The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)
Study of Latinos Investigation of Neurocognitive Aging Ancillary Study

Investigator Use Database Overview

Study Website: http://www.cscc.unc.edu/hchs/

Study Email: hchsadministration@unc.edu

SOL-INCA Analytics Team

Date: 2020-07-29

Contents

DATA RELEASE

This document accompanies the release of the **inca_mci_derv_inv3a** dataset distributed in July of 2020. Information about previously distributed datasets and the corresponding codebooks is available at http://www.cscc.unc.edu/hchs/.

INTRODUCTION

The Hispanic Community Health Study/ Study of Latinos (HCHS/SOL) Investigation of Neurocognitive Aging (SOL-INCA) Ancillary Study (AS) builds on the HCHS/SOL cohort to provide new and important public health knowledge about cognitive health and aging. Data for this AS were collected concurrently with HCHS/SOL examination Visit 2 (V2). Participants were administered a brief battery of neurocognitive tests that provided new information on cognitive health and decline. A limited number of design and demographic variables from the HCHS/SOL main study are included in the **inca_mci_derv_inv3a** dataset. Other variables from the main study can be obtained through the coordinating center at the University of North Carolina, Chapel Hill.

SOL-INCA repeated the cognitive tests administered at Visit 1 (V1), which included the Six-Item Screener (SIS), the Brief-Spanish English Verbal Learning Test (B-SEVLT), the Word Fluency Test (WFT), and the Digit Symbol Substitution (DSS). New questionnaires were introduced in SOL-INCA to provide sufficient information to yield research diagnoses of Mild Cognitive Impairment (MCI) (1, 2). The two new questionnaires concerned self-reported cognitive decline (eCog12) and instrumental activities of daily activity (IADL). To improve diagnostic specificity and control for educational bias, the NIH ToolBox Picture Vocabulary Test was administered. Additionally, tests were selected if they had been used in other NHLBI cohort studies (e.g., ARIC-NCS), to afford opportunities for cross-cohort studies. These quick and simple neurocognitive tests will give scientists important information to reduce the risk of cognitive decline and impairment.

HCHS/SOL SOL-INCA OBJECTIVES

HCHS/SOL is a multisite, prospective, observational cohort study of over 16,415 Hispanic/Latino persons aged 18-74 years in four Hispanic communities (Bronx, Chicago, Miami, and San Diego) in the United States. HCHS/SOL V2 includes 11,623 Hispanic/Latino persons followed on average 6 years later. SOL-INCA focuses on a subset of participants 50-years and older who participated in the neurocognitive module at the HCHS/SOL baseline. The aim of HCHS/SOL is to obtain longitudinal measures of cardiovascular function, metabolic status, and other health conditions. SOL-INCA builds on the design of the HCHS-SOL parent study and focuses on collecting longitudinal measures of cognitive aging.

STUDY DESIGN

Electronic copies of the SOL-INCA AS study protocol and manuals of operation for the HCHS-SOL main study, the SOL-INCA AS, and other ancillary studies associated with HCHS-SOL are available online at http://www.cscc.unc.edu/hchs/. Detailed descriptions of the framework underlying SOL-INCA and the methodology underlying the derivation of its main outcomes of interest have been published (1, 2).

Participants

Participants were classified as eligible for SOL-INCA if they completed neurocognitive testing during HCHS/SOL Visit 1 and were 50 years or older when completing the cognitive interview for SOL-INCA. A total of 7,168 individuals were screened from which 6,377 participated in the study and completed cognitive assessments. Among individuals who did not participate (n=791), 574 were eligible but refused to enroll in the study while the remaining 217 were not able to complete an assessment before recruitment for SOL-INCA concluded. Among the 6,377 participants, 2,166 completed an assessment on the same day as the HCHS/SOL V2 examination and 4,211 completed an assessment on a different day.

DATABASE STRUCTURE

Data Set Organization

Data from SOL-INCA forms are provided in separate datasets. Data values pertaining to each participant based on the completed forms are stored in one record and comprise a single observation (row). Each question included in the form is stored as a variable (column). A special derived variable dataset (**inca_mci_derv_inv3a**) was generated to augment the original data measurement values. The derived variable dataset included computed scores and values based on study specific algorithms generated to satisfy SOL-INCA aims (e.g. MCI).

Codebooks have been produced for all datasets. A careful review of the codebooks, in conjunction with the forms, is critical to understanding and interpreting the source data. The codebooks provide a description of every variable in the dataset as well as simple descriptive statistics and labels of variables. Analysts are strongly encouraged to use the codebook, paying attention to the data user notes contained in this document.

Forms and Data Set Naming Conventions

Each SOL-INCA data collection instrument (see PDF forms) has a unique three-letter mnemonic associated with it (e.g., DLE). The corresponding datasets begin with the same first three letters of the mnemonic for the English version, followed by the character string "INCA", for Investigation of Neurocognitive Aging (SOL-INCA) Investigator Use, Version 3 (e.g. "dle_inv3.sas7bdat"). The naming convention serves both to identify the originating form and provide version control when subsequent datasets are produced. Note, since the questionnaire battery for the ancillary study has both English and Spanish language

versions of the forms each has been merged into one common data record format which follows the main HCHS/SOL study conventions. Language of administration is retained as a data element of each record.

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. These items are:

1. HCHSID: A random 8-digit masked identification code, unique to each HCHS/SOL participant. The HCHS/SOL Coordinating Center created a unique ID for this ancillary study which is different purposely to the HCHS/SOL Study ID.
2. ID: Main Study Investigator Use ID, which is eight digits (including leading zeros) which can be used to merge with the HCHS main study INV3 series data.
3. VISIT: Contact year number, a two digit field, "02" for the examination year.

Variable Naming Conventions

While the key fields have the same name on each SAS record type, 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 dataset name, followed by the form version letter, and then the question number as indicated on the form.

For example, question 5 on the INCA Eligibility and Recruitment Screening form IER, "gender", is named IER5 on the corresponding SAS file. Similarly, question 2, "Remembering the current date or day of the week", from the everyday cognition (CGE) form is names CGE2.

Changes to Variables to Preserve Confidentiality

As part of the HCHS/SOL and SOL-INCA commitment to comply with HIPAA regulations for participant confidentiality and to follow current guidelines from NHLBI/NIH the HCHS/SOL Coordinating Center has made explicit modifications and/or deletions to some variables that could jeopardize respondents privacy across all datasets.

All participant ID values were transformed from the original ID to random identifier codes to produce Investigator Use datasets that protect the confidentiality of the individual. Analyst will only have access to de-identified datasets. For more detailed information about the rules and regulations governing data sharing and security users are advised to consult the HCHS/SOL manuals created by the coordinating center at the University of North Carolina, Chapel Hill.

It is important to note that even though shared datasets are de-identified, users have to comply with all language identified in their signed Data Use Agreements (DUAs) with the Coordinating Center or internally with their institution's PI. All authorized users are responsible for actively attending and ensuring to the security and confidentiality of these Investigator Use files as part of their DUAs.

Missing Values

The HCHS/SOL study database employs a standard set of special missing value codes (see study codebook) that have contextual meaning.

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

All missing values in SOL-INCA as are left as empty ".".

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.

DESCRIPTION OF DATA COLLECTION FORMS

Eligibility and Recruitment Form (IER)

Includes detailed information pertaining to study completion date, participants enrollment status, and dates and times, as well as other administrative and processing information deemed safe to provide context for data collection but preserve participant confidentiality.

The IER also includes limited demographic information on study participants (e.g. age, gender, and education).

Everyday Cognition-12 (CGE)

The eCog is a brief questionnaire that asks participants to rate any changes in memory and thinking over the past 10 years. It is very important that the participant understands that he/she is comparing his/her current cognitive functioning with their functioning 10 years ago. Following the script, the examiner briefly explains the eCog 10-year comparison purpose and then proceeds with the instructions of the actual eCog test.

Activities of Daily Living (DLE)

The IADL questionnaire provides a tool to assess overall individual functioning and can be used for both clinical purposes and in large community dwelling population surveys.

There are 7 IADL questions that are used to assess a participant's ability to function independently. In response to each question, study participants will report their ability to perform the specific task independently (i.e., without help), with some help, or not (able to perform the task) at all.

Neurocognitive Scoring Summary (NEE)

The NEE includes the battery of objective cognitive tests administered to study participants. A brief description of these tests is provided below.

Brief Spanish English Verbal Learning Test (B-SEVLT)

The B-SEVLT is a measure of learning and verbal memory. The participant is asked to recall a list of 15 household items over three learning trials. Recall of the words occurs after a short delay, during which a new (distracter) list of words is presented.

Word Fluency

The Word Fluency Test is a measure of verbal fluency and functioning. The participants are asked to produce as many words as possible that begin with the letters F and A within a time limit of 60 seconds for each letter, avoiding proper nouns (e.g., David or Dallas), variations (dance, dancing, danced), plurals (e.g., dances), numbers (e.g. four, five, etc.) and repetitions. English or Spanish words are acceptable.

Digit Symbol Substitution test

The Digit Symbol Substitution (DSS) test (also referred to interchangeably as DSST) is a measure of psychomotor speed and sustained attention. In this task, the participant is asked to translate numbers (1-9) to symbols using a key provided at the top of the test form.

Trails Making Test A and B

The TMT is a timed task in which participants connect letters and numbers in sequence as quickly as possible. TMT measures attention, sequencing, mental flexibility, and visual search and motor function. In TMT A, the participant is asked to draw a line and connect a series of numbers (from 1-25) as quickly as possible. In TMT B, the participant is asked to draw a line and connect a series of numbers and letters, alternating between a given number and letter (e.g., 1 to A, A to 2, 2 to B, B to 3, etc.) as quickly as possible. Prior to each test, the participant is given a sample test to demonstrate the task. Please note that TMT A and TMT B were not administered at V1.

Six Item Screener (SIX)

The Six-Item Screener (SIS or SIX) is derived from the Mini-Mental Status exam, which is a short mental status test that is used by many doctors and in several studies. Following the script on the paper form, the examiner briefly explains the purpose of the cognitive function portion of the HCHS/SOL examination and then proceeds with the instructions for the six-item screener.

Picture Vocabulary Test (PVT)

The PVT is a measure of receptive vocabulary that is administered in a computerized adaptive format (3). That is, the next question a participant receives depends on his/her response to the previous question. Computer Adaptive Testing (CAT) ensures a test that is tailored to the participant's needs. The respondent is presented with an audio recording of a word and four photographic images on the computer screen and is asked to select (click on) the picture that most closely matches the meaning of the word. This test takes

approximately four minutes to administer and is recommended for ages 3-85.

The PVT is part of the National Institute of Health Toolbox and was administered on a dedicated computer system. Field center staff were responsible to enter participants' age and education information when they administered the PVT. The electronic data generated by the PVT application includes three subsets of variables.

1. Registration Variables File (PVR)

Section	Name	N's
PVR2	Consent	6111
PVR9	DteApprch	6111
PVR11	RegDte	6111
PVR12	Baseline	6111
PVR27	StdyArm	6111
PVR28	Schdl	6111

2. Scores and Demographics Variable File (PVS)

Section	Name	N's
PVS		
PVS1	Age	6111
PVS5	Language	6111
PVS6	Education	6103
PVS8	Assmnt	6111
PVS9	Form	6111
PVS13	SE	6111
PVS15	Consent	6111
PVS21	Computed Score	6111
PVS22	Unadjusted Scale Score	6106
PVS23	Age Adjusted Scale Score	6106
PVS24	National Percentile (age adjusted)	6106
PVS25	Fully Adjusted Scale Score	masked

3. Assessment Variable File (PVA)

Section	Name	N's
PVA2	Assmnt	masked
PVA3	MdlOrdr	masked
PVA4	InstOrdr	masked
PVA5	InstSctn	masked
PVA6	ItmOrdr	masked
PVA7	Inst	masked
PVA8	locale	masked
PVA9	Mode	masked
PVA10	ItemID	masked
PVA11	PHI	masked
PVA12	Response	masked
PVA13	Score	masked
PVA14	Theta	masked
PVA15	Tscore	masked
PVA16	SE	masked
PVA17	Data Type	masked
PVA18	Postn	masked
PVA19	Time	masked
PVA20	DteCrted	masked
PVA21	InstStr	masked
PVA22	InstEnd	masked
PVA23	Consent	masked

Details on the PVT's methodology and generated content are available from the NIH.

ADDITIONAL DERIVED VARIABLES

The derived variables in the **inca_mci_derv_inv3a** dataset are not associated with any particular form because they were generated using variables from multiple datasets. There is one record per screened participant (n=6,377). This file is a cross-section of "derived variables" whose values are defined based on combinations of data items (e.g. thresholds for cognitive performance) derived using a combination of cognitive, health, and demographic records. Details on code and process used to generate these variables are provided in a Stata markdown document containing the following parts.

- Part 1
  - Rename Cognitive tests
  - Adjusts trail tests
  - Creates CESD-10 variable
  - Creates Normed Sample
  - Creates Age groups
  - Regression on normed samples
  - Calculates and graphs Z-Scores
  - CGE Domains Generation to create eCog indicators
  - Creates IADL indicators
- Part 2
  - Mplus Modelling to generate cognitive change in multiple scenarios
  - Generate cognitive decline indicator based on estimated factor scores
  - Mplus indicators loaded into Stata for later use
- Part 3
  - Create MCI Groupings based on latent class model and IADL indicators
  - Generate categorical MCI indicators
- Part 4
  - Variable management to decide which variables go into final dataset

Table 1. SOL-INCA AS Derived Variables

Variable Name	Name	N's
cog_test_signal	Objective Cog Test Signal flag	6272
ecog_signal	MCI eCog Signal	6369
ecog_average	eCog Average Score	6364
iadl_signal	Any IADL Limitation V2	6371
iadl_cog_specific	Any IADL Limitation Due to Attention, Concentration, Memory V2	6371
iadl_count	Count of IADL Limitations V2	6371
decline_signal	Decliners Signal flag	6373
global_cog_baseline	Fscore V1 Consistent battery: Fixed loading, means, and variances	6377
global_cog_inca	Fscore V2 Consistent battery: Fixed loading, means, and variances	6377

Variable Name	Name	N's
global_decline	Change from Visit 1 to Visit 2	6377
mci3	Cognitive Impairment 3 Category Classification	6265
Mci	MCI 2 Category Classification	6265

Table 2. SOL-INCA AS Derived Variables Descriptives

Variable Name	Name	Value	Description
ecog_signal	eCog Signal	0	no eCog limitation
		1	suspect eCog limitation
			Missing
iadl_signal	Any IADL Limitation V2	0	Answered question(s) without any assistance on the DLE
		1	Answered question(s) with assistance on the DLE
			Missing
iadl_cog_specific	Any IADL Limitation Due to Attention, Concentration, Memory V2	0	Answered question(s) without any assistance on the DLE
		1	Answered question(s) with assistance on the DLE
iadl_count	Count of IADL Limitations V2	Range	0-7 sum of DLE1-DLE7 indicators
			Missing
cog_test_signal	Objective Cog Test Signal flag	0	within 0 to 1 SD
		1	2 or more SD
			Missing
decline_signal	Decliners Signal flag	0	No decline (-0.055 SD per year)
		1	Decline of -0.55 SD per year or more
			Missing
mci3	Cognitive Impairment 3 Category Classification	0	No MCI
		1	MCI
		2	Suspect Severe Impairment
			Missing
mci	MCI 2 Cateogry Classification	0	No MCI
		1	MCI/Suspect Severe Impairment
ecog_average	eCog Average Score	Range	1-4 (median=1.33 mean=1.54 std=0.57)
global_cog_baseline	Fscore V1 Consistent battery: Fixed loading, means, and variances	Range	-2.73 - 3.008 (median=.004 mean=-3.49e-06 std=.88)
global_cog_inca	Fscore V2 Consistent battery: Fixed loading, means, and variances	Range	-3.41-3.1 (median= -0.136 mean=-0.17 std=0.96)
global_decline	change from visit 1 to visit 2	Range	-4.69-5.84 (median= .046 mean=8.86e-10 std=1.00)

IMPORTANT ANALYSIS NOTES

Statistical analyses involving SOL-INCA data must account for the complex sampling design by specifying strata (STRAT), primary sampling unit (PSU_ID) and sampling weights (WEIGHT_NORM_CENTER_INCA or WEIGHT_NORM_OVERALL_INCA). The sampling weights that accompany the data are calibrated to the 2010 US Census Population according to age, sex, and Hispanic background. The normalization procedure corresponds to multiplying all weights by a constant factor such that the sum of weights equals the cohort sample size (16,415) and the average weight equals 1. The normalized weight is primarily intended for use when analyzing data from all four field centers combined. A sampling weight that is normalized to each field center sample size separately is also provided. The center-specific normalization procedure corresponds to multiplying each weight in the center by a constant factor such that the sum of weights in the center equals the center sample size and the average weight across all persons in the center equals 1. With this procedure, the constant factors differ from center to center. Center-specific normalized weights are appropriate for use whenever data from single centers are being analyzed separately or when centers are being compared with respect to certain characteristics.

Unequal weighting, stratification, and cluster sampling can all impact analysis of data arising from a complex sample design such as that employed for HCHS/SOL. In addition to affecting point estimates, unequal weighting and cluster sampling tend to increase the variability of population estimates and reduce the power available for statistical tests while stratification has the reverse effect. For valid inference to the target population, all three aspects of the HCHS/SOL sample design need to be taken into account. Failure to do so typically results in over-stating both the precision of estimates and the statistical significance of hypothesis tests. Several statistical software packages can accommodate complex survey data such as SAS, Stata, SPSS, SUDAAN, R, and Mplus. Analysts are strongly encouraged to read the document "ANALYSIS METHODS FOR HCHS/SOL" in the HCHS/SOL Main Study to ensure that the study design is correctly specified prior to analysis.

REFERENCES

1. González, H. M., Tarraf, W., Fornage, M., González, K. A., Chai, A., Youngblood, M., ... & Gallo, L. C. (2019). A research framework for cognitive aging and Alzheimer's disease among diverse US Latinos: Design and implementation of the Hispanic Community Health Study/Study of Latinos—Investigation of Neurocognitive Aging (SOL-INCA). Alzheimer's & Dementia, 15(12), 1624-1632.

2. González, H. M., Tarraf, W., Schneiderman, N., Fornage, M., Vásquez, P. M., Zeng, D., ... & Kaplan, R. (2019). Prevalence and correlates of mild cognitive impairment among diverse Hispanics/Latinos: Study of Latinos-Investigation of Neurocognitive Aging results. Alzheimer's & Dementia, 15(12), 1507-1515.

3. Gershon, Richard C., Karon F. Cook, Dan Mungas, Jennifer J. Manly, Jerry Slotkin, Jennifer L. Beaumont, and Sandra Weintraub. "Language measures of the NIH toolbox cognition battery." Journal of the International Neuropsychological Society 20, no. 6 (2014): 642-651.