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

Investigator Use Incident Events 2008-2015 Database Overview

Prepared by the Collaborative Studies Coordinating Center
Version 1.0, June 2021

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)
Investigator Use Incident Events 2008-2015 Database
Version 1.0 June 2021

1. INTRODUCTION

This document describes the content and structure of the Investigator Use datasets created for HCHS/SOL endpoints files related to incident events. This database contains event eligibility for each qualifying emergency department and/or hospital admission, medical records abstraction of those suspected events, and derived event classification data for the 16,415 cohort members for the calendar years 2008-2015. The AFU reported hospitalization and emergency department visits are closed each year and the transferred records are abstracted and sent to outcome reviewers who are trained in the HCHS/SOL classification for study outcomes. The source data from interviews and medical records has been redacted and some elements transformed to preserve participant confidentiality by de-identifying the data.

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. Overall 16,415 adults of 18 to 74 years, were enrolled at four field centers over a 36 month period from 2008-2011 at the baseline visit and have been followed since that initial visit to assess health outcomes (see Sorlie et.al, 2010).

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 4,000 persons of Hispanic origin to participate in the study. The age range at baseline was 18-74. Study participants were 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 examined more than once. Electronic copies of the study protocol and manuals of operation are also included elsewhere for reference with this data release.

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 attend an 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 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 are not a reason for exclusion for Spanish speakers not proficient in English, since all contact with participants is done using the appropriate language.

3.2. Schedule of Participant Events Classification Data

Table 1 lists the endpoints data collection forms used in evaluating potential events for outcome classification, medical chart abstraction, and the derived variable files from the resulting reviewer adjudications in the outcome 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. Individual reviewer forms and closed event files for stroke and pregnancy related complications will be released at a later date in 2021.

Table 1. Assessment Battery

Endpoints Processing and Adjudication Forms	Form Code	Count
Adjudication Forms		
Event Eligibility	EEF	5,065
MI Record Abstraction	MIF	1,340
Heart Failure Record Abstraction	HTF	629
Pulmonary Record Abstraction	PUL	1,582
Derived Variable Files		
Events Participants Derived 2008-15	n/a	16,415
Incident HF Events 2008-15	n/a	127
Incident MI Events 2008-15	n/a	97
Incident Pulmonary Events 2008-15	n/a	196

4. DATABASE STRUCTURE

4.1. Data Set Organization

There is one table (SAS data set) in the database for each type of data collection form (provided as PDFs). The data values from one completed paper form are stored in one record in the corresponding table (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.

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 2008-2015 (INCIDENT_MI_2015) is an example of the type of derived variables included in the data release. These algorithms have been included in the outcome events 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 (PDF form) has a unique four-letter mnemonic associated with it (e.g., EEFA for the HCHS/SOL Event Eligibility Form, Version A). 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 2008-2015 release 1 is "EEF_2008_2015_INV1". The naming convention serves both to identify the originating form, event period, and provide version control when subsequent generations 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 (see HCHS/SOL Data Management Guide), and by the use of 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 using 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 a second 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 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, "source type for event", is named MIF1 on the corresponding SAS file, MIF_2008_2015_INV1.

Similarly, question 4, "Primary admitting diagnosis code", from the “A” version MI abstraction form is named MIF4.

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 2015 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 person organized by event ID and admission date. See the HCHS/SOL Manual 15 on Endpoints Ascertainment for more information on those algorithms which is located on the study web site.

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.

6. SPECIAL USE DERIVED FILES

6.1. Derived Events Variables (EVENT_PART_DERV_2008_2015)

The participant derived variable data sets are not associated solely with a specific form because they contain variables derived from many forms. There is one record per enrolled HCHS/SOL cohort participant (16,415 observations) at baseline in EVENT_PART_DERV_2008_2015_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. Indictor variables for death and withdrawal from the study and the related follow-up times are included in case composite endpoints need to be created by analysts that make use of reports of deaths from all causes. See, “HCHS Incident Events 2008-2015 Derived Variable Dictionary” document for the definitions of the variables included in this special purpose file.

6.2. Incident Heart Failure Events (INCIDENT_HF_EVENT_2008_2015)

For participants reviewed and classified for heart failure this derived file includes the incident and recurrent indicator variables and related follow-up times.

6.3. Incident Myocardial Infarction Events (INCIDENT_MI_EVENT_2008_2015)

For participants reviewed and classified for myocardial infarctions this derived file includes the incident and recurrent indicator variables and related follow-up times.

6.4. Incident Pulmonary Events (INCIDNET_PUL_EVENT_2008_2015)

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.

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.