HCHS/SOL Investigator Use Database Overview for Annual Follow-Up (AFU)

April 2025

INV Version 13

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

Daniela Sotres-Alvarez
Franklyn Gonzalez II

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)
ANNUAL FOLLOW UP YEAR 1-13
Investigator Use Database
April 2025

TABLE OF CONTENTS

1. INTRODUCTION

This document describes the content and data structure of the HCHS/SOL Annual Follow up (AFU) dataset which includes 13 years of follow-up, subject to constraints (described within) to preserve participant confidentiality by de-identifying the data. Participants are called annually around the date of their baseline anniversary plus a window of six months before it closes. Baseline data was collected between March 2008 to July 2011. The first AFU started on April 2009 and the last 13th AFU was administered in January 2025 for those who were enrolled in the last year of baseline. COVID-19 waves 1, 2, and 3 questionnaires funded by C4R are also included as these were administered during annual follow up calls from May 2020 to February 2025.

2. STUDY OBJECTIVES

This multi-center observational longitudinal health study is designed to document health status in four Hispanic/Latino communities around the United States. At baseline, 16,415 adults of 18 to 74 years who self-identified as Hispanic/Latino were enrolled at four field centers over a 36-month period and are being followed annually to assess cardiovascular and pulmonary outcomes (Sorlie et. al., 2010). Pirzada et. al. (JACC, 2023) summarized the aims/objectives, the first 15 years of data collection (clinic visits, annual follow-up calls, and endpoint adjudication) and highlighted findings and contributions of the HCHS/SOL parent study and its dozens of ancillary studies to date.

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 (LaVange et. al 2010). Each of four field centers recruited approximately 4,000 persons of Hispanic/Latino origin to participate in the study. The baseline age range is 18-74, and study participants aged 45-74 years were oversampled to accrue endpoints more rapidly. Recruitment was designed to occur in stable communities so that people can be contacted and examined over time. Visit 3 screening began January 2020 at the Bronx and stopped in March 2020 due to the COVID-19 pandemic. All sites continued annual follow up calls and some V3 questionnaires over the phone during 2020 and started seeing participants at the clinic in January 2021. V3 data collection ended in January 2025.

3.1. Participants

All study participants were 18-74 years of age at screening (2008-2011), self- identified as being Hispanic/Latino, and reported not planning to move from the community during the period of follow-up. There was no exclusion 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 were not a reason for exclusion for Spanish speakers not proficient in English, since all contact with participants was done by bilingual staff in the participant's preferred language.

3.2. Schedule of Participant Data Collection for Annual Follow-up

Study participants are eligible for their annual follow-up interview five weeks before the anniversary of their baseline examination. The window for completion of the interview is open for 6 months past the anniversary date. An AFU contact status is required at the conclusion of that time for all cohort members. For example, AFU-1 interview cycle started in March 2009 and ended in December 2012. See HCHS/SOL manual 3 on Retention and Follow-up for more details on how the randomly selected waves of participant recruitment align with the year of annual follow-up interview. Table 1 lists the number of data collection forms collected during the annual follow-up telephone contacts for the first 13 years of AFU. The food frequency instrument was originally planned to be administered at baseline but was moved to the first year of AFU to shorten the baseline examination time. The CPE and CVE COVID interview forms in this data release are not linked to a specific year of AFU because they are a cross-sectional survey forms.

Table 1. Annual Follow-up Assessments Years 1-13[a]

AFU Questionnaires	**Form Code**	**Count**
COVID Psychosocial Interview Wave 1	CPE	11,137
COVID Interview Wave 1	CVE	11,350
COVID Psychosocial Interview Wave 2	CPEB	7,443
COVID Interview Wave 2	CVEB	7,104
COVID Interview Wave 3	CVEC	7,820
General Health Status*	GHE	203,341
Hospitalizations and ER Events*	HOE	175,427
Outpatient Self-reported Conditions*	OPE	168,264
Genetic Testing Awareness (AFU Y8-11)*	GTE	8,940
General Interview Questions (AFU Y7-Y9)*	GEE	8,651
Medications, ver. A (AFU Y1-Y5)*	MEE	76,887
Self-reported Events (AFU Y2-Y5)*	EVE	54,024
Other Risk Factors (AFU Y3)	OTEA	14,109
Food Propensity Questionnaire		
Food Propensity-Long, ver. A (AFU Y1)	FPEA	2,387
Food Propensity-Short, ver. B (AFU Y1)	FPEB	11,493

* Multiple records per participant.

[a] All empty/permanently missing records removed from all files including those from previous releases..

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 annual follow-up. 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.

Since forms can be revised 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. For example, GHE (general health) used the same version named GHEA for AFU-1 and 2 interviews, but AFU-3 onwards used the GHEB version and item level responses AFU-1 and 2 have been remapped to this latest version.

A codebook has been produced and it includes all data sets. 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 (PDF form) has a unique three or four-letter mnemonic associated with it (e.g., OPE for the HCHS/SOL outpatient self-reported conditions). Corresponding data sets begin with the same first three letters of the mnemonic, followed by the character string "_AFUINV13 indicating AFU data up to Year 13. The naming convention serves both to identify the originating form and provide version control when subsequent generations of datasets are produced. Previous AFU data release in March 2024 included years 1 to 11 and the extension was "_AFUINV11". In the April 2025 data release, the naming convention was changed to "_AFUINV13" to better reflect the database includes up to AFU year 13.

Where different form / versions existed, questions were renamed to match the current version. For instances in which the item did not exist on the new form, the old variable name was retained. Table 2 shows the forms that changed version and that are available in each annual follow-up and, in some instances, indicating if the form was updated (e.g., from version A to version B). For convenience, all the forms except GEE listed in the table are combined into one single pdf called AF1, AF2… AF13 that has different sections. The question numbering is not reset for each section. For example, the GHE begins with question 1 (GHEA1) and the HOE begins with question 3 (HOEA3) from the overall AFU interview form.

Table 2. AFU Form Version from Year 1 to 13.

AFU Section Title	AFU Forms	AFU-YR1	AFU-YR2	AFU-YR3	AFU-YR4	AFU-YR5	AFU-YR6	AFU-YR7	AFU-YR8 to 11	AFU-YR12 to 13
Genetic testing Awareness	**GTE/GTS**	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Ver A	n/a
General Interview Questions	**GEE/GES**	n/a	n/a	n/a	n/a	n/a	n/a	Ver A	Ver A	n/a
General Health Status	**GHE/GHS**	Ver A	Ver A	Ver B	Ver B	Ver B	Ver B	Ver B	Ver B	Ver B
Hospitalization & ER Visits	**HOE/HOS**	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A
Out-Patient Self-Reported Conditions	**OPE/OPS**	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A	Ver A
Self-Report of Events Since Baseline Visit	**EVE/EVS**	n/a	Ver A	Ver B	Ver B	Ver B	n/a	n/a	n/a	n/a
Medications	**MEE/MES**	Ver A	Ver B	Ver C	Ver C	Ver C	n/a	n/a	n/a	n/a
Other Items	**OTE/OTS**	n/a	n/a	Ver A	n/a	n/a	n/a	n/a	n/a	n/a
Place of Birth	**CBE/CBS**	n/a	n/a	Ver A	n/a	n/a	n/a	n/a	n/a	n/a
Follow-Up Interview	**CIE/CIS**	Ver A	Ver B	Ver C	Ver C	Ver C	Ver C	Ver D	Ver E	Ver E

n/a Not administered.

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, and by the use of a sequencing field for the few forms that could occur many times per visit. These items are:

ID: A random 8-digit identification code, unique to each HCHS/SOL participant.

VISIT: Contact year number for a clinic visit, a two-digit field, “01” for baseline, “02” for Visit 2, and “03” for Visit 3.

AFU_YEAR: In the AFU interview form battery this variable is used to track the series of annual contacts, 1= AFU year 1, 2= AFU year 2, etc.

OCCURRENCE: Sequential counter for multiple forms per visit, such as the HOE.

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:

VERS: 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 VER are “A”, “B”, and “C”, representing the first, second, and third versions, respectively. Most forms have only one version, but a few have a second version (MEE, FPE). For the FPE the analyst needs to merge the files carefully to use all available information from both versions.

FORM: The original 3-letter form code that appears on the paper-based forms or on the form code selection menu in the Data Management System (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 5 on the Out-patient self-reported conditions form, "Diagnosed w/ serious lung condition", is named OPE5 on the corresponding SAS file, OPE_AFUINV11.

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. HCHS/SOL ID (ID) was re-derived for use in all data sets as a random identifier code for participants. This ID is the same as the one used for all ancillary studies.

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) Addresses, phone numbers, and SSN of the participants were omitted from these files.
2) STAFF ID codes were deleted across all forms and not substituted.
3) DATES were kept unaltered and separate month, day, year text strings preserved for each item in case incomplete information was collected. AFU dates have the year preserved separately to highlight the linkage with event year whenever months and day of the month are unknown.

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
.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. When using SAS, analysts are strongly encouraged to detect missing values by using **"≤ .Z"** which will detect these special missing values rather than **"= ."**, which will not. Alternatively, the SAS missing function can be used. In the clinic visits, 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). **Note:** Any records with completely empty/permanently missing records were removed from all files including those from previous releases.

5. DESCRIPTION OF DATA COLLECTION FORMS: ANNUAL FOLLOW-UP

The Annual Follow Up Interview is broken into sections for ease of real-time online administration. Each section of the instrument has its own dataset with distinct key field structure and version control (described below). **The AFU instrument is divided into several sections (forms). All AFU years have these three forms: GHE, HOE, and OPE. And for some AFU years there are additional forms such as OTEA (only Y3), EVE (Y2 to Y5) and MEE (Y3 to Y5). The question numbering is not reset for each**

section. For example, the GHE begins with question 1 (GHEA1) and the HOE begins with question 3 (HOEA3) from the overall AFU interview form. In 2015 the AFU interviews were integrated into a new data management system (CDART2). Extensive renaming of most forms occurred at that time with data values being converted into the appropriate item values in the current version of the AFU data collection instruments for AFU-4 onwards.

5.1. General Health Status (GHE)

A check on vital status and general health starts off the AFU interview sequence. The same version of the GHEA is used for AFU-1 and 2 interviews. AFU-3 onwards used the GHEB version and item level responses AFU-1 and 2 have been remapped to this latest version. Note that participants who were not contacted and interviewed (GHEA1=5) will have missing data for the remaining AFU battery except for hospitalizations (HOE) following the protocol-based rules for administration. There is one record for each AFU year contacted. The unique combination of key fields are ID + AFU_YEAR.

5.2. Hospitalizations and Emergency Department Visits (HOE)

All first reports of visits to hospitals or emergency departments for any reason are captured here and are the basis for initiation of a request for medical records so that events can be adjudicated. This form has not changed across the years of follow-up. Multiple hospitalizations per AFU year can be reported using this multi-line form. Each hospitalization will have a separate HOE entry. The unique combination of key are ID + AFU_YEAR + OCCURRENCE.

5.3. Outpatient Self-reported Conditions (OPE)

This form asks about five key areas of health conditions and if there are any new diagnoses, worsening, or changes in therapy. The areas covered are emphysema, chronic bronchitis, chronic obstructive pulmonary disease (COPD), asthma, high blood sugar, high blood pressure, high cholesterol. There is one record for each AFU year contacted. The unique combination of key fields are ID + AFU_YEAR.

5.4. Medications (MEE)

The use of prescribed medications in the past two weeks is captured in this brief survey instrument. Up to fifteen medication names, strength and units can be coded in this part of the AFU interview. All versions of the MEE have been mapped into the current form/version layout (see version C) question numbering scheme starting in AFU-3 and ending in AFU-5. There is one record for each AFU year contacted. The unique combination of key fields are ID + AFU_YEAR.

5.5. Self-reported Events (EVE)

This section began in AFU-2 and ended in AFU-5 and asks about new physician diagnosis of several cardiac and pulmonary diseases and their associated signs and symptoms. In AFU-2 the EVE form is version A and in AFU3-5 is version B (Table 2 of section 4.2). The numbering of some questions is different between versions A and B and these were mapped into EVE file as question numbers from EVE10 to EVE22 including data for all AFU2-5. Some questions are only available in AFU2 (version A) and were not included in version B. These are available in file EVE with variable names

EVEA21 to EVEA25. There is one record for each AFU year contacted. The unique combination of key fields are ID + AFU_YEAR.

5.6. Other Risk factors (OTEA)

There are two questions on current smoking status and one question on current marital status in this brief section only collected at AFU-2.

5.7. Food Propensity (FPE versions A, B)

The food propensity questionnaire (FPQ) was adapted from NHANES (version A) to assess intake frequency of specific foods and food groups during the past 12 months. On April 2010 the FPQ was shortened by dropping 40% of individual questions to form (version B) to reduce its time of administration. Note, the FPQ is only collected at the AFU-1 interview.

5.8. Genetic Testing Awareness (GTE version 1)

The genetic testing awareness questionnaire was used in a cross-sectional administration for everyone active in AFU years 8 through 11 administered from March 2018 to April 2021. The assessment domains in this version cover awareness and/or obtainment of certain genetic tests as well as potential interest in receiving them. There is one record for each AFU year contacted. The unique combination of key fields are ID + AFU_YEAR.

5.9. General Interview Questionnaire (GEE version 1)

The general interview questionnaire was used in a cross-sectional administration for everyone active in AFU years 7, 8 and 9 administered from March 2017 to April 2018. The assessment domains in this version cover stroke symptoms in past year, vigorous/moderate physical activity from GPAQ, and cannabis use. There is one record for each AFU year contacted. The unique combination of key fields are ID + AFU_YEAR.

5.10. COVID Psychosocial Interview Questionnaires (CPE, CVE C4R Wave 1)

In response to the SARS-CoV-2 pandemic in Spring 2020, the two form COVID Psychosocial Interview battery was developed as an adjunct to the AFU interviews across the current active years of follow-up. The administration interval was from April 2020 until June 2021 for these two questionnaires which are harmonized with the C4R ancillary study wave 1 questionnaire. A subset of the questions from these forms is available in the harmonized C4R wave 1 database in Bio-Data Catalyst.

5.11. COVID Psychosocial Interview Questionnaires (CPEB, CVEB C4R Wave 2)

A second questionnaire was administered from May 2021 until February 2023 which was harmonized with the C4R ancillary study wave 2 questionnaire. It was NOT only administered to those who had wave 1. A subset of the questions from these forms is available in the harmonized C4R wave 2 database in Bio-Data Catalyst.

5.12. COVID Interview Questionnaire (CVEC C4R Wave 3)

A third questionnaire was administered from May 2023 until October 2024 which was harmonized with the C4R ancillary study wave 3 questionnaire. It was NOT only administered to those who had wave 1 or 2. A subset of the questions from these forms is available in the harmonized C4R wave 3 database in Bio-Data Catalyst.

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 (and corresponding files in this database).

6. 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**

Pirzada A, Cai J, Heiss G, Sotres-Alvarez D, Gallo LC, Youngblood ME, Avilés-Santa ML, González HM, Isasi CR, Kaplan R, Kunz J, Lash JP, Lee DJ, Llabre MM, Penedo FJ, Rodriguez CJ, Schneiderman N, Sofer T, Talavera GA, Thyagarajan B, Wassertheil-Smoller S, Daviglus ML. Evolving Science on Cardiovascular Disease Among Hispanic/Latino Adults: JACC International. J Am Coll Cardiol. 2023 Apr 18;81(15):1505-1520. PMID: 37045521