5-20-2010

The NHANES III database: Design and a retrospective study to identify associations between vitamin D and hypertension

Brandon J. Marzullo

The NHANES III Database: Design and a Retrospective Study to Identify Associations between Vitamin D and Hypertension

Approved: ____

Michael V. Osier, Ph.D.
Thesis Advisor

____

Gary R. Skuse, Ph.D.
Thesis Advisor
Interim Head, School of Biological and Medical Sciences

Submitted in partial fulfillment of the requirements for the Master of Science degree in Bioinformatics at the Rochester Institute of Technology.

Brandon J. Marzullo

May 20, 2010

Abstract

The results from government surveys are often available wholly or in part to the general public. However the format of this information is not always optimal for distribution or analysis, the National Health and Nutrition Examination Survey III, NHANES III, is one such example. This survey is released as data sets containing a Statistical Analysis Software (SAS) program, a column delimited data file, and documentation. This limits access to the data by requiring a license to read in and query the data. By taking the data in its available format and parsing it into a MySQL database the wealth of information contained in the survey has become more widely accessible while maintaining its completeness and without any loss in ability to query the data. The programs used to create the database consist of tools available at no cost.

Recent studies have provided evidence that vitamin D may play a part in cardiovascular health. This evidence, along with the information included in the NHANES IIII, a wide range of survey and examination data, including cardiovascular health and blood serum vitamin D levels, was used to perform a retrospective study searching for a correlation between vitamin D and hypertension across age groups in white males, black males, white females and black females.

The analysis found statistically significant evidence mainly in younger individuals, and especially white females that those diagnosed with hypertension have lower serum levels of vitamin D than their normotensive counterparts.

Thesis Committee

1 Committee Co-Advisor
Dr. Michael Osier
Director of Bioinformatics
Assistant Professor of Biological Sciences
Rochester Institute of Technology

2 Committee Co-Advisor
Dr. Gary Skuse
Interim Head School of Biological and Medical Sciences
Professor of Biological Sciences
Rochester Institute of Technology

3 Committee Member
Mr. Scott Stearns
Bioinformatics Manager
Systems Research & Development
Ortho-Clinical Diagnostics

4 Committee Member
Dr. Leslie Kate Wright
Assistant Professor of Biological Sciences
Rochester Institute of Technology

Acknowledgements

I would like to express my deepest gratitude to all the members of my thesis committee for their help in the preparation of this thesis. Without their help none of this would be possible. I would like to thank every one of my teachers throughout my education. You rarely get to see the finished product of your hard work and dedication and may never get to see how you influence each one of your students.

Finally I would like to thank my family for their love and support in everything I do. It means so much to me.

List of Figures

List of Tables

Table of Contents

Introduction

The NHANES and NHANES III

Since the early 1960's the National Center for Health Statistics (NCHS), as a part of the Center for Disease Control and Prevention (CDC), has been responsible for the survey of the health and wellness of the citizens of the United States. One such survey, the National Health and Nutrition Examination Survey, or NHANES, includes demographic, dietary and many other health related questions for children and adults. It is unique in its use of a personal interview and survey, as well as physical examinations and laboratory diagnostics. Prior to 1999, when the survey became continuous with around 5000 individuals surveyed each year from 15 counties, large amounts of data were collected from written survey and physical examinations (U. D. Services 2007). The final survey prior to this change, the NHANES III, which ran from 1988 until 1994, surveyed persons two months of age and older. It constitutes the final large scale survey on the civilian population of the continental United States, excluding individuals who were institutionalized.

This final survey contains information on many heath conditions. These include cardiovascular health, diabetes, dietary habits and nutrition, and hearing loss. These are just a few of the areas covered, and variables exist for all manner of health related statistics. While some of its data is over 20 years old, the large sample size combined with the wide breadth of its questioner, physical examinations and diagnostics makes the NHANES III an excellent tool in the development of retrospective studies. Complete documentation on how the physical examinations were preformed (U. D. Services, NHANES III Examination Data File Documentation

1996), as well as laboratory procedures (U. D. Services, NHANES III Laboratory Data File Documentation 2006) are available, for free, from the CDC's web page along with most of the results from the previous studies. Information which was deemed problematic to privacy of the individuals, such as location, or any unique or extremely rare condition that could be used to identify a particular person was omitted from the publicly released data but may still be available by request. This allows information collected from these studies to be compared to, or appended to current studies to provide larger sample sizes.

While these surveys are available online at no cost they are distributed, in part, though proprietary formats. Data produced though the NHANES III is released periodically in Statistical Analysis Software (SAS) format, with each section of the survey being released in parts over a period of time, often years, in data sets. Each NHANES III data set is typically released using one SAS-file, one DAT-file and PDF documentation. The SAS-file contains human readable code defining metadata used by SAS about the contents of the DAT-file. Using SAS as an intermediate tool, the SAS-file contains instructions on how to read in the raw data contained in the DAT-file. While SAS has been, and continues to be, the standard format of release for survey data it limits those who can use the data.

The majority of the data released from the NHANES III was done so using this file-pair format. However, at some point around 2005, data sill being released from NHANES III began being released in XPT-files. This format, again requiring an SAS license, allowed for the transport of files quickly across UNIX and Windows platforms. It also combines the metadata of the SAS-file and the data of the DAT-files so that each data set required only one file in addition to the necessary documentation. Unlike the old format the XPT-files are not human readable further obfuscating, and limiting, who has the ability to use the data. Information released from the NHANES III before this switch however constitutes a majority of the total amount of data from that particular survey.

Data from the NHANES III, which is freely available to the public, but difficult to use due to its format, could better serve its purpose if it was distributed instead though a database format which is centralized, complete, and accessible without the need for a license. This can be accomplished by developing tools which take all of the available, and compatible, files released from the NHANES III and parsing them into a MySQL database. These tools themselves should also be freely available and be written in a high-level computer programming language, which is widely known and supported, as well freely available. In that way all the necessary tools available to read, update and query this data set is available at no cost. This dataset, once parsed into a database, would have all the tools built into MySQL for the mining of relationships of interest to any user.

Vitamin D – Role in Health

The biologically inactive forms of vitamin D, referred to as vitamin D_2 and vitamin D_3, are obtained though food or synthesized though the skin when it comes into contact with UVB radiation. Once synthesized it is activated though a two step process that takes place first in the liver and then the kidneys. Vitamin D, actually steroid derived, is activated by parathyroid hormone, PTH, and then works in conjunction with it to stimulate the intake of Ca^{2+} in the intestines (Campbell and Reece 2001). As a hydrophobic steroid hormone, vitamin D is able to enter the cell by way of diffusion across the plasma membrane. Vitamin D then is able to bind to receptors which are members of the nuclear receptor super family. This super family contains domains related to ligand binding, DNA binding, and transcriptional activation. This allows for vitamin D to play a direct role in the regulation of gene expression (Cooper and Hausman 2007). These two properties of vitamin D, its role in calcium uptake as well as its ability to permeate the cell membrane and bind to nuclear receptors, provide vitamin D with several very different roles in health.

Vitamin D is a fat soluble vitamin found naturally in some foods, added though a fortification process to others, and in supplement form. However, it is most commonly obtained when the skin is exposed to UVB radiation, mostly from the sun's rays. Vitamin D ingested, either though foods or supplements, as well as the form of vitamin D synthesized in the skin though UVB exposure is inactive or must undergo two processes to become active in the body (Supplements 2009). Vitamin D_2 (ergocalciferol), or vitamin D_3 (cholecalciferol) is produced in the skin though exposure to UV radiation. These are inactive and need the addition of two OH groups first in the liver, then the kidney, to produce the active form of vitamin D (1, 25-dihydroxycholecalciferol) (Voet, Voet and Pratt 2006).

While most individuals maintain healthy levels of vitamin D though sun exposure, those individuals who live north of 42° N latitude, an imaginary line which stretches approximately from the northern most boundary of California across to Boston, are not exposed to enough sunlight to produce sufficient amounts of vitamin D between the months of November and February, or even longer in the higher latitudes (Cranney, et al. 2007). Individuals who live south of 34° N latitude, below Los Angeles and Columbia, South Carolina, allow for sufficient synthesis year round (Shils, et al. 2005).

Bone Health

Vitamin D is most commonly associated with healthy bone development and maintenance. This role is exemplified though two common examples. First, in infants and adolescents, vitamin D deficiency is associated with stunted growth and deformed bones known as rickets. While rickets was first described as early as 1645, it was not until much later, in the early 20th century, that a method of treatment was developed. Due to vitamin D's hydrophobisity it tends to congregate in the fatty deposits in animals, particularly in fish oils. These fish oils, rich in vitamin D, were later used to treat the disease. Additionally treatment came in the form of exposing the adolescents to sunlight, more specifically UV light with a wavelength between 230 and 313nm (Voet, Voet and Pratt 2006). This treatment was first demonstrated in 1919. In 1930 a steroid, isolated from yeast, called ergosterol, was found to have antirachitic properties when irradiated. This irradiated form of ergosterol was vitamin D_2 (Solomons and Fryhle 2004).

The second example is most prevalent in the elderly, and especially women. Osteomalacia, also characterized through the softening of bones, is caused in adults totally or in part due to vitamin D deficiency. The causes are similar to those of rickets and include low levels of vitamin D, or calcium, or a lack of exposure to sunlight. Bone pain, especially in the hips is also common. In addition to calcium uptake, vitamin D's ability to bind to domains associated in gene expression has led to recent studies finding additional risks to low levels of vitamin D.

Blood Pressure

With its role in muscle contraction and relaxation it is no surprise that calcium plays a role in the regulation of blood pressure. Using the information from the first National Health and Nutrition Examination Survey (NHANES I) it was found that deficiencies in nutritional patters are more likely to characterize hypertensive individuals. Additionally low calcium was the most common nutrient whose lower intake was associated with hypertension (McCarron and Reusser 1999). Vitamin D, as an important regulator in calcium also plays an important role in blood pressure regulation (Gallagher, et al. 1979).

In another study, focusing on the short term effects of calcium on blood pressure and heart rate, found that elderly women who took a vitamin D_3 supplement in addition to a calcium supplement found both a decrease in systolic blood pressure of 9.3% and a decrease in heart rate of 5.4%, both results found to be statistically significant. Additionally 81% of the individuals in the calcium and vitamin D, compared to 47% in the calcium only group, showed a decrease in systolic blood pressure 5mm Hg or greater. No statistically significant results were found regarding diastolic blood pressure (Pfeifer, et al. 2001).

Heart Disease

Vitamin D has also been linked to heart disease in women. In a prospective study of 1484 non-hypertensive women aged 32 to 52, plasma 25(OH)D, the active form of vitamin D, was measured against the multivariable odds ratio. The case and controls were matched on age, race, and month of blood collection, with additional adjustments on family history, prescription drug use, and physical fitness. Evidence suggests that plasma 25(OH)D levels are inversely and independently associated with the risk of developing hypertension (Forman, Curhan and Taylor 2008).

Early thought was that vitamin D may perhaps be a risk factor in heart disease. One such example was a study from northern Norway focusing on whether or not vitamin D was a risk factor for myocardial infarction. This study took 30 individuals, 23 who had no known diseases but subsequently developed myocardial infarctions, and compared known risks factors to 60 controls paired on hemoglobin concentration, height and weight. While the mean concentration of vitamin D was lower in the patient population than the controls, 59.0(24.1) nmol/L to 63.4(27.2) nmol/L, it was not determined to be statistically significant. It was concluded that there was no reason to suggest that higher vitamin D intake, common to northern Norway due to high fish intake, put individuals at a higher risk for myocardial infarction (Vik, et al. 1979). In fact later studies found that diets high in fish, and particularly fatty fish, are protective against mortality cause by coronary heart disease. One study found consumption was associated with a reduction of 34% mortality in Finland, Italy, and the Netherlands (Oomen, et al. 2000).

Vitamin D has also been shown to have a role beyond that of heart disease risks but also cardiovascular diseases. Studies have shown that moderate to severe vitamin D deficiency is a risk factor for cardiovascular disease. In some cases the risk for cardiovascular disease was almost doubled between groups with 25 (OH) D levels below 15ng/mL and those equal or above that level (Wang, et al. 2008). Further studies also concluded that in addition to low serum 25(OH)D being associated with cardiovascular mortality, the differences between serum levels in blacks and white may contribute to higher levels of cardiovascular mortality in blacks (Fiscella and Franks 2010).

Cancer

There have also been links to vitamin D and particular cancers. The presence of a *Taq*I restriction fragment length polymorphism (RFLP) at codon 352 provided evidence, however not statistically significant, that vitamin D plays an important role in prostate cancer risk (Taylor, et al. 1996). Another study also linked men above the median age of 57 and lower levels of 1, 25-D to the risk of prostate cancer. However it was not linked in younger men, yielding similar results in both black and white men (Corder, et al. 1993). Other studies suggest that vitamin D, in conjunction with calcium supplement help to decrease the risk of all cancer types in post menopausal women (Lappe, et al. 2007).

Deficiency and Intoxication Prevention

These deficiencies can be mostly prevented, however, though the proper daily intake of vitamin D. The National Academy of Sciences (NAS) recommends that the adequate intake of vitamin D in normal infants, children and adolescents is 200 IU, 5mcg, per day, down from a previous recommended 400 IU, 10mcg, per day reported previously (Gartner and Greer 2003). The National Institute of Health's Office of Dietary Supplements recommends 200 IU per day for individuals from birth to 50 years old, 400 IU for individuals 51-70, and 600 IU, 15mcg, for individuals 71 years of age and older (Supplements 2009). Obtaining the recommended daily dose only through food can be difficult without the aid of vitamin D fortified foods such as milk and baby formulas. This is especially the case in the higher latitudes and during the winter months where direct sunlight is less common. Infants and adolescents that do not consume at least 500mL of fortified formula or milk each day are recommended to take supplements, especially those who are not regularly exposed to direct sunlight. While the infants, the elderly, and especially women, are at greater risk for vitamin D deficiency due to the mode by which vitamin D is synthesized in the skin, correlations exist between vitamin D and other variables as well.

While hypovitaminosis D, vitamin D deficiency, can cause health risks, vitamin D intoxication also poses risks and is caused by high levels of serum vitamin D over an extended period of time. High levels of vitamin D over stimulate the uptake of Ca^{2+} in the intestine and cause blood levels of Ca^{2+} to rise. This can cause hypercalcification of soft tissues and kidney stones. Evolutionary processes have developed methods of preventing this though skin pigmentation. Darker skin pigmentation prevents as much absorption of UV rays though the skin and decreases the synthesis of vitamin D in the skin. This is one explanation for the tendency for individuals who are indigenous to latitudes near the equator to have darker skin pigmentation than those who are indigenous to the higher latitudes (Voet, Voet and Pratt 2006).

Goals

By transforming the information provided in the NHANES III from its native SAS format into a MySQL database, using tools that are both free to download and widely used and understood, the number of individuals to which have access to the data would no longer be limited to those who have access to Statistical Analysis Software. The large number of individuals in the NHANES III allow for specific queries to produce sufficient sized samples. The data included in the NHANES III then could be used to perform retrospective studies on any number of interesting correlations, including potential links between vitamin D deficiency and high blood pressure, also called hypertension. As hypertension and vitamin D levels are related to many variables it is necessary to split up the data into specific groups based on gender, age, and race. All of these variables are present as individual characteristics in the NHANES and would be easily queried from a database.

By producing groups that are as homogeneous as possible, based on gender, race, and age, it was found that correlations do exist between hypertension and low levels of vitamin D. This correlation suggests hypertensive individuals having lower serum levels of vitamin D than normotensive individuals. These correlations exist at statistically significant levels ($\alpha = .05$) mainly in younger populations. The most significant results were found in young white females. These differences in mean vitamin D levels were found to be at probabilities far too small, less than .0001, to have occurred though random chance, even after conservative corrections for multiple testing.

Materials and Methods

Database Creation

The process of creating a database using the data from the NHANES III was first dependant on how the information was going to split up in the database. The releases from the CDC separate out the data into more than 50 data sets, with each of these sets containing at least three files. The first file contains a Statistical Analysis Software (SAS) program which is used in conjunction with their software to read in the data contained in the DAT-files. The SAS-files contained all the necessary information needed to create a table for each one of the sets, while the DAT-files contained all the data that composed the rows of each table. Any additional files, often PDF-files, were documentation on how the data was collected, summary of answers or results to each question or test, and general background information on the data set as a whole. Several Java programs, written using Java version 1.6.0_17, were written and compiled to perform all of the necessary steps to create the database. When the Java code directly altered or queried the database the Java Database Connectivity API, using the JDBC driver for MySQL, Connector/J version 5.0.8, was used.

In addition to the tables created from the released NHANES III, data two support tables were also necessary to store all the information stored in the SAS-Files. First, the DataDictionary table was populated using all the information from the SAS-files and then was used to create all the other tables. Additionally it provided a central location to search for variables of interest using a text based search on each of the variables short text descriptions provided by the CDC, the variable name itself, or the identification number of the SAS-file, sas_file_id, of the data set you are interested in or any combination of these.

The statement used to create a MySQL table consists of several parts. These parts were parsed out of the SAS-file that accompanied each of the data sets and the names of these files. First each table needed a name with it can be referenced by. Each data set contained two files which had the same name, but different file extensions. This common name was used for each of the tables in the database. For example information from the household youth file pair, which contains all information from the household interview for children two months to 16 years old, is split into two files, youth.sas and youth.dat. Thus the table in the database that refers to this information was named youth. Additionally the name of the files was also used to create a row id for each data set. This integer column was named after the table it appears, the row id column for the youth table was called youth_id.

The SAS-files, which are human readable, were then parsed into the remaining parts necessary to create a MySQL create table statement. Information was also needed on the variable, or column, name and the format of each one of the columns. This data was all obtainable from the human readable SAS-files.

The SAS-files contained either three or four sections, in addition to a heading, each terminated with a semicolon, detailing a particular piece of metadata that was used to define each variable present in that data set. All of the data present under these headers was stored in the DataDictionary, even when it was not necessary to create the MySQL tables. The first section lists the lengths, as an integer, of each of the variables in the set. This length is pertinent to how SAS would read in the file and is not necessary for the creation of the MySQL database directly. However not every section contained a full list of the variables in the set. Since each variable would contain a length, or else it would not exist, the length heading of the

SAS-file was used as the complete list of variables for the data set. These variable names are typically named in a consistent order thought each heading of the SAS-file; however, this is not always the case. Variable names from the length heading were used to create the column names for the table. While the same variable can appear in multiple tables, using the same name, the other information, such as its position in the DAT-file, differed. So while multiple records can have the same variable name the records as a whole are unique. In addition to these column names two additional column names were needed to link the information, through foreign keys, to both the SASFiles and DataDictionary table. The use of foreign keys required the use of MySQL's InnoDB storage engine.

In addition to the column names, the length heading provided the data type for the column as well. SAS allows for two data types, either numerical or character. It differentiates between the two by placing a dollar sign ($) before the variables length. Those variables that SAS defined to be of character type are converted to the MySQL VARCHAR data type. If the dollar sign is absent the variable defaults to numerical in SAS which corresponds to the DOUBLE MySQL data type.

The next heading, format, included in the SAS-file contained information that SAS uses to print the data to the screen while maintaining increased precision data internally. This data was not used in the creation of the database or its population but was included into the DataDictionary. This information could be used in the development of a graphical user interface in the future to define printing routines for each of the variables. The format defines how many numbers should be printed on each side of the decimal point using a "*w.d*" format. The number before the decimal point, *w*, represents the width, the maximum number of digits in the number. The number after the decimal point, *d*, represents the number of digits that appear after the decimal point. For example the format 5.2 would consist of five digits, two of which are to the right of the decimal point, as in 123.45. The "BEST" format is the default format in SAS. Not every variable has a format, and some SAS-files are missing this heading completely.

The input heading, a data range, provided the range of columns that were used to store each of the variables. Unlike many array numbering systems the numbering begins at one, in conflict with the zero based array indexing used by Java. This range refers to character indexes used in the DAT-file for each particular variable. For example many data sets first variable is SEQN, the sample person identification number, and its input is 1-5, meaning the first five characters, indexed one though five, in each row of the DAT-file correspond to that rows SEQN.

The label heading gives a short text description, surrounded in double quotes, of each of the variable names. These labels were stored in the data dictionary, without the double quotes, and provide a rough search method though the DataDictionary by using keywords rather than the often cryptic variable names themselves.

Each of the headings in each of the SAS-files was stored in the DataDictionary along with the primary key, and a reference to the foreign sas_file_id key. This provides an integer key to reference which of the SAS-files the DataDictionary information comes from.

Once all the information from the SAS-file was stored into the DataDictionary the empty tables were created. All the necessary information was parsed from the DataDictionary into the corresponding MySQL create table statement, excluding two variables created for each row entered into the database. These variables included the sas_file_id number which tells which

SAS-file contained the data, and an id number for each row. This number is used to link rows from the same individual that are split up between multiple tables.

The way in which the NHANES III is organized into datasets results in some data sets having many variables, such as those found in the exam data set, while others have only a few. Problems were encountered in those data sets that contained more than 1000 variables, the hard limit for the number of columns in a MySQL table. The data sets that fell into this category, adult, exam, and examse, all needed to be split up into two or more tables to accommodate the large number of columns. The adult dataset, for example, was split up into two tables, the first of which (adult) contains the first 998 variables, the first 998 questions to that section of the survey, plus the sas_file_id, and the table id, for a total of 1000 columns. The remaining 240 variables plus the sas_file_id, and the table id, are in a second table, adultI, which contains 242 columns. Exam and examse each contain 2368 and 2057 variables respectively. These questions were split up between three tables, the first two containing exactly 1000 columns and the third table containing the remaining variables. Identification numbers are included in every row with matching identification numbers between tables representing rows that are from the same individual. The individual whose data was stored with adult_id equal to ten, for example, would correspond to the remaining data stored in table adultI with adultI_id equal to ten.

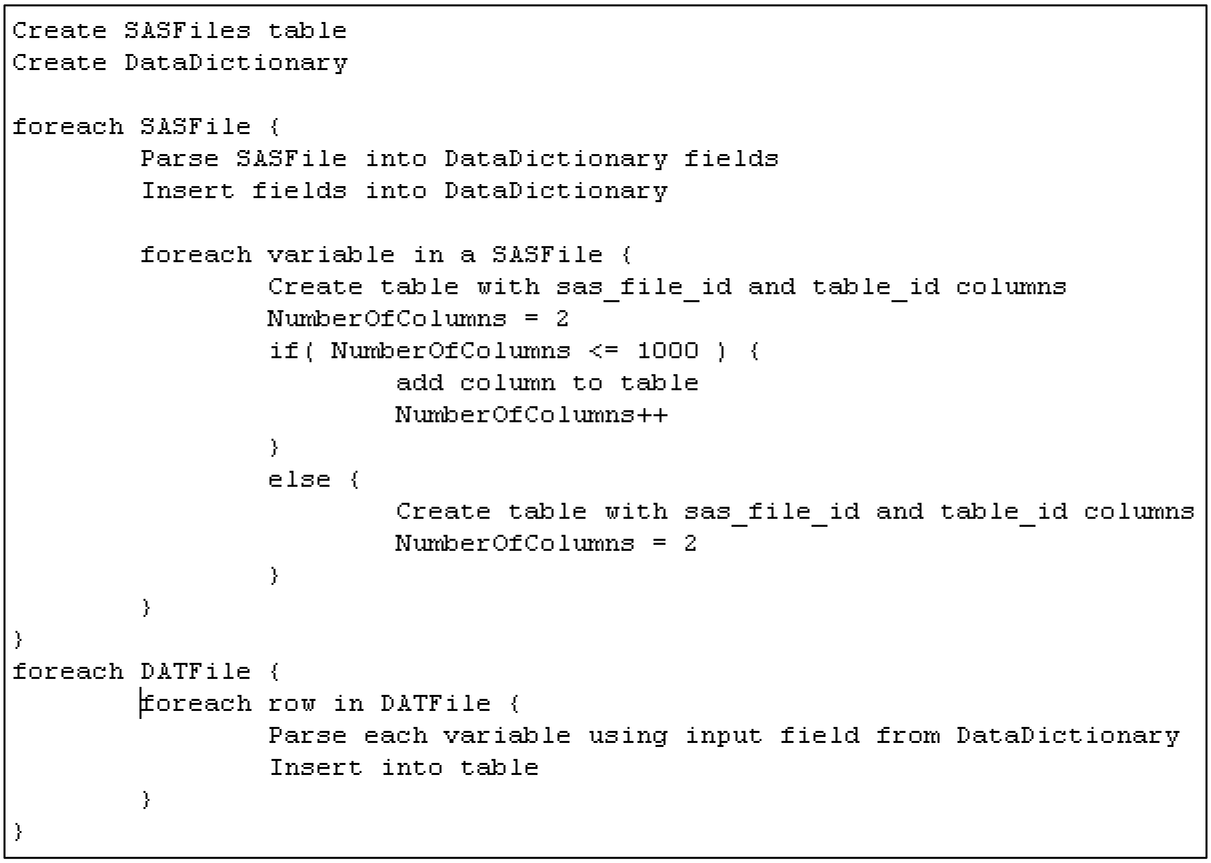
In addition to the data sets available in DAT-SAS file-pairs, some files were in a XPT-file format. This format, also written as XPORT file, is used as a transport file format for SAS. These files, unlike the SAS-files, are not human readable and cannot be parsed using the Java programs used to parse the SAS-files. These files were read into SAS and exported as commaseparated values, CSV, files. An additional Java program was written to read in these files into the DataDictionary and create an empty table based on the metadata of these files. The CSV-files, however, do not contain much of the important metadata contained in the SAS-files. The CSV-files produced by SAS contain only the variable names, and not their input, format, length, or label. For this reason DataDictionary entries for these files are missing this data, though it is still available in the PDF documentation. All the files from the NHANES III that are in XPT-file formats contained data that was in numerical form. For this reason every column created from these variables are of type DOUBLE. None of the CSV-files contained more than 1000 variables.

Population

Once the DataDictionary was created and populated using the information from the SAS-files and the empty tables were constructed, the tables were populated using the information from the DAT-files. The DAT-files contain far less text based formatting that their accompanying SAS-files but contain column delimited data. Using the DAT-file, with its corresponding metadata stored in the DataDictionary, each line of the DAT-file was split into the correct number of variables using each variables input range to split it apart from its neighboring values. For example in most data sets the first variable is the identifier sequence number (SEQN), the input range for the SEQN in each data set is 1-5. During population the first five characters are split from the row and entered into the corresponding table that shares the name of the DAT-file.

Problems only arose in particular rows where individuals left blank answers at the end of a survey. Some rows, which correspond to a particular participant, do not contain results to questions or measurements at the end of that section and the end of these lines do not have the appropriate white space added to make them as long as the input range would require. This results in some rows having a shorter length than others, with these differences varying. In order to cope with this lack of uniformity IndexOutOfBoundsExceptions thrown by Java's substring method, used to parse out the data, must be caught and ignored, inserting null values into all columns that throw those errors.

Data from the CSV-files were inserted without the aid of the DataDictionary and those tables were populated straight from their corresponding CSV-files.



```
Create SASFiles table
Create DataDictionary

foreach SASFile {
        Parse SASFile into DataDictionary fields
        Insert fields into DataDictionary

        foreach variable in a SASFile {
                Create table with sas_file_id and table_id columns
                NumberOfColumns = 2
                if( NumberOfColumns <= 1000 ) {
                        add column to table
                        NumberOfColumns++
                }
                else {
                        Create table with sas_file_id and table_id columns
                        NumberOfColumns = 2
                }
        }
}
foreach DATFile {
        foreach row in DATFile {
                Parse each variable using input field from DataDictionary
                Insert into table
        }
}
```

Figure 1 – Algorithm for Database Creation

Query Design

In order to be included in the statistical analysis individuals must have had serum vitamin D samples, variable VDP, present in the database. These samples of 25(OH)D were performed at the National Center for Environmental Health in Atlanta, GA using the DiaSorin RIA kit (National Center for Health Statistics 2009). The detection limit for the tests was 5.0ng/mL and the results were recorded to one decimal point (U. D. Services, NHANES III Laboratory Data File Documentation 2006). Two separate queries were developed to divide hypertensive and normotensive individuals and then determine average levels of serum 25(OH)D. This version of vitamin D is the active form of vitamin D and is used due to the fact that it is the predominant circulating form of vitamin D (Gunter, Lewis and Koncikowski 1996).

The NHANES III contained a survey question, variable HAE2, which asked the question, “Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?” This question was used to separate individuals into two groups, those who responded yes to the question, and those who responded no. Individuals who responded that they didn’t know, or left the field blank were not included. These two groups were further divided based on three other variables, race (DMARACER), sex (HSSEX), and age (HSAGEIR). Individuals who responded yes to variable HAE5A, in regards to if they were currently taking medication for their high blood pressure, were excluded. Individuals who reported that they were not on high blood pressure medications, no to HAE5A, or did not respond to that question were included. Individuals who left the question blank only did so as they were ineligible to answer questions about hypertension, not ever having been diagnosed with hypertension.

While vitamin D data was collected from individuals 12 years of age and older only individuals who were 18 to 89 years old were included. Information was present for several individuals 90 years of age and older, however the age of individuals 90 years of age and older were all recorded as 90 years old in an effort to protect their identity. The age ranges were split into three approximately equally sized groups based on the yes responses, of which there were a far fewer number than negative responses and would be a limiting factor in the statistical analysis.

The second query utilizes the blood pressure readings present in the NHANES III database. The average systolic (HAZMNK1R) and diastolic (HAZMNK5R) measurements included for each individual contains an average of blood pressure of all available readings. This average contains zero to three blood pressure readings from the in home survey and/or zero to three blood pressure readings done at the mobile examination centers. These readings were done on the same day, or within a few weeks of the blood being drawn. Each of these averages utilize six or less separate measurements with no more than three being taken at either the MEC or during the household interview. These measurements, being averages, may contain unusual numbers. The diagnosis for hypertension was an average systolic blood pressure reading over 140 mm Hg or an average diastolic blood pressure reading over 90 mm Hg (Chobanian, et al. 2003). This was the same diagnosis used to separate individuals into two groups, those whose blood pressure readings reflect a hypertensive state and those whose blood pressure readings do not reflect a hypertensive state. Again the age groups were separated into thirds based on the number of individuals who are hypertensive, and the attempt to fit an equally large number of total individuals into each group to maximize statistical power. Individuals on prescription drugs to treat their hypertension were excluded.

Statistical Analysis

To compare the two groups a non-paired two-sample t-test was preformed. This test is used to compare the means of two groups from a normal distribution. The information provided by the NHANES III provides a snapshot in time of a particular individual's health and wellness. As such comparisons were non-paired as each individual contributed only one data point. Using Minitab 15 to graph the frequencies of all available vitamin D levels revealed a slightly normal distribution. However steps were necessary to achieve a more normal distribution.

Normalization

To normalize all the available vitamin D variables Minitab 15 was used to perform a Box-Cox transformation on all serum vitamin D levels (VDP) to create the normalized transformed variables (tVDP) which were then inserted into the database.

In order to assure that the data had not been overly skewed from this transformation. Normal cumulative distribution functions were calculated using both the normalized and non-normalized data to ensure that approximately the same percentage of people were vitamin D deficient in both the normalized and non-normalized data.

Statistical Testing

Two sample t-tests were first performed using the CPAN module Statistics::TTest, version 1.10 using Perl version 5.10.0, using the average, sample standard deviation, and count values returned by the MySQL queries. This Perl module first calculates the F-statistic to test for equal variances with alpha equal to .05. This determines if the variances of the two samples can be pooled during the t-test. Using this information a two sample t-test was performed to determine if the average level of vitamin D in individuals who have been diagnosed with hypertension, either at any time by a health care provider or using their blood pressure information in the NHANES III database, was different than those who were normotensive. The test was preformed with Ho: $\mu1 = \mu2$ and Ha: $\mu1 \neq \mu2$, at a confidence level of 95% and the p-value is recorded. Additionally a t-test was preformed with Ho: $\mu1 = \mu2$ and Ha: $\mu1 < \mu2$. The Perl module does not support one sided alternate hypothesis testing so they were calculated by hand. If the p-value of the one sided t-test is less than the confidence level of .05, and the non-normalized mean of the hypertensive group is less than the non-normalized mean of the normotensive group, Ho is rejected in favor of Ha which shows that there is statistical evidence to support that individuals who are diagnosed with hypertension have a lower serum level of vitamin D than their normotensive counterparts. In addition to the normalized averages, normalized standard deviations, and p-values the non-normalized statistics were also collected for reference.

These tests were performed on four demographic groups, white males, black males, white females, and black females as recorded by the NHANES III. These four groups were then split into tertiles based on age. All statistics were preformed on the normalized (tVDP) vitamin D data, with $\mu1$ being the hypertensive group and $\mu2$ the normotensive group. Statistically significant was defined as less than .05.

Results

The NHANES III Database

Inserting all available files from the NHANES III resulted in the SASFiles table, which contains information on what files have been inserted into the database, containing 62 records. In addition to these files the two support tables, SASFiles and the DataDictionary, were created and populated. The tables range in size from a few hundred records to nearly 450,000 records. The DataDictionary, which contains information on each variable in the database, contains 10,281 records.

Normalization

The raw serum vitamin D levels (VDP) from all available individuals in the NHANES III database roughly fit a normal curve, as seen in figure 2. Vertical reference lines in figure 2 correspond to two normal cut-offs used to denote vitamin D deficiency. Levels above 15ng/mL typically are accepted as adequate levels for bone and overall health in most situations, with those between 10 and 15ng/mL considered inadequate for overall health. Individuals below 10ng/mL are typically associated with vitamin D deficiency and leads to rickets in children and osteomalacia in adults (Supplements 2009).

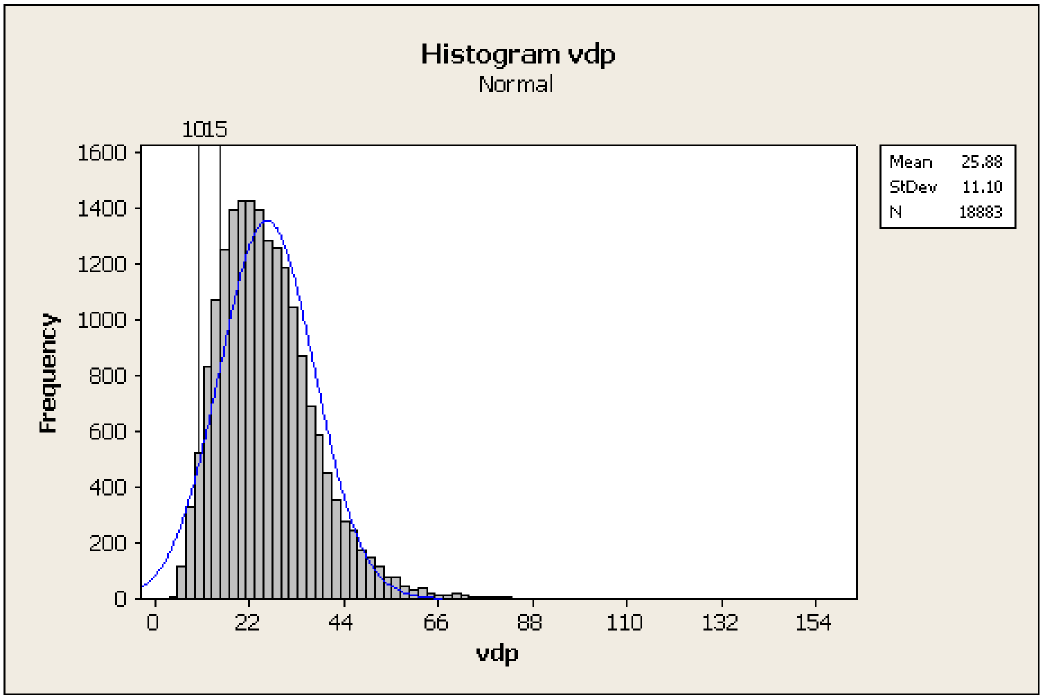


Figure 2 – VDP Frequency

Minitab 15 calculated the optimal lambda from every available vitamin D reading (VDP) in the NHANES III. The optimal lambda was determined to be .27 and each data point was transformed using the Box-Cox Transformation, figure 3. Upper specification limits, USL, and lower specification limits, LSL, correspond to the transformed values of 10 and 15ng/mL. A second histogram, figure 4, of the transformed vitamin D readings, tVDP, shows the improved normality of the data which was necessary for analysis using the t-test, and transformed values of 10 and 15ng/mL

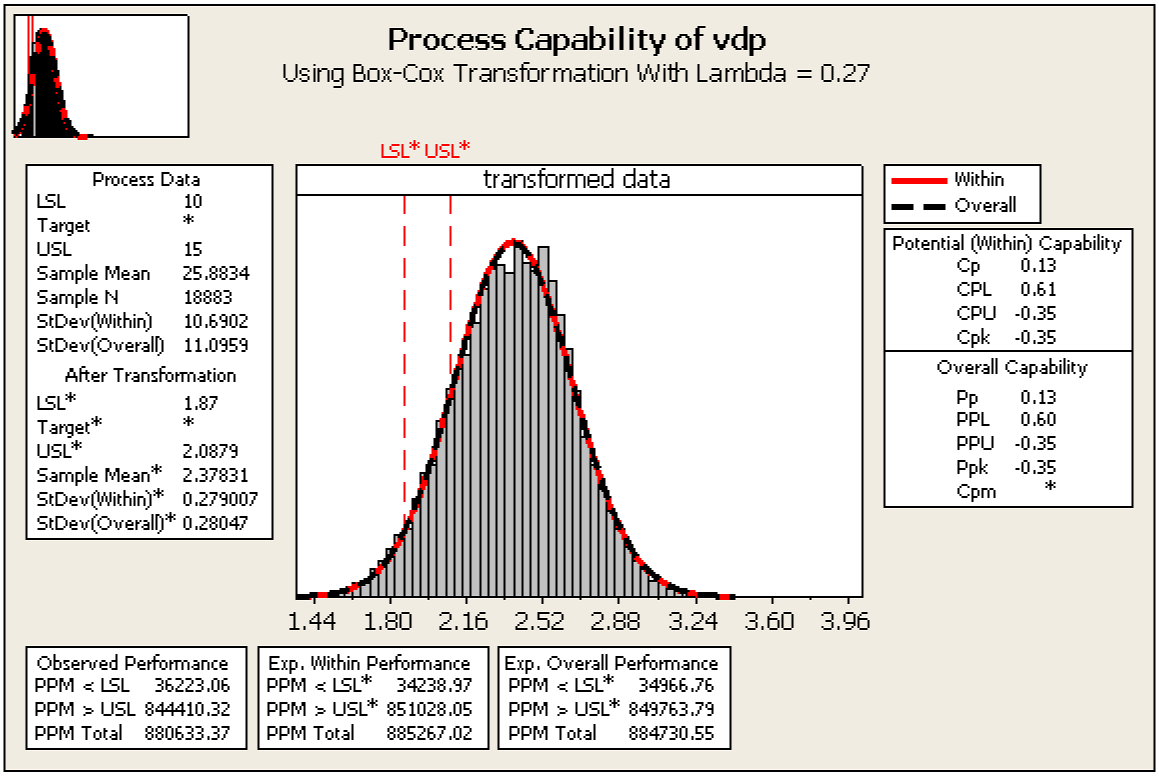


Figure 3 – Box-Cox Transformation on VDP

The cumulative distribution functions which were preformed defined having a vitamin D level of less than 15ng/mL as the cutoff for vitamin D deficiency. In the non-transformed group, using the VDP data, assuming its normality, the normal cumulative distribution function from negative infinity to less than 15 was about 16.34%. In the normalized data the normal cumulative distribution function from negative infinity to less than 2.088, the normalized equivalent of 15, was 15.05%.

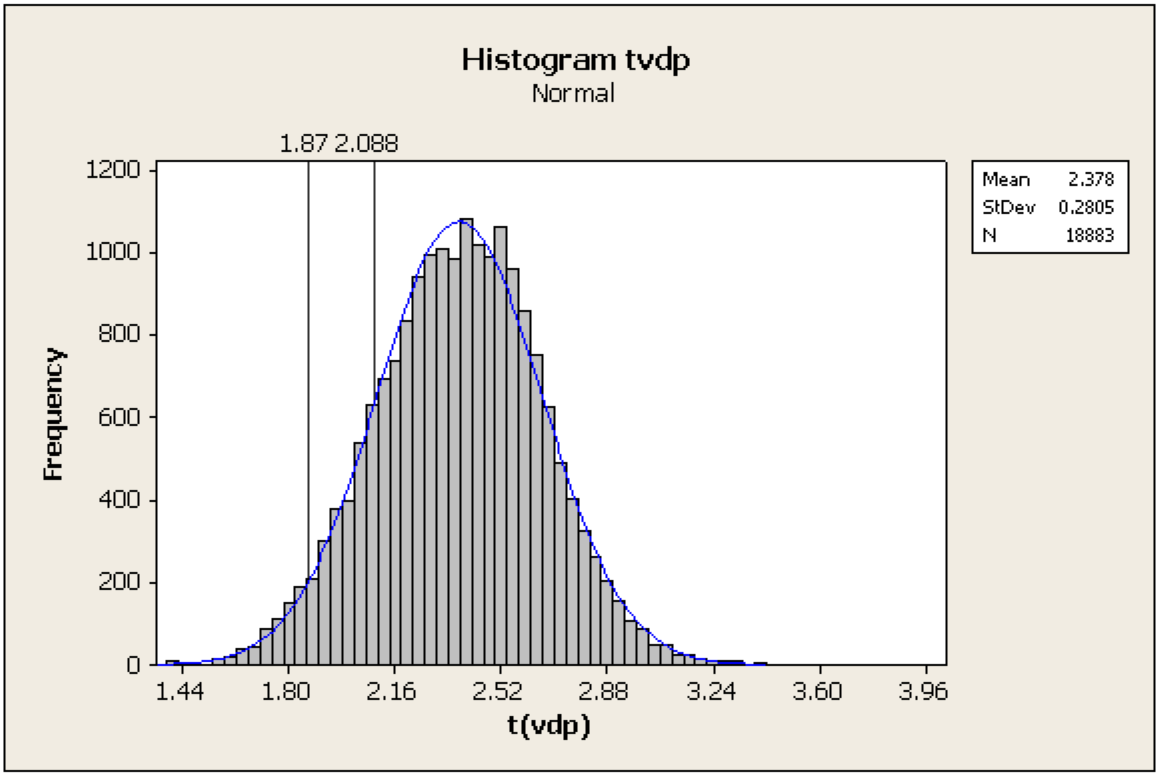
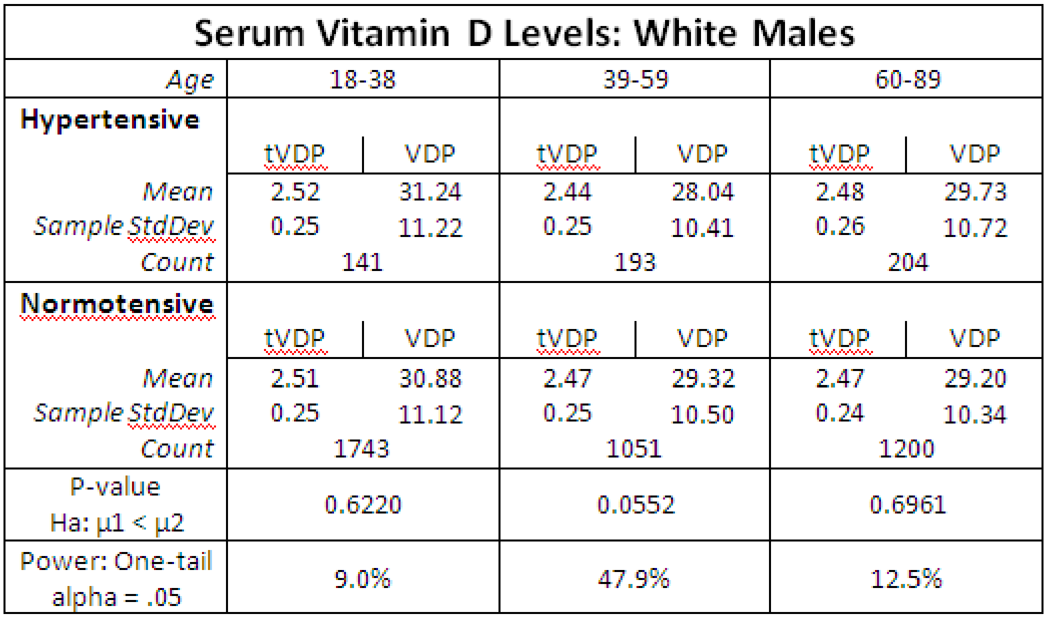


Figure 4 – tVDP Frequency

Survey Response to HAE2

The average level of serum vitamin D for diagnosed hypertensive white males, those who responded yes to the question, “Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?” was highest in the first, or youngest tertile, and lowest in the second tertile, as seen in table 1. For individuals who responded no to the survey question, having never been diagnosed with hypertension, the highest average vitamin D level was also the first tertile, however the lowest age group was the third tertile (60-89).

Only in the second tertile were average vitamin D levels lower in the hypertensive group than in the normotensive group. While it was not found to be statistically significant (P = .0552, α = .05) it was close to the cutoff.



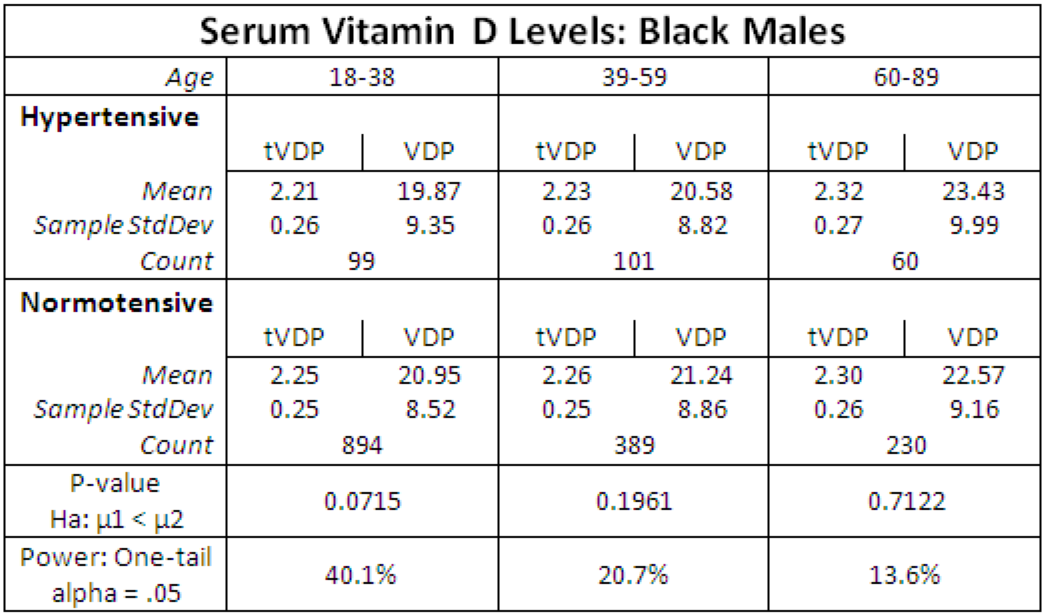
Serum Vitamin D Levels: White Males

Age	18-38		39-59		60-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.52	31.24	2.44	28.04	2.48	29.73
Sample StdDev	0.25	11.22	0.25	10.41	0.26	10.72
Count	141		193		204	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.51	30.88	2.47	29.32	2.47	29.20
Sample StdDev	0.25	11.12	0.25	10.50	0.24	10.34
Count	1743		1051		1200	
P-value Ha: $\mu1 < \mu2$	0.6220		0.0552		0.6961	
Power: One-tail alpha = .05	9.0%		47.9%		12.5%	

Table 1 – HAE2: White Males

The average level of serum vitamin D in black males was greatest in the third tertile and lowest in the first tertile for individuals who had been diagnosed as hypertensive (table 2). This was also the case for normotensive individuals.

In the first and second tertiles the average vitamin D level of hypertensive individuals were lower than those who were normotensive, however neither were statistically significant (P = .0715, P = .1961, α = .05).The results found that age was inversely associated with vitamin D levels; the oldest tertile in both the hypertensive and normotensive groups had the highest levels.

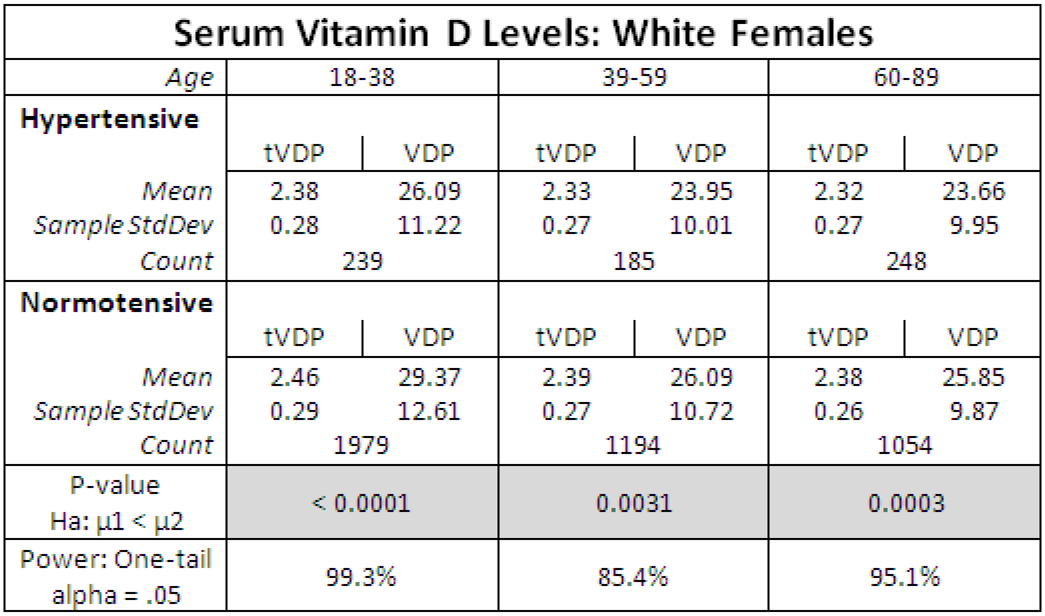


Serum Vitamin D Levels: Black Males

Age	18-38		39-59		60-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.21	19.87	2.23	20.58	2.32	23.43
Sample StdDev	0.26	9.35	0.26	8.82	0.27	9.99
Count	99		101		60	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.25	20.95	2.26	21.24	2.30	22.57
Sample StdDev	0.25	8.52	0.25	8.86	0.26	9.16
Count	894		389		230	
P-value Ha: µ1 < µ2	0.0715		0.1961		0.7122	
Power: One-tail alpha = .05	40.1%		20.7%		13.6%	

Table 2 – HAE2: Black Males

For both hypertensive and normotensive white women who were not on prescription medications for hypertension, the average vitamin D levels decreased as age increased (table 3). Additionally the average level of vitamin D was lower in hypertensive individuals of all three groups. In each of the three groups hypertensive individuals had statistically significantly lower levels of vitamin D than their normotensive counterparts (P < .0001, P = .0031, P = .0003, α = .05).

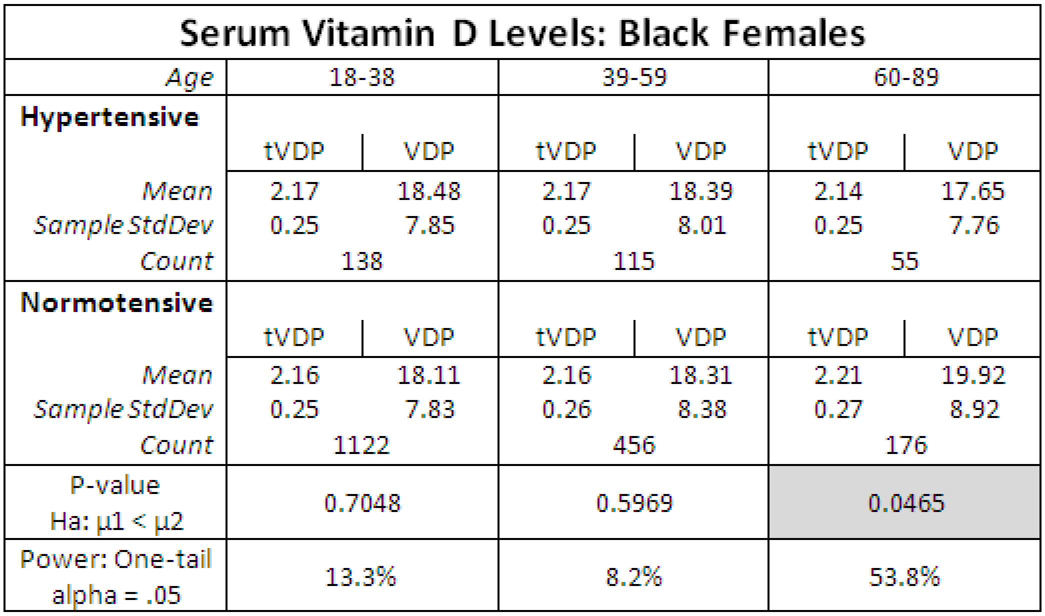


Serum Vitamin D Levels: White Females

Age	18-38		39-59		60-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.38	26.09	2.33	23.95	2.32	23.66
Sample StdDev	0.28	11.22	0.27	10.01	0.27	9.95
Count	239		185		248	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.46	29.37	2.39	26.09	2.38	25.85
Sample StdDev	0.29	12.61	0.27	10.72	0.26	9.87
Count	1979		1194		1054	
P-value Ha: µ1 < µ2	< 0.0001		0.0031		0.0003	
Power: One-tail alpha = .05	99.3%		85.4%		95.1%	

Table 3 – HAE2: White Females

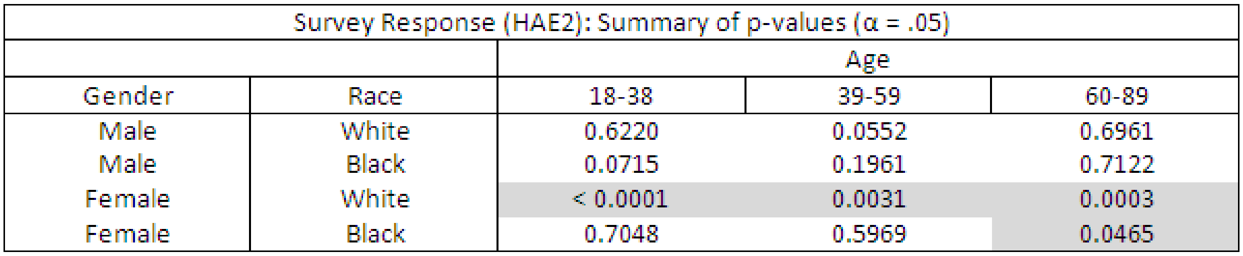
The average vitamin D level in hypertensive black females decreased as the age increased (table 4). In Normotensive individuals, however, the levels of vitamin D increased as age increased. Only in the third tertile was the average level of vitamin D lower in hypertensive individuals than in normotensive individuals. This difference was statistically significant (P = .0465, α = .05).



Serum Vitamin D Levels: Black Females						
Age	18-38		39-59		60-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.17	18.48	2.17	18.39	2.14	17.65
Sample StdDev	0.25	7.85	0.25	8.01	0.25	7.76
Count	138		115		55	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.16	18.11	2.16	18.31	2.21	19.92
Sample StdDev	0.25	7.83	0.26	8.38	0.27	8.92
Count	1122		456		176	
P-value Ha: µ1 < µ2	0.7048		0.5969		0.0465	
Power: One-tail alpha = .05	13.3%		8.2%		53.8%	

Table 4 – HAE2: Black Females

The results show that white individuals, in general, have higher levels of vitamin D than blacks, and males have higher levels than women. Of the four groups white males tended to have the highest levels of vitamin D, then white females, then black males, with black females having the lowest levels of vitamin D regardless of hypertension diagnosis. The strongest statistical evidence exists in white females of any age group (table 5).



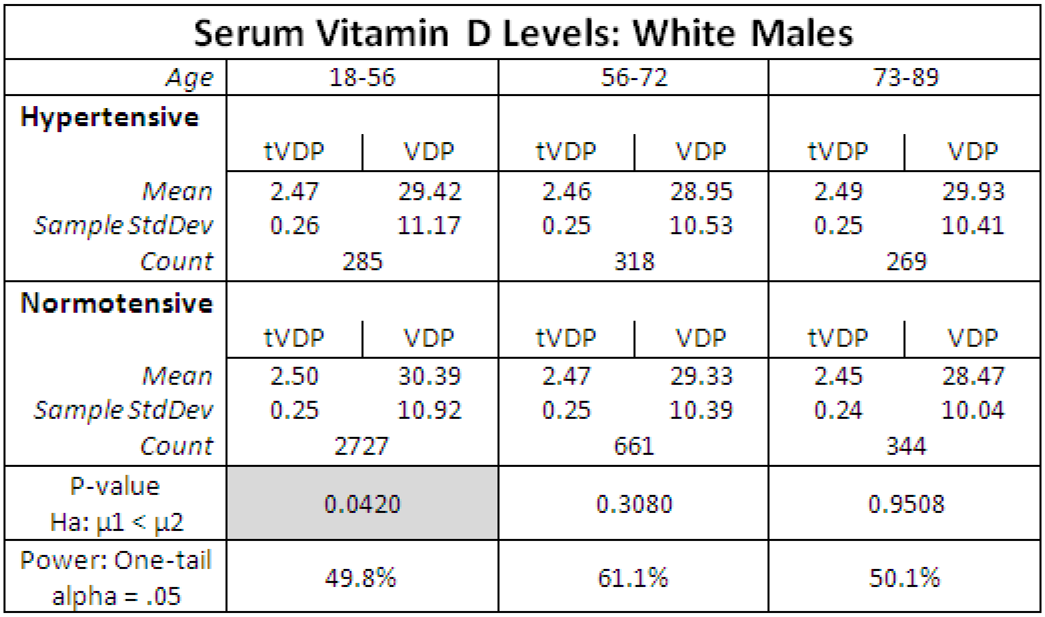
Survey Response (HAE2): Summary of p-values (α = .05)				
		Age		
Gender	Race	18-38	39-59	60-89
Male	White	0.6220	0.0552	0.6961
Male	Black	0.0715	0.1961	0.7122
Female	White	< 0.0001	0.0031	0.0003
Female	Black	0.7048	0.5969	0.0465

Table 5 – HAE2: P-value Summary

Blood Pressure Diagnosis

After separating the white males who blood pressure reflects that of a hypertensive individual, systolic blood pressure greater than or equal to 140 mm Hg or diastolic blood pressure greater than or equal to 90 mm Hg, from those who are normotensive, the average vitamin D level in hypertensive individuals was highest in the third tertile and lowest in the second tertile. The vitamin D levels of normotensive individuals decreased as age increased.

In the first and second tertile the average level of vitamin D in white males was lower in hypertensive individuals than normotensive individuals (table 6). This difference was statistically significant in the first tertile (P = .0420, α = .05). In the third tertile, however, the average level of vitamin D was high enough in hypertensive individuals to be statistically significant (Ha: µ1 > µ2, P = .04923, α = .05).



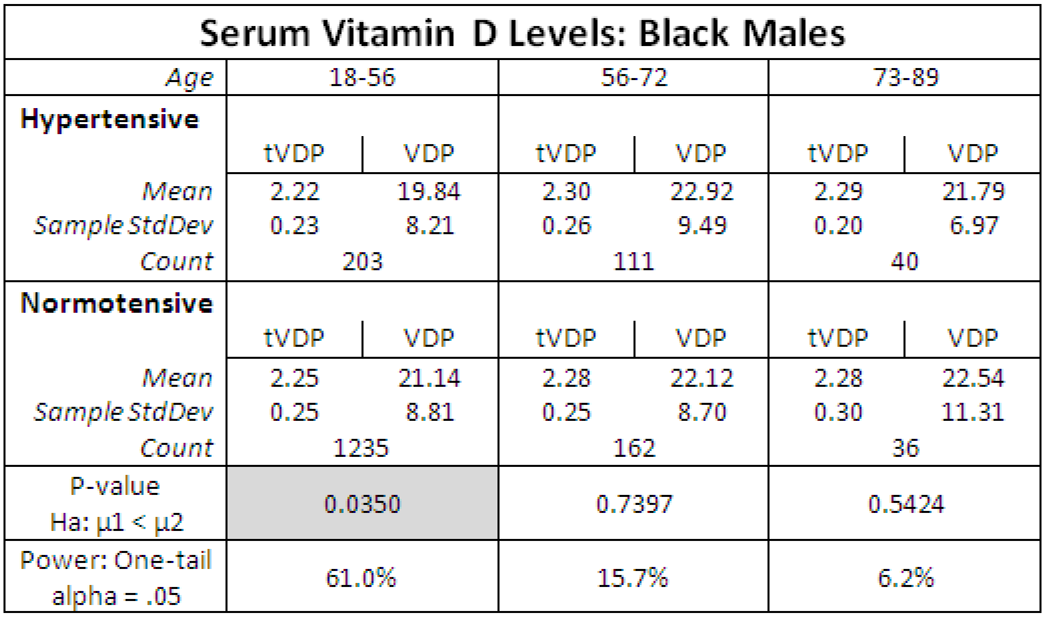
Serum Vitamin D Levels: White Males

Age	18-56		56-72		73-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.47	29.42	2.46	28.95	2.49	29.93
Sample StdDev	0.26	11.17	0.25	10.53	0.25	10.41
Count	285		318		269	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.50	30.39	2.47	29.33	2.45	28.47
Sample StdDev	0.25	10.92	0.25	10.39	0.24	10.04
Count	2727		661		344	
P-value Ha: µ1 < µ2	0.0420		0.3080		0.9508	
Power: One-tail alpha = .05	49.8%		61.1%		50.1%	

Table 6 – HBP Diagnosis: White Males

The average level of vitamin D in hypertensive black males was highest in the second tertile and lowest in the first tertile, as seen in table 7. The average level of vitamin D increased as age increased in normotensive individuals. In the first and third tertiles the average vitamin

D levels were lower in hypertensive individuals than their normotensive counterparts. This difference was only statistically significant in the first tertile (P = .0350, α = .05).

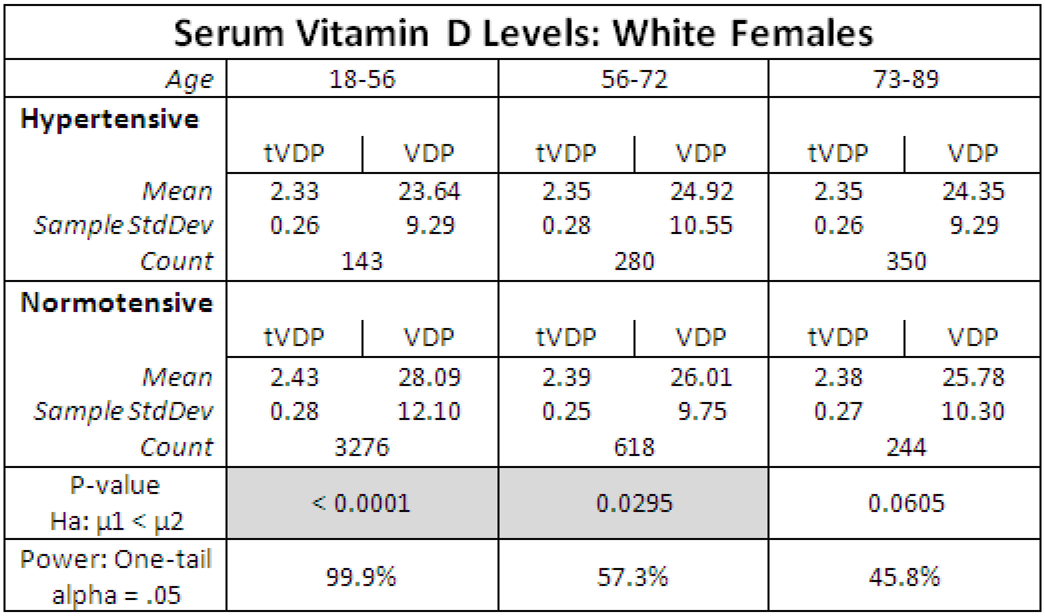


Serum Vitamin D Levels: Black Males

Age	18-56		56-72		73-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.22	19.84	2.30	22.92	2.29	21.79
Sample StdDev	0.23	8.21	0.26	9.49	0.20	6.97
Count	203		111		40	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.25	21.14	2.28	22.12	2.28	22.54
Sample StdDev	0.25	8.81	0.25	8.70	0.30	11.31
Count	1235		162		36	
P-value Ha: µ1 < µ2	0.0350		0.7397		0.5424	
Power: One-tail alpha = .05	61.0%		15.7%		6.2%	

Table 7 – HBP Diagnosis: Black Males

The level of vitamin D in hypertensive white females was highest in the second tertile and lowest in the first tertile, as seen in table 8. The average level of vitamin D in normotensive individuals decreased as age increased. In all three tertiles the average vitamin D levels were lower in hypertensive individuals than in normotensive individuals. This difference was statistically significant in the first and second tertiles (P < .0001, P = .0295, α = .05). The third quartile had a p-value of .0605.

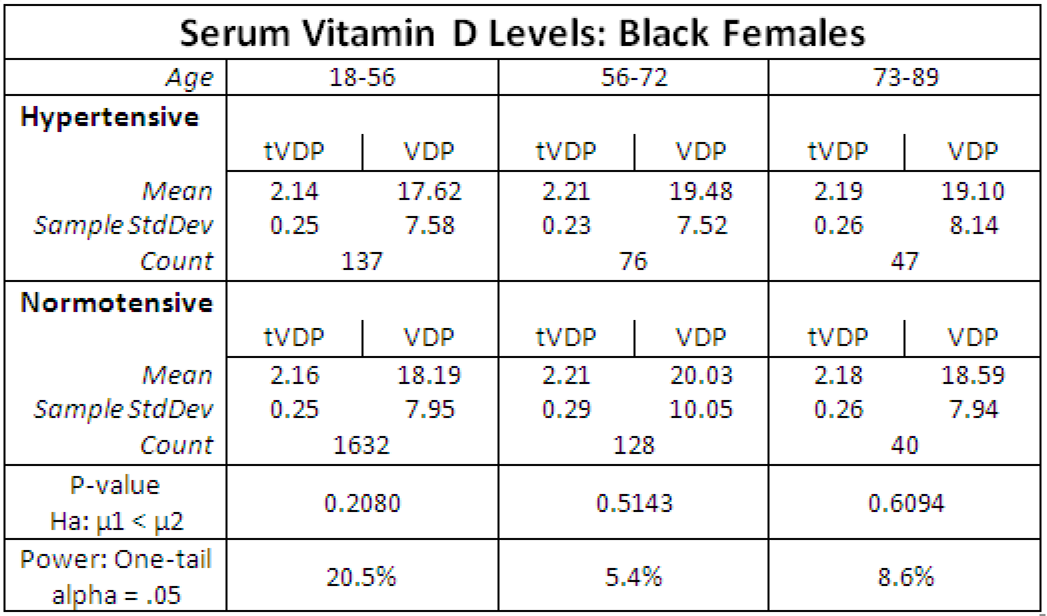


Serum Vitamin D Levels: White Females

Age	18-56		56-72		73-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.33	23.64	2.35	24.92	2.35	24.35
Sample StdDev	0.26	9.29	0.28	10.55	0.26	9.29
Count	143		280		350	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.43	28.09	2.39	26.01	2.38	25.78
Sample StdDev	0.28	12.10	0.25	9.75	0.27	10.30
Count	3276		618		244	
P-value Ha: µ1 < µ2	< 0.0001		0.0295		0.0605	
Power: One-tail alpha = .05	99.9%		57.3%		45.8%	

Table 8 – HBP Diagnosis: White Females

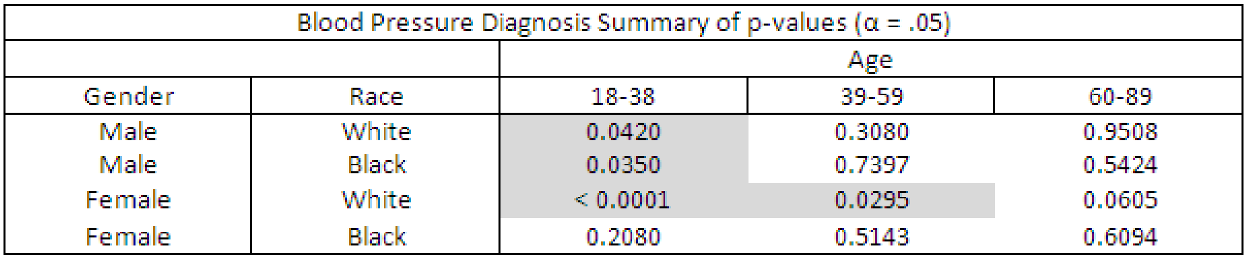
The average levels of vitamin D in both hypertensive, and normotensive black females (table 9) were lowest in the first tertile, and highest in the second tertile. The vitamin D levels of hypertensive individuals was lower than the level of normotensive individuals in the first and second tertiles, however this difference was not statistically significant (α = .05).



Serum Vitamin D Levels: Black Females						
Age	18-56		56-72		73-89	
Hypertensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.14	17.62	2.21	19.48	2.19	19.10
Sample StdDev	0.25	7.58	0.23	7.52	0.26	8.14
Count	137		76		47	
Normotensive	tVDP	VDP	tVDP	VDP	tVDP	VDP
Mean	2.16	18.19	2.21	20.03	2.18	18.59
Sample StdDev	0.25	7.95	0.29	10.05	0.26	7.94
Count	1632		128		40	
P-value Ha: μ1 < μ2	0.2080		0.5143		0.6094	
Power: One-tail alpha = .05	20.5%		5.4%		8.6%	

Table 9 – HBP Diagnosis: Black Females

The results, when separated using blood pressure measurements, show similar trends to that using the responses to question HAE2. In general whites tend to have higher levels of serum vitamin D than blacks, and men have higher levels than women. Again white males have the highest levels of vitamin D, then white females, then black males. Black females tended to have the lowest levels of vitamin D regardless of hypertensive diagnosis. The strongest statistical evidence that hypertensive individuals have lower levels of vitamin D than their normotensive counterparts exists in younger individuals, especially white females (table 10).



Blood Pressure Diagnosis Summary of p-values (α = .05)				
		Age		
Gender	Race	18-38	39-59	60-89
Male	White	0.0420	0.3080	0.9508
Male	Black	0.0350	0.7397	0.5424
Female	White	< 0.0001	0.0295	0.0605
Female	Black	0.2080	0.5143	0.6094

Table 10 – HBP Diagnosis: P-value Summary

Discussion

NHANES III Database

Features

By taking the NHANES III data and creating a database the main advantage is that one is no longer limited by the proprietary format and license costs of the Statistical Analysis Software. The software used to parse the SAS and DAT-files is written in Java. Java is available at no cost and is a common teaching language in most college based computer programming courses. This increases the ability of any future individuals, with knowledge of object-oriented programming in Java, who wishes to alter or expand the code base to do so. Additionally the JDBC connector, written in Java, which connects the Java programs to the MySQL database, is also free.

The MySQL software used as the database management system is also free and provides a wide array of built in tools for querying the database. Additionally once the database has been created the ability of the JDBC to connect Java code directly to the database allows for more complex query designs or statistical analysis to be preformed which are not natively supported by the set of MySQL functions. MySQL also provides user access control to reading from or writing to the database. Specialized tables or views containing combinations of the variables from one or more tables can be created by administrators and then can be accessed by users with read-only permissions preventing unwanted changes to the database once it has been constructed.

Potential Uses

With the vast quantity of information included in the NHANES III and the vast topics it covers the potential uses of the database are limited only by the user's knowledge of the questions and information that the survey contains. The information contained in the NHANES III has already been used in many retrospective studies, including studies on coronary heart diseases (Alexander, et al. 2003), obesity (Flegal, et al. 1998), and cavities (Vargas, Crall and Schneider 1998) as well as many others. With the broad goal of determining the health and wellness of individuals thought the United States the survey contains information on the general physical and mental medical history, eating and exercise habits, tobacco and alcohol use, as well as laboratory data. The large number of participants in the dataset allows for queries to be very specific and still yield sizeable result sets containing individuals fitting the query parameters. This data is able to provide baseline information on individuals between 1988 and 1994 using any one of the available variables.

Advantages of Design

The major advantage of the NHANES III database is that it provides all the information freely available in the NHANES III using tools that are also freely available. These tools are well known and easy to pick up and use for those with some object-orientated programming and SQL experience.

While the program was designed for MySQL specifically, it was also ported for use with Microsoft's SQL Server 2005. This required few changes to the program itself and most of the changes were preformed on the SQL statements contained in the Java code to perform inserts, updates, create statements etc. Anyone with experience using a particular database management system should be able to alter these SQL statements to the syntax for any number of database management systems.

The design of the database collects the information from each dataset of the survey in a way that is as similar to the manner in which it is provided by the CDC, and would be familiar to those who are familiar to the SAS version of the data. Those datasets that are limited by the number of columns allowed in a MySQL database are split into the fewest number of tables possible. Split tables conserve the order of the rows and provide id numbers to link these tables together if variables of interest are split between multiple tables from the same dataset.

Disadvantages of Design

While the amount of documentation that is provided by the CDC is very thorough the information is provided mostly in PDF-files. These files are not included in the database transfer and instead each variable is given a short text description in the label field of the DataDictionary. The PDF-file contains information on how and when during the survey a particular question was asked or laboratory procedure preformed. While the label text descriptions in the DataDictionary provide a way to quickly query the database for variables of interest using keywords this was not their intended use. As such the wording of the labels is inconsistent. For example an interested user wishing to find variable containing information on blood pressure would find variables containing the words blood pressure, but might miss those which use the word hypertension instead. Additionally one would miss variables whose labels include BP as the abbreviation of blood pressure if they were not aware that it was used, or not find HBP depending on how the query was preformed.

Even more difficult to overcome are survey questions that pertain to blood pressure but do not contain any form of that term in their label. As many of the questions came in the form of a written survey, there was a progression of questions that was fallowed depending on eligibility. Questions toward the end of this progression were also shortened as the reader would be familiar as to what the question was referring to and to prevent the questions from getting tedious. For example the final question in the blood pressure series of questions is in reference to if the individual is currently using a prescription drug to treat their hypertension (HAE5A). In the adult data file the label of HAE5A reads, “Now taking prescribed medicine for HBP” while the label for the variable of the same name in the core data file reads “Now taking prescr med for HBP.” The same question from the PDF documentation reads, “Are you now taking prescription medication?” with the survey taker knowing what the question was referring to from previous the questions. This inconsistency makes it difficult to be confident that one has found all variables pertinent to your area of interest using only a series of queries to the DataDictionary label field. This then requires the user to have access to the PDF documentation of each of the files and search each of the files, which can be several hundred pages long, for information of interest.

Also the conversion from a survey answer to a database entry requires knowledge of how blank answers were recorded in the survey section and how they are transferred to the database. In the survey a series of eights equal to the length of that variables width was recorded if an individual was eligible to receive a question but refused, lacked the time or staff, the data was lost, such as if a vial was broken, or if the language barrier prevented its completion or the answer was unreadable. These 8-fills are recorded as such in the database.

While this means that an answer is not available for that cell it differs from a blank response in the survey which is recorded as NULL in the database, which represents that an individual did not answer a question as they were not eligible to. Since MySQL automatically ignores null values in the calculation of rows a query for the average, or other math function, of a particular column could yield an incorrectly high value if MySQL is not explicitly told to omit the rows with 8-fills.

In addition to the use of 8-fills as representing special cases where there is no answer available six, seven, or nine-fills may be present and must be identified on a variable to variable basis. The meaning of these may vary from question to question as well. While these numerical fills increase the specificity of questions that can be asked of the data, it is important to have access to the PDF documentation for that particular variable, and have adequate knowledge of the survey design and format, as well as what numbers to expect from queries, before using the database.

Limitations

Currently the software used to create the database from the data sets has only been tested with files from the NHANES III. Differences between the NHANES III and past surveys may prevent the software from being compatible. Files released after NHANES III, including the continuous NHANES, are in the XPT-file format which requires conversion into CSV-files using SAS, and does not contain all of the information metadata for the DataDictionary table. While no formal testing has been done with either the NHANES or the NHANES II, the data sets of both surveys include the necessary SAS-files as well as an EXE-file used to extract DAT-files in formats similar to those used in the NHANES III. While the files appear to be similar, and any changes necessary to the programs, would appear to be minimal, the usefulness of the information from 1971-1975 may be limited, not only in its age but as to the questions the survey would ask individuals from the early 70's as opposed to questions asked in a more current survey. Additionally the data files in the NHANES III that were released in XPT-file format require the information be converted into CSV-files. This is not possible without an SAS license.

Limitations within MySQL also require that tables be no more than 1000 columns wide. As good practice would dictate, tables of this size are uncommon and generally considered poorly designed. As such, it would be unlikely that this limit be expanded much, if at all, to accommodate more than twice the current limit that would be necessary to fit particular datasets into a single table. While the software was programmed using MySQL statements other database management systems have similar hard limits, Microsoft's SQL Server 2008, for example, limits the total number of columns to 1024.

Potential Improvements

Currently for each data set that is entered into the database three connections are made to the database, while these connections are closed automatically when the process is complete these represent an opportunity for optimization. While this is a great improvement over early versions of the programs, which needlessly connected hundreds of times, this process of connecting is not necessarily noticeable when inserting one file but iterating though each file represents a waste of system resources and time. Many of the classes could be altered or rewritten to accept a MySQL database connection as an additional parameter. This would allow all classes and methods to share a single connection, improving the runtime of inserting the 50 or more datasets drastically, where connecting to the database takes three times as long as it does to perform an insert or query.

For each data set that contains over 1000 variables multiple tables must be created. The first columns of these tables contains the variables, sas_file_id, table_id, and then as many variables that would fit given the limit. The first of these variables is always the SEQN which is used as a unique five digit identifier. The second table then contains sas_file_id, table_id and the columns that would not fit in the first table. After performing the queries and using the table_id as a method to query data located in two or more split data sets it became apparent that it would be more intuitive if the start of each of the tables was instead the sas_file_id, the reference to the file id number in the SASFiles id, the SEQN, and then all the remaining variables, eliminating the table_id, which serves the same purpose as the SEQN. This would allow linkage between all tables using the SEQN and would not require the use of the table_id columns in addition to the SEQN when querying from multiple tables where one or more is not the first table from a split data set.

Analysis

Location Analysis

One variable that was not included in the queries was location. While latitude and season can greatly affect an individual's vitamin D level, and the NHANES III took place over the course of several years, the existing data from previous studies on the vitamin D information contained in the NHANES III suggested that the sample was, intentionally or unintentionally, distributed in such a way as to eliminate some of this bias. While the NHANES III contains variables used to determine the location where the interview and examination were taken these are removed to protect the identity of the participants in the study. In the publicly released version of the data the individuals location was only recorded when the individual's location was in a high population area where it would be more difficult to identify an individual based on a combination of location and other available traits. The use of individuals who fall into these groups would have resulted in sample sizes too small to be useful. Season also plays a role in the exposure and intensity of sunlight. Seasonality variables are only available though the NCHS Research Data Center and are not released in the publicly available online data sets.

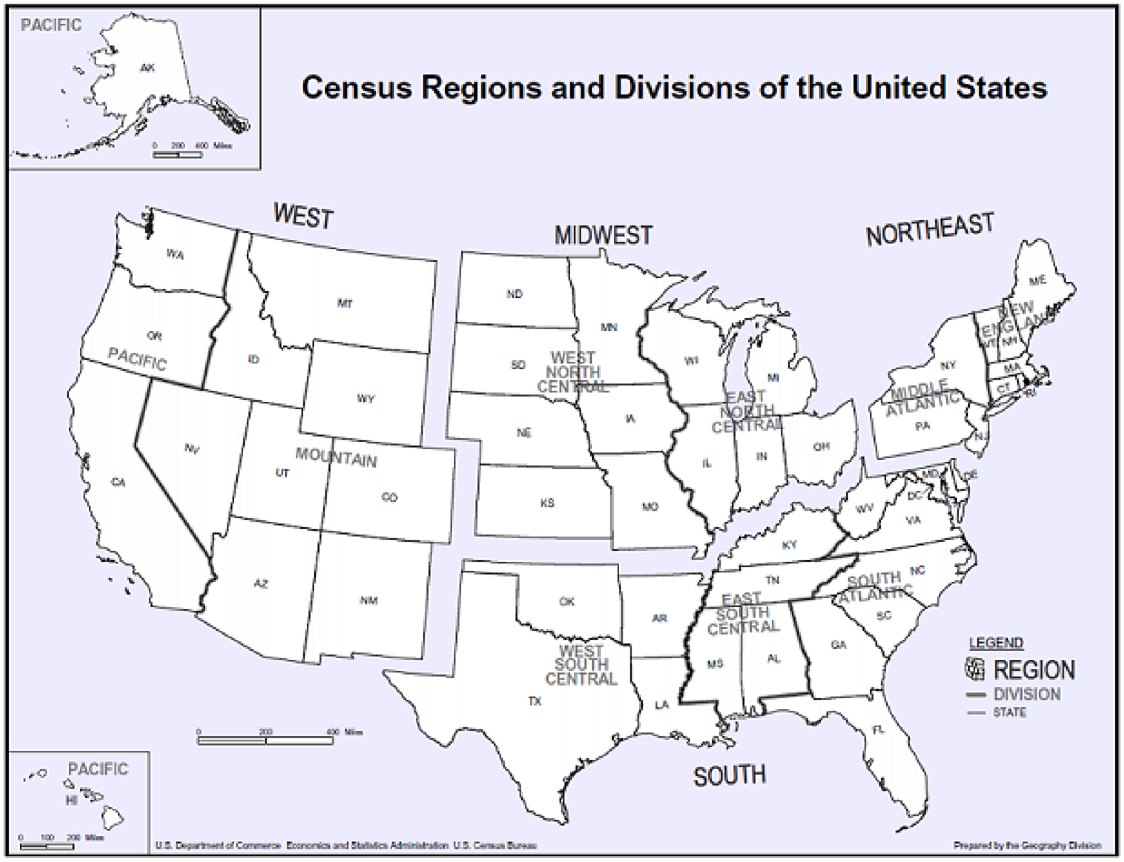


Figure 5 – Census Regions and Divisions of the United States

While the exact locations are not available in the public releases region codes are available for each individual in the survey. These region codes split the United States into four major regions, Northeast, Midwest, South, and West with several smaller sub regions, as seen in figure 5 (Division 2010). These regions, variable DMPCREGN, are number coded and provide for some localization to be preformed. However if north and south subpopulations were to be defined by this variable, individuals from the Northeast and Midwest would be defined as being from the north, and only individuals from the south region would be placed in the South group. Since the West census region spans a wide range of latitudes it would be incorrect to label it as either north or south and it would have to be omitted from the study, effectively eliminating more than 25% of peoples from the survey.

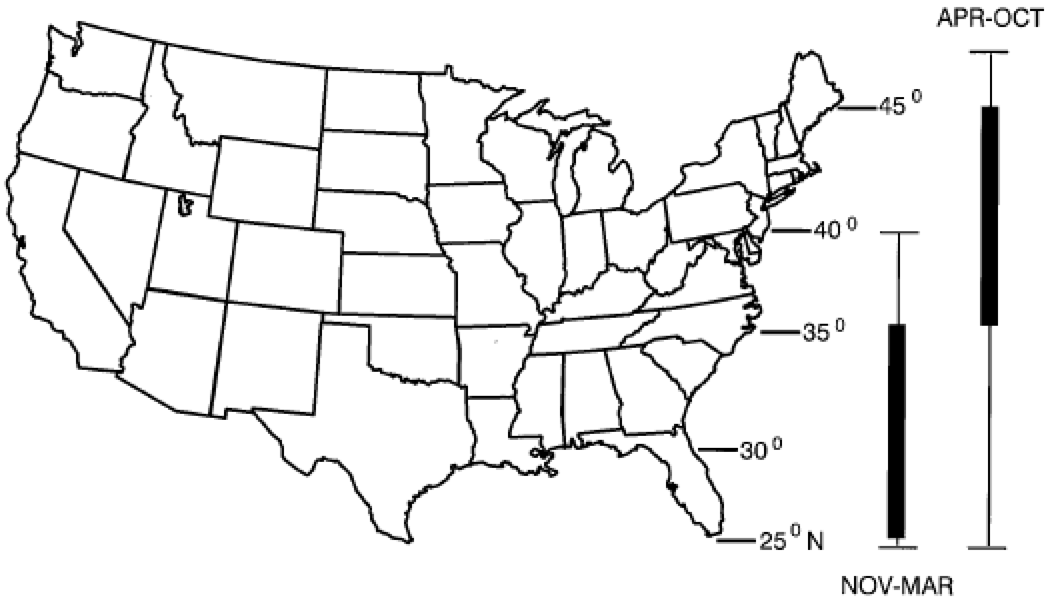


Figure 6 – Two Seasonal Subpopulations of NHANES III

A study of the two seasonal subpopulations present in the NHANES III and their distribution was previously preformed. In this study two subpopulations of the study were identified based on when vitamin D information was collected. The first of these seasonal subgroups was defined as all samples taken between November and March, what would be considered the winter months in the northern hemisphere. It was found that while samples from this subgroup were taken as far north as 40° N latitude, the top line of the NOV-MAR bar in figure 6, and as far south as 25° N, the bottom line on the bar. More than 75% of the samples in this seasonal subgroup were taken from the southern latitudes, south of 35° N latitude. The thick opaque bars indicate where greater than 75% of the data was collected during the indicated months. The second seasonal subgroup, consisting of samples taken between April and October, the summer months in the northern hemisphere, had samples taken thought the entire range of latitudes in the United States. Over 75% of the samples taken during these summer months were from the northern latitudes, north of 35° N latitude. These two subpopulations, the Winter-South and Summer-North groups were then compared to one another. In both subpopulations it was found that the prevalence of vitamin D deficiency, defined in the study as levels less than 17.5nmol/L, was less than one percent. Based on these findings it was decided that instead of separating the available data based on regions, resulting in much smaller sample sizes, that grouping them together would yield larger samples and still be reliable without introducing additional variability in the data (Looker, et al. 2002).

Cause and Effect Relationship

While the results lend additional support to existing studies between the link of vitamin D and hypertension, these existing studies are often primarily focused on vitamin D in conjunction with calcium supplements. While the effectiveness of calcium supplementation is increased with the addition of a vitamin D supplement (Pfeifer, et al. 2001), there has not been as many studies that have attempted to determine how effective vitamin D supplementation on its own would be in the treatment of hypertension. While this study contained two groups, one treated with only calcium and a second with calcium and vitamin D it did not contain a group being treated with only vitamin D. The results from such an experiment would be useful in further understanding the role of diet in hypertension.

Once able to determine the effectiveness of a treatment of only vitamin D one could assess what deficiencies in diet are more likely to be an underlying cause of hypertension. If a treatment of only vitamin D is equally, or more, as effective as that of a calcium only treatment, the dietary deficiency which underlies, in part, the cause of hypertension is not the total daily intake of calcium but rather the daily intake of vitamin D. The low levels of vitamin D would prevent the adequate amount of calcium to be effectively absorbed. Thus the low levels of calcium thought to play a role in hypertension may in fact be adequate for a healthy diet. This adequate level, however, is not being properly absorbed due to a vitamin D deficiency. The supplementation of vitamin D, would then be treating the true cause of the hypertension, low levels of vitamin D, instead of an additional symptom, that of poor calcium absorption.

There is also a probable correlation between exercise and sun exposure, a primary source of vitamin D. Exercise is known to decrease the risk for heart disease. An added benefit of outdoor activity would be increased vitamin D production from the sun exposure which may also contribute to the reduced risk for cardiovascular disorders. Additionally individuals who are frequently sick, and/or bedridden, are indoors much more frequently and away from the sunlight adding to their risk.

Bonferroni Corrections

The NHANES III includes sampling bias in both the proportion of individuals based on race and based on age. This prevents the direct comparison of the counts of individuals between groups. The correction factors are available in the documentation by the CDC.

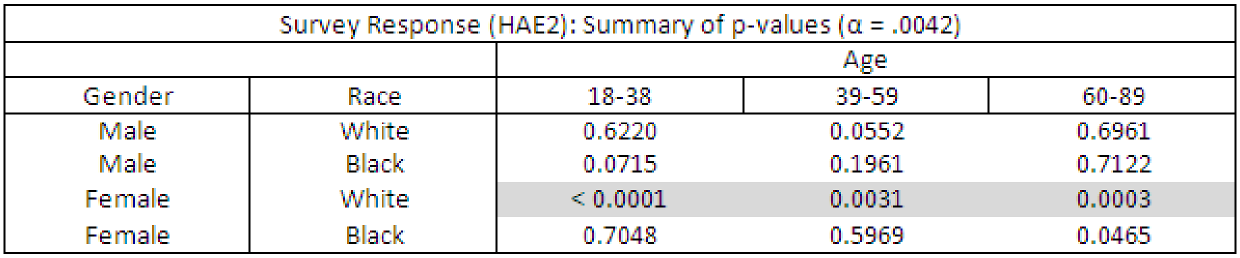
It is believed by some that a correction factor must be introduced when many statistical tests are preformed during a particular analysis. Others argue that the application of such statistical corrections are, often times, incorrect. Arguments presented against their use cite the correction corrects the wrong problem, decreases Type I errors by increasing Type II errors, or simply defies common sense (Perneger 1998). While Bonferroni corrected statistics are included, and multiple tests performed on stratified analysis are often corrected in some method, all statistics should be interpreted keeping in mind *a priori* knowledge as to what results one expects based on ones previous experience.

In a Bonferroni correction, a conservative form of correction, the alpha value, in this case .05, is divided by the total number of tests performed. The null hypothesis is then rejected or accepted based on the comparison of this alpha value rather than .05. For each of the two methods of query 12 statistical tests were performed resulting in a p-value of .05/12 ≈ .0042. All statistical tests are preformed in the manner in which was described in the materials and methods. The rejection of H_o only occurs when their p-values are less than .0042.

HAE 2

Using the uncorrected alpha resulted in few significant results; however those that were significant were extremely small. So much so that even using the correction method three groups were deemed statistically significant. Using the more restrictive alpha still results in three groups, white females of any age tertile, where individuals with hypertension have statistically significantly lower levels of vitamin D than their normotensive counterparts.

Due to the wording of the survey question, “Have you ever been told by a doctor or other health care professional that you had hypertension, also called high blood pressure?” it is impossible to know when the person was diagnosed with hypertension and whether or not they were treated or if they are currently diagnosed. Older individuals are more likely to have been told at one point that had hypertension but may have since been treated. Results from younger individuals should be more heavily weighted as it is more likely that they have been diagnosed recently, in the past 20 years, rather than the second or third tertile as they could have been diagnosed up to 41 to 71 years prior to the survey when their blood samples were taken for vitamin D analysis.



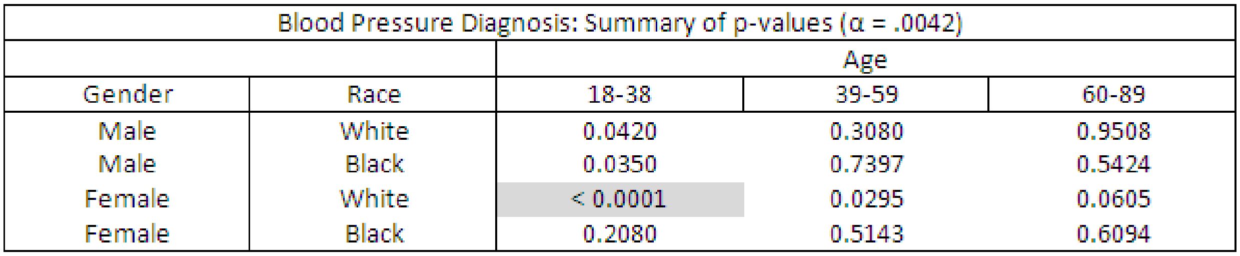
Survey Response (HAE2): Summary of p-values (α = .0042)				
		Age		
Gender	Race	18-38	39-59	60-89
Male	White	0.6220	0.0552	0.6961
Male	Black	0.0715	0.1961	0.7122
Female	White	< 0.0001	0.0031	0.0003
Female	Black	0.7048	0.5969	0.0465

Table 11 – HAE2: P-value Summary w/Bonferroni Correction

Diagnosis

While the statistical tests preformed based on a blood pressure diagnosis without correction for multiple testing yielded the same number of groups that differed statistically significantly the p-values of the values were not as small. Thus, when using the correction, only one group was found to be statistically significant. This group, white females ages 18-38, was found in to have significant levels in both testing methods. Additionally the case where white males in the third tertile with hypertension had statistically significantly higher levels of vitamin D than their normotensive counterparts was not found to be statistically significant using the correction.

This form of analysis provides a better snapshot in time than the analysis based on the survey question. It eliminates sampling bias of individuals who forgot or responded incorrectly. It also gives a hypertension diagnosis that is closer to the time at which the blood samples were drawn, and up to three of the six blood pressure readings were taken the same day at the same location. The averaging of up to six blood pressure readings attempts to limit the effect of incorrect or artificially high or low readings due to nervousness or other environmental variables.



Blood Pressure Diagnosis: Summary of p-values (α = .0042)				
		Age		
Gender	Race	18-38	39-59	60-89
Male	White	0.0420	0.3080	0.9508
Male	Black	0.0350	0.7397	0.5424
Female	White	< 0.0001	0.0295	0.0605
Female	Black	0.2080	0.5143	0.6094

Table 12 – HBP Diagnosis: P-value Summary w/Bonferroni Correction

Conclusion

The NHANES III database provides access to the wide range of information contained in the NHANES III without the need for an SAS licenses. The use of MySQL and Java, two well known and free standards, increases the potential user base and provides all necessary tools to query any combination of the many variables contained in the database.

The use of this database in a retrospective study on the correlation between serum vitamin D, specifically 25-hydroxyvitamin D (25(OH)D), found that several groups of hypertensive individuals, based on gender, age, and race, had statistically significantly lower levels of vitamin D than their normotensive counterparts. These differences were especially prevalent in younger white women.

References

Alexander, Charles M., Pamela B. Landsman, Steven M. Teutsch, and Steven M. Haffner. "NCEP-Defined Metabolic Syndrome, Diabetes, and Prevalence of Coronary Heart Disease Among NHANES III Participants Age 50 Years and Older." *Diabetes*, 2003: 1210-1214.

Campbell, Neil A., and Jane B. Reece. *Biology 6th Edition.* Glenview, Il: Benjamin Cummings, 2001.

Chobanian, Aram V., et al. "Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure." *Hypertension*, 2003: 1206-12552.

Cooper, Geoffrey M., and Robert E. Hausman. *The Cell: A Molecular Approach 4th Edition.* Sunderland Massachusetts: Sinauer Associates Inc., 2007.

Corder, Elizabeth H., et al. "Vitamin D and Prostate Cancer: A Prediagnostic Study with Stored Sera." *Cancer Epidemiology, Biomarkers & Prevention*, 1993: 467-472.

Cranney, Ann, et al. *Effectiveness and Safety of Vitamin D in Relation to Bome Health.* Rockville, MD: Agency for Healthcare Research and Quality, 2007.

Division, Geography. "Census Regions and Divisions of the United States." *U.S. Census Bureau.* February 25, 2010. http://www.census.gov/geo/www/us_regdiv.pdf (accessed April 7, 2010).

Fiscella, Kevin, and Peter Franks. "Vitamin D, Race, and Cardiovascular Mortality: Findings From a National US Sample." *Annals of Family Medicine*, 2010: 11-18.

Flegal, KM, MD Carrol, RJ Kuczmarski, and CL Johnson. "Overweight and obesity in the United States: prevalance and trends, 1960-1994." *International Journal of Obesity and Related Metabolic Disorders*, 1998: 39-47.

Forman, John P., Gary C. Curhan, and Eric N. Taylor. "Plasma 25-Hydroxyvitamin D Levels and Risk of Incident Hypertension Among Young Women." *Hypertension*, 2008: 828-832.

Gallagher, J.C., B. Lawrence Riggs, John Eisman, Alan Hamstra, Sara B. Arnaud, and Hector F. DeLuca. "Intestinal Calcium Absorption and Serum Vitamin D Metabolites in Normal Subjects and Osteoporotic Patients." *The American Society for Clinical Investigation*, 1979: 729-736.

Gartner, Lawrence M., and Frank R. Greer. "Prevention of Rickets and Vitamin D Deficiency: New Guidelines for Vitamin D Intake." *Pediatrics* (American Academy of Pediatrics), no. 111 (2003): 908-910.

Gunter, Elaine W., Brenda G. Lewis, and Sharon, M. Koncikowski. "The Third National Health and Nutrition Examination Survey (NHANES III, 1988-94) Reference Manuals and Reports." *National Center for Health Statistics.* 1996. http://www.cdc.gov/nchs/data/nhanes/nhanes3/cdrom/nchs/manuals/labman.pdf (accessed April 9, 2010).

Lappe, Joan M, Dianne Travers-Gustafson, K Michael Davies, Robert R Recker, and Robert P Heaney. "Vitamin D and Calcium Supplementation Reduces Cancer Risk: Results of a Randomized Trial." *The American Journal of Clinical Nutrition*, 2007: 1586-1592.

Lips, Paul: Duong, Tu, Anna Oleksik, Dennis Black, Steven Cummings, David Cox, and Thomas Nickelsen. "A Global Study of Vitamin D Status and Parathyroid Function in Postmenopausal Women with Osteoporosis: Baseline Data from the Multiple Outcomes of Raloxifene Evaluation Clinical Trial." *The Journal of Clinical Endocrinology & Metabolism* 86, no. 3 (2001): 1212-1221.

Looker, A.C., B. Dawson-Huges, M.S. Calvo, Gunter E.W., and N.R. Sahyoun. "Serum 25-Hydroxyvitamin D Status of Adolescents and Adults in Two Seasonal Subpopulations from NHANES III." *Bone*, 2002: 771-777.

McCarron, David A., and Molly E. Reusser. "Finding Consensus in the Dietary Calcium-Blood Pressure Debate." *Journal of the American College of Nutrition*, 1999: 398S-405S.

National Center for Health Statistics. "Analytical Notes for NHANES 2000-2006 and NHANES III (1988-1994)." *National Center for Health Statistics.* 2009. http://www.cdc.gov/nchs/data/nhanes/nhanes3/VitaminD_analyticnote.pdf (accessed April 9, 2010).

Oomen, Claudia M., et al. "Fish Consumption and Coronary Heart Disease Mortality in Finland, Italy, and the Netherlands." *American Journal of Epidemiology*, 2000: 999-1006.

Perneger, Thomas V. "What's wrong with Bonferroni adjustments." *British Medical Journal*, 1998: 1236-1238.

Pfeifer, Michael, Bettina Begerow, Helmut W. Minne, Detlef Nachtigall, and Corinna Hansen. "Effects of a Short-Term Vitamin D and Calcium Supplementation on Blood Pressure and Parathyroid Hormone Levels in Elderly Women." *The Journal of Clinical Endocrinology and Metabolism*, 2001: 1633-1637.

Services, U.S Department of Health and Human. "NHANES III Examination Data File Documentation." *Center for Disease Controll and Prevention.* December 1996. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHANES/NHANESIII/1A/exam-acc.pdf (accessed March 16, 2010).

—. "NHANES III Laboratory Data File Documentation." *National Center of Health Statistics.* Center for Diesease Controll and Provention. September 2006. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHANES/NHANESIII/1A/lab-acc.pdf (accessed March 16, 2010).

Services, U.S. Department of Health and Human. "National Health and Nutrition Examination Survey, 2007-2008 Overview." *Centers for Disease Control and Prevention.* 1 2007. http://www.cdc.gov/nchs/data/nhanes/nhanes_07_08/overviewbrochure_0708.pdf (accessed March 16, 2010).

Shils, Maurice E, Moshe Shike, A. Catherine Ross, Benjamin Caballero, and Robert J. Cousins. *Modern Nutritionin Health and Disease.* Philadelphia: Lippincott Williams & Wilkins, 2005.

Solomons, T.W. Graham, and Craig B Fryhle. *Organic Chemistry 8th Edition.* Hoboken, NJ: John Wiley & Sons, 2004.

Supplements, Office of Dietary. *Dietary Supplement Fact Sheet: Vitamin D.* November 13, 2009. http://dietary-supplements.info.nih.gov/factsheets/vitamind.asp (accessed March 19, 2010).

Taylor, Jack A., Ari Hirvonen, Mary Watson, Gary Pittman, James L. Mohler, and Douglas A. Bell. "Association of Prostate Cancer with Vitamin D Receptor Gene Polymorphism." *Cancer Research*, 1996: 4108-4110.

Vargas, Clemencia M., James J. Crall, and Donald A. Schneider. "Sociodemographic Distibution of Pediatric Dental Caries: NHANES III 1988-1994." *Journal of the American Dental Association*, 1998: 1229-1238.

Vik, Torstein, Kenneth Try, Dag S. Thelle, and Olav H. Forde. "Tromso Heart Study: Vitamin D Metabolism and Myocardial Incarction." *British Medical Journal*, 1979: 176.

Voet, Donald, Judith G Voet, and Charlotte W. Pratt. *Fundamentals of Biochemistry: Life at the Molecular Level 2nd Edition.* Hoboken, NJ: John Wiley & Sons, 2006.

Wang, Thomas J., et al. "Vitamin D Deficienfy and Risk of Cardiovascular Disease." *Circulation*, 2008: 503-511.