# HCHS/SOL Investigator Use Database Overview

## July 2020

### Visit 2 INV Version 3.1

**Prepared by**<br>
**HCHS/SOL Coordinating Center**<br>
Collaborative Studies Coordinating Center<br>
UNC Department of Biostatistics

Marston Youngblood<br>
Daniela Sotres-Alvarez<br>
Franklyn Gonzalez II<br>
Yanping Teng

**The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)**
**Visit 2 Investigator Use Database**
**Version 3.1**
**July 2020**

**Table of Contents**

**1. INTRODUCTION ........ 1**

**2. STUDY OBJECTIVES ........ 1**

**3. STUDY DESIGN ........ 1**

3.1. Participants ........ **1**

3.2. Schedule of Participant Data at Visit 2 ........ **2**

**4. DATABASE STRUCTURE ........ 2**

4.1. Data Set Organization ........ **2**

4.2. Form and Data Set Naming Conventions ........ **3**

4.3. Key Fields for Data Records ........ **3**

4.4. Common Variables Across Data Sets ........ **4**

4.5. Variable Naming Conventions ........ **4**

4.6. Changes to Variables to Preserve Confidentiality ........ **4**

**5. DESCRIPTION OF VISIT 2 EXAM DATA COLLECTION FORMS: ........ 5**

5.1. Acculturation (ACE) ........ **5**

5.2. Alcohol Use (ALE) ........ **5**

5.3. Family Cohesion (FCE) ........ **6**

5.4. Health Care Use (HCE) ........ **6**

5.5. Medical History (MHE) ........ **6**

5.6. Medication Use (MUE) ........ **6**

5.7. Pregnancy Complications (PCE) ........ **6**

5.8. Participant Disability (PDE) ........ **6**

5.9. Participant Feedback (PFE) ........ **7**

5.10. Reproductive History (RME) ........ **7**

5.11. Socioeconomic (SEE) .......... 7

5.12. Social Support (SSE) .......... 7

5.13. Stress (STE) .......... 7

5.14. Tobacco Use History (TBE) .......... 7

5.15. Well Being (WBE) .......... 7

5.16. Anthropometry (ANT) .......... 8

5.17. Biospecimen Collection (BIO) .......... 8

5.18. Sitting Blood Pressure (SBP) .......... 8

5.19. Informed Consent Checklist (ICT) .......... 8

**6. SPECIAL USE DERIVED FILES .......... 8**

6.1. Participant Derived Variables (PART_DERV_V2) .......... 8

6.2. Pregnancy Complications Derived Variables (PCE_DERV_V2) .......... 9

6.3. Laboratory Results Derived DATA SET (LAB_DERV_V2) .......... 9

6.4. Echocardiograhy Record (ECH) .......... 9

**7. REFERENCES .......... 9**

## 1. INTRODUCTION

This document describes the content and structure of the Investigator Use datasets created for HCHS/SOL. This database contains data collected at the Visit 2 clinic visit for the approximate three-year examination cycle (Oct. 2014 through Dec. 2017). This V2_INV3 release has data from 11,623 of the original cohort of 16,415 enrolled participants who had a 2nd visit. The contents of the release is limited by constraints (described within) to preserve participant confidentiality by de-identifying the data. No ancillary study data are included in this Visit 2 database.

## 2. STUDY OBJECTIVES

This multi-center observational longitudinal health study is designed to document health status in four Hispanic communities around the United States and to obtain baseline measures of pulmonary function, cardiovascular function, metabolic status, oral health, and measures of neurocognitive and psychological functioning. At baseline, 16,415 adults of 18 to 74 years, were enrolled at four field centers over a 36 month period, and are being followed annually to assess health outcomes (see Sorlie et.al, 2010). In the Visit 2 clinic visit, there were 11,623 cohort members re-examined to again collect data predictive of cardiopulmonary outcomes and the onset of diabetes. A comprehensive reproductive history of women of childbearing age was assessed as well. Follow-up and ascertainment of endpoints outcomes is planned to continue until the end of 2024.

## 3. STUDY DESIGN

To address the study objectives the prospective follow-up cohort study was conducted in 4 field centers (Bronx, Chicago, Miami, and San Diego) as described in Sorlie, et al. Ultimately, 16, 415 participants were enrolled from a randomly selected set of household postal addresses in the target communities (see LaVange et. al 2010). Each of four field centers recruited approximately 4,000 persons of Hispanic origin to participate in the study. The baseline age range is 18-74, and study participants are selected to obtain approximately 2,500 persons age 45-74, and approximately 1,500 persons age 18-44. Recruitment was designed to occur in stable communities so that persons can be contacted over time, and possibly examined more than once. Electronic copies of the study protocols for both Visits and manuals of operation are also included elsewhere for reference with this data release. Visit 2 screening began late Summer 2014 and the initial wave of exams for Visit 2 started in October 2014 and concluded in December 2017.

### 3.1. Participants

All study participants were 18-74 years of age at baseline (2008-2011), self- identified as being Hispanic/Latino, and not planning to move from the community during the period of follow-up. The recruited individuals attended a first examination to assess cardiovascular and other disease risk factors, both known and potential. The risk factors of particular interest are occupational exposure, nutrition, oral health, physical activity, family structure, and acculturation. The study strives to make the percent of identified persons who actually attend the examination high, to reduce bias from non-response. There is no exclusion of persons based on existing health status but the following persons were not recruited at baseline: those who plan on moving away in the next 3 years; those who have health problems, disabilities, or mental problems so severe as to prohibit informed consent and actual clinic attendance. Language barriers are not a reason for exclusion for Spanish speakers not proficient in English, since all contact with participants is done using the appropriate language. The Visit 2 examination was a call-back of the HCHS/SOL cohort starting 5 to 6 years after baseline for the group being surveilled through annual follow-up interviews. Average time between visits was 6 years (median time 5.9 yrs).

### 3.2. Schedule of Participant Data at Visit 2

Table 1 lists the number of data collection forms collected during the second HCHS/SOL examination among the 11,623 participants included in this data release. Ancillary study forms are not included in this distribution.

**Table 1. Visit 2 Assessment Battery**

| Questionnaires | Form Code | Count |
|---|---|---|
| Acculturation | ACE | 11,191 |
| Alcohol Use History | ALE | 11,611 |
| Family Cohesion | FCE | 11.182 |
| Health Care Use | HCE | 11,191 |
| Medical History | MHE | 11,610 |
| Medication Use | MUE | 11,608 |
| Pregnancy Complications | PCE* | 693 |
| Participant Disability | PDE | 11,453 |
| Participant Feedback | PFE | 8,707 |
| Reproductive History | RME | 7,204 |
| Socioeconomic | SEE | 11,212 |
| Social Support | SSE | 11,180 |
| Stress | STE | 11,188 |
| Tobacco Use | TBE | 11,611 |
| Well-being | WBE | 11,185 |
| **Procedure(s) Forms** | | |
| Anthropometry | ANT | 11.609 |
| Bio-specimen Collection | BIO | 11,592 |
| Sitting Blood Pressure | SBP | 11,614 |
| **Derived Variable Files** | | |
| Echocardiography record | ECH | 6,949 |
| Laboratory Results Derived | n/a | 11,588 |
| Participant Derived | n/a | 11,623 |
| Pregnancy Comp. Derived | n/a | 693 |
| **Administrative Forms** | | |
| Informed Consent Checklist | ICT* | 11,639 |

*Multiple-record form.
See section 5 for key fields that uniquely identify each record.

## 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 at baseline. The data values from one completed paper form are stored in one record in the corresponding table (observation in the SAS data set). Each data item on a paper form is stored as one or more columns (variables) in the data set. Collection of direct measurements during examination procedures can also result in the creation of a data file. Similarly, sitting blood pressure measurements are recorded on the SBP form while the technician uses the Omron HEM-907XL sphygmomanometer.

Since forms can be revised at during the course of the study, the version of the paper form used to collect the data is also included on each record (e.g., versions A or B). The SAS data set is a composite of the data items required to accommodate all versions of the corresponding data form. Some version specific data items will be missing in a given record depending upon which version was completed at time of data acquisition in the field.

Special derived variable datasets have been created to augment the original data measurement values. The participant derived variable file has computed outcomes and summary score values based on standard algorithms for some of the instruments in question (e.g. Current_Smoker_V2, CESD_V2). These algorithms have been included in the derived variable dictionary and can be found in the documents issued with this volume.

A codebook has been produced for each data set. A careful review of the codebooks, in conjunction with the forms, is critical to interpreting the data. The codebook provides a description of every variable in the data set 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 data collection instrument in the Visit 2 CDART entry and data management & reporting system has a unique three-letter mnemonic associated with it (e.g., SBP for the HCHS/SOL Sitting Blood Pressure form. The corresponding data sets begin with the same first three letters of the mnemonic, followed by the character string "_V2_INV3 for Visit 2 Investigator Use, Version 3. For example, the Sitting Blood Pressure data set for the second Visit 2 release is "SBP_V2_INV3". The naming convention serves both to identify the originating form and provide version control when subsequent generations of datasets are produced. Specialized reading center records like the echocardiography interpretation (ECH_V2_INV3) can deviate from this general study naming convention.

### 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 for the baseline exam datasets (see HCHS/SOL Data Management Guide), and by the use of a sequencing field for the few forms that could occur more than once per visit. These items are:

1) ID: A random 8-digit identification code, unique to each HCHS/SOL participant.
2) VISIT: Contact year number, a two digit field, "02" for Visit 2 examination.
3) OCCURRENCE: Form sequencing number, a two digit sequencing number (01-99) for multiple forms per visit (see variable FSEQNO in the baseline release).

### 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:

1) VERS: Version of the data collection form. Version can be a numeric or character variable indicating which version of the paper form was used to collect the data. Starting with the Visit 2 exam, the CDART versions are numeric. Data files are combined across versions. Knowing the version number helps map back to the original form based items and responses. See the included forms for this release for the current version in production for data collection.
2) FORM: The original 3-letter form code that appears on the paper-based forms or on the form code selection menu in CDART uses the convention of having the third letter designate the language version in use. Use this variable to detect changes in language of administration ("E" for English language forms versus "S" for the Spanish language version).

### 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, OCCURRENCE, and VERS), 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 1 on the Acculturation form, "languages read or spoken", is named ACE1 on the corresponding SAS file, ACE_V2_INV3. Similarly, question 2, "languages as a child", from the Acculturation form is named ACE2.

### 4.6. Changes to Variables to Preserve Confidentiality

As part of the study commitment to complying with HIPAA regulations for participant confidentiality and in following guidelines from NHLBI/NIH the 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 values to produce Investigator Use data files 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.

A HCHS/SOL ID (ID) was re-derived for use in all data sets as a random identifier code for participants and is the same masked identifier used in the baseline release.

1) Addresses, phone numbers, immigration status, date of birth, and SSN of the participants were omitted from these files.
2) CENTER, is a real code to distinguish among participating field centers was created for the Investigator Use database and is included in the Participant derived variable set, PART_DERV_V2_INV3 but removed from the ID string.
3) STAFF ID codes were deleted across all forms and not substituted.
4) DATES were kept unaltered and separate month, day, year text strings preserved for each item in case the linkage with event year whenever months and day of the month are unknown.
5) DATE OF BIRTH was converted to age at Visit 2 (AGE_V2) and appears in the derived variable data set, PART_DERV_V2_INV3.

### 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 |
| .U | Unknown |
| .Q | Don't know / refused |
| .S | Skipped field |
| .L | Below lower limit of analysis |
| .H | Above higher limit of analysis |
| .N | Not applicable/ not available |

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. Using SAS, analysts are strongly encouraged to detect missing values by using code that employs "**≤ .Z**" which will detect all of these special missing values rather than "**= .**", which will not. Alternatively the SAS missing function can be used to return a TRUE/FALSE value (1/0) for the presence of missing values. 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 Biospecimen Collection and Processing Manual of Procedures (MOP 7) for HCHS/SOL Visit 1 has Appendix 1 with the limits of detection for lab measurements (e.g. serum glucose, total cholesterol, LDL-C, HDL-C, triglycerides).

## 5. DESCRIPTION OF VISIT 2 EXAM DATA COLLECTION FORMS:

### 5.1. Acculturation (ACE)

A questionnaire designed to assess language use and ethnic preferences for social interactions. The acculturation scale will assess participant's degrees id integration into American culture or the "host" culture. Additional items assess visits to country of origin in participants who were not born on the U.S. mainland.

### 5.2. Alcohol Use (ALE)

A brief screen for history of alcohol use was administered that collects data on lifetime use, current use, and former use of alcohol. Limited metrics on at risk drinking can be derived from the quantity/frequency measures and appear in the derived variable file.

### 5.3. Family Cohesion (FCE)

This instrument based on 9 items is a subset of the copyright protected Family Environment Scale which measures the degree of commitment and support family members provide for one another. Do not re-distribute this form. These items are used in HCHS/SOL for research by permission of the copyright holder.

### 5.4. Health Care Use (HCE)

The purpose of this questionnaire is to understand patterns of health services use in the preceding 12 months, utilization of screening and preventive services, and health insurance status. This new version to be administered during Visit 2 has been expanded in order to assess health utilization patterns, use of adult preventive or screening services, and health insurance coverage and eligibility. Most of these questions were obtained from different national surveys, including the Behavioral Risk Factor Surveillance System (BRFSS), the National Health Interview Survey (NHIS), the Medical Expenditure Panel Survey (MEPS), the Health Information National Trends Survey (HINTS) and the 2010 U.S. Census Survey.

### 5.5. Medical History (MHE)

The medical history form inquires about personal medical history. Participants are asked to provide information since the first HCHS/SOL visit, from the last telephone interview, within the last 12 months, etc. So the responses have to be interpreted and used within that temporal context. This instrument contains general questions on self-reported cardiovascular disease, pulmonary diseases, stroke, hypertension, hypercholesterolemia, metabolic problems, cancer, and continence.

### 5.6. Medication Use (MUE)

The medication use questionnaire captures the self-report of medication use and an inventory of both medications and supplements used during the last four weeks. Since participants may or may not know the actual indication for a specific medicine, there is embedded list of conditions for which medications could be prescribed. Medications are not automatically coded and classified in Visit 2, unlike Visit 1.

### 5.7. Pregnancy Complications (PCE)

The pregnancy complications history questions are designed to ask about pregnancies that lasted 6 or more months that occurred after the first HCHS/SOL visit date. Each PCE form captures information all pregnancies that last more than 6 months. The form can capture information on more than one baby in the case of childbirths with multiple babies (twins, triplets, quadruples). This form will only be administered conditional on qualifying questions on the Reproductive Medical History form (see below). The key fields that uniquely identify each record are ID and OCCURRENCE (identical to PCE1) which is the order of the pregnancies.

### 5.8. Participant Disability (PDE)

The disability screening form is based on a set of six questions from the American Community Survey (ACS) and other major national surveys to gauge disability. These items are the minimum necessary subset to compute a standardized measure of disability.

### 5.9. Participant Feedback (PFE)

The feedback questionnaire was designed to assess a variety of factors that have motivated and/or discouraged SOL participants to stay connected to the study. It will also help capture information that will allow the study staff improve and create new strategies to keep the SOL cohort interested in the study.

### 5.10. Reproductive History (RME)

This questionnaire is administered to ALL women and has two sections: (A) Hormone and Menstrual History, and (B) Pregnancy History (pregnancies ever had before or after HCHS/SOL visit 1). Pregnancy complications before V1 are documented in this RME form, whereas pregnancies AFTER V1 (lasted 6 or more months) are documented in more depth in the PCE form (see above).

### 5.11. Socioeconomic (SEE)

The questionnaire updates information about additional education since baseline, current income, and occupational status.

### 5.12. Social Support (SSE)

The questionnaire asks about social resources such as being part of a social network, receiving support from others, and relationships with family. The goal is to determine the types of relationships participants have and their impact on health and well-being. The ISEL 11-item scale is embedded in this questionnaire (SSE1-SSE12).

### 5.13. Stress (STE)

The Chronic Stress Scale is a measure of ongoing stress in several life domains. The measure asks about ongoing stress related to health problems in self or others, job or ability to work, finances, personal relationships, alcohol or drug use, and one un-specified domain. For each domain if the respondent indicates that the stressor did occur, there are two additional questions about whether the stressor has persisted for six months or more, and how stressful the participant found the stressor to be to them.

### 5.14. Tobacco Use History (TBE)

The tobacco use instrument contains items on current and/or former use of tobacco products as well as exposure to secondhand tobacco smoke. The 23-items could be used to derive variables on current/former/any use of tobacco. The data elements are present for computing pack-years of exposure to cigarettes. Smoking cession items are also included in the instrument.

### 5.15. Well Being (WBE)

The Well-being questionnaire consists of two brief measures that assess depressive and anxiety symptoms respectively, specifically, the Center for Epidemiological Study measure of depression (10- item version) and the GAD-7 Scale (7-item version). These measures are administered to evaluate the levels of depression and anxiety symptoms in the Hispanic/Latino population and their association with health. The participant derived file has the summary score for the CESD-10.

### 5.16. Anthropometry (ANT)

The direct measurements obtained at the anthropometry station are recorded on these entry screens.  See Visit 2 Manual 2 for a full description of the procedures and measurements. Computed variables BMI and waist-hip ratio are in the derived file.

### 5.17. Biospecimen Collection (BIO)

The timing and condition at the time of collection for the complete set of Visit 2 blood and urine specimens is recorded on this form.  Reasons for exclusion from the OGTT or any other conditions that would affect specimen collection are also noted here.  The derived variable for fasting time is based on the elapsed time since the participant reporting eating or drinking anything and the start of the sample collection process.

### 5.18. Sitting Blood Pressure (SBP)

Sitting blood pressure measurements are directly recorded while the technician uses the Omron HEM-907XL sphygmomanometer.  A series of three systolic and diastolic measures is automatically recorded along with the Omron determined average.  See Visit 2 Manual 2 for a detailed description of these procedures. NOTE: ALL THREE MEASUREMENTS  were AVERAGED by the Omron sphygmonmanometer. This is unlike other studies which used only the $2^{ND}$ and $3^{RD}$ readings.

### 5.19. Informed Consent Checklist (ICT)

The elements of informed consent are tracked in this file for the administrative form that was explained and executed at Visit 2.  This file can be updated during follow-up to reflect changes in permissions levels for use of participant data. The key fields that uniquely identify each record are ID and OCCURRENCE.

## 6. SPECIAL USE DERIVED FILES

### 6.1. Participant Derived Variables (PART_DERV_V2)

The participant derived variable dataset is not associated solely with any particular form because it contains variables from many forms and files.  There is one record per enrolled participant (11,623) at Visit 2 in PART_DERV_V2_INV3.  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, or body mass index from height and weight, waist-hip ratio from girth measurements), primarily from the anthropometry, demographics, respiratory history, clinical laboratory analysis and pulmonary function records. Important study design variables like sample weight and strata identifiers are also found here.  See the separate document, *“HCHS/SOL Visit 2  Derived Variable Dictionary”* for the definitions of the variables included in this special purpose file. Statistical analysis using HCHS/SOL data must account for the complex sampling design by specifying strata (STRAT), primary sampling unit (PSU_ID) and sample weights (WEIGHT_NORM_OVERALL_V2). Analysts are strongly encouraged to read the documents “ANALYSIS METHODS FOR BASELINE FOR HCHS/SOL” and “ANALYSIS METHODS – VISIT 2 FOR HCHS/SOL” in the HCHS/SOL Main Study to ensure that the study design is correctly specified prior to analysis.

### 6.2. Pregnancy Complications Derived Variables (PCE_DERV_V2)

These variables are derived from items on the RME and PCE forms to provide derived variables from pregnancies and births since Visit 1 (e.g. birth weight for gestational age z-score, preterm).

### 6.3. Laboratory Results Derived DATA SET (LAB_DERV_V2)

The central clinical chemistries laboratory for the study at University of Minnesota Fairview Hospital Advanced Research Diagnostic Laboratory (ARDL) provides the HCHS/SOL study with data for this laboratory derived results data set. The Visit 2 clinical chemistry assay values are included in the LAB_DERV_V2 data set. Gender and age specific reference ranges were supplied by the laboratory and appear in both the Visit 2 Examination manual (Manual 2) and in the Laboratory and Biospecimen Processing manual (Manual 7). See the appendix for Manual 2 for an example of clinical laboratory results and reference ranges.

### 6.4. Echocardiograhy Record (ECH)

The processing of the Visit 2 ultrasound cardiac scans for participants who are age 45 and above at Visit 2 and have not also participated in the Echo-SOL ancillary study. See Manual 17a and 17b for echocardiography procedures. Values for outliers have undergone QC checks and recoding before inclusion in the _INV3 version of the data.

**IMPORTANT ANALYSIS NOTE**: In a few cases, inconsistencies or omissions in the information required to define these variables could not be corrected on the original data forms for this interim release. When values of variables are in question, these idiosyncratic cases were adjudicated by the HCHS/SOL Coordinating Center and their resolutions are included in the derived variable files.

## 7. REFERENCES

See the HCHS/SOL Study web site for a full list of current publications in print with active PUBMED citation hyperlinks.

**Here are the two design papers for HCHS/SOL:**

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**

For a current list of published HCHS/SOL manuscripts see the study website: https://sites.cscc.unc.edu/hchs/publications-in-print