2010

NHANES VARIABLE NORMALIZATION:

THE IHANES DATA SET

Shih-Fan Lin, DrPH.
Audrey Beck Ph.D.
Aimee Bower, B.S.
Brian K. Finch, Ph.D.

Center for Health Equity Research and Policy

SAN DIEGO STATE UNIVERSITY

CHERP

TABLE OF CONTENTS

Section Page

I. Project Description 2

II. IHANES Variable Selection 3

III. NHANES Database Retrieval and Data Extraction 4

IV. Variable Normalization 7

A. Normalization Processes 7
B. Creation of Age, Period, and Cohort Variables 9
C. Division of Labor for Variable Normalization 11

V. Merging of Data Files 13

VI. Data Availability and Dissemination 13

I. Project Description

The primary goal of our NIMHD funded study, *A Social Demography of Racial Health Disparities*, is to understand how Black-White health disparities changed across time and over the life-course. We focus on age, period, and cohort simultaneously while (a) examining the temporal change in health disparities and (b) exploring the correlates and potential causes of Black-White health disparities. To accomplish this goal, we rely on the analysis of two core domains of data: National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES). These datasets contain individual-level measures including: health related variables, age, interview year (period), birth year (cohort), and a broad set of socio-demographic control variables. In addition, a second domain of data containing cohort and period characteristics will be merged with both the NHIS and NHANES data to model specific period and cohort characteristic and their mediating and moderating effects on race.

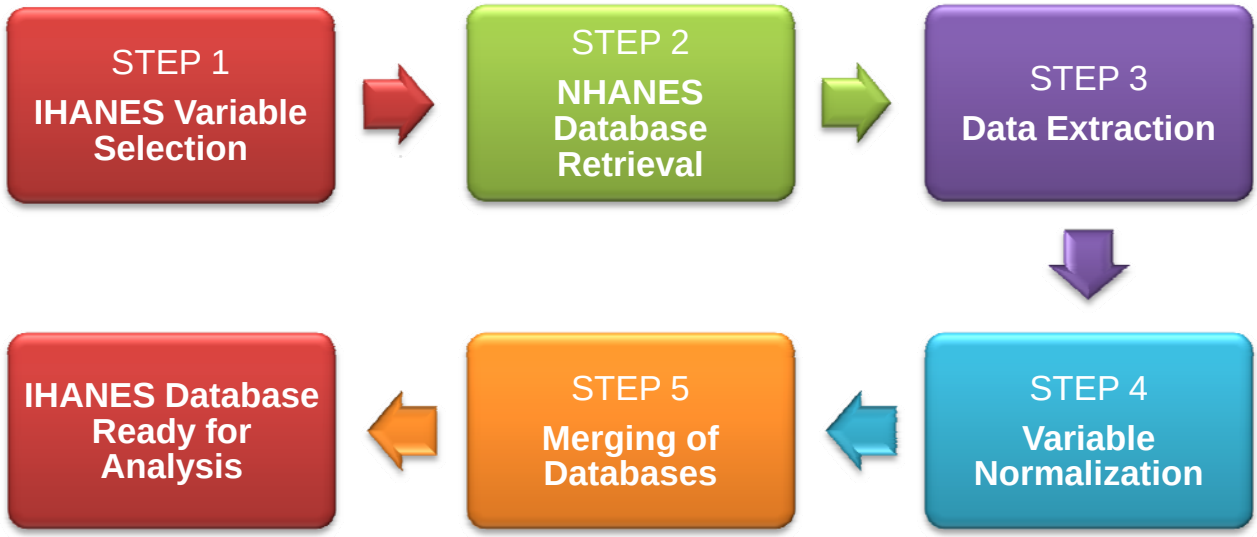
As both core datasets contain multiple cross-sectional waves, a major daunting task for us is to normalize variables across each wave. Fortunately, the integrated Health Interview Survey (IHIS) project at the University of Minnesota has already accomplished this goal and the integrated dataset contains 40 years of data and over seven thousand conformed variables that are ready to be analyzed. However, an integrated dataset for NHANES is currently not available. Thus, the focus of this paper is to describe the protocol used to compile an integrated NHANES (heretofore IHANES) dataset which contains a set of normalized variables across six *waves* of NHANES data collections:

- NHANES 1 (NH1)
- NHANES 2 (NH2)
- NHANES 3 (NH3)
- NHANES 99-00 (NH4)

- NHANES 01-02 (NH5)
- NHANES 03-04 (NH6)

Figure 1 below illustrates the order of processes involved to generate the IHANES database.

Figure 1. A Five-Step Sequence to generate the IHANES database.



II. IHANES Variable Selection

As the project personnel had extensive experience working with NH3, we began to search for variables of interest (IHANES variables, hereafter) from three domains of the NH3 interview and examination components: (a) adult household data, (b) laboratory data, and (c) examination data. All electrocardiogram data were also included in our initial selection of IHANES variables. Following the completion of variable selection, the project personnel systematically searched for the variable names, labels, and descriptions across other waves that matched the IHANES variables.

An excel spreadsheet (see the file entitled “IHANES variable list.xlsx”) was created to record the specific catalog numbers (NH1 & NH2 only), tape positions (NH1 & NH2 only), and

NHANES variable names (NH3-NH6) for all IHANES variables. This allowed the project personnel to easily identify the location of IHANES variables in each NHANES wave. All IHANES variables were also given new variable names to be used in the IHANES database. Column B of the IHANES variable list includes the newly assigned IHANES variable name and column C represents the variable labels for each corresponding variable. Columns D and E indicate the tape position/catalog number for each IHANES variable to be retrieved from NH1 and NH2, respectively. Columns F, G, H, and I indicate the original NHANES variable name for each IHANES variable to be retrieved from NH3, NH4, NH5, and NH6, respectively.

III. NHANES Database Retrieval and Data Extraction

Multiple databases in each wave containing the IHANES variables were downloaded from the NHANES website. The names of the databases that were downloaded from each NHANES wave were summarized in **Table 1**. Subsequently, IHANES variables were extracted from these downloaded databases. The extracted variables were given new variable names as indicated in the IHANES variable list (column B) to conform the variable names across each NHANES wave. After the extraction and renaming process, several data files were created to combine variables from multiple databases in each wave. The names of these files are shown in **Table 2**.

Table 1. A list of NHANSE databases downloaded from NHANES website

NHANES Wave	Databases Retrieved
NH 1	http://www.cdc.gov/nchs/nhanes/nhanesi.htm 1. Medical History Questionnaire, Ages 12-74 years, 4081 2. Electrocardiogram 3. Health Care Needs, General Medical History Supplement, and Respiratory and Cardiovascular Supplements, Ages 25-74 years, 4091 4. Medical Examination, Ages 1-74 years, 4233

NHANES Wave	Databases Retrieved
	5. Anthropometry Goniometry, Skeletal Age, Bone Density, and Cortical Thickness, Ages 1-74 yeas, 4111 6. Spirometry-Best Trials Only, Ages 25-74 years, 4250 7. Biochemistry, Serology, Hematology, Blood Slides, Urine Dipst., 4800
NH 2	http://www.cdc.gov/nchs/nhanes/nhanesii.htm 1. Medical History Questionnaire, Ages 12-74 years, 5020 2. Electrocardiogram 3. Physician's Examination, 5302 4. Anthropometry, 5301 5. Health History Supplement Ages 12-74 years, 5305 6. Hematology and Biochemistry, Ages 6 months-74 years, 5411
NH 3	http://www.cdc.gov/nchs/nhanes/nh3data.htm#NHANES%20III%C2%A0%20Series%2011,%20No.%201a 1. Household Adult File 2. Laboratory File 3. Examination File 4. Electrocardiogram (under Series 11, No. 2A)
NH 4 (1999-2000)	http://www.cdc.gov/nchs/nhanes/nhanes99_00.htm 1. Demographics 2. Examination a. Blood pressures b. Blood measures c. Body measures 3. Laboratory a. Phlebotomy b. Lab 10 c. Lab 11 d. Lab 13 e. Lab 13am f. Lab 18 4. Questionnaire a. Health insurance (HIQ) b. Medical conditions (MCQ) c. Acculturation (ACQ) d. Smoking and tobacco use (SMQ/SMD) e. Family smoking (SMD) f. Food security (FSD/FSQ) g. Hospital utilization (HUQ/HUD) h. Blood pressure (BPQ) i. Occupation (OCQ/OCD) j. Physical activity (PAQ) k. Social support (SSQ/SSD)

NHANES Wave	Databases Retrieved
	l. Alcohol use (ALQ) m. Diabetes (DIQ)
NH 5 (2001-2002)	http://www.cdc.gov/nchs/nhanes/nhanes01-02.htm 1. Demographics 2. Examination a. Blood pressures b. Blood measures c. Body measures 3. Laboratory a. Phlebotomy b. Lab 10 c. Lab 11 d. Lab 13 e. Lab 13am f. Lab 40 4. Questionnaire a. Health insurance (HIQ) b. Medical conditions (MCQ) c. Acculturation (ACQ) d. Smoking and tobacco use (SMQ/SMD) e. Family smoking (SMD) f. Food security (FSD/FSQ) g. Hospital utilization (HUQ/HUD) h. Blood pressure (BPQ) i. Occupation (OCQ/OCD) j. Physical activity (PAQ) k. Social support (SSQ/SSD) l. Alcohol use (ALQ) m. Diabetes (DIQ)
NH 6 (2003-2004)	http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/nhanes03_04.htm 1. Demographics 2. Examination a. Blood pressures b. Blood measures c. Body measures 3. Laboratory a. Phlebotomy fasting questionnaire b. Lab 10 c. Lab 11 d. Lab 13 e. Lab 13am f. Lab 40

NHANES Wave	Databases Retrieved
	4. Questionnaire a. Health insurance (HIQ) b. Medical conditions (MCQ) c. Acculturation (ACQ) d. Smoking and tobacco use (SMQ/SMD) e. Family smoking (SMD) f. Food security (FSD) g. Hospital utilization (HUQ/HUD) h. Blood pressure (BPQ) i. Occupation (OCQ/OCD) j. Physical activity (PAQ) k. Social support (SSQ/SSD) l. Alcohol use (ALQ) m. Diabetes (DIQ)

Table 2. Combined files for each NHANES wave

NHANES Wave	File Names
NH1	NHANES 1 Combined Data.sav
NH2	NHANES 2 Combined Data.sav
NH3	NHANES 3 Adult.sav NHANES 3 ECG.sav NHANES 3 Exam and lab.sav
NH4 (1999-2000)	NHANES 99-00 Demographics.sav NHANES 99-00 Exam and lab.sav NHANES 99-00 Questionnaire combined.sav
NH5 (2001-2002)	NHANES 01-02 Demographics.sav NHANES 01-02 Exam and lab.sav NHANES 01-02 Questionnaire Combined.sav
NH6 (2003-2004)	NHANES 03-04 Demographics.sav NHANES 03-04 Exam and lab.sav NHANES 03-04 Questionnaire Combined.sav

IV. Variable Normalization

A. Normalization Processes

Before the extracted variables can be used for analysis, we must normalize all IHANES variables across each NHANES wave in order to merge the data files listed in **Table 2**. The

normalization process is necessary given that variable questions and response categories are not always identical across the NHANES waves. Thus, variables were normalized to maintain original data integrity as well as to maximize comparability across waves of data collection. To normalize each variable, we tabulate the value labels for all the categorical variables and the unit of measurements and missing values for all continuous variables across six waves. A spreadsheet was compiled (see the file entitled "IHANES Variables Normalization.xlsx") to facilitate this process. This spreadsheet enabled us to determine the normalized or collapsed categories for all categorical variables and verify whether conversions were necessary to conform the unit of measurement for all continuous variables. A new variable entitled "WAVE" was created for all respondents to denote the specific NHANES wave to which each respondent belonged. In addition, new categorical variables for the majority of biomarker variables (e.g. blood pressure, serum cholesterol, plasma glucose) were generated to indicate respondents' health risk levels.

While there was inconsistency in data collection between the earlier waves (NH1 and NH2) and the later waves (NH3 onward) of NHANES, variables that were available for less than 3 waves or too inconsistent to normalize were dropped from our IHANES variable list. When variable recoding or transformation required special assumptions, flag variables were created to note these assumptions. Assumptions, coding logics, and formulas used to normalize each IHANES variable are described in our IHANES codebook (see file entitled "IHANES Codebook.xlsx"). The following information regarding each IHANES variable was also included in the codebook:

- IHANES variable name
- NHANES variable name
  - Tape positions and catalog number were recorded for NH1 and NH2 only
- IHANES variable label
- Normalized value labels
- Normalization and analytical notes (including assumptions and coding logics).

B. Creation of Age, Period, and Cohort Variables

Another important issue to note is that the key component of our study is to simultaneously estimate age, period, and cohort effects. While age variable is readily available in all waves of the NHANES data, period (survey year) and cohort (birth year) are de-identified in order to protect the anonymity of respondents. For this reason, we are only able to precisely specify age, period, and cohort for NH1-2, while less precisely estimating period and cohort for NH3 onward. The following sections provide detailed descriptions on how period and cohort variables were derived for NH3, NH4, and NH5-6.

Derivation of NH3 Period and Cohort Variables. NH3 period variables were derived from the phase and month of interview. Phase one was conducted from October 18, 1988, through October 24, 1991. Respondents interviewed in October of phase one could have been interviewed as early as 1988 and as late as 1991. Therefore PERIODLO is set equal to 1988 and PERIODHI is set equal to 1991 for respondents in phase one interviewed in October. For respondents interviewed in November and December, PERIODLO=1988 and PERIODHI=1990. For respondents interviewed in January through September, PERIODLO=1989 and PERIODHI=1991.

Phase two was conducted from September 20, 1991 through October 15, 1994. Respondents interviewed in October of phase two could have been interviewed as early as 1991 and as late as 1994. Therefore, these respondents were assigned values of 1991 for

PERIODLO and 1994 for PERIODHI. Respondents interviewed in November and December of phase two have PERIODLO=1991 and PERIODHI=1993. For respondents interviewed in January through September, PERIODLO=1992 and PERIODHI=1994.

Cohort ranges are derived using a similar methodology. The date of birth is estimated by decrementing the date of interview by the age of the respondent. Looking again at respondents interviewed in October in phase one, the earliest possible date of birth is calculated by subtracting the age of the respondent from October 1988 (PERIODLO). (Since the date of the interview is not provided, we assume interviews occur on the 15th of the month). The latest possible date of birth is calculated by decrementing October 1991 (PERIODHI) by the age of the respondent. Respondents interviewed in months other than October are calculated using the same methodology. Once the date of birth range has been defined, the cohort variables are simply the year of the date of birth: COHORTLO equals the year of the earliest date of birth and COHORTHI equals the year of the latest date of birth.

Derivation of NH4 Period and Cohort Variables. For NHANES 4, the period variables are calculated using the period of examination: respondents were identified as having been examined in one of two time periods: from November 1, 1999 through April 30, 2000 and from May 1, 2000 through October 31, 2000. Respondents examined in the first period were assigned values of 1999 for PERIODLO and 2000 for PERIODHI. Respondents examined in the second period were assigned values of 2000 for both PERIODLO and PERIODHI.

Cohort ranges were calculated based on the period of examination and the respondent's age at the time of the examination. First, the range of the respondent's date of birth range was calculated. The "low" range of the date of birth was estimated by decrementing

the respondent's age from the start of the period of examination. For respondents examined in period one, the earliest date of birth would be his or her age decremented from November 1, 1999 (PERIODLO). The latest possible date of birth would be his or her age decremented from April 30, 2000 (PERIODHI). For respondents examined in period two, the earliest date of birth would be his or her age decremented from May 1, 2000 (PERIODLO). The latest possible date of birth would be his or her age decremented from October 31, 2000 (PERIODHI).

As in NH3, once the date of birth range has been defined, the cohort variables are equal to the year of the date of birth: COHORTLO equals the year of the earliest date of birth and COHORTHI equals the year of the latest date of birth.

Derivation of NH5-6 Period and Cohort Variables. For NH5-6, the methodology for calculating cohort and period variables is identical to the methodology used for NH4. The periods of examination for NH5-6 are as follows:

	Period 1	**Period 2**
NH5	November 1, 2001 through April 30, 2002	May 1, 2002 through October 31, 2002
NH6	November 1, 2003 through April 30, 2004	May 1, 2004 through October 31, 2004

C. Division of Labor for Variable Normalization

The division of labor for programming the variable normalization was among three project staff and three statistical software packages; SAS (version 9), STATA (version 10), and SPSS (version 15.0), were utilized to generate the normalized variables[1]. The type of programming codes used to normalize each corresponding variable is shown in **Table 3**. The actual programming codes are also available upon request from the project staff

[1] Three different statistical analysis/data management software programs were used given the disparate expertise of the three staff primarily responsible for the IHANES normalization coding. Data files were converted to a common database prior to merging, however, programming syntax remains in the preferred software coding language of choice. This allowed the project to gain efficiencies in distributing workload, but clearly makes things more difficult for interested end-users to both decipher the code and to replicate our normalization.

(http://cherp.sdsu.edu). Three separate data files were created after running these codes in separate statistical packages.

- IHANES normalization_AB.sas7bdat
- NHANES _AB2.dta
- IHANES Exam and Lab Variables.sav

Table 3. Types of programming codes use to normalize the IHANES variables

SAS Codes		
AGE_EXAM	AGE_INTERW	INTER_WEIGHT
EXAM_STAT	RACE*	FEMALE
HHSIZE	PHASE	MAR
MILITARY	BRITH_PLACE	EDU_YEAR
SMKER_HOM	POVERTYR	PSEUDOSTRA
PSEUDOPSU	DOE_12	DOE_456
DOB	COHORT	COHORTLO
COHORTHI	PERIOD	PERIODLO
PERIODHI	DOBLO	DOBHI
AGE_INTERW_MO	AGE_INTERW_MOT	AGEFLAG
AGE_INTERWT	AGE_EXAM_MO	AGE_EXAM_MOT
AGE_EXAMT	INTER_WEIGHT_2YR	INTER_WEIGHT_4YR
EXAM_WEIGHT_2YR	EXAM_WEIGHT_4YR	EXAM_WEIGHT
STATA Codes		
INS_ANY	INS_STATUS	W1_INSFLAG
EMP_STATUS	W1_2_UNEMPFLAG	PRES_SCORE*
FATH_BPL	MOM_BPL	FOOD_INSEC
W6_FINSFLAG	SMOKER_ST	SR_HEALTH
ACT_SELF	ACT_OTHER	YRSUS_CAT
SPSS Codes		
PULSE	PULSE_CAT	PULSE_IRREG
PULSE_IRR_FLAG	SBP	SBP_CAL_FLAG
DBP	DBP_CAL_FLAG	HYPER_CAT
HYPER_CAT_FLAG	BP_MBL	WEIGHT_KG
WEIGHT_LBS	HEIGHT_CM	WAIST_CM
WAIST_MBL	SKIN_SUB	SKIN_TRI
SKIN_FLAG	SUB_TRI_SUM	SUB_TRI_RATIO
BMI	BMI_CAT	DRK5_LAST_12MO
PREG	PREG_FLAG	CHOL_SERUM
CHOL_SERUM_SI	CHOL_CAT	TRIG_SERUM
TRIG_SERUM_SI	TRIG_CAT	TRIG_MBL
LDL_SERUM	LDL_SERUM_FLAG	LDL_SERUM_SI
WTSAF99_00_TRI_LDL	WTSAF01_02_TRI_LDL	WTSAF99_02_TRI_LDL

WTSAF03_04_TRI_LDL	LDL_CAT	HDL_SERUM
HDL_SERUM_SI	HDL_CAT	HDL_MBL
CHOL_HDL_RATIO	CHOL_HDL_RATIO_FLAG	FIBRINOGEN
FIBRINOGEN_SI	FIBRINOGEN_CAT	CRP_SERUM
CRP_CAT	ALBU_SERUM	ALBU_SERUM_SI
ALBU_CAT	GLYC_HEMO	GLYC_HEMO_CAT
GLU_PLAS	GLU_PLAS_SI	WTSFA99_00_GLU
WTSFA01_02_GLU	WTSFA99_02_GLU	WTSFA03_04_GLU
GLU_CAT	GLU_MBL	METABOLIC
METABOLIC_FLAG1	METABOLIC_FLAG2	

*Supplemental documents for creating the RACE and PRES_SCORE variables can be found in the "Race Recode" and "Prestige Score Supplement" tabs respectively in the IHANES codebook.

V. Merging of Data Files

The last step to generate the finalized IHANES database is to merge the three files created after running the program codes. Two files (IHANES normalization_AB.sas7bdat & IHANES Exam and Lab Variables.sav) were converted by StatTransfer to make files readable by the statistical software, Stata. All three files were transferred into Stata and they were then merged together using two variables: SEQN (respondent sequence number) and WAVE (NHANES wave). Our finalized IHANES database consists of 126 variables including demographic variables, age, period, cohort, self-rated health and many important biomarker variables. These variables are normalized across the six waves of NHANES and they are ready to be analyzed.

VI. Data Availability and Dissemination

The IHANES database is available to download from our server. If you are interested in acquiring the dataset, please contact the CHERP director, Dr. Brian K. Finch, at 619-594-6502 or bfinch@mail.sdsu.edu.