NHLBI grant: Ancillary to HCHS/SOL: Sociocultural Factors and CVD risk/prevalence in Hispanics (RC2 HL101649, LC Gallo & FJ Penedo)

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)

Sociocultural Ancillary Study

Investigator Use Database Overview

Version 5, November 2012

Prepared by the HCHS/SOL Coordinating Center
Collaborative Studies Coordinating Center
UNC Department of Biostatistics

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)
Sociocultural Ancillary Study
Investigator Use Database Overview Table of Contents
Version 5 (November 2012)

Version 5 data release (November 2012) has the following updates:

- **Missing item rule for scores.** The Coordinating Center computed subscale or summary scores if all the items used in the score are present, *or* a particular scale has an author designated algorithm for handling missing values. Only 9 of 64 scores had established algorithms for missing values that were cited (see section 6 for details). Overall, the derived scores in the ancillary study now have less than 5% missing values using this conservative approach.
- **Screened participants.** We have excluded from the Screening Form (ANE) those who have missing participation status (ANEA3) as agreed by Drs. Gallo, Penedo, and Sotres-Alvarez on 9/11/2012. Hence ANEA will have 7,321 participants that the study recruiters attempted to reach as specified in submitted manuscript MS61.
- **Sociocultural Ancillary Study derived variables.** The SES_PART_DERV_SOCIO includes derived variables from Sociocultural AS and not variables from main study. Ancillary study participants must have been part of the final 16,415 main study cohort in order to be included in this data release.
- **Each dataset has two identifiers available for use:** HSCAS_ID which is the same identifier used in the last release of the sociocultural files (INV3 &4), and ID which is the HCHS main study masked identifier that can be used to link observations to any form in that investigator use SAS library.

1. INTRODUCTION

This document describes the content and structure of the Investigator Use datasets created for the Hispanic Community Health Study/ Study of Latinos (HCHS/SOL) Sociocultural Ancillary Study. This database contains all the data collected at this interview which occurred within 6-9 months of the HCHS/SOL baseline examination. There were 7,321 screened and 5,313 participants who agreed to participate in the study and completed questionnaires. Data for prospective participants who screened out because they were not eligible/screening closed/no contact (n=2,467) are not included in the interview battery. HCHS/SOL main study data included in this database release is a limited number of demographic, health outcomes (diabetes, metabolic syndrome) and "acculturation" variables which appear in the derived variable file.

2. HCHS/SOL SOCIOCULTURAL ANCILLARYSTUDY OBJECTIVES

HCHS/SOL is a multi-center observational longitudinal study of over 16,000 Hispano/Latino persons aged 18-74 years in four Hispanic communities (Bronx, Chicago, Miami, and San Diego) in the United States. The aim is to obtain baseline measures of pulmonary function, cardiovascular function, metabolic status, oral health, neurocognitive and psychological functioning, and to follow participants for 36 months to assess health outcomes (Sorlie et.al, 2010, LaVange et. al 2010).

The HCHS/SOL Sociocultural Ancillary study seeks to thoroughly characterize the intersection of socioeconomic and cultural contextual influences, intermediate social and psychological processes, and cardiovascular disease and metabolic conditions in U.S. Hispanics from HCHS/SOL.

3. STUDY DESIGN

Electronic copies of the study protocol and manuals of operation for the main study are available online at the following URL.

http://www.cscc.unc.edu/hchs/utilities/docfilter.php?study=hchs&filter_type=forms

3.1. Participants

Participants in the HCHS/SOL Sociocultural Ancillary Study includes any HCHS/SOL participant who completed their baseline exam within six months of the Ancillary Study enrollment start date, or was recruited into the parent study following the onset of Ancillary Study sample enrollment. No additional eligibility criteria are imposed. Language barriers were not a reason for exclusion for Spanish speakers not proficient in English, since all contact with participants was done using the appropriate language.

3.2. Data Collection Forms

Table 1 lists the number of data collection forms collected during the HCHS/SOL Sociocultural Ancillary Study interview for the 5,313 participants. Screening information and demographics are included for 7,835 main study participants who were actively recruited for this ancillary study. HCHS main study forms are not included in this distribution of the ANSOL database.

Table 1. Assessment Battery

Questionnaires	Form Code	Count
Acculturation Stress	ACE	5,312
Cancer Screening	CNE	5,313
Discrimination & Neighborhood Stress	DCE	5,312
Emotions	EME	5,311
Familism & Fatalism	FME	5,312
Gender Roles	GNE	5,312
Immigration & Ethnicity	IME	5,313
Interpersonal Resources	SOE	5,313
Intrapersonal Resources	IPE	5,313
Life, Chronic & Perceived Stress	STE	5,309
Personal Relations	PRE	5,313
Religion	RLE	5,312
Socioeconomic Assessment	SEE	5,312
Derived Variable Files		
Participant Derived File	n/a	5,313
Administrative Forms		
Ancillary Study Screening	ANE	7,315

4. DATABASE STRUCTURE

4.1. Data Set Organization

There is one table (SAS data set) in the database for each type of data collection form (provided as PDFs). The data values from one completed paper form are stored in one record in the corresponding table with information from each participant comprising a single observation (row) in the SAS data set. Each data item on a paper form is stored as one or more variables (columns) in the data set.

A special derived variable dataset (SES_PART_DERV_SOCIO_INV5) has been created to augment the original data measurement values. The sociocultural participant derived variable file has computed score values based on standard algorithms for some of the instruments in question (e.g. STAXI- trait anger). These algorithms are included in a separate document called "HCHS Sociocultural AS Derived Variable Dictionary INV5".

A codebook has been produced for each data set. A careful review of the codebooks, in conjunction with the forms, is critical to understanding and interpreting the source data. The codebook provides a description of every variable in the database as well as the frequency and meaning of variables' values. **Analysts are *strongly* encouraged to use the codebooks, paying attention to the data user notes contained in this document.**

4.2. Form and Data Set Naming Conventions

Each HCHS/SOL Sociocultural Ancillary Study data collection instrument (see PDF forms) has a unique four-letter mnemonic associated with it (e.g., CNEA for the Cancer Screening form in English, CNSA for Cancer Screening in Spanish, Version A). The corresponding data sets begin with the same first three letters of the mnemonic for the English version, followed by the character string "SOCIO_INV5", for Sociocultural Investigator Use, Version 3 (e.g. "CNEA_SOCIO_INV5.sas7bdat"). The naming convention serves both to identify the originating form and provide version control when subsequent datasets are produced. Note, since the questionnaire battery for the ancillary study has both English and Spanish language versions of the forms each has been merged into one common data record format which follows the main HCHS/SOL study conventions. For example, the CNE and the CNS both map to the CNEA_SOCIO_INV5 in this data release.

4.3. Key Fields for Data Records

The unique identification of a participant data record within a file is determined by three primary key fields for forms that are collected once per visit (see HCHS/SOL Investigator Use Database Overview), and by the use of a sequencing field for the few forms that could occur many times per visit (e.g. 24 hr. diet recall in DTIA of the Main Study which is included in this data release). These items are:

1) ID_HSCAS: A random 8-digit masked identification code, unique to each HCHS/SOL participant. The HCHS/SOL Coordinating Center created a unique ID for this ancillary study which is different purposely to the HCHS/SOL Study ID.
   ID: Main Study Investigator Use ID, which is eight digits (including leading zeros) which can be used to merge with the HCHS main study INV3 series data.
   VISIT: Contact year number, a two digit field, "01" for baseline examination year. NOTE: For HCHS/SOL Sociocultural Ancillary Study, all forms have visit "01".

2) FSEQNO: From sequence number, a two digit sequencing number (01-99) for multiple forms per visit (e.g. such as when a repeat assessment might be performed). This field is effectively constant at "01" when forms are administered only once.

4.4. Common Variables Across Data Sets

An additional variable appears in every data set, and may be useful in identifying particular subsets of the data:

3) VERSION: Version of the data collection form. A one character variable indicating which version of the paper form was used to collect the data. Possible values for VERSION are “A”, “B”, and “C”, representing the first, second, and third versions, respectively. Version remained a constant in the ancillary study.

   NOTE: For HCHS/SOL Sociocultural Ancillary Study, all forms are version “A”.

4) FORM: The original 3-letter form code that appears on the paper-based forms or on the form code selection menu in the DMS uses the convention of having the third letter designate the language version in use. Use this variable to detect changes in language of administration. The standard taken from the main study uses “E” for English language forms versus “S” for the Spanish language version (CNE vs. CNS for cancer knowledge and screening).

4.5. Variable Naming Conventions

While the key field and sort variables (see Sections 4.3 and 4.4) have the same name on each SAS record type (ID, VISIT, FSEQNO, and VERSION), other SAS variables are unique to a specific form. To predictably and uniquely link data items to forms, these form-specific variable names begin with the same three characters as the data set name, followed by the form version letter, and then the question number as indicated on the form. For example, question 5 on the Ancillary Screening form ANEA, "gender", is named ANEA5 on the corresponding SAS file, ANEA_SOCIO_INV5. Similarly, question 7, "Hispanic/ Latino Background", from the “A” version Ancillary Screening form is named ANEA7. These demographic variables were automatically populated from source data in the main study. Since screening occurred for the ancillary study, some corrections have been applied in version 3.1 of the HCHS main study derived variable file as described elsewhere so merging those latest updated demographics with these analysis files would assure use of those corrections.

4.6. Changes to Variables to Preserve Confidentiality

As part of the study commitment to comply with HIPAA regulations for participant confidentiality and to follow current guidelines from NHLBI/NIH the HCHS/SOL Coordinating Center has made explicit modifications and/ or deletions to variables that were common across all forms. All participant ID values were transformed from the original ID to random identifier codes to produce Investigator Use datasets that protect the confidentiality of the individual. However, the authorized user will need to actively attend to the security and confidentiality of these Investigator Use files as part of the end user agreement.

1) A HCHS/SOL Main Study Investigator Use ID (ID) was re-derived for use in all data sets as a random identifier code for participants.
2) A secondary identifier called, ID_HSCAS, which is the same as used in the INV3 data release to the Sociocultural AS Investigators is also included.
3) Addresses, phone numbers, and SSN of the participants were omitted from these files.

4) CENTER, is a real code to distinguish among participating field centers was created for the database and is included in the Participant derived variable set, SES_PART_DERV_SOCIO_INV5 but removed from the ID string.
5) STAFF ID codes were deleted across all forms and not substituted.
6) DATES were kept unaltered and separate month, day, year text strings preserved for each item in case incomplete information was collected.

4.7. Missing Values

The study database employs a standard set of special missing value codes (see study codebook) that have contextual meaning. Since SAS allows numeric variables to assume up to 27 unique missing values, ".A to .Z, and ." the Coordinating Center uses several of these special missing codes to convey additional meaning to the analyst. Here is a table that describes that usage of missing values in HCHS/SOL.

Missing value	Meaning
. or blank	Empty field, missing
.Q	Don't know / refused
.S	Skipped field
.L	Below lower limit of analysis
.H	Above higher limit of analysis

Selective recodes may need to be made to make use of known refusals, or to account for skip patterns in coding derived variables based on multiple items in a form. Laboratory variables with results reported as "< number", or "> number" for values below or above the assay limits are set to the special values of ".L" or ".H". The Quality Control manual for HCHS/SOL has an appendix with the limits of detection for lab measurements (e.g. serum glucose, total cholesterol, LDL-C, HDL-C, triglycerides).

5. DESCRIPTION OF DATA COLLECTION FORMS / DATABASE TABLES

5.1. Acculturation Stress (ACE/ACS)

This form has 34 items that focus on the Hispanic Stress Inventory. Scales and subscale scores are found in the participant derived variable file.

5.2. Cancer Screening (CNE/CNS)

This form has 37 items that focus on screening, knowledge and cognitions about cancer. Some items are gender specific.

5.3. Discrimination and Neighborhood Stress (DCE/DCS)

This 32-item instrument records responses to perceptions of ethnic discrimination and neighborhood sources of stress.

5.4. Emotions (EME/EMS)

The instrument is a 48-item form that contains several commonly used instruments in the domains of depression, anxiety, anger, loneliness, and hopelessness. Derived variable scores for the CESD-10, Spielberger trait anxiety and trait anger scores, the Cook-Medley cynicism score appear in the derived variable file.

5.5. Familism and Fatalism (FME/FMS).
The 22-item questionnaire focuses on the constructs of familism (14 items) and fatalism (8 items). See the derived variable file for overall score on familism and fatalism and the subcomponents of family obligations and support.

5.6. Gender Roles (GNE/GNS)
The 34 item questionnaire is split between male (10 items) and female (24 items) related to gender specific cultural roles. The derived variable file has one summary score for male roles and 5 individual subscale scores for female roles.

5.7. Immigration and Ethnicity (IME/IMS)
The immigration and ethnicity questionnaire was adapted and extend from NHANES. The 37 items cover information about origins and ethnic identity. The derived variable files has an ethnic experiences score and the constructs of ethnic identity, social experiences and comfort, and ethnic discrimination.

5.8. Interpersonal Resources (SOE/SOS)
The well-known scales on Interpersonal Support and the Social Network Index are found in this 55 item questionnaire. There are additional sections on family cohesion and family conflict. See the derived variable file for ISEL total score and the SNI and related index subcomponents, and the family cohesion and conflict scores.

5.9. Intrapersonal Resources (IPE/IPS)
The 25 item questionnaire assess resiliency in the areas of self-esteem, optimism (LOT-R), and life-engagement. The derived variable file contains the scores for these personality dimensions.

NOTE: In the Intrapersonal form (IPE and IPS) administered to the participants had the Likert scale for Optimism (items 11 to 18) in the opposite direction in the Spanish version (IPS form). Specifically,

- English version (IPE) has 1 for “I disagree a lot”
- Spanish version (IPS) has 1 for “I agree a lot” (“Estoy de acuerdo completamente”)

The direction was reversed for those who completed the form in Spanish when creating the dataset for data release to match the English version scaling. No additional recoding needs to be performed to correct for this inconsistency between language versions.

5.10. Life, Chronic, and Perceived Stress (STE/STS)
This questionnaire measures lifetime stress and trauma including childhood stress (ACE) and perceived stress scales (PSS 10-item). See the derived variable file for the traumatic life schedule score, lifetime burden, past year stress score, childhood stress exposure, and total perceived stress scale .

5.11. Personal Relations (PRE/PRES)
The 20 item questionnaire is divided into sections assessing Simpatia, and social desirability. The derived variable file has the simpatia total score and the Marlow-Crowne social desirability scale.

5.12. Religion (RLE/RLS)
Dimensions of spiritual well-being and spirituality were assessed in this 28 item questionnaire. The DUREL and FACIT-SP scores from this form are in the derived variables file.

5.13. Socioeconomic (SEE/SES)
This 14 item socioeconomic form extends the information in this area to wealth, assets, immigrant mobility, and material deprivation.

5.14. Sociocultural Ancillary Study Screening (ANE/ANS)
Screening HCHS study participants to assess their willingness to participate in the sociocultural ancillary study was recorded using this form. Each person contacted for the ancillary study had one of these forms completed so that demographics could be compared for those agreeing to participate vs. those who did not agree and come in for an interview. This data was collected so that the investigators could more accurately report response rates for the study.

6. SPECIAL USE DERIVED FILES
6.1. Derived Variable DATASET (SES_PART_DERV_SOCIO_INV5)
The participant derived variable dataset is not associated solely with any particular form because it contains variables from many forms. There is one record per screened (7,321 observations) participant in SES_PART_DERV_SOCIO_INV5. This file is a cross-section of “derived variables” whose values are defined based on combinations of data items (e.g. age from date of birth) from socio-economic and demographics, records. The file is also a repository for all of the subscale and total summary scores from the ancillary study questionnaire battery (like the CES-D and STAS scores from the emotions questionnaire). Missing item counts for constituent questions that contribute to a subscale are also present here to inform the user about data completeness and its impact on scoring an instrument. The following 9 scores have a tolerance level for missing items defined by the author. All other scores in the derived variable file are based on totally complete item components for scoring. See the separate document, “HCHS Sociocultural Ancillary Study Derived Variable Dictionary” for the definitions of the variables included in this special purpose file.

Racism/discrimination scores and Neighborhood scores from DCE form

RACISM- Racism/discrimination Scale Score
R_EXCL- Exclusion subscale/Stigma subscale Score
R_DISCRIM- Discrimination Subscale Score
R_THREAT- Threat Subscale Score
NEIGHBOR_COHESION- Neighborhood Social Cohesion Score
NEIGHBOR_PROBLEM- Neighborhood Problem Score

Chronic stressor scores from STE form

CHR_STR_TOT- Total Number of Chronic Stressors
M_V_TOT- Total Moderate to Very Stress Score for Chronic Stressors

Hostility-Cook Medley Cynicism Scale total score from EME form

CYN_HOST_TOT- Hostility-Cook Medley Cynicism Scale

Statistical analysis using HCHS/SOL Sociocultural Ancillary Study data must account for the complex sampling design by specifying strata (STRAT), primary sampling unit (PSU_ID) and sampling weights (WEIGHT_FINAL_NORM). Analysts are strongly encouraged to read the document “ANALYSIS METHODS FOR HCHS/SOL” in the HCHS/SOL Main Study to ensure that the study design is correctly specified prior to analysis.

IMPORTANT ANALYSIS NOTES:

(1) In a few cases, inconsistencies or omissions in the information required to define these variables could not be corrected on the original data forms (and corresponding files in this database). These idiosyncratic cases were adjudicated by the HCHS/SOL Coordinating Center and their resolutions are included in the derived variable files.

(2) Data records were kept if participants had an ancillary study screening form present where the eligibility status was recorded as "Agrees to Participate" (ANEA3=4), the participant was "Full AFU eligible", *and* a sample weight could be computed. When sub-setting an analysis on screened vs. eligible and enrolled study participants the derived variable, HSCAS_PART can be used to include or exclude cases. The final maximum N=5313 for this ancillary study.

(3) In the first data release, the INV1 files contained observations for everyone who agreed to participate and who had forms. However, in INV5 Sociocultural Ancillary Study participants must be a HCHS/SOL cohort participant which requires AFU eligibility criteria so that a sample weight can be computed. Hence, individuals with missing final sample weights have been dropped from the INV5 files for this ancillary study.

Table 2. An **"X"** appearing below indicates whether the instrument or domain is assessed in the 1) HCHS/SOL baseline; 2) the Sociocultural Ancillary study and; 3) the supplement study.

Instrument	Administered in HCHS/SOL Baseline	Administered in Sociocultural Ancillary Study	Administered in Psychometric Supplement
Demographics	**X**		
Health Factors			
Diabetes, METS, IFG, IGT, BMI, Hypertension	**X**		
Stroke, CHD, and family history of disease	**X**		
Short-Form Health Survey (SF-12)	**X**		X
Cultural Factors			
Acculturation: Short Acculturation Scale for Hispanics (10 items)	**X**		X
Familism (14 items)	**X**	X	X
Fatalism: Multiphasic Assessment of Cultural Constructs-Short Form (MACC-SF): Fatalism subscale (10 items)		X	X
Ethnic Identity: The Scale of Ethnic Experience (32 items)		X	X
Functional Assessment in Chronic Illness Therapy-Spiritual Well-Being Scale (FACIT-Sp-12) (12 items)		X	X
Cognitive-Emotional Factors			
Depression: Center for Epidemiological Study-10 (10 items)	**X**	X	X
Patient Health Questionnaire (PHQ-9) (9 items)			X
Spielberger Trait Anxiety Scale (10 items)	**X**	X	X
Spielberger Trait Anger Scale (10 items)		X	X

7. REFERENCES

Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, Kaplan RC, Daviglus ML, Giachello AL, Schneiderman N, Raij L, Talavera G, Allison M, Lavange L, Chambless LE, Heiss G. *Design and implementation of the Hispanic Community Health Study/Study of Latinos.* Ann Epidemiol. 2010 Aug; 20(8):629-41. **(http://www.sciencedirect.com/science/article/pii/S1047279710000724)**

Lavange LM, Kalsbeek WD, Sorlie PD, Avilés-Santa LM, Kaplan RC, Barnhart J, Liu K, Giachello A, Lee DJ, Ryan J, Criqui MH, Elder JP. *Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos.* Ann Epidemiol. 2010 Aug; 20(8):642-9. **http://www.sciencedirect.com/science/article/pii/S1047279710001171**