SOL
STUDY OF LATINOS
Hispanic Community Health Study

HCHS/SOL Visit 3
Investigator Use Database Overview

INV Version 2.0 (FINAL DATA)
September 2024

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

Franklyn Gonzalez II
Nicole Slye
Daniela Sotres-Alvarez
Jianwen Cai

HCHS/SOL Visit 3 Investigator Use Database
INV Version 2.0 (FINAL DATA), September 2024

Table of Contents

TRACKING TABLE OF DATA RELEASE (VERSION CONTROL)

Version	Date	Description	Documentation
_INV1	7/6/2023	Interim Data (as of 5/3/2023; N=7,458 participants)	V1.0 (May 2023)
_INV2	9/19/2024	Final Data - N=9,090 participants had in-person exam visit; 774 additional participants have only V3 phone questionnaires - Two sets of V3 sampling weights: exam participants and ALL participants - PART_DERV_V3 has 85 derived variables	V2.0 (September 2024)

1. INTRODUCTION

This document describes the content and structure of the Investigator Use datasets created for HCHS/SOL Visit 3. This final database contains data collected as part of Visit 3 from January 2020 to January 2024. This _INV2 release has data from 9,864 participants from the original cohort of 16,415 participants who participated in Visit 3. Average time between baseline and Visit 3 is 12.3±0.97 years (median time 12.2 years; range: 9.4-15.8), and between Visits 2 and 3 is 6.4±0.77 years (median time 6.3 years; range: 2.8-9.2). The content 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 3 database.

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 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, 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, 2010. 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. 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 over time.

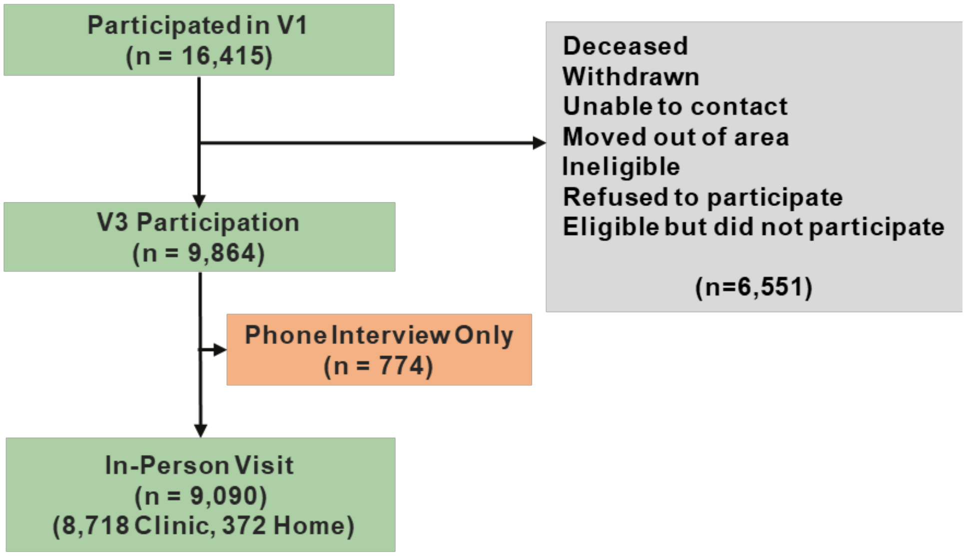
3.1. Participants

At baseline, all study participants were 18-74 years of age at screening (2008-2011), self-identified as being Hispanic/Latino, and not planning to move from the community during the period of follow-up. There was 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 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.

Visit 3 screening began January 2020, but only at the Bronx and it stopped in March 2020 due to the COVID-19 pandemic. As a result, HCHS/SOL Steering Committee modified the protocol for the main study and ancillary studies to enable remote data collection when possible. By June 2020, following IRB approval, the study began administering some visit 3 questionnaires over the phone. Additional questionnaires to evaluate the health-related, psychosocial, and economic effects of the pandemic were collected during annual follow-up (see AFU forms CPE and CVE). In-person activities resumed in early 2021 after approval from both the single and local IRBs, incorporating extensive safety modifications in line with current federal and local guidelines. See Appendix III of the HCHS/SOL Manual 2 "Field Center Procedures for Visit 3" for the modification to the protocol due to COVID-19 pandemic, including Table 8 with the questionnaires administered over the telephone. **Visit 3 data collection ended in January 2024.**

Of the original 16,415 cohort participants, a total of 9,864 cohort members participated in Visit 3 (Figure 1). Overall, 9,090 participants had an in-person visit including 372 home visits, and 774 participants only had the Visit 3 questionnaires administered over the phone. Appendix I shows that of the 7,179 participants who started Visit 3 data collection by phone, 89% (n=6,405) eventually had an in-person visit (6,226 Clinic,179 Home).

Figure 1. HCHS/SOL Visit 3 Consort Diagram.



3.2. Schedule of Participant Data at Visit 3

Table 1 lists the number of data collection forms collected during the Visit 3 (n=9,864). See section 5 for a brief description of each procedure and questionnaire form. Except for ICT (informed consent), administrative forms are not released such as AQC (anthropometry quality control), CHK (clinic checklist), ELE (eligibility screening), ETF (enrollment tracking), IDE (personal identifiers), MAE (minor adverse events), PSE (participant safety screening), RET (tracking release of results to physician or other health care), and SIB (cognitive screening). Ancillary study forms are also not included in this distribution.

Table 1. Visit 3 Data Files

Assessments	**Form codes** (English, Spanish)	Count of unique IDs
Administrative		
Informed Consent[£]	ICT	9,864
Procedures		
Anthropometry	ANT	9,076
Biospecimen Collection and Processing	BIO	9,058
Lab	LAB	9,064
Sitting Blood Pressure	SBP	9,081
Questionnaires		
Behavior and Finance	BFE/BFS	9,002
Health Care	HCE/HCS	9,039
Health Use	HUE/HUS	9,803
Personal Medical History	MHE/MHS	9,063
Neighborhood Description	NDE/NDS	9,800
Participant Disability Screener	PDE/PDS	9,089
Reproductive and Medical History	RME/RMS	5,871
Socio Economic Status/Occupation	SEE/SES	8,996
Sensitive Topics	SIE/SIS	8,982
Tobacco Use	TBE/TBS	9,062
Derived Files		
Participant derived variables	PART_DERV_V3	9,864
Medication Use	MUE_DERV_V3	9,072
Sexual Orientation and Gender Identity, Part 1*	SME_DERV_V3	8,902
Sexual Orientation and Gender Identity, Part 2*	SOE_DERV_V3	8,500

£ Multiple records per participant (16,335 observations).

* SME_DERV combined SME and SGE variables. SOE_DERV combined SOE and SGE variables. The sexual orientation and gender identity forms (SME and SOE) contain sensitive variables, and a few have cell counts < 10. Thus, these variables have been excluded from this final data release. Investigators with approved manuscripts using these variables please contact the Coordinating Center.

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 visit 3. The data values from one completed form are stored in one record in the corresponding table (observation in the SAS data set). Each data item on a form is stored as one or more columns (variables) in the data set. Collection of direct measurements during examination procedures can be entered in a form and as a result in the creation of a data file. For example, sitting blood pressure measurements are recorded on the SBP form while the technician uses the Omron HEM-907XL sphygmomanometer.

Since forms can be revised during the study, the version of the form used to collect the data is also included on each record (e.g., versions 1 or 2). 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 currently has computed outcomes and summary score values based on standard algorithms for some of the instruments (e.g. CURRENT_SMOKER_V3, CESD_V3). These algorithms are defined in the Visit 3 Derived Variable Dictionary.

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 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 3 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 "_V3_INV2" for Visit 3 Investigator Use, Version 2. For example, the Sitting Blood Pressure file name is "SBP_V3_INV2". The naming convention serves both to identify the originating form, visit and provide version control when subsequent versions of datasets are produced.

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, and using a sequencing field for the few forms that could occur more than once per visit. These items are:

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

VISIT: Contact year number, a two-digit field, "03" for Visit 3 examination.

OCCURRENCE: Form sequencing number, a two-digit sequencing number (01-99) for multiple occurrences per visit (e.g., ICT consent form as it can be changed over time). Visit 3 files have their latest consent.

4.4. Common Variables Across Data Sets

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

VERS: Version is a numeric variable indicating which version of the paper form was used to collect the data. Data files are combined across versions. In Visit 3, SIE form changed version.

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. This variable is useful to identify the language of administration ("E" for English versus "S" for Spanish).

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 medical history form MHE, "Since the last SOL visit, has a doctor said you had a heart attack", is named MHE1 on the corresponding SAS file, MHE_V3_INV2. Similarly, question 2, "How well do you speak English?", from the HUE form is named HUE2. **The exception to this numbering convention is the NIMHD questionnaire which has six forms (HUE, NDE, BFE, SIE, SME, and SOE) and variables have a consecutive number from 1 to 114 ordered from**

the 1st form (HUE) to the sixth form (SOE). For example, the first variable in the NDE form is NDE35 and not NDE1.

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 masked 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. HCHS/SOL ID (ID) is the same masked identifier used in all visits and ancillary studies.

1) Addresses, phone numbers, immigration status, date of birth, 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 and year text strings preserved for each item in case the linkage with event year whenever months and day of the month are unknown.**
4) **DATE OF BIRTH was converted to age at Visit 3 (AGE_V3) and appears in the derived variable data set, PART_DERV_V3_INV2.**

4.7. Missing Values

The study database employs a standard set of special missing value codes (see HCHS/SOL V3 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. This table describes the different 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 3 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 3 EXAM DATA COLLECTION FORMS:

5.1. Administrative Form

5.1.1. Informed Consent Checklist (ICT)

The elements of Visit 3 informed consent are tracked in this form. This file can be updated over time to reflect changes in permissions levels for use of participant data.

5.2. Procedure Forms

5.2.1. Anthropometry (ANT)

The direct anthropometric measurements are recorded on the ANT form. See Visit 3 Manual 2 for a full description of the procedures and measurements. Derived variables BMI and waist-hip ratio will be in the participant derived file (PART_DERV_V3).

5.2.2. Biospecimen Collection (BIO)

The timing and condition at the time of collection for the complete set of Visit 3 blood and urine specimens is recorded on this form. 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 and will be in the participant derived file (PART_DERV_V3).

5.2.3. 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 3 Manual 2 for a detailed description of these procedures. NOTE: ALL THREE MEASUREMENTS were AVERAGED by the OMRON sphygmomanometer. This is unlike other studies which used only the 2nd and 3rd readings.

5.3. Questionnaire Forms

The NIMHD questionnaire has six forms listed below and variables have a consecutive number from 1 to 114 ordered from the 1st form (HUE) to the sixth form (SOE). For example, the 1st variable in form NDE is NDE35 and not NDE1 had it followed HCHS/SOL numbering convention.

1. HUE – Health Use. Questions go from 1 to 34.
2. NDE - Neighborhood Description. Questions go from 35 to 49.
3. BFE – Behavior and Finance. Questions go from 50 to 67.
4. SIE – Sensitive information. Questions go from 68 to 102.
5. SME – Sexuality and Gender Identity (part 1). Questions go from 103 to 107.
6. SOE – Sexuality and Gender Identity (part 2). Questions go from 108 to 114.

Forms in this section are ordered alphabetically.

5.3.1. Behavior and Finance (BFE)

The form is part of the NIMHD questionnaire battery and collects data on: sleep disturbance, alcohol use, physical activity, financial insecurity, housing insecurity, and food insecurity. Questions are numbered from BFE50 to BFE67.

5.3.2. Health Care (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 administered during Visit 3 was expanded 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. **NOTE: Although the health care form HCE/HCS has the same name on all three visits, not all variables are the same.** Some questions were asked on all visits, but the order can be different. Analysts need to pay careful attention to map questions across visits.

5.3.3. Health Use (HUE)

This form is part of the NIMHD questionnaire battery and collects data on spoken English proficiency, interpersonal processes of care, health information seeking, patient-clinician communication through HER portals, and health literacy. Questions are numbered from HUE1 to HUE34.

5.3.4. Personal Medical History (MHE)

This form inquiries about personal medical history. Participants are asked whether a doctor has ever told them of a condition, or to provide information since the last HCHS/SOL visit, from the last telephone interview, within the last 12 months, etc. 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, and metabolic problems.

5.3.5. Medication Use (MUE_DERV)

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 a list of conditions for which medications could be prescribed.

5.3.6. Neighborhood Description (NDE)

This form is part of the NIMHD questionnaire battery and collects data on neighborhood description, stress, safety, and cohesion. Questions are numbered from NDE35 to NDE49.

5.3.7. Participant Disability Screener (PDE)

This questionnaire screens for disability and is completed after informed consent and before the participant has their examination.

5.3.8. Reproductive and Medical History (RME)

This questionnaire is administered to ALL women and has two sections: (A) hormone and menstrual history, and (B) pregnancy history (pregnancies since the last visit). **NOTE: Visit 3 RME/RMS has less questions than the same forms in Visit 2.** Analysts need to pay careful attention to map questions across visits.

5.3.9. Socio Economic/Occupation (SEE)

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

5.3.10. Sensitive Topics (SIE)

This form is part of the NIMHD questionnaire battery and collects data on cognitive function, social isolation, chronic stress, stress of immigration, ethnic discrimination, resilience, and a depression screener. Questions are numbered from 68 to 102.

5.3.11. Sexual Orientation and Gender Identity, Part 1 (SME)

This form is part of the NIMHD questionnaire battery and has the screening questions administered in person to all Visit 3 participants for the Sexual and Gender Minority (SGM) ancillary study (R01HL149778 by Poteat and Perreira). Questions are numbered from 103 to 107. THIS FORM IS NOT BEING DISTRIBUTED AS PART OF V3 FINAL DATA RELEASE. Instead, file SME_DERV is released with some variables recoded; see section 7.2.

5.3.12. Sexual Orientation and Gender Identity, Part 2 (SOE)

This form is part of NIMHD questionnaire battery and the Sexual and Gender Minority (SGM) ancillary study (R01HL149778 by Poteat and Perreira). Questions go from SOE108 to SOE114. THIS FORM IS NOT BEING DISTRIBUTED AS PART OF V3 FINAL DATA RELEASE. Instead, file SOE_DERV is released with some variables recoded; see section 7.2.

5.3.13. Tobacco Use (TBE)

The tobacco use form contains items on current and/or former use of tobacco products as well as exposure to secondhand tobacco smoke. The 17 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. **NOTE: Although the tobacco form TBE/TBS has the same name on all three visits, not all variables are the same. Some questions were asked in all visits, but the order can be different.** Analysts need to pay careful attention to map questions across visits.

6. LAB DATA

Under Visit 3 contract, the Central Laboratory was funded to perform six blood analytes for participants who fasted for at least 8 hours.

- **Total Cholesterol (LABA66)**
- **Triglycerides (LABA107)**
- **HDL-Cholesterol (LABA68)**
- **LDL-Cholesterol (LABA 69)**
- **Fasting Glucose (LABA70)**
- **Glycosylated Hemoglobin (LABA72)**

7. SPECIAL USE DERIVED FILES

7.1. Participant Derived Variables (PART_DERV_V3)

The participant derived variable dataset is not associated solely with any particular form because it contains variables derived from many forms and files. There is one record per enrolled participant (9,864) at Visit 3 in PART_DERV_V3_INV2. This file is a cross-section of “derived variables” whose values are defined based on combinations of data items. See *“HCHS/SOL Visit 3 Derived Variable Dictionary”* for the definitions.

7.2. Combined SME/SOE and SGE Variables (SME_DERV_V3 & SOE_DERV_V3)

The first version of the Sexual Orientation and Gender Identity (SGE/SGS) form was replaced by the SME/SMS and SOE/SOS forms on 02/18/2021. The following table provides a concordance between variable names in the SGE (version 1, 3/3/2020) and SME and SOE forms. To create common variables across the two forms, we created new variables using the new form name. A few variables with some cell counts <10 are not released but descriptive statistics are included in the codebook. Investigators with approved manuscripts using these variables please contact the Coordinating Center.

Concept	SGE	SME/SOE	Recoded Variable
Sex assigned at birth	SGE103	SME103	SME103R
Gender identity	--	SME104	SME104
Intersex/DSD	--	SME106	SME106
Same Sex Attraction	--	SME107	SME107
Sexual orientation identity	SGE104	SOE110	SOE110R
Transgender identity	SGE105	SOE111	SOE111R
Sexual Behavior Past 12 months	SGE106	SOE112	SOE112R
Sexual Attraction	SGE107	SOE108	SOE108R
Romantic Partners	--	SOE109	SOE109

Note: SGE105 (age first identify with current gender) is a follow-up question to SGE104.

It is recommended to consult with Drs. Krista Perreira and Tonia Poteat on the correct use of terms and the appropriate analyses for variables about sexuality and gender identity. They have the expertise and were the original investigators funded to collect and derive variables as part of the SGM SOL ancillary study (R01HL149778).

8. REFERENCES

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. PMID: 20609344; PMCID: PMC2921622.

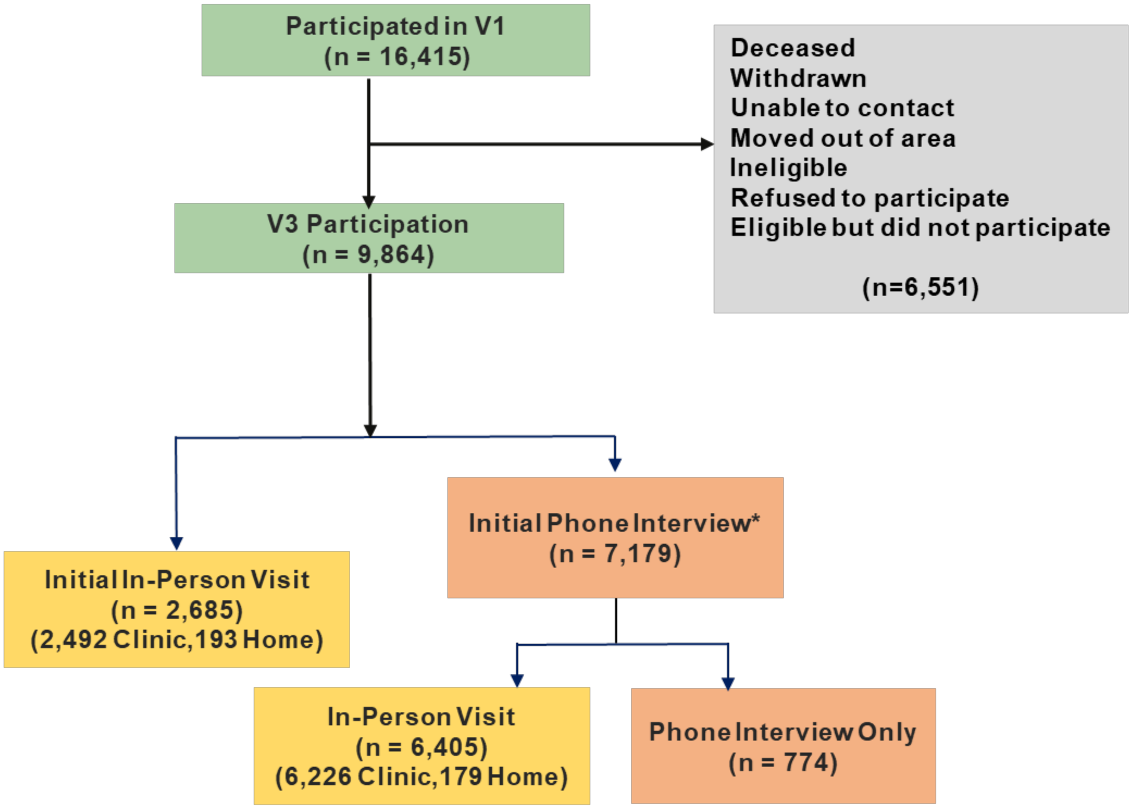
Pirzada A, Cai J, Heiss G, Sotres-Alvarez D, Gallo LC, Youngblood ME, Avilés-Santa ML, González HM, Isasi CR, Kaplan R, Kunz J, Lash JP, Lee DJ, Llabre MM, Penedo FJ, Rodriguez CJ, Schneiderman N, Sofer T, Talavera GA, Thyagarajan B, Wassertheil-Smoller S, Daviglus ML. Evolving Science on Cardiovascular Disease Among Hispanic/Latino Adults: JACC International. J Am Coll Cardiol. 2023 Apr 18;81(15):1505-1520. PMID: 37045521; NIHMSID: NIHMS1877960.

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. PMID: 20609343; PMCID: PMC2904957.

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

APPENDIX I.

Appendix I. HCHS/SOL Visit 3 Consort Diagram Differentiating Initial V3 Type (Phone or In-Person) and Differentiating In-Person Visit (Clinic or Home)



*Phone interviews were conducted for Visit 3 when clinics were closed to participants due to COVID-19.