     SIAPROXY    SIAINTRP
      FIALANG     FIAPROXY    FIAINTRP    MIALANG     MIAPROXY    MIAINTRP
      AIALANG;
                set demographics2;
                keep SEQN SDDSRVYR RIDSTATR RIDEXMON DMDHRGND2  DMDHRAGE
      DMDHRBRN    DMDHREDU    DMDHRMAR    DMDHSEDU    SIALANG     SIAPROXY
      SIAINTRP    FIALANG     FIAPROXY    FIAINTRP    MIALANG     MIAPROXY
      MIAINTRP    AIALANG;
                run;

                *Demographics;
                data demographics;
                retain SEQN SDDSRVYR;
                /*general cat characteristics*/ retain    DMQMILIT RIDRETH1
DMDMARTL  RIAGENDR      RIDEXPRG    ;
                /*household*/ retain DMDHHSIZ DMDFMSIZ;
                /*age*/     retain RIDAGEEX    RIDAGEMN    RIDAGEYR;
                /*socioeconomic progress*/ retain INDHHINC      INDFMINC
      INDFMPIR  IND235  INDFMMPI    INDFMMPC ;
                /*education*/ retain DMDEDUC3 DMDEDUC2    DMDSCHOL;
                set demographics2;
                drop /*Weights*/  WTINT2YR WTMEC2YR SDMVPSU     SDMVSTRA
                      /*Respondent Info*/ RIDSTATR RIDEXMON  DMDHRGND
      DMDHRAGE    DMDHRBRN    DMDHREDU    DMDHRMAR    DMDHSEDU    SIALANG
      SIAPROXY    SIAINTRP    FIALANG     FIAPROXY    FIAINTRP    MIALANG
      MIAPROXY    MIAINTRP    AIALANG
                      /*immigration*/ DMDBORN DMDCITZN DMDYRSUS
                      ;
                      /*General Characteristics*/
                      rename      DMQMILIT    =     veteran     ;
                      *RIDRETH1; *DMDMARTL; *RIAGENDR; *RIDEXPRG;
                      rename      DMDHHSIZ    =     hs_sz ;
                      rename      DMDFMSIZ    =     fam_sz      ;
                      *Age;
                      /*Socioeconomic Progress*/
                      rename INDHHINC  = hs_inc;
                      rename INDFMINC  = fam_inc;
                      rename      INDFMPIR    =     PIR   ;

                      /*Education*/
                      rename DMDEDUC3  = child_ednlvl;
                      rename DMDEDUC2  = adult_ednlvl;
                      rename DMDSCHOL  = inschool;
                run;
      data work.demographics; set work.demographics;
      *Format for General Charistics variables;
            attrib veteran format=standardf.;
            attrib RIDRETH1 format=racef.;
            attrib DMDMARTL format=marriedf.;
            *attrib RIAGENDR;
            attrib RIDEXPRG format= pregnantf.;
            *household;
            attrib hs_sz format=sizef.;
```

```
            attrib fam_sz format=sizef.;

      *Used screening age in years (namely as it is more accurate);
                  rename RIDAGEYR=age;
                  attrib RIDAGEYR label="Screen Age in Years"
format=Comma5.1;
                  *RIDAGEMN does not go from 0 to 12, but 0 to 185;

      *Formats for Socioeconomic Status;
      *     attrib hs_inc format=incf.;
      *     attrib fam_inc format=incf.;
            *attrib PIR;

      *Formats for education;
            attrib child_ednlvl format=ednlvlf.;      *Children education;
            attrib adult_ednlvl     format=adultednf.;      *Adult education;
            attrib inschool format=inschoolf.;  *Attending school;
      RUN;
      proc contents data=demographics varnum; run;

*Tables before alteration;
      title "Orginal Data Before Management";
      proc means; var age PIR; run;

      title2 "# of Indvs under 17"; *needed as I will change veteran;
      proc means N Nmiss; where age<17; var veteran; run;
*Follows order above, except for age;
 data demographics;
 set demographics;
      /*Age*/
                  *Age Difference: Calculate difference for it to be
meaniful;
                  age_mean=48;      *Average age of adults--really 47.5;
                  agediff=age-age_mean;
                  attrib agediff label="Adjusted Screen Age in Years"
format=Comma5.1;
                  exam_age=RIDAGEEX/12;               *Exam Age (Took it in
years);
                  attrib exam_age label="Exam Age in Years";

                  *Adult? defined as age>=20;
                        if age>=20 then adult=1;            *yes;
                        else if 0<=age && age<20 then adult=2; *no;
                        else adult=.;
                        attrib adult label="Adult? (age>=20)"
format=standardf.;

                  *Literature suggests that Age>65 increases risk. So create
dummy variable;
                        if age>=65 then old=1;        *yes;
                        else if age NE . then old=0; *no;
                        else if age=. then old=.;
                        attrib old label="Older than 65?";
*format=standardf.;

            *Veteran: ALTERED VARIABLE;
```

```
                  *Ask to individuals who are 17 and over. It's against the
law to join army under 17
                              *Cf: http://www.military.com/join-armed-
forces/join-the-military-basic-eligibility.html;
                              *Obviously, people 17 or under are two young;
                  if 0<=age && age<17 then veteran=-1;
                        *or could say didn't fight(veteran=2);
                  *Run a proc means to make sure this is okay;

            *Race;
            *Combined Hispanics into one category;
                  *potential issue: race is different in study design;
                  if RIDRETH1=3 then race=1;    *white;
                  else if RIDRETH1=1 or RIDRETH1=2 then race=2; *hispanic;
                  else if RIDRETH1=4 then race=3;           *black;
                  else if RIDRETH1=5 then race=4;     *other/multiracial;
                  attrib race format=race2f.;   *label "Race/Ethnicity";

                  *Black has highest risk of developing hypertension.
                  As the odds of whites and others are similar, combine them
;
                  if race>0 then do;
                        if race=3 then black=1;
                        else black=0;
                  end;

            *Marital Status;
                  *Combine widowed,divorced, and separated as "once married"
                  *Combine married with "lived with partner" -I think this is
what "living with partner means";
                  *Treat not married as "single";
                  *ignore refused, don't know (treat as missing);
                  if DMDMARTL=1 OR DMDMARTL=6 THEN married=1;     *married;
                  else if DMDMARTL=2 oR DMDMARTL=3 OR DMDMARTL=4 THEN
married=2; *once married;
                  else if DMDMARTL=5 then married=3; *single;
                  else if DMDMARTL=77 OR DMDMARTL=99 then married=7;
*refused;

                  attrib married format=married2f.;   *label="Marital
Status";

            *Gender;
                  *: Adjust so that Male=0 and Female=1. Females known to
have greater risk for hypertension;
                  *RIAGENDER  1     =  Male .. 2 =    Female;

                        if RIAGENDR=1 then male=1;    *MALES;
                        else if RIAGENDR=2 then male=0;
                        attrib male format=malef.;

            /*Pregnancy*/;
                  pregn=RIDEXPRG;

                  /*RIDEXPRG Pregnancy status at the time of the health
examination was ascertained for females 8-59 years of age. Due to
```

```
                  disclosure risks pregnancy status will only be released for
women 20-44 years of age. The information used to code RIDEXPRG
                  values included urine pregnancy test results and self-
reported pregnancy status. Urine pregnancy tests were performed prior
                  to the dual energy x-ray absorptiometry (DXA) exam. Persons
who reported they were pregnant at the time of exam were assumed
                  to be pregnant.
                        *1    Yes, positive lab pregnancy test or self-
reported pregnant at exam     ;
                              *if the urine test was negative, but the
subject reported they were pregnant, the status was coded as 'pregnant at
exam'
                        *2    'not pregnant at examination' ; ;
                                                    *If the urine pregnancy
results were negative and the respondent stated that they were not pregnant,
                        *3    Cannot ascertain if SP is pregnant at exam;
                                          *if the urine pregnancy results
were negative and the respondent did not know her pregnancy status, the
                                          respondent was coded 'could not be
determined' (RIDEXPRG=3).
                                          *Also include persons who were
interviewed, but not examined
                  */
                  *Age Categories for Pregnancy: Used to preview where
RIDPREX is present;
                              if 0<=age<8 then pregnage=-1; *Children cannot
be pregnant & not asked;
                              else if 8<=age<20 then  pregnage=0;
      *Teens may be pregnant but due to exposure, unable to determine;
                              else if 20<=age<=44 then pregnage=1;
*Data for pregnancy given;
                              *Menopause usually occurs in late 40s and 50s;
                              else if 44<age<55 then pregnage=2;
      *Menopause Women may be pregnant, but due to exposure, cannot
determine;
                              *The last group is >=55, but the data stops
collecting at age 59,
                              so made two "subgroups". Those older cannot be
pregnant;
                              else if 55<=age<59 then pregnage=2.5;
                              else if 59<=age then pregnage=3;
                              else if age=. then pregnage=.;
                  *Adjustments: Gender and Age are used to modify pregnancy
variable.;
                        *if male, you are not pregnant;
                              if male=1 then pregn=2;
                        *Age Adjustments for Females: ;
                        else if male=0 && RIDEXPRG=. then do;
                              if 0<=age && age<8 then  pregn=2;*Children
cannot be pregnant & not asked;
                              else if 8<=age && age<20 then  pregn=3;
      *Teens may be pregnant but due to exposure, unable to determine;
                              *else if 20<=age && age<=44 then Data for
pregnancy given;
                              *Menopause usually occurs in during their late
40s or early 50s (i.e. 55);
```

```
                              else if 44<age && age<55 then pregn=3;    *Old
Women may be pregnant, but due to exposure, cannot determine;
                              else if 55<=age then  pregn=2; *Those older
cannot be pregnant;
                        end;
                  *Gender and Pregnancy together;
                        gpregn=.;

                  *Formats and Labels;
                        attrib Pregn  format= pregnantf.; *label="Pregnancy
Status at Exam";
                  attrib gpregn label="Pregnancy and Gender"
format=gendpregnf.;
                        *Option 2: Have two different variables.  ;

            /*Size of House: Use Household Size*/
                  * The NHANES defines a family as “a group of two people or
more (one of whom is the householder)
                  related by birth, marriage, or adoption and residing
together”;
                  *Households have unrelated individuals, while familes are
related. eg. a household of 5 with a family of 4, incidicating that 1 person
is unrelated to the family.   Generally, this is not the case;

                  *If size of the home is important, use HS_SZ.   (For
hypertension, it might be an indicator of stress, which is a risk factor for
high Blood Pressure.;

                  *Summarizing if the household size=family size;
                        if HS_SZ NE . && FAM_SZ NE . then do;
                              if HS_SZ=FAM_SZ  then family=1;     /**The size
of household is the familysize*/
                              else if HS_SZ<FAM_SZ then family=2; /*The
Household Size is smaller than family size--cnanot exist*/
                              else if HS_SZ>FAM_SZ then family=3;       /*The
Household Size is bigger than family size--i.e. 1 or more unrelated people
are residing in the same house*/
                        end;
                        else family=.;
                        *?????????Use: Importance of creating family?
            /*Age*/ *see above;

            /*Socioeconomic progress-PIR*/
                  *Add differences to make the intercept term more
meaningful;
                        PIR_lev=1;  *PIR=1 is definition of poverty;
                        pirdiff=pir-pir_lev;

            /*Education*/
                              /*Education*/
            *Age Categories;
                  if 0<age && age <6 then do;
                        agecat=1;
                        if adult_ednlvl=. then adult_ednlvl=-1;   /*too
young*/
                        if child_ednlvl=. then child_ednlvl=-1;   /*too
young*/
```

```
                        if inschool=.   then inschool=-1;         /*too
young*/
                        end;
                  *Children (agecat=2);
                  else if 6<age && age<20 then do;
                        agecat=2;
                        if child_ednlvl=. then child_ednlvl=111; *Children
with unknown education levels (not said);
                        if inschool=.    then  inschool=11;  *Unsure if
children are still in school or not;
                        *Missing education values for children are given a
number;

                        *Other age groups;
                        if adult_ednlvl=. then adult_ednlvl=113;  *Children
who cannot have an education level;
                        end;
                  *Adults (agecat=3);
                  else if age >= 20 then do;
                        agecat=3;
                        if adult_ednlvl=. then adult_ednlvl=111; *Adults with
unknown education levels (not said);

                        *Other age groups....;
                        if child_ednlvl=. then child_ednlvl=113; *Done with
education;
                        if inschool=. then inschool=13;     *Done with
education;
                        end;
                  attrib agecat format=agecatf.;


                  *An education variable: CHILDREN->In school;
                        *Are too young;
                  if inschool=-1 then edu=-1;   /*too young */
                        *Are not in school;
                  else if inschool=3                     then edu=0;
      /*Not Educated (Dropped out/not attend school)*/
                        *Are in school or on vacation;
                  else if inschool=1 or inschool=2  then edu=1;
      /* currently in school*/
                        *.............................. edu=2,3,4;
                        *Refused/don't know/children unsure;
                  else if inschool=7 or inschool=9 or inschool=11 then
edu=11;           /*unsure where to classify rightnow*/
                        *Too old to be in school;
                  else if inschool=13     then do; *edu=13;
      /*wrong age group, nothing left*/

                        *Adult Education level;
                        if adult_ednlvl=-1 then edu=-1;
      *Are too young;
                        else if adult_ednlvl=2  then edu=0;
      *Not Educated (Dropped out/not attend school);
                        *else if    XXXXXXXXXXX  then  edu=1;
      /*currently in school*/
```

```
                        else if adult_ednlvl=1 or adult_ednlvl=3 then edu=2;
*High School Degree or less (<9th Grade or High School Grad);
                        else if adult_ednlvl=4 then edu=3;  *some college;
                        else if adult_ednlvl=5 then edu=4;  *more than
college, highly educated;
                        else if adult_ednlvl=7 or adult_ednlvl=9 or
adult_ednlvl=111 then edu=11;   /*unsure right now  (Refused/don't know/
unknown education levels)*/
                        else if     adult_ednlvl=113  then edu=13;
      /*WRong Age Group, *none left*/
                  *;
                  end;
                  attrib edu format=eduf.;

                  *Indicator variable if educated?;
                  *Too young if under 5, and is not educated if dropped
out/did not attend school. Any education (High School, College, in school,
etc. is considered "educated");
                        if edu=-1 then edu2=-1;
                        else if edu=0  then edu2=0; /*not educated*/
                         else if 0< edu && edu < 11 then edu2=1; /*educated*/
                         else if edu=11 then edu2=.; *ignore the missing;
                         else edu2=.;

                  attrib edu2 label="EDUCATED?" format=edu2f.;
      run;

      *Dropping...Old Variables;
            *drop DMDMARTL (combined into married) RIDRETH1 (combined into
race)
            RIDAGEEX RIDAGEMN RIDAGEYR (all three combined into age)
            RIDEXPRG (combined into pregnancy);

      proc contents data=demographics varnum; run;
*Overall: Both the new variables along with their data;
      proc means data=demographics N NMiss  mean std min max;
            title "Demographics";
            var veteran
                  race black RIDRETH1
                  married DMDMARTL
                  male RIAGENDR
                  pregn RIDEXPRG

                  age adult old exam_age  RIDAGEEX RIDAGEMN
                  PIR PIR_lev pirdiff  hs_inc fam_inc
                  child_ednlvl adult_ednlvl inschool edu edu2
                  ;
      run;
      proc freq data=demographics;
      *General Characterisitics (Cat);
                  tables veteran/list missing; title2 "Veterans";
                  tables RIDRETH1 race black/list missing; title2 "Race";
      *Covariate selected;
                  tables DMDMARTL married/list missing; title2 "Martial
Status";
                  tables male*RIAGENDR/missing; title2 "Male";  *Covariate
selected;
```

```
                  tables RIDEXPRG pregn pregnage gpregn /list missing; title2
"Pregnancy";
                  tables hs_sz fam_sz/list missing; title2 "Household Size";
      *Age;
                  *tables age exam_age;
                  tables adult old/list missing; title2 "Age";  *Covariate
selected;
      *Socioeconomic Status;
                  *tables pir;tables hs_inc fam_inc/list missing; title2
"Socioeconomic Status"; *Covariate selected;
      *Education;
                  tables child_ednlvl adult_ednlvl inschool edu edu2; title2
"Education";
            run;
            ods text="okay if the race categories not equal?";
            ods text="the other race category is too small--combine based on
odd ratios";
      /*Covariates ready for analysis (# of levels if categorical for
analysis after modification if necc (denoted by *)
                                    , cont for continuous) :
                        veteran*(2) black(2) married*(3) male(2)--OR--
pregn*(2),  hs_sz(cont),  age(treat as cont) OR old(2),
                        pir(cont), edu*(5) OR edu2*(2);
            *: Data modification needed: Remove refused/don't know or
uncertain*/
proc freq;
/*covariates used in analysis*/ tables                black
male                                adult /list missing;
/*extra covariates not used*/ tables  VETERAN           married         pregn
pregnage   HS_SZ  old    edu edu2/list missing;
run;

*Subset Looking At Hypertension Analysis;
proc means data=demographics N NMiss  mean std max;
            where age>=20;
            var race black RIDRETH1
                  married DMDMARTL
                  male RIAGENDR
                  pregn RIDEXPRG
                  age adult old exam_age  RIDAGEEX
                  PIR PIR_lev pirdiff ;
            title "Demographics for Hypertension and Arsenic Dataset";
      run;

      proc freq data=demographics;
            where age>=20;
            tables adult old race black /*RIDRETH1*/        married
/*DMDMARTL*/      male /*RIAGENDR*/ pregn /list;
            run;

*Specifically;
*Looking at Age;
            proc means data=work.demographics n nmiss mean std min q1 median
q3 max maxdec=2;
            var age;
            where adult=1;
            run;
```

```
            *Percentiles for Age;
            proc univariate data=work.demographics noprint;
            var age;
            output out=data1 pctlpre=P_ pctlpts=42,43,44,  25 to 50 by 5;
            run;
            proc print data=data1;run;

*Pregnancy;
      title "Pregnancy";
      *Pregnancy Subset to play;
            data demo2;
            set demographics;
                  keep SEQN male age pregnage  RIDEXPRG pregn;
            run;
      proc freq data=demographics;
      tables pregnage/list;
      tables male*pregnage*(pregn RIDEXPRG)/missing norow nocol nopercent;
      tables pregn RIDEXPRG gpregn/list;
      run;

/*Socioeconomic Progress*/
      proc means data=demographics N NMISS MEAN STD MIN MAX;
*     where SDDSRVYR>4;
      var PIR IND235 INDFMMPI INDFMMPC;
      run;
      proc freq data=demographics; tables hs_inc fam_inc/list missing; run;
      proc freq data=demographics;
      where PIR=. and age>=20;
      tables hs_sz hs_inc fam_inc/list missing;
      tables hs_sz*hs_inc;
      title "When PIR is missing, use hs_inc?";
      run;
      *Depends on # of people in household:
http://aspe.hhs.gov/poverty/14poverty.cfm;

      *Condense hte HS_Inc to a smaller bit after formatting it;

*Education;
      *Correlation between education and pir--used in some variables;
      *See edu;

      *Data summary;
                  *My variables created with orginal data variables
underneath;
            data work.demographics;
                  retain SEQN SDDSRVYR;
                  /*general cat characteristics*/ retain veteran race black
married   male    pregn pregnage gpregn

            RIDRETH1   DMDMARTL  RIAGENDR     RIDEXPRG      ;
                  /*household*/ retain hs_sz family fam_sz;
                  /*age*/     retain age exam_age adult old age_mean agediff
                                          RIDAGEEX  RIDAGEMN
RIDAGEYR;
                  /*socioeconomic progress*/ retain hs_inc fam_inc pir
PIR_lev pirdiff;
                  /*education*/ retain edu2 edu agecat
```

```
                                                  child_ednlvl adult_ednlvl
inschool ;
                  set work.demographics;
                  *drop DMDMARTL (combined into married) RIDRETH1 (combined
into race)
                        RIDAGEEX RIDAGEMN RIDAGEYR (all three combined into
age)
                  *Checking code variables no longer needed;
      run;

      proc contents data=demographics varnum; run;
      /*Covariates ready for analysis (# of levels if categorical for
analysis after modification if necc (denoted by *)
                                    , cont for continuous) :
                        veteran*(2) black(2) married*(3) male(2)--OR--
pregn*(2),  hs_sz(cont),  age(treat as cont) OR old(2),
                        pir(cont), edu*(5) OR edu2*(2);
            *: Data modification needed: Remove refused/don't know or
uncertain*/
            proc freq; tables veteran black married male pregn hs_sz old edu2
edu/list; run;

            *I ALREADY CODED ALL VARIABLES: In my work.demographics analysis,
I also made major modifications to age, education, household/family members
in demographics set
            (noticed that Age is missing for some education analysis);
            *Also edited immigration as well;

*********************************************************************
**********************************;
```

BMI

```
*Categories for BMI--clinical significance;
            %InputnHanes (filename=BMX_D.xpt);
            %InputnHanes (filename=BMX_E.xpt);
            %InputnHanes (filename=BMX_F.xpt);
            data body2;
            set BMX_D BMX_E BMX_F;
                  rename      BMXBMI       =    bmi    ;
                  attrib BMIWT format=weightcommentf.;
                  attrib BMIHT format=heightcommentf.;
                  one=1;
            run;
            data body;
                  merge work.demographics(keep=SEQN SDDSRVYR AGE ADULT MALE
PREGN) body2; * demog.percentiles(keep=SEQN BMIPCT);
                  by SEQN;
                  if one=1; drop one;           *Merging variable so that
body has nonmissing variables from body2;
                  *if BMDSTATS=1 or BMDSTATS=2;       /*component code--1 or
2 indicates height, weight data accessed*/
            run;

      proc contents data=body varnum; run;
            data body;
            retain SEQN SDDSRVYR ADULT AGE BMIAGE MALE PREGN BMI BMI2 BMXHT
BMXRECUM BMXWT;
```

```
            set body;
            /*Age Categories*/
                  if 0<=age && age<2 then bmiage=-3;        *Infants;
                  else if 2<= age && age<5 then bmiage=-2;  *Preschool;
                  else if 5<=age && age<18 then bmiage=-1;  *children;
                  else if 18<=age && age<20 then bmiage=0;   *Young Adults;
                  else if 20<=age                then bmiage=1;
      *Eligible;

            /*Calculate BMI Manually: Any different than BMI*/
            BMI2=BMXWT/(BMXHT**2);

            *Clinically categorizing the BMI: Adjusting variables;
            **Literature suggests that BMI>=25 increases risk of
hypertension--So create dummy variable;
                  if adult=1 then do;
                        if BMI>=25 then overweight=1;
      /*overweight/obese*/
                        else if 0<BMI && BMI<25 then overweight=0;
                        else overweight=.;
                  end;
            *(2)bmicat is uneven--combine. Huh? 3% underweight seems not
representative...;
                  if bmicat=1 or bmicat=2 then bmicat=1.5;
      /*normal/underweight*/
      run;

      *Note: BMIPCT, when calculated, has already been adjusted for male and
age.;
            /*    For Children and Teens:
                        The percentile indicates the relative position of the
child's BMI number among children of the same sex and age

                  Weight Status Category  Percentile Range
            Underweight       Less than the 5th percentile
            Healthy weight    5th percentile to less than the 85th percentile
            Overweight  85th to less than the 95th percentile
            Obese       Equal to or greater than the 95th percentile

                  *Source:
http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/#Definition
                  *source:
http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/;
      */
*Data Analysis;
      proc means data=body n nmiss mean std min max maxdec=2;
            class bmiage/missing;
            var age bmxrecum bmxht bmxwt bmi bmi2;
            run;
      proc freq;
            *BMI Status;
            tables BMDSTATS/list;
            *Age Categories for BMI;
            tables bmiage/list;
            *Comments for Height and Weight (IMPORTANT TO INCLUDE);
            tables BMIRECUM BMIHT BMIWT BMIHT*BMIWT   /list;
            *BMI itself;
```

```
            tables BMIcat overweight;
      run;
      *Demographics important?;
            data  body2;
            set body;
            *Ignore cannot ascertain;
            if pregn=3 then pregn=.;
            run;
            proc genmod;
            class male pregn (ref="no (at time of exam)" param=ref);
            model bmi= MALE AGE PREGN;
            run;

*Why is BMI missing?;
            title "Why is BMI missing?";
            proc means data=body N Nmiss  ;
            class bmiage/missing;
            where BMI=.;
            var age bmxht BMXRECUM bmxwt bmi bmi2;
            run;

*BMI & Age;
      proc means data=body n nmiss mean std min q1 median q3 max;
            class bmiage;
            var bmi;
      run;
      proc freq;
            table bmiage/list missing;
            table bmiage*bmicat/missing norow nocol nopercent;
            *table adult*bmiage*bmicat/missing norow nocol nopercent;
      run;
      *For adult=age>=18, 1166 are between 18 and 20, so they can be dropped.
?? (6.95%);
      *Summary;
      data body;
                  retain SEQN SDDSRVYR
                              BMDSTATS    /*overall component code*/
                              bmiage MALE AGE ADULT PREGN         /*age
descriptions*/
                              overweight BMI    BMI2 BMICAT BMIPCT
      /*bmi*/

                              BMXWT BMIWT             /*weight*/
                              BMXRECUM BMIRECUM BMXHT BMIHT  Height
Height_comm       /*height (recumbent,adjusted (<5 years of old), standing
(>=5)) */
                              ;
      set body;
      drop              /*Length, circumference of body parts.
                              *eg. waist arm circumference, tricep skin
circumference, leg length, arm length, thigh circumference, etc.
                              **(not used here...focused on height/weight
dataset) */
                              BMXHEAD     BMIHEAD
      /*head circumference*/
                              BMXLEG      BMILEG      BMXWAIST BMIWAIST
      /*leg, waist circumference*/
```

```
                              BMXARML     BMIARML     BMXARMC     BMIARMC
                  /*arm length, arm circumference*/
                              BMXTRI      BMITRI      BMXSUB      BMISUB
                  /*tricep skin circumference, subcutaneous fat*/
                              BMXCALF     BMICALF     BMXTHICR BMITHICR
      /*calf circumference, thigh circumference*/
                        ;
      run;

      *Check contents;
      proc sort data=body; by SEQN;run;
      proc contents data=body varnum; run;

*************************************************************************
**********************************;
```

Smoking

```
*Smoking;   *very important confounder: Interested in: do people smoke?
      ;
                  *A smoker is defined as individual who smoked at least 100
cigarettes in life
                  or used tobacco and nicotine in the last 5 days. If only
one question contains a refused,
                  the other question is used to determine smoking status. If
both questions are don't know,
                  refuse, or missing, smoking status is missing.;

                        %InputnHanes (filename=SMQ_D.xpt);
                        %InputnHanes (filename=SMQFAM_D.xpt);  *Secondhand
smoke;
                        %InputnHanes (filename=SMQRTU_D.xpt);  *recent tabaco
use--used to validate dataset (A);
                        data smoking_d; merge smq_d smqfam_d smqrtu_d; by
SEQN; run;

                        %InputnHanes (filename=SMQ_E.xpt);
                        %InputnHanes (filename=SMQFAM_E.xpt);  *Secondhand
smoke;
                        %InputnHanes (filename=SMQRTU_E.xpt);  *recent
tobacco use--used to validate dataset (A);
                        data smoking_e; merge smq_e smqfam_e smqrtu_e; by
SEQN; run;

                        %InputnHanes (filename=SMQ_F.xpt);
                        %InputnHanes (filename=SMQFAM_F.xpt);  *Secondhand
smoke;
                        %InputnHanes (filename=SMQRTU_F.xpt);  *recent
tobacco use--used to validate dataset (A);
                        data smoking_F; merge smq_f  smqfam_f smqrtu_f; by
SEQN; run;

            data smoking2;
                  set smoking_d smoking_e smoking_f;
                  /*Important variables to indicate whether people smoke*/
                  rename SMQ020=smoke100;   attrib SMQ020 format=standardf.;
*SMQ020: Smoked at least 100 cigarettes in life;
```

```
                  rename SMD410=smokehome;   attrib SMD410 format=standardf.;
*SMD410 - Does anyone smoke in home?  ***SECOND-HAND SMOKE;
                  rename SMQ680=smokefive;    attrib SMQ680
format=standardf.;  *SMQ680: Used tobacco/nicotine last 5 days?:  Check for
validation to SMQ020 --- ;
                  run;
      *NOTE:      *All other questions are 0 for nonsmokers, to those who
refuse, and to those who don't know.;

                  data smoking;
                  merge demographics(keep=SEQN SDDSRVYR age adult) smoking2;
                  by SEQN;
                  *Create an age category for smoking--useful in analysis--
from NHANES documentation;
                        attrib sagecat label="Smoking Age Category";
                        if 0<=age && age<12 then sagecat=-1; /*too
young/ineligible*?*/
                        else if 12<=age && age<20 then sagecat=0;  /*data for
SMOKEFive only*/
                        else if age>=20 then sagecat=1;     /*eligible*/
      *data for SMOKE 100 and SMOKEFive;
                  run;
                  data smoking;
                  retain SEQN SDDSRVYR age adult sagecat smokefive smoke100
smokehome;
                  set smoking;
                  run;

                  *Note: For some, smokefive=. and smoke100=-1;
                  proc means data=smoking N Nmiss;
                        var SEQN sagecat age smoke100 smokefive smokehome;
                  run;

                  proc freq data=smoking;
                  where age=.;
                  tables SDDSRVYR*sagecat/missing;
                  tables smokefive*smoke100 smokefive*smoke100/norow nocol
nopercent missing;
                  title "Smoking with Missing Age";
                  run;
                  proc freq data=smoking;
                  where sagecat=0;
                  table smokefive;
                  title "Age between 12 & 20";
                  run;

                  proc freq data=smoking;
                  tables sagecat*smokefive*smoke100/norow nocol nopercent
missing;
                  table sagecat/list missing;
                  table sagecat*adult/ missing norow nopercent;
                  table sagecat*(smokefive smoke100)/norow nocol missing;
                  title "Age Categories and Two Variables used";
                  run;

            *Interested in: Do people smoke--yes/no?, which depends on age;
            data work.is_smoking;
```

```
                  set smoking (keep=SEQN SDDSRVYR age sagecat SMOKE100
SMOKEFIVE SMOKEHOME);
                  *Interested in: Do people smoke--yes/no?: Depends on age;
                  if sagecat=-1 then do;
                        smokeFive=-1; smoke100=-1; is_smoke2=-1;
      *SageCat=-1;
                    end;
                  *SmokeFive only exists between ages 12 & 19--so use data
extensively. If smokefive=. what should I do with them?;
                  else if sagecat=0 then do;
                        SMOKE100=-1;  /*ineligible*/
                        if smokefive=1 then is_smoke2=1;
                        else if smokefive=2 then is_smoke2=2;
                        else if smokefive=7 OR smokefive=9 OR smokefive=.
then is_smoke2=.;
                        end;

                  *Combined missing ages with sagecat=1--since data is
present, it must indicate that these people are the appriopriate age;
                  else if sagecat=1 OR sagecat=. then do; /*eligible*/

                        if SMOKE100=1 OR SMOKEFIVE=1 then is_smoke2=1;
                        else if SMOKE100=2 AND SMOKEFIVE=2 then is_smoke2=2;
                        else if SMOKE100=. OR SMOKE100=7 OR SMOKE100=9 then
do;
                              if SMOKEFIVE=1 then is_smoke2=1;
                              else if SMOKEFIVE=2 then is_smoke2=2;
                              else is_smoke2=.; *else if SMOKEFIVE=7 OR
SMOKEFIVE=9 OR SMOKEFIVE=.;
                        end;
                        else if SMOKEFIVE=. OR SMOKEFIVE=7 OR SMOKE100=9 then
do;
                              if SMOKE100=1 then is_smoke2=1;
                              else if SMOKE100=2 then is_smoke2=2;
                              else is_smoke2=.;
                        end;
                  end;
            attrib is_smoke2 format=standardf. label="Do you smoke? yes/no";
            rename is_smoke2=is_smoke;

            *Second-Hand smoke--anyone in your home?;
                        *if someone in your home is a smoker, you are a
second-hand smoker;
                        if SMOKEHOME=1 then SecHnd=1;
                        *if there is no one in your home who is not a smoker,
then not a second hand smoker;
                        else if SMOKEHOME=2 then SecHnd=0;
                        *if you don't know or refuse to tell, you may be
somewhat a second-hand smoker;
                        else if SMOKEHOME=7 OR SMOKEHOME=9 then SecHnd=1.5;
                        else if SMOKEHOME=. then SecHnd=.;
                        attrib sechnd format=sechndf. label="Are you exposed
to second-hand smoke?";
 run;

            title  "The relationship of smokers and nonsmokers";
            proc freq;
```

```
            tables SMOKEFIVE*SMOKE100 is_smoke/missing norow nocol nopercent;
             run;

            proc freq;
            where sagecat=1;
            tables SMOKEFIVE*SMOKE100 is_smoke/missing norow nocol nopercent;
             run;

            proc freq data= is_smoking;
                  title2 "For Analysis-Hypertension";
                  where age>=20;
                  tables Smoke100 * SmokeFive/missing norow nopercent nocol;
                  tables is_smoke/list;
            run;
            ods text="NOTE: SmokeFive is only recorded between ages 12 & 19--
so data is used extensively to detect smoking";
            ods text="comments: ignoring don't know/refused.";

            proc freq data=work.is_smoking;
                  title2 "For Analysis-Diabetes";
                  where age>=8;
                  tables is_smoke/list;
            run;

            *Second Hand Smoking;
            proc freq data=work.is_smoking;
                  title2 "Second-hand smoke";
                  tables SMOKEHOME  sechnd/list;
            run;


      *Summary of is_smoking;
      data work.is_smoking;
            retain SEQN SDDSRVYR
                        sagecat     age
                        /*summary*/ is_smoke
                        /*orginal data*/smoke100 smokefive
                        /*second-hand smoke*/ sechnd smokehome;
            set work.is_smoking;
            run;
      proc contents data=work.is_smoking varnum; run;
      proc means N NMISS Mean; where sagecat=1; run;

      /*Summary of Variables and Orginal Data*/
      proc freq;
            tables sagecat
                        /*summary*/ is_smoke
                        /*orginal data*/smoke100 smokefive smoke100*smokefive
                        /*second-hand smoke*/ sechnd smokehome/list missing;
            run;
      *Datasets making secondhand smoke and smokers is included in
demographics;

*************************************************************************
**********************************;
```

Alcohol Use

```
*This code creates the alcohol dataset in the demog library. THe main
level of alcoholism is denoted by alchlusg --a variable derived from ALQ120Q
and ALQ120U

      *Almost identical as for smoker;
      *Participants aged 12 years and older were eligible.
      Only data from participants aged 20 years and older are included in
this data file.;

      *Logic of creating alcoholic:
      last year=yes(1)   -> alcoholic (1) (by design)

      last year=no
        & lifetime=yes(1)  -->somewhat alcholic (2)
        & lifetime=no (2)  -->totally sober  (4)
        &  lifetime=don't know (9) -->borderline alcholic (3)
        & lifetime=refused/missing (7) --> missing (.)

      lifetime=don't know (9)
        & last year=yes -->borderline alcoholic (3)
        &l last year=no -->properly sober (4)

      lifetime=refused OR last year=refused (7)  =same as missing.;
      *lifetime=dont know and last year=don't know then borderline alcoholic
(3)
      *If not yes to either ALQ101 or ALQ110  then no more questions asked;
*/;
libname data "C:\VCU Biostatistics\research\Arsenic Project\Data\Arsenic
Mixture Project Data (Final_Fall 2013)";              /*The output dataset*/
title ""; title2 "";
      *OPTIONS FMTSEARCH=(data. nHANES);  /*Find format*/

*Formats;
      proc format; *library=data.nhanes;
      *alcholic;
                  value alcoholicf  1="alcoholic"
                                    2="somewhat alcoholic"
                                    3="borderline alcoholic"
                                    4="non-alcoholic"
                                  ;
            *////////;
                  value isalcholicf  -1="too young (age<12)"
                                        -0.5="classified (12<=age<20)"
                                        1="alcoholic"   /*Defined
(08/09/13: Had at least 12 alcoholic drinks last year OR in lifetime*/
                                     0="_non-alcoholic_"
                                         ;
             *Alcoholic age category;
            value aagecatf  -1="inegilble (age<12)"
                                  0="classified (12<=age<20)"
                                  1="eligible (age>20)"
                                  2="Missing age yet values"
                                  ;
            *refused/don't know category;
            value responsef         1="Have response"
```

```
                                        7="Refused"
                                        9="Don't know"
                                        ;
run;
      *Days Alchl Lifetime;
*refused1000f;

      /*ALD020 - # days drank alcohol over lifetime
                        During your life, on how many days have you had at
least one drink of alcohol?

                        Code or Value     Value Description
                        1     0 days
                        2     1 or 2 days
                        3     3 to 9 days
                        4     10 to 19 days
                        5     20 to 39 days
                        6     40 to 99 days
                        7     100 or more days
                        77    Refused
                        99    Don't know
                        .     Missing

            ALD030 - # days drank alcohol over past 30 days
            During the past 30 days, on how many days did you have at least
one drink of alcohol?
            Code or Value     Value Description
            1     0 days
            2     1 or 2 days
            3     3 to 5 days
            4     6 to 9 days
            5     10 to 19 days
            6     20 to 29 days
            7     All 30 days
            77    Refused
            99    Don't know
            .     Missing

            ALD040 - # days w/5 or more drinks/past 30 days
            During the past 30 days, on how many days did you have 5 or more
drinks of alcohol in a row, that is, within a couple of hours?
            Code or Value     Value Description
            1     0 days
            2     1 day
            3     2 days
            4     3 to 5 days
            5     6 to 9 days
            6     10 to 19 days
            7     20 or more days
            77    Refused
            99    Don't know
            .     Missing
            */

      /*Heavy episodic drinking is defined as: consumption of six or more
drinks in one
```

```
      sitting at least once a week for Lebanon; consumption of five or more
drinks in
      one sitting at least once a week for Malaysia and Czech Republic.
(WHO).http://www.who.int/violence_injury_prevention/violence/world_report/fac
tsheets/fs_youth.pdf;
      According to this: */
            *if 2<=ALD040 && ALD040<=7 then is_alchly=1;
*1: Data INPUT;
      %InputnHanes (filename=ALQ_D.xpt);
      %InputnHanes (filename=ALQ_E.xpt);
      %InputnHanes (filename=ALQ_F.xpt);
      %InputnHanes (filename=ALQY_F.xpt);       *Include Alcohol youth?;
      run;

      data alq;
      set ALQ_D ALQ_E ALQ_F;
      rename ALQ101=lastyr;   attrib ALQ101 label="Had at least 12 alcohol
drinks/1 yr?" format=standardf.;
      rename ALQ110=lifetime; attrib ALQ110 label="Had at least 12 alcohol
drinks/lifetime?" format=standardf.;
      run;

proc sort data=demographics; by SEQN; run;
data alcohol;
            merge alq demographics(keep=SEQN age);
            by SEQN;
      *Levels of alcoholism--see logic below--based on age;
            if 0<age && age<12 then aagecat=-1;  *too young/ineligible--can
we assume that the answer is "NO" FOR THESE?";
            else if 12<=age && age<20 then aagecat=0;  *classified;
            else if  20<=age then aagecat=1;
            else aagecat=.; *missing ages;
            attrib aagecat format= aagecatf.;

      /*ALQ120Q: How often drink alcohol over past 12 mos-->#  of REFUSED*/
            /*    ALQ120Q: 0-365 days; 777=refused; 999=don't know;*/
            if ALQ120Q=777 or ALQ120U=777 then alchlusgr=7;
            else if ALQ120Q=999 or ALQ120U=999 then alchlusgr=9;
            else if ALQ120Q=.  AND ALQ120U=. then alchlusgr=.;
            else if ALQ120Q>=0  AND ALQ120U=. then alchlusgr=11;  *Those who
responded, but units unknown";
            else if ALQ120Q=.  AND ALQ120U>=0 then alchlusgr=12;  *Those who
responded, but no number";
            else if ALQ120Q>=0 AND ALQ120U>=0 then alchlusgr=1;
            attrib alchlusgr label="Refused/Don't know for # of average
drinks??"  format=responsef.;

      *ALQ130 - Avg # alcoholic drinks/day -past 12 mos;
            rename ALQ130=avgdrinks;
            attrib ALQ130  format=refused100f.;       /*...r stands for
refused*/

            *Seperate the refused/don't know;
            if ALQ130=777 then avgdrinksr=7;
            else if ALQ130=999 then avgdrinksr=9;
            else if ALQ130=. then avgdrinksr=.;
            else avgdrinksr=1;
```

```
            attrib avgdrinksr label="Refused/Don't know for Average #
Alcoholic Drinks per day" format=responsef.;

      *ALQ140: #days have 5 or more drinks/past 12 mos per day, week, or
year;
            /*    ALQ140Q: 0-365 days; 777=refused; 999=don't know;*/
            if ALQ140Q=777 then alchl5r=7;
            else if ALQ140Q=999 then alchl5r=9;
            else if ALQ140Q=.  then alchl5r=.;
            else alchl5r=1;
            attrib alchl5r label="Refused/don't know for 5 or more drinks..."
format=responsef.;

      *ALQ150 - Ever have 5 or more drinks every day?;
            attrib ALQ150 format=standardf.;
      run;

*B: Basic Analysis;
      proc contents varnum; run;
      title "AFTER";
      proc means data=alcohol n nmiss mean std min max;
      run;
      proc means data=alcohol n nmiss mean std min max;
      where lifetime>1;       *Those no/refused/don't know are excluded.
Matches what documentation says;
      run;
      proc means data=alcohol n nmiss mean std min max;
      where lastyr=1;               *Those who said yeas to last year are not
asked about lifetime;
      run;
      proc freq data=alcohol;
        tables lastyr lifetime alchlusgr ALQ120U avgdrinksr alchl5r ALQ140U
ALQ150/list missing;
        tables alchlusgr*ALQ120U     alchl5r*ALQ140U/missing norow nocol
nopercent;
        tables aagecat*(lastyr lifetime alchlusgr ALQ120U avgdrinksr alchl5r
ALQ140U ALQ150)/missing norow nocol nopercent;
        table aagecat*alchlusgr*ALQ120U/missing norow nocol nopercent;;
       run;


/******C: DATA MODIFICATION--Add data to those that are unstructured*****/;
      data alcohol;
      set alcohol;
      *for the categorical variables, add -1 or 0 for those too young;

      *Those who answer yes to last year are not asked about lifetime. It's
assumed to be a yes;
      *if LASTYR=1 then LIFETIME=1;

      *those who answer no, refused, or don't know to lifetime are not asked
any other questions.
      Assume that the answers are 0 days;
      if lifetime=2 OR lifetime=7 OR lifetime=9 then do;
            ALQ120Q=0;        *0 drinks over past 12 months;
            ALQ120U=1;        *per day;
```

```
            avgdrinks=0;            *0 avg alchl drinks per day for past 12
moths;
            ALQ140Q=0;        *0 days having 5 or more alcoholic drinks;
            ALQ140U=1;
            ALQ150=2;         *NO;
      end;

/*D Alcoholic Consumption Measurements--A better measurement;
      /*ALQ120Q: How often drink alcohol over past 12 mos per week*/
      if ALQ120U=1 then alchlusg=ALQ120Q*1;     /*weeks*/
      else if ALQ120U=2 then alchlusg=ALQ120Q*12/52;        /*months*/
      else if ALQ120U=3 then alchlusg=ALQ120Q*1/52;     /*years*/
      else if ALQ120U=. then do;
            if ALQ120Q=0 then alchlusg=0;
            else if 0<=ALQ120Q && ALQ120Q<777  then alchlusg=.;
      /*Ignore those who have #'s but no units--don't know how to interpret

      ALQ120Q*365.25/52;  ??: days */
            else if ALQ120Q=777 or ALQ120Q=999 then alchlusg=.;   *Ignore
those who refuse/don't know;
            else if ALQ120Q=. then alchlusg=.;
            end;
      else if ALQ140U=7 OR ALQ140U=9 then alchlusg=.;   /*Ignore Units that
are refused or don't know*/
      attrib alchlusg label="# of drinks per week in the past year";

*ALQ140: #days have 5 or more drinks/past 12 mos per day, week, or year;
      if ALQ140U=. then  do;
                  if ALQ140Q=0 then alchl5=0;
                  *else if ALQ140Q>0 then alchl5=ALQ140Q* 365.25/52   ;
      /*days*/
                  else ALQ140Q=.;
                  end;
      else if ALQ140U=1 then alchl5=ALQ140Q*1 ;       /*weeks*/
      else if ALQ140U=2 then alchl5=ALQ140Q*12/52;    /*months*/
      else if ALQ140U=3 then alchl5=ALQ140Q*1/52  ;         /*years*/
      else if ALQ140U=7 OR ALQ140U=9 then alchl5=.;
      attrib alchl5 label="# of times have 5 or more drinks in a single day
per week";

      *There are some who answered 0 to ALQ120Q and ALq140Q;
run;
      proc freq;
      tables alchlusgr/list;
      run;
proc freq;
      where aagecat=1;
      tables alchlusgr/ norow nocol nopercent missing;
      run;
proc means N Nmiss;
      class alchlusgr/missing;
      var ALQ120Q ALQ120U;
      run;

*////////////////////////////////////////////////////////////////////////////
//////////////////////////;
*F) Summary;
```

```
      proc sort data=alcohol; by SEQN; run;
      data alcohol;
            retain SEQN
                        /*demographics*/ age    aagecat
                        /*intermediate/summary variables*/ alchlusg
                        lastyr      lifetime
                        ALQ120Q     alchlusgr ALQ120U
                        avgdrinks   avgdrinksr
                        alchl5      alchl5r                ALQ140Q
      ALQ140U
                        ALQ150
                        alcoholic is_alchl            /*Are you alcholic?
(old)*/;
      set alcohol;
      drop chk;
      run;
      proc contents data=alcohol varnum; run;

*G) Save dataset;
      *data data.alcohol;
      *set alcohol;
      *run;
 *-------------------------------------------------------------------------------
-----------------------------------
```

```
*******************************************************************************
*********************************;
```

Outcome variable: hypertension;

```
*Outcome variable: hypertension;
*From Blood Pressure, Pulse, and Hypertension;

*Read in Data: Blood Pressure and Pulse;
      %InputnHanes (filename=BPX_D.xpt);
       %InputnHanes (filename=BPX_E.xpt);
       %InputnHanes (filename=BPX_F.xpt);
*Read in Data--Questioniare Chlosterol&BloodPressure;
      %InputnHanes (filename=BPQ_D.xpt);
      %InputnHanes (filename=BPQ_E.xpt);
      %InputnHanes (filename=BPQ_F.xpt);

data blood3;
set BPQ_D BPQ_E BPQ_F;

/*Hypertension Questions*/
rename      BPQ020       =    ToldHyper    ; attrib BPQ020 format=standardf.;
      ***;
rename      BPQ030       =    Toldhyper2plus     ; attrib BPQ030
format=standardf.;
rename      BPD035       =    Toldhyper_age      ;
rename      BPQ040A      =    Prescp       ; attrib BPQ040A format=standardf.;
      ***;
```

```
rename      BPQ050A      =    Prescp2     ; attrib BPQ050A format=standardf.;
      ***;
rename      BPQ052       =    prehyper    ; attrib BPQ052 format=standardf.;
rename      BPQ057       =    prehyper2   ; attrib BPQ057 format=standardf.;
rename      BPQ056       =    bp_home     ; attrib BPQ056 format=standardf.;
rename      BPD058       =    bp_home2    ;
rename      BPQ059       =    Toldbphome; attrib BPQ059 format=standardf.;

/*Cholestrol Questions*/
*rename     BPQ055 - CHECK ITEM      =          ;
BPQ055=.;   *AGE>=20;
rename      BPQ060       =    ckchol      ;
rename      BPQ070       =    ckchol_time ;
rename      BPQ080       =    toldchol    ;
rename      BPQ090A      =    toldfat     ;
rename      BPQ090B      =    toldweight  ;
rename      BPQ090C      =    toldexc     ;
rename      BPQ090D      =    toldpres    ;
BPQ095=.;   *Yes to BPQ090A OR BPQ090B OR BPQ090C OR BPQ090D--otherwise, end
section;
rename      BPQ100A      =    dofat ;
rename      BPQ100B      =    doweight    ;
rename      BPQ100C      =    doexc ;
rename      BPQ100D      =    dopres      ;
run;

data blood2;
      set BPX_D BPX_E BPX_F;
      *Blood Pulse;
      *BPXDB * - # of dropped beats in 30 seconds     (Session D only) *12-
150;
      *BPXCHR *- 60 sec HR (30 sec HR * 2);           *0-7 years;
      *BPXPLS *- 60 sec. pulse (30 sec. pulse * 2): *8-150 years;
      *BPXCHR     BPXPLS BPXDB;

      *Blood Pressure Quality Control;
            rename      PEASCST1     =    bpstatus    ; attrib    PEASCST1
      label="     Blood Pressure Data Completeness    " format=   missingf.
;                 *If the data is missing or not;
            rename      PEASCTM1     =    bptime      ; attrib    PEASCTM1
      label="     Blood Pressure Time in Seconds      " format=   Comma6.
;
            rename      PEASCCT1     =    bpcomm      ; attrib    PEASCCT1
      label="     Blood Pressure Comment  " format=   bpcommentf.  ;

      *Blood Pressure Enhancements;
            attrib BPAEN1 format=enhancef.;
            attrib BPAEN2 format=enhancef.;
            attrib BPAEN3 format=enhancef.;
            attrib BPAEN4 format=enhancef.;

      *Blood Pressure Covaraiates;
            rename      BPAARM       =    arm   ; attrib          BPAARM
      label="     Arm selected:     " format=   armslf.      ;
            rename      BPACSZ       =    cuff_sz     ; attrib    BPACSZ
      label="     Coded cuff size   " format=   cuffsizef.   ;
```

```
            rename      BPXML1       =    inflation   ; attrib    BPXML1
label="     MIL: maximum inflation levels (mm Hg)     " format=
      refused800f.       ;

      *Both Blood Pulse and Pressure Covariates;
            rename      BPQ150A      =    bpfood30    ; attrib    BPQ150A
      label="     Had food in the past 30 minutes?    " format=   standardf.
;                       *BPComment may indicate where the data is missing;
            rename      BPQ150B      =    bpalchl30   ; attrib    BPQ150B
      label="     Had alcohol in the past 30 minutes? " format=   standardf.
;
            rename      BPQ150C      =    bpcoffee30  ; attrib    BPQ150C
      label="     Had coffee in the past 30 minutes?  " format=   standardf.
;
            rename      BPQ150D      =    bpcig30     ;      attrib
      BPQ150D           label="     Had cigarettes in the past 30 minutes?
      " format=   standardf.   ;

      *Blood Pulse Covariates;
            rename      BPXPULS      =    pulsereg    ; attrib BPXPULS
      label="     Pulse regular or irregular?   " format=   pulseregularf.
;
            rename      BPXPTY       =    pulsetype   ; attrib BPXPTY
      label="     Pulse type: " format=   pulsetypef.  ;
run;
proc contents data=blood2 varnum; run;

  *Include other useful demographics;
      data blood;
      merge demographics(keep=SEQN SDDSRVYR Male age) body (keep=SEQN BMI)
blood2 blood3;
      by SEQN;

      *Adjust questions to reflect NHANES design;
      if toldhyper>1 then do;
      toldhyper2plus=2;
      toldhyper_age=.;
      prescp=2;
      if prescp>1 then prescp2=2;
      prescp2=2;
      end;
      if prehyper=1 then prehyper2=.;

      run;

*BloodPressure Dataset;
      data bloodpressure;
      set blood;
      drop  BPXCHR      BPXPLS BPXDB pulsereg pulsetype;
      drop   ckchol     ckchol_time toldchol
             toldfat    toldweight  toldexc      toldpres
             dofat      doweight    doexc dopres BPQ095     BPQ055;
      *Age Category;
      if 0<=age && age<8 then bpage=-2;   *INFANTS. Data not collected;
      else if 8<=age && age<16 then bpage=-1;   *Children. Data on BP is
collected, questions not asked.;
```

```
      else if 16<=age && age<18 then bpage=0;   *Teens.     Data on BP and
Question on BP asked;
      else if age>=18 then bpage=1;       *Adults;    *Data on BP collected,
Question of BP asked, hypertension defined;
      else if age=. && BPXSY1>0 then bpage=2;
      else bpage=.;
run;

*Data Altertation ahead--Look at data;
proc means data=bloodpressure N NMiss Mean Std Min Max;
class bpage/missing;
var SEQN BPXSY1-BPXSY4 BPXDI1-BPXDI4 toldhyper prescp  prescp2;
run;

data bloodpressure;
retain  SEQN SDDSRVYR bpstatus bpage
            /*Demographics*/ Male age Height BMI
      /*Summary Measures*/ avgSP avgDP  totEn hypertension
      /*Questions*/     ToldHyper Toldhyper2plus      Toldhyper_age
                              Prescp Prescp2
                             prehyper    prehyper2
                             bp_home     bp_home2    Toldbphome
      /*Blood Pressure*/      BPXSY1      BPXDI1      BPAEN1      BPXSY2
      BPXDI2      BPAEN2      BPXSY3      BPXDI3      BPAEN3      BPXSY4
      BPXDI4      BPAEN4
      /*Blood Pressure Covaraiates*/      bpcomm      bptime      arm
      cuff_sz     inflation
            bpfood30 bpalchl30 bpcoffee30 bpcig30     ;
set bloodpressure;

*Blood pressure is measured on 8 year-olds and higher--take average;
      if age>=8 then do;
                  avgSP=MEAN(of BPXSY1-BPXSY4); /*The mean takes the average
of the nonmissing data*/
                        attrib avgSP label="Average Systolic Blood Pressure
(mm Hg)" format=Comma4.0;
                  avgDP=MEAN(of BPXDI1-BPXDI4); /*doesn't matter if colmns
are out of order*/
                        attrib avgDP label="Average Diastolic Blood Pressure
(mm Hg)" format=Comma4.0;
                  totEn=SUM((BPAEN1-1),(BPAEN2-1),(BPAEN3-1),(BPAEN4-1));
      /*# of No's in 4 readings*/
                              attrib     totEn label="     Number of No
Enhancements used in the four readings    " ;
                  *Some way to keep track of nonmissing numbers in
average???;
                  * SP_n=N(of BPXSY1-BPXSY4);    *DP_n=N(of BPXDI1-BPXDI4);
            end;
      else do; *If age<8 or is missing then ignore;
                  avgSP=.; avgDP=.;
            end;

      *If an individual has not been told to have hypertension (or
refused/don't know) (toldhyper=2,7, or 9), assume not currently taking
medicines(and not taking prescription)  (prescp2=2);
      *If not taking prescribed medicine or refused or don't know (prescp=2,
7, or 9), assume not currently taking medicines (prescp2=2);
```

```
            if toldhyper>1 OR prescp>1 then prescp2=2;
      *     if toldhyper>1 then prescp=2;
            *It's an or due to design;

      *Adjust bloodpressure above by exlcuding those who drank coffee or
smoked 30 mintues before:
                  code: *if bpcoffee30 NE 1 OR bpcig30 NE 1 then do;
*Hypertension;
      *Adults: age>=18;
      if bpage>=1 or bpage=. then do;
            *For adults, Hypertension is defined as avgSP>=140 OR avgDP>=90
OR current use of antihypertensive medicine;
            if avgSP>=140 OR avgDP>=90 OR prescp2=1 then hypertension=1;
      *hypertension;

            *Being normal is having avg SP<140 AND avgDP<90  OR not taking
current antihypertensive medicine. Assume refusal/don't know
            also means being normal;
            else if (0<=avgSP && avgSP<140 AND 0<=avgDP && avgDP<90) OR
prescp2=2 then  hypertension=0;     *does not have hypertension;

            *If missing SP OR DP OR refused/don't know for current use of
antihypertensive medicine then hypertension is missing;
            else if (avgSP=. AND avgDP=.) OR prescp2>2 then hypertension=.;

            end;

      *Children and Teens:8<=age<18;
      else if bpage=-1 or bpage=0 then do;

            *Teens who currently take antihypertension medicine have
hypertension;
            if prescp2=1 then hypertension=1;
            else if prescp2=2 then hypertension=0;
            else do;
                  hypertension=-1;
             *For missing medication(children): Those aged 3 and over--
Hypertension is determined by:
            Hypertension is defined as an average systolic or diastolic blood
pressure on three or more occasions
                  equal or higher than the 95th percentile appropriate for
the sex, age and height of the child
            Prehypertension is average SP or DP > 90th percentile;
            *For now we will ignore;
            *see SAS code at:
http://www.cdc.gov/nccdphp/dnpao/growthcharts/resources/sas.htm;
                  end;
      end;
      *Blood Pressure is not measured on infants nor questions asked;
      else if bpage=-2 then hypertension=.;

      /*ALTER: TOO YOUNG*/
      /* if age<16 then do; prescp2=-1; prescp=-1; toldhyper=-1; end;  */

      attrib hypertension label="Have Hypertension?"; * format=diseasef.;
      run;
proc contents varnum; run;
```

```
*Data ANalysis;
 proc means data=bloodpressure N NMiss Mean Std Min Max;
 var SEQN avgSP avgDP prescp2 hypertension prescp toldhyper;
run;


proc freq data=bloodpressure;
tables bpage/list missing; title2 "ages";
tables hypertension/list missing;
tables prescp2 prescp toldhyper/list missing;   title2 "How hypertension
defined: Question";
run;

proc means data=bloodpressure N NMiss Mean Std Min Max;
class bpage/missing;
*where bpage>=1 or bpage=.;
var SEQN avgSP avgDP prescp2 hypertension prescp toldhyper;
run;

*For Hypertension & Arsenic study;
proc freq;
where age>=20;
tables hypertension/list;
run;

*Blood Pressure Quality Control;
proc freq;
      tables bpstatus bpcomm/list missing;
      tables BPAEN1-BPAEN4 totEn/list missing;
run;
      proc means N NMiss;
      title2 "Overall";
      var BPXSY1-BPXSY4 avgSP;
      run;

      proc means N Nmiss mean std min max;
      class bpstatus/missing;
      var BPXSY1-BPXSY4 avgSP;
      run;

      proc means N Nmiss;
      class bpcomm/missing;
      var avgSP avgDP;
      run;

*Blood Pressure Covariates;
      title "BP Covariates";
      proc means data=bloodpressure N Nmiss Mean Std Min Max;
      var age Height BMI bptime inflation BPXSY1-BPXSY4 ;
      run;
      proc freq;
      tables bpstatus bpcomm/list missing;
      tables Male arm cuff_sz bpfood30 bpalchl30 bpcoffee30 bpcig30/list
missing;
      tables BPAEN1-BPAEN4/list missing;
      run;
```

```
      *BPAEN1: 1=Yes, 2=No, 8=Could Not Obtain;

      *Very messy regression--didn't clean up;
            *Why don't we exclude those who said yes to bpfood30, bpalchl30,
etc.? ;
            *Include adults;
            *Include BP Questions;
      proc genmod;
      where age>=20;
      class male arm cuff_sz bpfood30 bpalchl30 bpcoffee30 bpcig30/ref=first;
      model avgSP=male arm cuff_sz bpfood30 bpalchl30 bpcoffee30 bpcig30 age
bmi bptime inflation /type3;
      run;
      proc genmod;
      where age>=20;
      class male arm cuff_sz bpfood30 bpalchl30 bpcoffee30 bpcig30/ref=first;
      model avgDP=male arm cuff_sz bpfood30 bpalchl30 bpcoffee30 bpcig30 age
bmi bptime inflation /type3;
      run;

*BP Question DataSet;
title "Do you have hypertension? (Asked to age>=16)";
proc means data=bloodpressure N NMiss Mean Std;
class SDDSRVYR;
var  ToldHyper    Toldhyper2  Toldhyper_age
                              Prescp      Prescp2
                              prehyper    prehyper2
                              bp_home     bp_home2    Toldbphome;
run;

proc freq data=bloodpressure;
      where bpage=1 OR bpage=2;
      tables ToldHyper  Toldhyper2
            Prescp      Prescp2
            prehyper    prehyper2
            bp_home           Toldbphome/list missing;
      *Three confusing variables;
               table prehyper*prehyper2/missing norow nocol nopercent ;
             *If have prehypertension, not asked about borderline
hypertension;
             *Prehypertension and borderline hypertension are the same thing;

             table            prescp*prescp2/missing norow nocol nopercent;
                  *prescp2 is asked only if answered yes to prescp. Prescp2 i
don't think is useful...;
                  *prescp2--now taken medicine; *prescp: if you have taken
prescribed;

             *Going home?;
             table bp_home*Toldbphome/missing norow nocol nopercent;
             *Does this table make sense;
              table toldhyper*prescp/missing norow nocol nopercent;
run;
proc means N Nmiss Mean Std Min Max;
      where bpage>=1;
      var Toldhyper_age bp_home2;
run;
```

```
title "# of Indvls controlling BP";
*--that is avgSP and avgDP say they are normal but there is evidence of
hypertension";
proc freq;
where bpage>=1 or bpage=.;
where avgSP<140 AND avgDP<90;
tables toldhyper toldhyper2 prehyper prehyper2  Prescp prescp2 bp_home
      Toldbphome/list;
run;

proc freq;
where bpage>=1 or bpage=.;
tables prescp2/missing;
tables hypertension/list;
run;


 *Cleanup and Summary;
data bloodpressure;
retain  SEQN SDDSRVYR bpstatus      bpcomm bpage
      /*Demographics*/ Male age Height BMI
      /*Summary Measures*/ hypertension avgSP avgDP   totEn
      /*Blood Pressure*/      BPXSY1      BPXDI1      BPAEN1      BPXSY2
      BPXDI2      BPAEN2      BPXSY3      BPXDI3      BPAEN3      BPXSY4
      BPXDI4      BPAEN4
      /*Blood Pressure Covaraiates*/      bptime      arm   cuff_sz
      inflation
            bpfood30 bpalchl30 bpcoffee30 bpcig30
      /*Questions*/  ToldHyper      Toldhyper2plus    Toldhyper_age
                           Prescp      Prescp2
                           prehyper    prehyper2
                           bp_home     bp_home2    Toldbphome
      ;
set bloodpressure;
      run;


      proc contents varnum; run;
      proc means N Nmiss; run;

 *Blood Pressure Covariates (SEE ABOVE);
      title "BP Covariates";
      proc means data=bloodpressure N Nmiss Mean Std Min Max;
      var age Height BMI bptime inflation BPXSY1-BPXSY4 ;
      run;
      proc freq;
      tables bpstatus bpcomm/list missing;
      tables Male arm cuff_sz bpfood30 bpalchl30 bpcoffee30 bpcig30/list
missing;
      tables BPAEN1-BPAEN4/list missing; *BP MEASUREMENT COMMENTS;
      run;
```

```
*******************************************************************************
*********************************;
```

Main variable: Arsenic;

```
*Copied from Arsenic SAS CIDE;
/*Variables:

SEQN - Respondent sequence number
WTSA2YR - Environmental A two year weights
URXUCR - Urinary creatinine (mg/dL)
URXUAS - Urinary total arsenic (ug/L)           URDUASLC - Urinary Arsenic
comment code
URXUAS3 - Urinary Arsenous acid (ug/L)          URDUA3LC - Urinary Arsenous
acid comment code
URXUAS5 - Urinary Arsenic acid (ug/L)           URDUA5LC - Urinary Arsenic
acid comment code
URXUAB - Urinary Arsenobetaine (ug/L)           URDUABLC - Urinary
Arsenobetaine comment code
URXUAC - Urinary Arsenocholine (ug/L)           URDUACLC - Urinary
Arsenocholine comment code
URXUDMA - Urinary Dimethylarsonic acid (ug/L)         URDUDALC - Urinary
Dimethylarsonic acid comment
URXUMMA - Urinary Monomethylacrsonic acid (ug/L)      URDUMMAL - Urinary
Monomethylacrsonic acid comment
URXUTM - Urinary Trimethylarsine Oxide (ug/L)         URDUTMLC - Urinary
Trimethylarsine Oxide comment
*/


*Arsenic: Creating Dataset;
            %InputnHanes (filename=UAS_D.xpt, outlib=work);
            %InputnHanes (filename=UAS_E.xpt, outlib=work);
            %InputnHanes (filename=UAS_F.xpt, outlib=work);

            data arsenic2;
                  set UAS_D UAS_E UAS_F;
                  one=1;
                  *make it easier to read;
                              rename URXUAS=tAs;
                              rename URDUASLC=tAsc;
                              rename URXUCR=creatinine;
      /*adjusts for dilution effect--creatinine*/
            run;

      data arsenic;
      merge weights(keep=SEQN SDMVPSU SDMVSTRA) arsenic2
demographics(keep=SEQN SDDSRVYR age);
      by SEQN;
      if one=1;
      *Formats for Comments;
                  attrib  tAsc format=limitf.;
                  attrib  URDUA3LC format=limitf.;
                  attrib  URDUA5LC  format=limitf.;
                  attrib  URDUDALC  format=limitf.;
```

```
                  attrib  URDUMMAL  format=limitf.;
                  attrib  URDUABLC format=limitf.;
                  attrib  URDUTMLC format=limitf.;
      run;

*Data Management;
      data arsenic;
            set arsenic;
                  if 6<=age OR age=. then do;
                        *Total inorganic arsenic: arsenic acid (AS(V)O4),
arsenite acid (AS(III)O3), DMA, MMA;
                        iAs=sum(URXUAS3,URXUAS5,URXUDMA,URXUMMA);
                              attrib iAs label="Total Urinary Inorganic
Arsenic (ug/L)";
                        iAsc=sum(URDUA3LC,URDUA5LC,URDUDALC,URDUMMAL);
                              *The variable named LBD__LC indicates whether
the result was below the limit of detection (DL).
                        There are two values: “0”, and “1”. “0” means that
the result was at or above the (DL).
                        “1” indicates that the result was below the limit of
detection, and in these cases the value is DL/sqrt(2). ;
                        attrib iAsc label="Number of Times Inorganic Urinary
Arsenic is Below Limit of Detection";

                        *Total organic arsenic: arsenobetaine, arsenocholine,
trimethyl oxide, arsenosugars. ;
                        oAs=sum(URXUAB,URXUAC,URXUTM);
                              attrib oAs label="Total Urinary Organic Arsenic
(ug/L)";

                        oAsc=sum(URDUABLC,URDUACLC,URDUTMLC);
                              attrib oAsc label="Number of Times Organic
Urinary Arsenic is Below Limit of Detection";
                  end;
                  else do; *ineligible;
                              iAs=.; iAsc=-1;  oAs=.; oAsc=-1; /*n/a and
ineligible*/
                  end;

                  *log_transformations needed to make distributions look more
normal;
                  lniAs=log(iAs);
                  lnoAs=log(oAs);

                  *useful for merging.;
                  one=1; attrib one label="Useful in selecting Arsenic filled
data";

                  *Drop organic and inorganic arsenic along with comments;
                  *drop URXUAS3     URDUA3LC    URXUAS5     URDUA5LC
      URXUAB      URDUABLC    URXUAC      URDUACLC    URXUDMA     URDUDALC
      URXUMMA     URDUMMAL    URXUTM      URDUTMLC          ;;
      run;

            %Categorize(var=iAs);
            %Categorize(var=oAs);
```

```
*Summary;
            data arsenic;
                  retain SEQN  SDDSRVYR
                  /*weights*/  SDMVPSU SDMVSTRA  WTSA2YR
                  /*Covariates/Demographics*/
                        creatinine  /*important covariate for dilution
effect*/
                        age         /*demographics*/
                  ./*total arsenic concentration (not used)*/
                          tAs   tAsc
                   /*inorganic arsenic*/
                        iAs lniAs iAsc  iAsq
                  /*Orginal Data--Inorganic Arsenic*/
                        URXUAS3     URDUA3LC    URXUAS5     URDUA5LC
      URXUDMA     URDUDALC URXUMMA  URDUMMALC
                  /*organic arsenic*/
                        oAs  lnoAs oAsc oAsq
                  /*orginal data--organic arsenic*/
                        URXUAB      URDUABLC    URXUAC      URDUACLC
      URXUTM      URDUTMLC
                  /*For merging*/
                         one;
            set arsenic;
            run;

      proc contents data=arsenic varnum; run;
      proc means data=arsenic N NMiss; run;
      *Covariates: COMMENTS--LIMITS OF DETECTION;
      proc freq  data=arsenic;
            tables  tAsc /list missing sparse;
            tables  iAsc URDUA3LC   URDUA5LC    URDUDALC /list missing
sparse;
      *     tables      oAsc URDUMMAL  URDUABLC URDUTMLC  /list missing
sparse;
            run;

*************************************************************************
*********************************;
```

Other toxic metals

```
      *See Toxic Metals. Copied and left largely unaltered: ln 118-253;
*Part D;
%InputnHanes (filename=PbCd_D.xpt, outlib=work); **;
%InputnHanes (filename=IHg_D.xpt, outlib=work); **;

/*Part E*/
%InputnHanes (filename=PbCd_E.xpt, outlib=work);      /*   */
%InputnHanes (filename=IHg_E.xpt, outlib=work);       /*   */

/*Part F*/
%InputnHanes (filename=PbCd_F.xpt, outlib=work);      /* Pd, Cd, total Hg  */
%InputnHanes (filename=IHg_F.xpt, outlib=work);       /*  iHg  */
*%InputnHanes (filename=UHG_F.xpt, outlib=work);      /*Urinary Heavy Metal*/
      *Also urnary metals too and more metals found;


*For PbCd;
```

```
      *Part D:  *The Household Questionnaire Data Files also contain all
survey design
      variables and sample weights required to analyze these data. The
Phlebotomy Examination file includes auxiliary information on duration of
fasting, the time of day of the venipuncture, and the conditions precluding
venipuncture
;
      data one;
      set IHg_D IHg_E IHg_F;
            rename LBXIHG=iHg_ug;
            rename LBDIHGSI=iHg;
            rename LBDIHGLC=iHg_comm;

            *For merging;
            one=1; attrib one label="For merging";
      run;

      data two;
      set PbCd_D  PbCd_E PbCd_F;
            *cd;
            rename LBXBCD=Cd_ug;
            rename LBDBCDSI=Cd_nmol;
            rename LBDBCDLC=Cd_comm;

            *If i have to convert to same units, convert Cd to
micromols/liter (although nanomoles per liter is better);
            Cd=LBDBCDSI/1000;
            attrib Cd label="Cadmium (umol/L )";

            *lead;
            rename LBXBPB=Pb_ug;
            rename LBDBPBSI=Pb;
            rename LBDBPBLC=Pb_comm;
            *tHg;
            rename LBXTHG=tHg_ug;
            rename LBDTHGSI=tHg;
            rename LBDTHGLC=tHg_comm;

            *For merging;
            one=1;
      run;
            *The fastquest dataset includes auxiliary information on duration
of fasting,
      the time of day of the venipuncture (That is when the metals were
taken), and the conditions precluding venipuncture. ;
      *important???;

      data toxicmetals;
            merge demographics(keep=SEQN SDDSRVYR age adult)  weights one
two; *fastquest;
            by SEQN;
            if one>0;
      *Merging variable not needed;  drop one;

      *Log transformation of metals to avoid influential observations: only
used for individual analysis ;
      *toxicologists actually interpert the log units;
```

```
      lnCd=log(Cd);
      lnPb=log(Pb);
      lniHg=log(iHg);
      lntHg=log(tHg);

      *Formats & Labels;
            attrib Cd_comm label="Number of Times Cd is Below Limit of
Detection" format=limitf.;
            attrib Pb_comm label="Number of Times Pb is Below Limit of
Detection" format=limitf.;
            attrib tHg_comm label="Number of Times tHg is Below Limit of
Detection" format=limitf.;
            attrib iHg_comm label="Number of Times iHg is Below Limit of
Detection" format=limitf.;

      *Fixing Comments: errors--got the number(DL/sqrt(2) from the class
statement on proc means when Cd_comm=1;
            if Cd_comm=. then do;
                  if Cd NE 0.0012500 then Cd_comm=0;
                  else Cd_comm=1;
                  end;
            if Pb_comm=. then do;
                  if Pb NE 0.0090000 then Pb_comm=0;
                  else Pb_comm=1;
                  end;
      run;

*LImits of Detection;
      proc means data=toxicmetals N Nmiss sum;
            var Cd_comm   Pb_comm iHg_comm tHg_comm;
      run;
      proc freq data=toxicmetals; tables Cd_comm        Pb_comm iHg_comm
tHg_comm          /list missing; run;
      proc means data=toxicmetals N Nmiss mean std min max;
            var Cd Pb iHg tHg;
            var lnCd lnPb lniHg lntHg;
      run;
*Summary;
      data toxicmetals;
      retain         SEQN SDDSRVYR
      /*weights*/    WTINT2YR WTMEC2YR SDMVPSU SDMVSTRA
      /*
                           Cd                Pb          iHg   tHg
                           lnCd    lnPb            lniHg lntHg
                           Cd_comm       Pb_comm iHg_comm tHg_comm
                           Cdq               Pbq         iHgq   tHgq
                                             Pb_nmol
            */
      /*Demographics*/  Age Adult

      /*Cadmimum (Cd)*/ Cd_nmol     Cd_ug       Cd    lnCd  Cd_comm     Cdq
      /*Lead (Pb)*/                       Pb_ug Pb    lnPb  Pb_comm     Pbq
      /*inorganic Hg*/              iHg_ug      iHg   lniHg   iHg_comm  iHgq
      /*total (Hg)*/                      tHg_ug      tHg   lntHg   tHg_comm
      tHgq
      ;
      set toxicmetals;
```

```
            *For now & for analysis, I will use the ummol/L quantity and no
covariates, and ignore the weights
      I will just assume there is no problem with any of the data, and use
the iHg, Cd, Pb, and Hg concentrations";
      *iHg Cd Pb tHg  & their quartiles;
      drop
      /*mcg/L measurement (not used)*/    Cd_ug Pb_ug iHg_ug tHg_ug
      /*conditions relating to vacupuncture*/
                  /*top conditions*/          food_fast
                  /*other conditions for vanciputure(lab work)*/
                        coffee coffee_fast
                        alchl  alchl_fast
                        gum    gum_fast
                        antacid antacid_fast
                        dietsuppl suppl_fast
                        ask concern
      /*for merging*/         one
      ;
run;
      proc contents data=toxicmetals varnum; run;
      *SEE ABOVE;

      *Lines of data analysis and macros left in Toxic Metals;
**********************************************************************
**********************************;
**********************************************************************
**********************************;
```

Making data.isAshyper

```
*From making data.isAshyper;
proc sort data=body; by SEQN; run;
proc sort data=bloodpressure; by SEQN; run;
data isAshyper;
merge       work.demographics
            body
            work.is_smoking
            alcohol

            bloodpressure (drop=Toldhyper2plus Toldhyper_age      prehyper
      prehyper2
                                    bp_home     bp_home2 Toldbphome)
/**Remove questions not used*/
            arsenic (drop=SDMVPSU SDMVSTRA  WTSA2YR)  /*Drop the weights*/
            toxicmetals
            ;
 by SEQN;
*if one=1;

      highfat=.;
run;
*Anyway to have more Arsenic subjects?  It is so much less than all my other
datasets;

*Ordering of each dataset;
proc contents varnum data=demographics; run;
proc contents varnum data=body; run;
proc contents varnum data=is_smoking; run;
```

```
proc contents varnum data=alcohol; run;

proc contents varnum data=bloodpressure; run;
proc contents varnum data=arsenic; run;
proc contents varnum data=toxicmetals; run;

*Full Dataset;
proc contents varnum data=isAshyper; run;

*Check if age is right;
proc means data=bloodpressure N Nmiss;
var age;
run;
proc means data= alcohol N Nmiss;
var age;
run;

*Reorganize: Please see Excel sheet;
*Saving the full dataset: isAshyperfull;
data isAshyper isAshyperfull;
retain      SEQN  SDDSRVYR    SESSION     ;
set isAshyper;
*drop merging variables, integer values of age;
            drop cnt one;
            drop note;

*(A)-(B) See line 88: Data Management: Dataset isAshyper. Add variables what
the literature suggests/adjusts;
      *(C) Formatting;
            *clear formatting of seqn variable;
            attrib SEQN format=best6.;
            attrib overweight format=standard2f.;
            attrib black format=standard2f.;

      *(D)Set refused/don't know as missing;
                              if veteran=7 OR veteran=9 then veteran=.;
                              if married=7 then married=.;
                              if pregn=3 then pregn=.;      *cannot ascertain
pregnancy;
                              if hs_inc=77 or hs_inc=99 then hs_inc=.;
                              if fam_inc=77 or fam_inc=99 then fam_inc=.;

run;

proc contents data=work.isAshyper varnum; run;

***FLOW CHART: How to get missing data;
            %put &x &theta &cat2;
            proc contents data=body varnum; run;

            title "Total # of Individuals ";
            *Full Dataset--All covariates and variables where they come from;
            proc means N Nmiss maxdec=3;
            /*&cat2 variables calculated*/
            var SEQN creatinine
            age exam_age old
            pir
```

```
            male
            black race
            alchlusg ALQ120Q  ALQ120U
            is_smoke smoke100 smokefive
            overweight bmi bmxwt bmiwt bmxrecum bmirecum bmxht bmiht
            /*extras*/VETERAN        married          pregn pregnage   HS_SZ
      old     hs_inc fam_inc edu edu2
            ;
            run;
            *Full Dataset: Covariates Only;
            proc means data=isAshyper N Nmiss maxdec=3;
            var SEQN SDDSRVYR
                        lniAs &lnx creatinine
                        pirdiff agediff male black overweight
                        is_smoke alchlusg
                        hypertension avgSP avgDP prescp2
                        highfat ;
            /*extra covariates not used*/ var  VETERAN        married
      pregn pregnage   HS_SZ
                                                                  old
hs_inc fam_inc edu edu2;
            run;
            proc freq;
            tables SDDSRVYR  &y &cat2 race
            pregn VETERAN  married  old HS_SZ edu edu2/list missing;
            run;

            %put &y &theta &cat2;
            *See line 238:  Subsetting the data, drop unneeded variables, and
remove missing values;

            *(D)Subset the data;
                        *Only include adults;
            data isAshyper;
                  set isAshyper;
                        if adult=1;             *if age>=18;
            run;

            *Decided to spell out &theta and &cat in mean statement;
            proc means data=isAshyper N Nmiss maxdec=3;
            var SEQN SDDSRVYR
                        lniAs &lnx creatinine
                        pirdiff agediff male black overweight
                        is_smoke alchlusg
                        hypertension avgSP avgDP prescp2
                        highfat ;
            /*extra covariates not used*/ var  VETERAN        married
      pregn    HS_SZ
                                                                  old
hs_inc fam_inc  edu2;
            title "Adults Only";
            run;

            *(F)Remove missing values--thinking i should do this each
time???;
            data isAshyper;
                  set isAshyper;
```

```
                        *(F) remove missing values;
                        if iAs=. then delete;
            run;
            proc means data=isAshyper N Nmiss maxdec=3;
            var SEQN SDDSRVYR
                        lniAs &lnx creatinine
                        pir pirdiff age agediff male black overweight
                        is_smoke alchlusg
                        hypertension avgSP avgDP prescp2
                        highfat ;
            /*extra covariates not used*/ var  VETERAN        married
      pregn    HS_SZ
                                                                    old
hs_inc fam_inc edu2;
            title "Arsenic Subsample in Adults";
            run;
            data isAshyper;
            set isAshyper;
                        if creatinine=. then delete;
                        if Cd=. then delete;
                        if Pb=. then delete;
                        if tHg=. then delete;
            run;
             proc means data=isAshyper N Nmiss maxdec=3 mean;
            var SEQN SDDSRVYR
                        lniAs &lnx creatinine
                        pirdiff agediff male black overweight
                        is_smoke alchlusg
                        hypertension avgSP avgDP prescp2 hypertension
                        highfat ;
            /*extra covariates not used*/ var  VETERAN        married
      pregn   HS_SZ
                                                                    old
hs_inc fam_inc  edu2;
            title "Complete Toxic Metals";
            run;
            data isAshyper;
            set isAshyper;
                        if agediff=. then delete;
                        if pirdiff=. then delete;
                        if male=. then delete;
                        if black=. then delete;
            run;
            proc means data=isAshyper N Nmiss maxdec=3 mean;
            var SEQN SDDSRVYR
                        lniAs &lnx creatinine
                        pirdiff agediff male black overweight
                        is_smoke alchlusg
                        hypertension avgSP avgDP prescp2
                        highfat;
            /*extra covariates not used*/ var  VETERAN        married
      pregn    HS_SZ
                                                                    old
hs_inc fam_inc  edu2;
            title "And Complete Demographics";
            run;
```

```
            data isAshyper;
            set isAshyper;
                              if overweight=. then delete;
            run;
            proc means data=isAshyper N Nmiss maxdec=3 mean;
            var SEQN SDDSRVYR
                        lniAs &lnx creatinine
                        pirdiff agediff male black overweight
                        is_smoke alchlusg
                        hypertension avgSP avgDP prescp2
                        highfat ;
            /*extra covariates not used*/ var  VETERAN        married
      pregn    HS_SZ
                                                                    old
hs_inc fam_inc  edu2;
            title "And Complete Demographics";
            run;

            data isAshyper;
            set isashyper;
                              if is_smoke=. then delete;
                              if alchlusg=. then delete;
            run;
            proc means data=isAshyper N Nmiss maxdec=3 mean;
            var SEQN SDDSRVYR
                        lniAs &lnx creatinine
                        pirdiff agediff male black overweight
                        is_smoke alchlusg
                        hypertension avgSP avgDP prescp2
                        highfat;
            /*extra covariates not used*/ var  VETERAN        married
      pregn   HS_SZ
                                                                    old
hs_inc fam_inc edu2;
            title "Removing Alcohol and Smoking";
            run;
            data isAshyper;
            set isAshyper;
                                    if &y=. then delete;          *Remove
missing y values;
            run;

            proc means data=isAshyper N Nmiss maxdec=3 mean;
            var SEQN SDDSRVYR  &lnx &theta age pir &y &cat2 lniAs ;
            /*hypertension*/ *var SEQN avgSP avgDP prescp2 hypertension ;
            var highfat;
            /*extra covariates not used*/ var pregn   VETERAN     married
HS_SZ edu2;
            run;


      *Categorizing the Continuous Variables;
                        *Biological/Important Variables (&bio);
            *           %Categorize (data=work.isAshyper, var=highfat);
*normal;
            *           %Categorize (data=work.isAshyper, var=highfat2);
*normal;
```

```
                        *Continuous Covariates (&theta);
                        %Categorize(DATA=work.isAshyper, var=creatinine);
                        %Categorize(data=work.isAshyper, var=PIR);
                        %Categorize(data=work.isAshyper, var=age);
                        %Categorize(data=work.isAshyper, var=BMI);
                        %Categorize(data=work.isAshyper, var=alchlusg);

                        *Continuous Metals (&x);
                        %Categorize(data=work.isAshyper, var=iAs);
                        %Categorize(data=work.isAshyper, var=Cd);
                        %Categorize(data=work.isAshyper, var=Pb);
                        %Categorize(data=work.isAshyper, var=iHg);
                        %Categorize(data=work.isAshyper, var=tHg);


*Save final dataset;
      data isAshyperfinal; *data.isAshyperfinal;
      set isAshyper;
      run;

      /*proc contents data=data.isAshyperfinal    varnum; run;*/
            *Subset the variables you need and drop the rest;
```

A2.3 Logistic Regression

```
*Introduction;
      **libraries;
*     libname mymacros "C:\SAS and R Help\SAS Macros (Data Analysis)";
*                      options mstored=yes sasmstore=mymacros;
*     libname data "C:\VCU Biostatistics\research\Arsenic
Project\Data\Arsenic Mixture Project Data (Final_Fall 2013)";
      /*The output dataset*/
*           OPTIONS FMTSEARCH=(data.nHANES );   /*Find format*/
*     run;
      **Getting a nice rtf output;
*      %RTFOptions(file="C:\VCU Biostatistics\research\Working on Journal
Article\03_Results\Final_Logisitic_SAS.doc");
       /*These options allow for better quality rtfs*/

*************************************************************************
************;
*Define common data and response variable for univariate calculations;
                  %let data=isAshyperfinal;
                  %let y=hypertension(event="1");
            %let theta=creatinine pirdiff age alchlusg;          /*list main
effects here*/
            %let cat2=is_smoke male black overweight;
      /*list any categorical variables here*/
            %let xextra=veteran;

proc contents data=isAshyperfinal varnum; run;

title ""; title2 "";
      data &data;
```

```
      set &data;
      *Add polynomial terms for modelfit;
      age2=age**2;
      age3=age**3;
      lnage=log(age);
      pir2=pir**2;
      alchlusg2=alchlusg**2;
      highfat=.;

      *Add id number (useful later for influential observations);
      n=_n_;
      run;

*check if missing values are removed;
      proc means N Nmiss  maxdec=5;
            var hypertension lniAs  &cat2 &theta highfat &xextra;
      run;
      proc means maxdec=2; var iAs; run;

*************************************************************************
************;
*Descriptive Statistics;
            proc means data=&data maxdec=2;
            class hypertension;
            var lniAs  &theta age highfat;
            run;
            proc freq data=&data;
            /*vs. hypertension*/ tables (&cat2 &xextra)*hypertension/nocol
nopercent relrisk;
            /*covariates used in analysis*/
            tables  hypertension VETERAN   black  married  male   pregn
pregnage  HS_SZ   adult old
                        hs_inc edu edu2 agecat overweight is_smoke/list
missing;
            run;
            ods text="Covariates: &x &cat2";

            title "correlation between lniAs and SP/DP";
            proc corr data=isAshyperfinal nosimple noprob;
              var avgSP avgDP lniAs;
            run;

*************************************************************************
************;
title "Step 0: Comparison--Does logistic work the way you think it is";
            proc logistic data=&data;
                        class &cat2/param=ref ref=first;
      *Categorical variables;
                        model &y=&cat2;
                        title "Is logistic working"?;
                  run;
*-------------------------------------------;
      *DATE OF MODEL--Modified the covariate structure Sept 29, 2013;
      title "Manual Model Building (Sept 29 2013)";

            *QUADRATIC EFFECTS;
            %BackwardElim(n=c);
```

```
                title3  "Orginal Covariate Model";
            %BackwardElim (cat=&cat2, x=creatinine pir age alchlusg, n=cov);

                ods text="CONCLUSIONS: although creatinine and is_alchl are
not significant at the 0.25 significance level, they are
                            important biological variables. Thus,  ALL
SIGNIFICANT AT THE 0.25 level , the final covariate model.";

            *Add quadratic terms;
            %BackwardElim (cat=&cat2, x=creatinine pir age alchlusg
alchlusg2, n=1q);             *AlchlUsg Quadratic?;
            %BackwardElim (cat=&cat2, x=creatinine pir pir2 age alchlusg,
n=2q);                  *PIR Quadratic?;
            %BackwardElim (cat=&cat2, x=creatinine pir age age2 alchlusg,
n=3q);                  *Age Quadtratic?;
            %BackwardElim (cat=&cat2, x=creatinine pir age age2 age3
alchlusg, n=4q);              *Age Quadtratic?;
            ods text="Reference—A non-white non-smoking and non-alcholic
(alchlusg=0)
                        20-year old male with normal/underweight BMI
(BMI<25), a PIR of 1";

            ods html;
*SEE TABLE 2
1q COVARIATES ONLY
R-Square 0.2913 Max-rescaled R-Square 0.4011
Hosmer and Lemeshow Goodness-of-Fit Test
Chi-Square DF Pr > ChiSq
49.8845 8 <.0001        *MODEL IS POOR FIT

2q
R-Square 0.2913 Max-rescaled R-Square 0.4011
Hosmer and Lemeshow Goodness-of-Fit Test
Chi-Square DF Pr > ChiSq
51.9740 8 <.0001        *MODEL IS POOR FIT


3q
R-Square 0.2914 Max-rescaled R-Square 0.4012
Hosmer and Lemeshow Goodness-of-Fit Test
Chi-Square DF Pr > ChiSq
47.8701 8 <.0001        *MODEL IS POOR FIT

4q
R-Square 0.2989 Max-rescaled R-Square 0.4116
Hosmer and Lemeshow Goodness-of-Fit Test
Chi-Square DF Pr > ChiSq
10.3426 8 0.2418        *SUGGESTS THAT MODEL IS GOOD FIT
;
```

```
            data summary;
            set parmcov parm1q parm2q parm3q parm4q;
            run;

            data LRT;
            *retain n;
            set summary;
                  drop _LINK_ _TYPE_ _STATUS_ _NAME_  _ESTTYPE_ /* Intercept
lniAs creatinine pirdiff agediff is_smokesmoker is_alchlalcoholic malefemale
black1 overweight*/;
                  constant=-2085.90;
                  LRT=-2*(constant-_LNLIKE_);
                  Critvalue=5.991464547; *2 df: 6.00, Significance
level=0.05;
                  if LRT>Critvalue then decision="reject"; else decision=.;
            run;
            proc print label width=full; title "Finding Covariate Model";
run;
            ods text="Adding quadratic term to age is sigfinicant, indicated
by sigmodal rhsp. The cubic term does not make it fit better;
            The others don't change the model very much in the three
messures.";

            *Final Covariate Model;
            *Center PIR, leave age as is, square age; *Although not
significant at 0.25 level, keep alchlusg, male, and smoking because
biologically relevant;
            %BackwardElim (cat=&cat2, x=creatinine pirdiff age age2 alchlusg,
n=cov);

            title "Logisitic Regression";
            title3 "Step 2: Add Arsenic to Covariate Model";

            proc logistic data=&data outest=parmfinal plots(only label)=all;
*(phat leverage);
                        class &cat2/param=ref ref=first;
      *Categorical variables;
                        model &y=lniAs creatinine pirdiff age age2 alchlusg
&cat2/rsq lackfit; *iplots;
                        contrast "age" age 1 age2 1 /estimate=exp;
                        contrast "lniAs" lniAs 1/estimate=exp;
                        output out=data.parmpred predicted=yhat H=leverage
ResDev=residual reschi=pearson;
                        effectplot slicefit;
                  run;

      /*What is effectplot; Confused because plots option and iplots seem to
be repetitive; */

      *R-Square 0.2991 Max-rescaled R-Square 0.4119 ;
      *Hosmer and Lemeshow Goodness-of-Fit Test
      Chi-Square DF Pr > ChiSq
      11.1453 8 0.1936
      ;
      *Fail to reject: good model;
```

```
      data data.parmpred;
      set data.parmpred;
            n=_n_;
            logit=log(yhat/(1-yhat));
      run;

      *Assumed that each covariate is linear with respect to the log odds of
hypertension;
            title2 "Checking for linearity of final model";
            proc freq data=data.parmpred;
                  tables (iAsq Cdq Pbq tHgq creatinineq pirq ageq alchlusgq)
*hypertension /nopercent norow nocol;
            run;

            *See Excel;


      *Check for Influential Points and outliers;
      proc sgscatter;
      plot(pearson residual leverage)*(SEQN)/datalabel=n group=hypertension;
      run;

*The phat and leverage, c detects outliers; *iplots plot the influence
poinits;
*The index plots
                  The vertical axis of an index plot represents the value of
the diagnostic,
                  and the horizontal axis represents the sequence (case
number) of the observation. The index plots are useful for identification of
extreme values.

*Pearson Residuals =( yi-mean) /sqrt(var(yi)) . One cna show that
sum(ri^2)~chi^2_n-p
                  , deviance residauls    can be used to detect outliers. ;

*Leverage (diagonal elements of hat matrix): extreme points
      (DFFITS, COko's distance)
*DFbetas: remove each observation, useful in finding influentail observation;


      *Note of influential statistics;

      *Leverage: obs 332**, 1876**, 3789**,     566, 1610,**2192, **3504,
4245, 2688
      *Dfbetas: ignore intercept, age2, nothing for categorical variables;

      *Look at influential observations;
      data influential;
      set isAshyperfinal;
      if n=332 or n=1876 or n=3789 or n=2192 or n=3504;
      run;
      proc print data=influential;
      var n SEQN SDDSRVYR lniAs creatinine pirdiff age age2 alchlusg &cat2
hypertension;
      run;

      *Run senstivity analysis with them out;
```

```
      data isAshyperfinal_sens;
      set isAshyperfinal;
            if n=332 or n=1876 or n=3789 or n=2192 or n=3504 then delete;
      run;

      proc logistic data=isAshyperfinal_sens outest=parmfinal;
                        class &cat2/param=ref ref=first;
      *Categorical variables;
                        model &y=lniAs creatinine pirdiff age age2 alchlusg
&cat2/rsq lackfit; *iplots;
                        contrast "age" age 1 age2 1 /estimate=exp;
                        contrast "lniAs" lniAs 1/estimate=exp;
                  run;

*Hosmer and Lemeshow Goodness-of-Fit
Test
Chi-Square DF Pr > ChiSq
12.8992 8 0.1154        *STILL SUGGESTS A GOOD MODEL;

            *To Table 4--Model Selection Criteria for various
models...Combine above;
************************************************************************
************************;
            *Sensitivity Analysis: Arsenic Alone (adjusted for creatine)
(Figure 1);
            proc logistic data=&data outest=parmfinal ;
                        model &y=lniAs  creatinine;
                        output out=parmpred predicted=yhat H=leverage
ResDev=residual reschi=pearson;
                  run;

            title "Effect of Arsenic on Hypertension, adjusted for
creatinine";
            proc sgplot data=parmpred;
                  scatter x=lniAs y=hypertension;
                  scatter x=lniAs y=yhat;
                  loess x=lniAs y=hypertension;
                  xaxis values=(0 to 7 by 1) grid minor minorcount=1;
                  yaxis values=(0 to 1 by 0.1) label="Prob of Hypertension"
grid;
            run;

            *The Effect of Age (Figure 2);
            title "Effect of age";
            proc logistic data=&data outest=parmfinal ;
                        model &y=age age2;
                        output out=parmpred predicted=yhat H=leverage
ResDev=residual reschi=pearson;
            run;

            proc sgplot data=parmpred;
                  scatter x=age y=hypertension;
                  loess x=age y=hypertension;
                  scatter x=age y=yhat;
                  xaxis values=(20 to 90 by 10) grid minor minorcount=1;
                  yaxis values=(0 to 1 by 0.1) label="Prob of Hypertension"
grid;
```

```
            run;

            *Pearson Residual Plot;
            proc sgplot data=parmpred;
                  scatter x=age y=pearson;
            run;

            *#2:; title3 "Senestivity Analysis Excluding Smoking,Eating, etc.
Right Before Hypertension Reading";
            *Looking at these variables that affect hypertension and
excluding them;
            proc freq data=isAshyperfinal;
            tables bpfood30 bpalchl30 bpcoffee30 bpcig30/list missing;
            run;

            *Exclude those who said yes;
            data isAshyperfinal2;
            set isAshyperfinal;
            if bpfood30=1 OR bpalchl30=1 OR bpcoffee30=1 OR bpcig30=1 then
delete;
            run;

            *A check: Total Sample Size: 3183;

            /*Reran final logistic model. Notes: */
            proc logistic data=isAshyperfinal2 outest=parmfinal;
                        class &cat2/param=ref ref=first;
      *Categorical variables;
                        model &y=lniAs creatinine pirdiff age age2 alchlusg
&cat2/rsq lackfit; *iplots;
                        contrast "age" age 1 age2 1 /estimate=exp;
                        contrast "lniAs" lniAs 1/estimate=exp;
                        output out=parmpred predicted=yhat H=leverage
ResDev=residual reschi=pearson;
                  run;



ods rtf close;
```

A2.4 Weighted Quantile Method Sum (WQS)

```
*Inorganic Arsenic and Hypertension among other chemicals.

*Author: Paul Hargarten
*Purpose: Often times, chemicals exist in mixtures. Does Arsenic along with
Pb, Cd, and Hg, affect hypertension?
*Dataset: data.isAshyperfinal;

*Used logistic regression analysis final & complete/one dataset;
*Split data into two parts: (1) test and (2) validation
      *Test Dataset: Do bootstraps on the correlated variables: Estimate
regression coefficients;
      *The model:
            *y=Alpha+Beta_1*Sum(w_j*q_j)+Theta'z;
            *Unknown parameters: Alpha Beta w_1 w_2 w_3 theta_1 theta_2
theta_3,
                  where alpha=beta_0+theta_0;
      *Use nonlinear regression, by constraining beta>0. After all, we expect
that the chemical concs have a direct
relationship with hypertension.

*Modifications;

*6/30/14: WQS method updated. To address the concern that choosing
significant regression coefficients--by some statisticians call this
is a bias since ..........  , the average weight becomes a weighted average,
with the weight being a signal function.
*************************************************************************
************************************************;
*Libraries;
*libname mymacros "C:\SAS and R Help\SAS Macros (Data Analysis)";
*                 options mstored=yes sasmstore=mymacros;
*libname WQSF "C:\VCU Biostatistics\research\Arsenic Project\Data\WQS_Final";
*libname data "C:\VCU Biostatistics\research\Arsenic Project\Data\Arsenic
Mixture Project Data (Final_Fall 2013)";              /*The output dataset*/
*     OPTIONS FMTSEARCH=(data.nHANES);    /*Find format*/
      options spool;

*%RTFOptions(file="C:\VCU Biostatistics\research\Working on Journal
Article\WQS_ReportSASOutput.doc" );             /*These options allow for
better quality rtfs*/
run;

*Easy to implement WQS;
            %let data=isAshyperfinal; *Work dataset used throughout to
calculate;
            %let y=hypertension;
            /*metals*/
            %let x=     iAs Cd      Pb    tHg   ; *iHg;           %let xn=4;
                  %let lnx=lniAs  lnCd    lnPb  lntHg;   *lniHg;
                  %let x_comm=      Cd_comm     Pb_comm     iHg_comm
      tHg_comm;
                  %let xq=    iAsq Cdq    Pbq   tHgq  ;
                  %let lnxq=&xq;
```

```
            /*Covariates
            %let thetan=8;
             /*# of covariates*/
            %let theta=creatinine pirdiff age age2 alchlusg;
      /*continuous covariates*/
            %let cat =is_smoke male black overweight;
      /*categorical covariates*/
            %let thetaextra=VETERAN
                  married          pregn  HS_SZ hs_inc fam_inc edu2; *edu;
/*Demographic covariates not selected*/

            *Size of Dataset (notes);
                  %let N=4386; *see proc means below;
                  *if high fat is included:     the sample size considered
becomes 3098;
                  *Including all missing values:      Ntotal=16557;

            *Common Variables (Report all these);
            %let seed=506079;
            %let quantiles=4;       /*THe division of the data into
quartiles*/
            %let proportion=0.4;    /*the proportion of data to the test*/
            %let method=RANDOM;     /*The method to split up the data. Either
RANDOM or START*/

            *The bootstrap samples;
            %let B=100;  *or 1000 (or 10: to play around to see if no
errors);  *L the number of bootstrap samples, usually 100 or 1000;
            %let nT=1743;     *The size of the bootstrap is the same as the
test dataset, since the method is random, it is &proportion*&N;
            %let alpha=0.1; *Significance level to determine if the slope
parameter of the WQS (beta) is significant;

            proc format;
            value datasetf  1="Test" 2="Validation";
            run;


************************************;
*I. Data Management;
      *(IA) Create local dataset;
            data &data;
            set data.isAshyperfinal;
            age2=age**2;
            run;

      *(B) Check if missing data is removed;
            proc means data=&data N Nmiss min max  maxdec=1;
                  var &y &x &theta &cat highfat;
                  title "Missing Data currently in dataset";
            run;
            *If missing data needs to be removed see below;
      *(C) Categorize the continuous variables that are highly correlated--
the "&x" and their transformations as appropriate;
            *already completed.--may need to edit &xq though;
                        *Use &xq for analysis;
            %Categorize(data=&data,var=iAs,quantiles=4);
```

```
            %Categorize(data=&data,var=Cd,quantiles=4);
            %Categorize(data=&data,var=Pb,quantiles=4);
            %Categorize(data=&data,var=tHg,quantiles=4);

            proc freq data=&data; tables iAsq Cdq Pbq tHgq; run;

      *(D) Reorder Data...see BELOW;
      *(E) Summary of Contents;
      title "Summary of Contents";
      proc contents data=&data varnum; run;

      *(F) Descriptive Statistics;
      title "Descriptive statistics";
proc freq data=&data ;
      tables &y &cat/list;
      title "";
      run;
proc means data=&data N Nmiss mean std min max maxdec=2;
      var &x &theta;
      run;

*-------------------------------------------------------------------------
-------------------------------------------------;
*II. Data Exploration;
*(A) Covariate Model (Used Whole Dataset as starting values for covariate
parameters & intercepts);
      *From logistic regression;
            proc logistic data=&data outest=covariate descend;
                  class &cat/param=ref ref=first;           *Categorical
variables;
                  model &y(event="1")=&theta &cat;
                  title "Covariate Model";
            run;

*(B) Correlations between metals;
      *The code to create a new SAS dataset is to reduce the # of decimal
places from proc corr;
      ODS OUTPUT pearsoncorr=corr;
      proc corr data=&data nosimple; *PLOTS(MAXPOINTS=NONE)=SCATTER;
      var avgSP avgDP &lnx ; with avgSP avgDP &lnx;
      run;
      ods output close;

      data corr; set corr; format _NUMERIC_ Best6.3; title "Correlation";
      run;
      proc print; run;

*(C) Plots;

proc sgscatter data=&data;
matrix lnPb lntHg lnCd lniAs;
run;

                                    axis1 color=black width=1
minor=(number=1);
                                    axis2 color=black label=(a=90) width=1
minor=(number=1);
```

```
%Macro SmoothPlot2(value=70, data=&SYSLAST, y=, x= )        ;
                        proc sort data=&data; by &x; run;
                        *correlation;
                        symbol1 i=sm&value  mode=include;
                                    *only include values within the symbol
statement that you will change;
                                    *The smooth function allows things to be
smoothed out. Play with number, make sure it sorted by horizontal axis
though!;
                        proc gplot data=&data;
                        plot &y*&x  /haxis=axis1 vaxis=axis2 grid noframe;
                        *title "Plot of &y by &x";
                        run; quit;
            %Mend SmoothPlot2;


      %SmoothPlot2(value=70, y=lnPb, x=lniAs);
      %SmoothPlot2(value=70, y=lntHg, x=lniAs);
      %SmoothPlot2(value=70, y=lnCd, x=lniAs);
      %SmoothPlot2(value=70, y=lnPb, x=lntHg);
      %SmoothPlot2(value=70, y=lnCd, x=lntHg);
      %SmoothPlot2(value=70, y=lnPb, x=lnCd);


*(D) Check distributions;
      *Check distribution of response;
      %DataLook (data=&data, var=&y);

      *Check distribution of coorrelated metals;
      %put &x;
      %DataLook(data=&data, var=Pb);
      %DataLook(data=&data, var=tHg);
      %DataLook(data=&data, var=Cd);
      %DataLook(data=&data, var=iAs);
      ods text="Pb, iHg, and Cd look skewed. Take the logarithm to avoid
influential observations.
            A: Influential Observations don't matter because WQS uses the
quantiles & logarithmic is monotonic so the 4th quartile from lniAs is the
same as the 4th from iAs.
            ";

*(E) Odds Ratios for Metals in Mixture Individually (Use in report)
      *(1) See what the individual effects would be--look at coefficients;
      %put "lnx: &lnx ; quartiles, &xq";

      *Individual Logistic Regressions;
      ods output OddsRatios(Persist=PROC) =parm;
      proc logistic data=&data; *ignore output;
      title "Effect of iAsq";
      class &cat iAsq Cdq Pbq tHgq/param=ref ref=first;
      model &y(event="1")=iAsq &theta &cat;
      run;
```

```
      proc logistic data=&data; *ignore output;
      title "Effect of Cdq";
      class &cat Cdq/param=ref ref=first;
      model &y(event="1")=Cdq &theta &cat;
      run;

      proc logistic data=&data;  *ignore output;
      title "Effect of Lead";
      class &cat Pbq/param=ref ref=first;
      model &y(event="1")=Pbq &theta &cat;
      run;

      proc logistic data=&data;
      title "Effect of tHgq";
      class &cat tHgq/param=ref ref=first;
      model &y(event="1")=tHgq &theta &cat ;
      run;  quit;

      ods output close;

      *(3)  Odds Ratios for Metals in Mixture Individually (Use in report);
            *Include only the metals;
      data parm;
      set parm;
      where anydigit(effect)>0;
      if effect~="age2";
      run;

      proc print data=parm;
            *Trying to select only the mixtures;
            title "Adjusted Odd Ratios for Metals in Mixture using Logistic
Regression Individually";
      run;
            ods text=" *Adjusted for creatinine, PIRdiff, agediff, is_smoke,
is_alchl, gpregn, race, and bmicat ";
            ods text=" *Reference:  Non-smoker, non-alcoholic, Male,
Nonblacke, Underweight/normal";

      *(4) Save permanently as odds ratio;
      /*   data WQS.oddsratio; set parm; run;  */

*(F) Data Summary/reset;
      title "";
      proc sort data=&data; by SEQN; run;

*-------------------------------------------------------------------------------
-------------------------------------------------;
*III. Split into two datasets;
            %put %str (Method: &method, Seed: &seed,
                        Proportion:&proportion, N: &N);

      *(A-B) Split depending on method above;
            data &data test validation error;
            set &data;
                  replicate=1; *useful for merging later;
                  method="&method";
                  seed=&seed;
```

```
                  if compress(lowcase(method))="random" then do;
                        *Split the data assuming a random uniform
distribution;
                        *Randomly assign an observation to either set;
                        u=ranuni(seed);
                        if u<&proportion then do;
                              dataset=1;
                              output &data test;
                              end;
                        else do;
                              dataset=2;
                              output &data validation;
                              end;
                        drop u;
                        end;
                  else if compress(lowcase(method))="start" then do;
                        /*The first N observations are test*/
                        *The number of observations to select from;
                        Ntest=&N*&proportion;
                        if _n_< Ntest then do;
                              dataset=1;
                              output &data test;
                              end;
                        else do;
                              dataset=2;
                              output &data validation;
                              end;
                        drop Ntest;
                        end;
                  else put "Method is error. Check to make sure if right";
*ADD STATEMENT TO LOG;
                  drop seed method;
            run;

      *(C) Determine if there is any difference in the two datasets (use in
report);
            title "Comparison of Covariates between Test and Validation
Datasets";
            data temp;
            set test validation;
            attrib dataset format=datasetf.;
            run;

            proc means data=temp N Mean Std maxdec=2;
            class dataset;
            var &theta &x &xq;
            output out=sample N=N;
            run;

            proc freq data=temp;
            tables (&y &cat)*dataset/nopercent norow;
            run;
*-------------------------------------------------------------------------
-------------------------------------------------;
*IV. Test Dataset;
      /*As &x is correlated, and using bootsampling, we randomly generate
with replacement the # of samples. We
```

```
      construct a nonlinear regression on y of the form:
            *y=Alpha+Beta_1*Sum(w_j*q_j)+Theta'z;
            *Unknown parameters: Alpha Beta w_1 w_2 w_3 theta_1 theta_2
theta_3,
                  where alpha=beta_0+theta_0
      */

*(A) Bootsampling;
      /*The number of bootstrap samples L, each with sample of size n;
            Get L repetitive samples of size n:
            Sample   1   2    ... L
            Observ.  1   1
                         2   2
                        ...  ...
                        n   n
            Note: The observations are selected from the data with
replacement.
            Usually L=100, 1000, or 10: play around to see if no errors.
n=same # of observations as sample size of dataset*/
            %put &L, &n &prop &nb "nb should be &prop*&n, the size of test
dataset";
                  proc surveyselect data=test method=urs outhits
                  seed=&seed  n=&nT reps=&B out=bootsamp outseed stats;
                  run;
                  *STATS  includes selection probabilities and sampling
weights in the OUT= output data set for equal probability selection methods
when you do not specify a STRATA statement. This option is available for the
following equal probability selection methods: METHOD=SRS, METHOD=URS,
METHOD=SYS, and METHOD=SEQ. For PPS selection methods and stratified designs,
the output data set contains selection probabilities and sampling weights by
default. For more information about the contents of the output data set, see
the section Sample Output Data Set.;
                  *SAMPSIZE=n:  specifies the sample size, which is the
number of units to select for the sample.
                  *ID STATEMENT: If you do not want to include all variables,
you can use the ID statement to specify which variables to copy from the
input data set to the output (sample) data set.;

*(B) A start dataset-initial values (Covariate Model used);
      *(1) covariates & intercept from covariate logistic model in full
dataset.;
      proc transpose data=covariate out=covariate2; run;
      data covariate2;
            retain n _NAME_ hypertension variable ;
            set covariate2;
            n=_n_;
            variable=cat("t",n);
            rename hypertension=ParmEstimate;
            drop n;
      run;
      proc print data=covariate2; run;
      ods text="Reference—a non-black non-smoking, and non-alcoholic obese
male with a PIR of 1 at 20 years of age. ";
      %put "N=&N";
      %put %str(Strings and Weights);     %put &x;     %put %str(w1, w2, w3,
w4);
```

```
      proc contents data=covariate varnum; run;
*EDIT THIS each time covariates (&theta &cat) changes AND CHANGE cov below in
proc nlp;

      data start;
      set covariate(drop=_LINK_ _TYPE_ _STATUS_ _NAME_ _LNLIKE_  _ESTTYPE_);
            _Type_="PARMS";

            *Initial Estimate for beta_i=0 or pretty small;
            beta=0.05;
            *Play around with this value;

            *From a covariate only model--the intercept & covariate terms are
estimated (using excel as a site to copy & paste);
            rename      Intercept   =     alpha ;
            *rename highfat=t1      ;
            rename      creatinine  =     t2    ;
            rename      pirdiff     =     t3    ;
            rename      age   =     t4    ;
            rename  age2 = t5;
            rename      alchlusg=   t6    ;
            rename      is_smokeyes       =     t7    ;
            rename malemale=t8;
            rename  blackyes=t9;
            rename overweightyes=t10;

            *Estimate the weights as being equal across the set (->each
chemical is non-informative);
            array inwts w1-w4;      *here: inwts w1-w4;
            do over inwts;
                  inwts=1/4;
            end;

            *For merging;
            SamplingWeight=1;
      run;
      data start;
      retain _type_  beta alpha t2-t10 w1-w&xn;
      set start; run;


      title "Starting Values";
      proc print data=start; run;

*(C) Use Proc NLP to estimate variables;
      %put &xq; %put "Covariates: &theta &cat";

      proc nlp data=bootsamp nomiss noprint
                  technique=TRUREG instep=0.15 maxiter=10000 maxfunc=10000
                  inest=start(type=est) outest=two; *cov=M;
            by replicate;           /*for bootstrap samples*/
            *define parameters;
*ADJUST;                parms  beta alpha  t2-t10 w1-w4;
            *Function--logisitic log liklihood with binomial likelihood;
                  WQS=w1*iAsq+w2*Cdq+w3*Pbq+w4*tHgq ;   *define WQS;
```

```
*ADJUST;    cov=alpha+ t2 * creatinine + t3 * pirdiff + t4 * age+t5*age2 + t6
*alchlusg+t7*is_smoke + t8 * male + t9 * black + t10 * overweight; *define
covariate;
                  z=beta*WQS+cov;
                  pi=(1+exp(-z))**(-1);          *logistic cdf;
                  logL=&y*log(pi)+(1-&y)*log(1-pi);  *binomial likelihood:
the binomial coefficient is a constant and can be omitted;
            *objetive function;
                  max logL;
            *constriants;
                  lincon w1+w2+w3+w4=1;
                  bounds 0<w1<=1, 0<w2<=1, 0<w3<=1, 0<w4<=1,   beta>0.000001;

       run;
                  *Created the cov function by copying the diccat variables
from proc logistic and the starting names from data start,
                        added the necessary signs. Copied into Word, merged.
Copied into excel, transposed. Copied into word, convert table->text, use
space. Then copied in SAS.
                  *NEWRAP: Newton-Raphson Algorithm;
            *Gradcheck: determine if the derivative difference is small or
not   ;
                  *maxiter=10000 for proc nlp, the maximum # of iterations to
avoid an infinite loop;
                  *maxfunc=10000 for proc nlp, the maximum # of correction
steps. change if get error;

      *(IV-D) Subset the data into only the parameter estimates and print;
             %put &x;
            data two;
            set two;
                  where _TYPE_="PARMS"; * _TYPE_="INITIAL" OR;
                  drop _RHS_ _ITER_ ;
                  attrib w1 label="weights for iAs";
                  attrib w2 label="weights for Cd";
                  attrib w3 label="weights for Pb";
                  attrib w4 label="weights for tHg";
                  format _numeric_ Best6.3;     *_numeric_ formats all
numeric variables;
                  sum=w1+w2+w3+w4;
            run;
            title "";
            *proc contents data=two varnum; run;
             proc print data=two;
            var replicate beta w1-w4;
            run;

*(E) Keep the betas that are "significant"—do nothing;
*(F)Mean weights;
      proc means data=two maxdec=4;
            var beta w1 w2 w3 w4;
            title "Raw Dataset";
      run;
      **CHANGE DATA SO THAT WEIGHTS ARE REALLY 0;
      data two;
      set two;
                  if -0.01<w1<0 then w1=0;
```

```
                  if -0.01<w2<0 then w2=0;
                  if -0.01<w3<0 then w3=0;
                  if -0.01<w4<0 then w4=0;
      run;
      proc means data=two min p25 mean std p75 max maxdec=4;
            var beta w1 w2 w3 w4;
            title "Adjust weights";
            output out=meanweight;  *Output the mean weights so I can
directly put them into WQS;
      run;
      data meanweight;
            set meanweight;
            where _STAT_="MEAN";
            drop _TYPE_ _FREQ_;
            dataset=2;
            *sum=w1+w2+w3+w4;
      run;
      title "";
      *proc contents data=meanweight varnum;run;


*(G) Histograms of Significant Weights
            *Note: have axis between 0 and 1;
            *Modified my datalook macro to force the range to be between 0
and 1 by 0.05 (midpoint option under histogram).
                  Deleted normal curve and nmidpoint options from DataLook;
            %Macro DataLook2 (data=&SYSLAST, var= ,format=best4.);
                  proc univariate  data=&data noprint;
                        var &var;
                        histogram /midpoints=0 to 1 by 0.05;  *The midpoint
option forces the range;
                        /*INSET Statement Adds a text box inside of the axes
of the plot. This provides a summary statistics*/
                        inset
                              n="N" min="min" q1="Q1" median="median"
      mean="mean"
                              std="std"   q3="Q3"     max="max"
                              /     position= ne            header= "Summary
stats"      format= &format ;
                              /*Takes the plot in top right (eg. NE*/ /*Title
of summary*/

                        title3 "Summary Statistics for &var";
                  run;
            %Mend DataLook2;

      title "Histogram of Significant Positive Weights (Beta)";
      title2 "Metals: &x";
      %DataLook2(data=two, var=w1);
      %DataLook2(data=two, var=w2);
      %DataLook2(data=two, var=w3);
      %DataLook2(data=two, var=w4);

      ods text="Metals: &x";
      title"";  title2"";

*A panel histogram to look at data;
```

```
      proc transpose data=two(keep=replicate w1-w4) out=plottwo
prefix=weight;
        by replicate;
      run;

      data plottwo;
      set plottwo;
    num=cat(trim(_NAME_), "_", substr(_LABEL_,13));   *A nice label for the
graph.
                  *The cat function contrates three character expressions
                  *Trim removes extra spaces
                  *Substring extracts the elements in remainder of string";
      rename weight1=weight;
      run;

      proc sgpanel data=plottwo;
      panelby num;
      histogram weight/scale=count;
      run;

*-------------------------------------------------------------------------------
-----------------------------------------;
*V. Validation Set-Testing WQS;
      title" ";
      %put &x &n;

*(A) Using the weight estimates from NLP, I would calculated the Weighted
Quantile Sum WQSb, ;
      data validation2;
      merge validation meanweight(drop=_STAT_);
            by dataset;
            WQS=w1*iAsq+w2*Cdq+w3*Pbq+w4*tHgq ;
            *The weights are the averages here;
run;

      proc print data=meanweight; run;
      %DataLook(data=validation2, var=WQS);

*(B) Determining whether WQS is significant;
      ods output FitStatistics (PERSIST=PROC)=liklihood;
      proc  LOGISTIC data=validation2;
        class &cat/param=ref ref=first;
        model &y(event="1")=WQS &theta &cat;
       contrast "age" age 1 age2 1 /estimate=exp;
       contrast "WQS" WQS 1  /estimate=exp;
      run;

      /*base model--only need likelihood*/
      proc  LOGISTIC data=validation2;
        class &cat/param=ref ref=first;
        model &y(event="1")=&theta &cat;
      run;
      ods output close;

      *Compare the WQS model to covariate model;
            data liklihood;
            retain n;
```

```
            set liklihood;
            *subset-keep only -2LL;
                  rename InterceptandCovariates=MOI;
                  drop _Run_ _PROC_ InterceptOnly;
            if 1<=_n_<=3 then n="covariate";
            else n="WQS+covariate";
            run;

      *Likelihood Ratio Test;
            data liklihood2;
            set liklihood;
            *reduced liklhd; Reduced= 2477.952  ;
                        LRT=Reduced-MOI;  *i.e. MOI=InterceptandCovariates
(MOI=Model of Interest);
                        Pvalue=1-CDF("CHISQ",abs(LRT),1);
                        if Pvalue<0.05 then decision="reject: WQS
significant"; else decision=.;
            run;
            proc print data=liklihood2;
            where TRIM(Criterion)="-2 Log L";
            title2 "Likelihood Ratio Test: Sig=0.05";
      *     var n  Reduced      MOI    LRT  Pvalue  Decision;
            run;

            *Conclusions;
            ods text="Conclusions: WQS not significant";


/*(C) Interactions between WQS and covariates
A very elegant way to test for interactions using Wald's Test and LRT"*/

      title "For Interaction Table";
            %put &y WQS &theta &cat;

ods text="Since WQS is not significant (after modifying the defintion of
hypertension and alchl usage),
there are no interactions. So the final model is the covariate model. :(";

/* (D) Final Model */
proc means data=validation2 mean std;
var WQS &theta &cat;
run;

proc  LOGISTIC data=validation2 outest=parmfinal; *plots(only label)=all;
*(phat leverage);
        class &cat/param=ref ref=first;
        model &y(event="1")=WQS &theta &cat /rsq lackfit;
       contrast "age" age 1 age2 1 /estimate=exp;
       contrast "WQS" WQS 1  /estimate=exp;
      output out=parmpred predicted=yhat H=leverage ResDev=residual
reschi=pearson;
                        effectplot slicefit;
      run;


 *-------------------------------------------------------------------------
-----------------------------------------;
```

```
/*VI: Assumptions and Senestivity Analyses*/

*Graph of prob(hypertension) with WQS; *Gennings graph;
      proc sgplot data=parmpred;
                  loess x=WQS y=yhat/nomarkers smooth=99;
                  scatter x=WQS y=hypertension;
                  loess x=WQS y=hypertension;
*                 xaxis values=(0 to 3 by 1) grid minor minorcount=1;
*                 yaxis values=(0 to 1 by 0.1) label="Prob of Hypertension"
grid;
            run;

*Principle Component Analysis of the Correlated Metals with hypertension;

*Linearity of WQS with log odds of hypertension in Validation Dataset (see
excel);
      title2 "Checking for linearity of final model";
            %Categorize(data=validation2, var=WQS);
            proc freq data= validation2;
                  tables WQSq *hypertension /nopercent norow nocol;
                  *tables (creatinineq pirq ageq
alchlusgq)*hypertension/nopercent norow nocol;
                              *Assumed to have same relationship with
hypertension in validation as it does in whole data (see logistic);
            run;
      title2 "";

*Run senstivity analysis without influential observations;
      data isAshyperfinal_sens;
      set isAshyperfinal;
            if SEQN=33908 or SEQN=45264 or SEQN=47567 or SEQN=56443 or
SEQN=58221 then delete;
      run;

      *See above for analysis;

*#2:; title2 "Senestivity Analysis Excluding Smoking,Eating, etc. Right
Before Hypertension Reading";
      *Exclude those who said yes;
            data isAshyperfinal2;
            set isAshyperfinal;
            if bpfood30=1 OR bpalchl30=1 OR bpcoffee30=1 OR bpcig30=1 then
delete;
            run;

            *A check: Total Sample Size: 3183;

            /*Reran WQS model. Notes: */
title2 "";


**Sample Size Calculation;

proc  LOGISTIC data=validation2 outest=parmfinal; *plots(only label)=all;
*(phat leverage);
        class &cat/param=ref ref=first;
        model &y(event="1")=WQS /rsq lackfit;
```

```
       *Prob(Hypertension=1|Mean values of Covariates);
       estimate "Mean prob" WQS     1.5045868   / ilink  exp;
       *Prob(Hypertension=1|Mean of WQS+1 std, Mean values of Covariates);
       estimate "Mean prob+1 std" WQS     2.286666 / ilink exp;   *Mean/std
err in table;
      run;



*Estimation of the squared multiple correlation coefficient;
proc glm data=validation2;
        class &cat/param=ref ref=first;
        model WQS=&theta &cat ;
run;
      *R-Square 0.349982 ;

*Estimation of P(hypertension|mean iAs);
proc ttest data=validation2 ci=equal umpu;
      class hypertension;
      var WQS;
run;
```