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

Investigator Use Incident Events 2008-2021 Database Overview

Prepared by the Collaborative Studies Coordinating Center
Version 1.0, April 2025

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)
Investigator Use Incident Events 2008-2021 Database
Version 1.0 April 2025

Table of Contents

1. INTRODUCTION

This document describes the content and structure of the Investigator Use datasets created for HCHS/SOL incident events. This database contains event eligibility for each qualifying emergency department and/or hospital admission, medical records abstraction of those HCHS/SOL suspected primary events (stroke, myocardial infarction, heart failure), and a derived event classification data file for the 16,415 cohort members during the calendar years 2008-2021 for MI, HF, and stroke and 2008-2018 for pulmonary events. The Annual Follow-Up (AFU) reported hospitalization and emergency department visits are closed each year and the transferred records are abstracted and sent to endpoint reviewers. Cumulative data from incident events are released when there have been at least ~95% endpoints adjudicated in a calendar year. Because pregnancy complications (gestational diabetes/pre-eclampsia/eclampsia) were only investigated and abstracted for pregnancies from baseline to visit 2, these files are static over time (i.e., are NO longer accumulating new events). Nevertheless, these files are described in this document and part of HCHS/SOL incident events database.

The source data from interviews and medical records has been redacted and some elements transformed to preserve participant confidentiality by de-identifying the data.

Note: These are only hospitalization-based events (MI, HF, Pul, Str, PRC, based on medical record obtainment, review, etc.) and does NOT include any fatal events.

2. STUDY OBJECTIVES

The aims of this multi-center observational longitudinal study, HCHS/SOL, are to identify the prevalence of and risk/protective factors for cardiovascular disease (CVD) and pulmonary disease in four diverse Hispanic/Latino communities in the United States; to determine the role of acculturation in their prevalence and development; and to quantify all-cause mortality, fatal and nonfatal CVD and pulmonary disease, and exacerbation of asthma and COPD over time (primary endpoints). Overall, 16,415 adults of 18 to 74 years were enrolled at four field centers from 2008-2011 at the baseline visit and have been followed annually since that initial visit to assess health 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). Ultimately, about 16,000 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 adults of Hispanic/Latino origin to participate in the study. The age range at baseline was 18-74 years. Study participants were selected to obtain approximately 2,500 persons age 45-74, and approximately 1,500 persons age 18-44 within each field center. Recruitment was designed to occur in stable communities so that people can be contacted over time and examined more than once.

3.1. Participants

All study participants were 18-74 years of age at screening, self-identified as being Hispanic/Latino, and not planning to move from the community during the period of follow-up. The recruited individuals attended an examination to assess cardiovascular and other disease risk factors, both known and potential. There was no exclusion of persons based on existing health status, but the following persons were not recruited: 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 is done by bilingual staff in the participant's preferred language.

Table 1. Assessment Battery

Endpoints Processing and Abstraction Forms	Form Code	Count
Event Eligibility*	EEF	7,420
MI Record Abstraction	MIF	2,278
Heart Failure Record Abstraction	HTF	1,134
Pregnancy Complication Record Abstraction	PRC	127
Pulmonary Record Abstraction	PUL	2,372
Stroke Record Abstraction	STR	452
Adjudication Forms		
Heart Failure Reviewer 1	HFD1	1,123
Heart Failure Reviewer 2	HFD2	1,123
Heart Failure Reviewer 3	HFD3	473
Pregnancy Complication Reviewer 1	PCD1	127
Pregnancy Complication Reviewer 2	PCD2	127
Pregnancy Complication Reviewer 3	PCD3	124
Myocardial Infarction Reviewer 1	MID1	2,238
Myocardial Infarction Reviewer 2	MID2	2,238
Myocardial Infarction Reviewer 3	MID3	564
Pulmonary Reviewer 1	PLD1	2,333
Pulmonary Reviewer 2	PLD2	2,329
Pulmonary Reviewer 3	PLD3	468
Stroke Reviewer 1	STD1	450
Stroke Reviewer 2	STD2	450
Stroke Reviewer 3	STD3	66
Derived Variable Files		
Events Participant Derived 2008-2021	n/a	16,415
Incident HF Events 2008-2021	n/a	16,415
Incident MI Events 2008-2021	n/a	16,415
Incident Pulmonary Events 2008-2018	n/a	16,415
Incident Stroke Events 2008-2021	n/a	16,415

* Multiple records per participant.
Pregnancy complications were only identified and abstracted between 2008 and 2018. These files do NOT change as calendar years are added as these are no longer endpoints.

3.2. Schedule of Participant Events Classification Data

Table 1 lists the data abstraction and adjudication forms used in evaluating potential events for endpoint classification, medical chart abstraction, and the derived variable files from the resulting reviewer adjudications in the endpoint areas for myocardial infarction, heart failure, and chronic lower respiratory disease. Each event is reviewed independently by two reviewers and disagreements are adjudicated by a third reviewer.

4. DATABASE STRUCTURE

4.1. Data Set Organization

There is one table (SAS data set) in the database for each type of data abstraction form collected at annual follow-up calls, or adjudication forms used by endpoints reviewers. 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.

Since forms can be revised during 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.

Special derived variable datasets have been created to augment the original data measurement values. The incident events file has computed follow-up times to incident and/or recurrent events. Incident MI during the period 2008-2021 (INCIDENT_MI_2021) is an example of the type of derived variables included in the data release. These algorithms are defined in the outcome events 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 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 four-letter mnemonic associated with it (e.g., EEFB for the HCHS/SOL Event Eligibility Form, Version B). Corresponding data sets begin with the same first three letters of the mnemonic, followed by the character string "_INV1 for Investigator Use Version, Version 1. For example, the Events Eligibility data set for 2008-2021 calendar years in release 1 is "EEF_2008_2021_INV1". The naming convention serves both to identify the originating form, event period, 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 many times per visit (like the Event Eligibility EEF). 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, "01" for baseline examination year.
3) OCCURRENCE: Form sequence number, a two-digit sequencing number (01-99) for multiple forms per visit (e.g. see EEF where two or more events have been abstracted).
4) EVENT_ID: A concatenation of ID | Visit | Occurrence from the originating HOE/HOS form used in AFU to report the event. When events are reviewed, they are tracked by EVENT_ID and can be reviewed for more than one outcome.

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:

5) 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. Most forms have only one version, but a few have more than one version. Possible values for VERSION are “A”, “B”, and “C”, or "1", "2", and "3", representing the first, second, and third versions, respectively.

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 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 1 on the MI Abstraction form (MIF), "source type for event", is named MIF1 on the corresponding SAS file, MIF_2008_2021_INV1. Similarly, question 4, "Primary admitting diagnosis code", from the “A” version MI abstraction form is named MIF4.

The reviewer diagnosis forms use a variation on the numbering scheme because they are routinely used in duplicate for paired reviews, or on occasion triplicate when an third review for adjudication occurs. For example, the MI diagnosis forms are records MID1, MID2, MID3 which contain three identical schema. The reviewer sequence number immediately precedes the item number for all the diagnosis forms described below. In this example, question #1 from MI review form 1 is variable “MID11” and question #11 from that form is “MID111”.

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.

1) A HCHS/SOL ID (ID) was re-derived for use in all data sets as a random identifier code for participants.
2) Addresses, phone numbers, and SSN of the participants were omitted from these files.
3) 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, but removed from the ID string.
4) STAFF ID codes were deleted across all forms and not substituted.
5) DATES were transformed into follow-up time since the baseline examination visit, or restricted to month and year of the event.
6) DATE OF BIRTH was converted to age at the end of 2021 or the censoring event.

5. DESCRIPTION OF DATA COLLECTION FORMS / DATABASE TABLES

5.1. Event Eligibility (EEF)

The EEF is used to perform a preliminary abstraction from de-identified medical records submitted to the HCHS/SOL Coordinating Center. Trained clinical staff use the case medical charts to extract and record the ICD-9 and/or ICD-10 codes and keywords used in the discharge summary and treatment report. A computer algorithm based on the ICD codes and qualifying keywords used in the medical chart determines the outcome area(s) covered by event and if the case requires adjudication, or no further abstraction. Each hospitalization admission is linked to one reported event evaluated for review using the EEF and the qualifying algorithms for processing. Consequently, the EEF will have multiple records per participant organized by event ID and admission date (EEF0D). However, in the endpoint INV release file, date variable was converted as follow-up time. See the HCHS/SOL Manual 15 on Endpoints Ascertainment for more information on those algorithms which is located on the study website.

5.2. Heart Failure Abstraction (HTF)

Medical records for cases eligible for heart failure review are abstracted by trained and certified clinical staff at HCHS/SOL Coordinating Center. See HCHS/SOL Manual 15 on Endpoints and Outcomes Ascertainment for a description of procedures and the heart failure abstraction form. Each hospitalization evaluated for HF has one record per EVENT_ID.

5.3. Myocardial Infarction Abstraction (MIF)

The medical records for cases eligible for myocardial infarction review are abstracted by trained and certified clinical staff at HCHS Coordinating Center in this area. See HCHS Manual 15 on Endpoints and Outcomes Ascertainment for a description of procedures and a copy of the MI abstraction form. Each hospitalization evaluated for MI has one record per EVENT_ID.

5.4. Pulmonary Abstraction (PUL)

Medical records for events eligible for pulmonary review are abstracted by trained and certified clinical staff at HCHS Coordinating Center in this area. See HCHS Manual 15 on Endpoints and Outcomes Ascertainment for a description of procedures and a copy of the pulmonary outcomes abstraction form. Each hospitalization evaluated for pulmonary related events has one record per EVENT_ID.

5.5. Stroke Abstraction (STR)

Medical records for events eligible for stroke review are abstracted by trained and certified clinical staff at HCHS Coordinating Center in this area. See HCHS Manual 15 on Endpoints and Outcomes Ascertainment for a description of procedures and a copy of the stroke outcome abstraction form. Each hospitalization evaluated for stroke events has one record per EVENT_ID.

5.6. Pregnancy Related Complications (PRC)

Medical records for events eligible for gestational diabetes/pre-eclampsia/eclampsia review are abstracted by trained and certified clinical staff at HCHS Coordinating Center in this area. See HCHS Manual 15 on Endpoints and Outcomes Ascertainment for a description of procedures and a copy of the pregnancy complications outcome

abstraction form. Each hospitalization evaluated for pregnancy related events has one record per EVENT_ID.

5.7. Myocardial Infarct Diagnosis (MID1, MID2, MID3)

There are up to three MI diagnosis records for each event reviewed for MI adjudication. Separate records represent the diagnosis from reviewer #1 (MID1) and reviewer #2 (MID2). If there is a disagreement on whether an event occurred, then a third review for adjudication is present (MID3). Each hospitalization admission evaluated for MI related events has 2 or 3 records per EVENT_ID for an admission date. Item numbering on the diagnosis forms include the record number so questions 1 to 6 of the MID form for reviewer #1 are named MID11 to MID16.

5.8. Heart Failure Diagnosis (HFD1, HFD2, HFD3)

There are up to three heart failure diagnosis records for each event reviewed for adjudication. Separate records represent the diagnosis from reviewer #1 (HFD1) and reviewer #2 (HFD2). If there is a disagreement on whether an event occurred, then a third review for adjudication is present (HFD3). Each hospitalization evaluated for heart failure related events has 2 or 3 records per EVENT_ID for an admission date. Item numbering on the diagnosis forms include the record number so questions 1 to 6 of the HFD form for reviewer #1 are named HFD11 to HFD16.

5.9. Pregnancy Complication Diagnosis (PCD1, PCD2, PCD3)

There are up to three pregnancy complication diagnosis records for each event reviewed for adjudication from AFU hospitalization admissions for pregnancy related causes. Separate records represent the diagnosis from reviewer #1 (PCD1) and reviewer #2 (PCD2). If there is a disagreement on whether an event occurred, then a third review for adjudication is present, PCD3, for an admission date). Item numbering on the diagnosis forms include the record number so questions 1 to 6 of the PCD form for reviewer #1 are named PCD11 to PCD16.

5.10. Pulmonary Diagnosis (PLD1, PLD2, PLD3)

There are up to three pulmonary event diagnosis files for each event reviewed for adjudication. Separate records represent the diagnosis from reviewer #1 (PLD1) and reviewer #2 (PLD2). If there is a disagreement on whether an event occurred, then a third review for adjudication is present (PLD3). Each hospitalization evaluated for asthma or COPD related events has 2 or 3 records per EVENT_ID for an admission date. Item numbering on the diagnosis forms include the record number so questions 1 to 6 of the PLD form for reviewer #1 are named PLD11 to PLD16.

5.11. Stroke Diagnosis (STD1, STD2, STD3)

There are up to three HF diagnosis files for each event reviewed for stroke adjudication. Separate records represent the diagnosis from reviewer #1 (STD1) and reviewer #2 (STD2). If there is a disagreement on whether an event occurred, then a third review for adjudication is present (STD3). Each hospitalization evaluated for stroke related events has 2 or 3 records per EVENT_ID for an admission date. Item numbering on the diagnosis forms include the record number so questions 1 to 6 of the STD form for reviewer #1 are named STD11 to STD16.

6. SPECIAL USE DERIVED FILES

The participant derived variable data sets are not associated solely with a specific form because they contain variables derived from many forms.

6.1. Derived Events Variables (EVENT_PART_DERV_2008_2021)

There is one record per enrolled HCHS/SOL cohort participant (16,415 observations) in EVENT_PART_DERV_2008_2021_INV1. This file is a collection of derived incident event variables whose values are defined based on combinations of data items (e.g., baseline data, date of admission for an event, withdrawal status, adjudication summary indicators), primarily from the myocardial infarct, heart failure, and pulmonary reviewer records. Both incident event and recurrent event indicators and follow-up times are provided for classified outcomes. In case composite endpoints need to be created by analysts utilizing reports of deaths from all causes, indicator variables for death and withdrawal from the study, along with their respective follow-up times, are included. Note, the death certificate form (DTH) will appear in a later data release once the death classification committee has had an opportunity to work with this data and has been combined and harmonized with results from the National Death Index (NDI) searches. See, “HCHS Incident Events 2008-2021 Derived Variable Dictionary” document for the definitions of the variables included in this special purpose file.

6.2. Incident Heart Failure Events (INCIDENT_HF_EVNT_2008_2021)

For participants reviewed and classified for heart failure this derived file includes the incident and recurrent indicator variables and related follow-up times. It has a short format structure, meaning there is only record per participant. Because there can be up to five events per participant, these are all in the same record with variable names differentiated with a consecutive number at the end. For example, for recurrent heart failure events there are five variables, RECUR_HF_2021_1 to RECUR_HF_2021_5.

6.3. Incident Myocardial Infarction Events (INCIDENT_MI_EVNT_2008_2021)

For participants reviewed and classified for myocardial infarctions this derived file includes the incident and recurrent indicator variables and related follow-up times. It has a short format structure, meaning there is only record per participant.

6.4. Incident Pulmonary Events (INCIDENT_PUL_EVNT_2008_2018)

For participants reviewed and classified for chronic lower respiratory disease and the outcomes of asthma or COPD this derived file includes the incident and recurrent indicator variables and related follow-up times. It has a short format structure, meaning there is only record per participant.

6.5. Incident Stroke Events (INCIDENT_STR_EVNT_2008_2021)

For participants reviewed and classified for ischemic or hemorrhagic stroke this derived file includes the incident and recurrent indicator variables and related follow-up times.

Transient ischemic attacks (TIA) are not an adjudicated outcome in HCHS/SOL. It has a short format structure, meaning there is only record per participant.

6.6. Pregnancy Complications Events (HCHS_INCIDENT_PRC_EVNT_ALL)

This outcome area was active only during the Visit 2 for events reported in the period 2008-2018 through annual follow-up reported hospitalizations and is now closed to further adjudications. Admissions were reviewed for the possible complications of gestational diabetes or pre-eclampsia/ eclampsia. Use of the Visit 2 exam related women's medical history (RME) and pregnancy complications history (PCE) may be needed to interpret the life course context of these events. It has a short format structure, meaning there is only record per participant.

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). These idiosyncratic cases were adjudicated by the HCHS/SOL Coordinating Center and their resolutions are included in the derived variable files.

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.

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. doi: 10.1016/j.jacc.2023.02.023. PMID: 37045521