NIDDK grant: Ancillary to HCHS/SOL: Neighborhood Environments and Cardiometabolic Disorders in Hispanic/Latinos (R01 DK106209, LC Gallo & M Allison)

SOL CASAS Ancillary Study Investigator Use Database Overview

December 2019

(Documentation V2.0)

INV2 data

Prepared by

HCHS/SOL Coordinating Center

Collaborative Studies Coordinating Center

UNC Department of Biostatistics

Daniela Sotres-Alvarez

Franklyn Gonzalez II

ACKNOWLEDGMENTS

SOL CASAS Study would not have been possible without the participation of HCHS/SOL participants and the support of study member staff. SOL CASAS Study (R01 DK106209) was funded by the National Institute of Diabetes and Digestive and Kidney Diseases.

Special thanks to:

Allison Matthew A, MD, MPH
SOL CASAS Principal Investigator
University of California San Diego

Carlson Jordan A, PhD
Co-Investigator
Children's Mercy Kansas City

Gallo Linda C, PhD
SOL CASAS Principal Investigator
San Diego State University

Geremia Carrie M, BA
Program Specialist
University of California San Diego

Gonzalez II, Franklyn, MS
CC Project Manager
University of North Carolina at Chapel Hill

Jankowska Marta M, PhD
Co-Investigator
San Diego State University

Rodriguez Tasi M, MPH
Clinic Manager
San Diego State University

Sallis James F, PhD
Co-Investigator
University of California San Diego

Sotres-Alvarez, Daniela, DrPH
Co-Investigator
University of North Carolina at Chapel Hill

Talavera Greg, MD, MPH
HCHS/SOL Principal Investigator
San Diego State University

SOL CASAS Investigator Use Database Overview INV2

Table of Contents

Updates to SOL CASAS Data Release or Documentation

DATA Version	Date	Description	Datasets	Documentation Version
INV1	2/16/18	1st data release	Suffix: _INV1 NSE, PAE, CASAS_PART_DERV, CASAS_PA_EPOCH, CASAS_PA_DAYS, CASAS_PA_DERV	V1 (Feb. 2018)
INV1	4/13/18	- CASAS sampling weights (CASAS_PART_DERV_INV1) - Micro environment data - Macro environment data	Suffix: _INV1 CASAS_PART_DERV CASAS_MICRO CASAS_MACRO	V1.1 (April 2018)
INV1	9/13/19	- **Accelerometer data scored at the minute level** (i.e. original epoch of 15sec was collapsed at the minute level and then scored). This allows comparison to baseline data which was measured at the minute level. - **New sedentary bout variables**. - **Macro data** was updated with additional macro variables for visit 1 (n=1,682) and a new data file has macro variables for visit 2 (n=1,682). - **Micro data** was updated with additional micro variables for visit 1 (N=1,680) and a new data file has micro variables for those who moved between baseline and visit 2 (n=599). - **Participant Derive Variables.** A minor error was discovered and now corrected in the derivation of the sampling weights.	Suffix: _INV1 **Replacement:** CASAS_GIS_MACRO_BUFFER_V1 (baseline) CASAS_PART_DERV **NEW datasets:** CASAS_PA_MIN CASAS_MIN_SEDBOUTS CASAS_PA_MIN_DAYS CASAS_PA_MIN_DERV CASAS_GIS_MACRO_BUFFER_V2 (visit 2) CASAS_MICRO_PARTICIPANT_V1 (baseline) CASAS_MICRO_PARTICIPANT_V2 (movers only)	V1.2 (Sept. 2019)
INV1.2	12/3/19	**Sedentary bout variables.** We corrected mean bout length in CASAS_PA_MIN_DAYS and CASAS_PA_MIN DERV. Means are only computed if there is at least one bout of the length of interest. - **Micro data. We are only releasing** a single data set with Timepoint 2 data (n=1,680). - **Macro data.** We include ALL San Diego participants with geocoded address available (N=3,854), and not limited to CASAS participants as before (INV1) - **Participant Derive Variables.** We include GPAQ_TRSPORT which was inadvertently not included in INV1 (9/12/19) data release	Suffix: _INV2 **Updated datasets:** CASAS_PA_MIN_DAYS CASAS_PA_MIN_DERV CASAS_MICRO_PARTICIPANT_V2 (visit 2) HCHS_GIS_MACRO_BUFFER_V1 (baseline) HCHS_GIS_MACRO_BUFFER_V2 (visit 2) CASAS_PART_DERV **Identical datasets with new extension:** NSE, PAE, CASAS_PA_EPOCH, CASAS_PA_DAYS, CASAS_PA_DERV, CASAS_PA_MIN, CASAS_MIN_SEDBOUTS	V2.0 (Dec. 2019)

1. INTRODUCTION

This document describes the content and structure of the Investigator Use database created for SOL CASAS Ancillary Study. This database contains all the data collected for the 1,776 enrolled participants, subject to constraints (described within) to preserve participant confidentiality by de-identifying the data. Data for prospective participants who screened out are not included.

2. STUDY OBJECTIVES

The scientific aims of the SOL CASAS Ancillary to the HCHS/SOL are to: (1) determine if HCHS/SOL Visit 1 macro-scale social and built neighborhood environments are associated with 6-year changes in metabolic health, physical activity, and depression, and whether changes in physical activity and depression mediate the association between the environment and metabolic health change; (2) determine if 6-year changes in macro-scale neighborhood environments between HCHS/SOL Visits 1 and 2 are associated with 6-year changes in metabolic health, physical activity, and depression, and whether changes in physical activity and depression mediate the association between neighborhood environment change and metabolic health change; and (3) using variables available at HCHS/SOL Visit 2 only, to investigate cross-sectional relations of macro- and micro- scale social and built neighborhood environments to metabolic health, physical activity, and multiple psychosocial factors (chronic stress, depression, and social support), and whether physical activity and psychosocial factors mediate the association between the environment and metabolic health.

3. STUDY DESIGN

SOL CASAS is an ancillary study only at the San Diego HCHS/SOL cohort, rather than participants from all four sites (San Diego, CA, Chicago, IL, Miami, FL, and Bronx, NY). From March 2008 to June 2011 (Visit 1, “baseline”). The HCHS/SOL cohort enrolled 16,425 male and female adult participants between the ages of 18 and 74 years from four field centers located across the U.S. The purpose of the parent study is to determine the prevalence and incidence of cardiovascular disease and other chronic conditions, and to identify risk factors playing a protective or harmful role in these conditions, in Hispanics/Latinos. Participants eligible for SOL CASAS include those who participated in the HCHS/SOL Visit 1 at the San Diego Field Center (N = 4,097) who returned for HCHS/SOL Visit 2.

3.1. Participants

The study aimed to recruit 1,750 Hispanic/Latino adults aged 18-74 at the San Diego Field Center (South Bay Latino Research Center, SBLRC). A HCHS/SOL participant is considered eligible to participate in SOL CASAS if they have completed both HCHS/SOL visits and can walk one block without help. Refer to the CEL QxQ (question by question) for further clarification on the type of assistance that would determine an individual ineligible for participation in SOL CASAS.

4. DATABASE STRUCTURE

4.1. Data Set Organization

There is one SAS data set in the database for each type of data collection form. 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.

Special derived variable datasets have been created to augment the original data measurement values. The participant derived variable file has computed score values based on standard algorithms (e.g. GPAQ). These algorithms have been included in the **SOL CASAS 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.

Table 4.1. SOL CASAS Ancillary Study datasets

SOL CASAS Dataset (_INV2 extension)	Key fields (identify each record uniquely)	Dataset description
PAE	ID	Self-reported physical activity (GPAQ) (N=1,775)
NSE	ID	Neighborhood selection (N=1,775)
CASAS_ PART_DERV	ID	Participant derived variables and sampling weights (N=1,776)
CASAS_ PA_EPOCH	ID + STARTDATE+ EPOCH	- Multiple records per participant; one record per each epoch (15 seconds) with the counts. There are 95,248,422 records. - Multiple devices per participant if wore it more than once (82 participants worn it two times and three participants worn it three times). - Nonwear time is identified with SAS special missing (.C) determined using Choi's algorithm.
CASAS_ PA_DAYS	ID + STARTDATE+ DAY	WARNING: IT IS RECOMMENDED TO USE CASAS_PA_MIN_DAYS that collapsed the 15 sec EPOCH to 1 min. - Multiple records per participant; one record per worn DAY per ID and STARTDATE (some participants have multiple wears). There are 17,507 records. - Each ID has as many days as the device was worn (range from 4 to 12 days). DAY 1 is the next day after the clinic visit day (protocol).
CASAS_ PA_DERV	ID	WARNING: IT IS RECOMMENDED TO USE CASAS_PA_MIN_DERV that collapsed the 15 sec EPOCH to 1 min. Also, note that using the 15sec data the adherence was defined with at least 4 days with 10+ hrs each. In contrast, in the 1 min data adherence was defined with at least three days 10+ hrs each. - ONE record per participant (n=1,725); - There is an indicator variable (ADHRENTYN) that identifies participants with at least four adherent days and an indicator variable (WEEKEND_INCLUDED) that identifies whether an adherent weekend day was included or not. - The dataset includes those participants with less than three adherent days, but summary variables are missing for them due to not meeting the minimum number of adherent days.
ACTICAL DATA COLLAPSED AT THE MINUTE LEVEL FIRST, the 15 sec epoch were collapsed into 1 minute and afterwards the counts were classified into the four intensities (sedentary, light, moderate and vigorous).		
CASAS_PA_MIN	ID + STARTDATE+ DAY + MIN	- Multiple records per participant (one record for each minute per day and per actical worn) - Multiple devices (identified by STARTDATE) per participant if wore it more than once (82 participants worn it two times and three participants worn it three times); - There are 23,811,380 records. - Nonwear time is identified with SAS special missing (.C) determined using Choi's algorithm.
CASAS _ MIN _SEDBOUTS	ID + STARTDATE + NDAY + BOUT NDAY is a consecutive day of Actical wear, redefined to tract	- Multiple records per participant; one record for each sedentary bout per day and per actical worn; - There are 896,051 records. - A sedentary bout is defined based on the sedentary intensity level (<100 counts/min) for any amount (>= 1 min) of an unbroken period of sedentary time, with its duration being the length (in minutes) of this consecutive period. - Sedentary bouts were measured separately for each participant and wear, across worn days. Thus, a sedentary bout can cross worn days, but neither participant nor wear. A new NDAY variable is created to tract bouts across two worn days, with the bout being accessed on the previous day.

SOL CASAS Dataset (_INV2 extension)	Key fields (identify each record uniquely)	Dataset description
	bouts across two days (midnight)	- Bouts are ordered by time of the day. For example, NDAY=1 has the first sedentary bout in day 1, NDAY=2 has the second, and sequentially.
CASAS _ PA_MIN _DAYS	ID + STARTDATE + DAY	- Multiple records per participant; one record per worn DAY per ID and STARTDATE (some participants have multiple wears). - There are 17,507 records. - Each ID has as many days as the device was worn (range from 4 to 12 days). DAY 1 is the next day after the clinic visit day (protocol).
CASAS_MIN_ PA_DERV	ID	- ONE record per participant (n=1,725). - Dataset includes those participants with less than three adherent days, but summary variables are missing for them due to not meeting the minimum number of adherent days. - There is an indicator variable (ADHRENTYN) that identifies participants with at least three adherent days and an indicator variable (WEEKEND_INCLUDED) that identifies whether an adherent weekend day was included or not.
HCHS_GIS_ MACRO_BUFFER_V1 (baseline)	ID	- ONE record per participant (n=3,854). - Dataset includes ALL San Diego participants with geocoded addresses at HCHS/SOL baseline. - Dataset with macro derived variables using two different buffers (600m and 1,600m) for geocoded addresses from HCHS/SOL baseline (baseline time point data only). - There is an indicator variable (OUT_COUNTY_T1) that identifies participants who moved out of the San Diego County area between recruitment and the baseline examination.
HCHS_GIS_ MACRO_BUFFER_V2 (visit 2)	ID	- ONE record per participant (n=3,854). - Dataset includes ALL San Diego participants with geocoded addresses at Visit 2. - Dataset with macro derived variables using two different buffers (600m and 1,600m) for geocoded addresses from HCHS/SOL baseline (visit 2 time-point data only). - There is an indicator variable (OUT_COUNTY_T1) that identifies participants who moved out of the San Diego County area between recruitment and the baseline examination and (OUT_COUNTY_T2) that identifies participants who moved out of the San Diego County area between baseline examination and Visit 2. - There is an indicator variable (MOVER_T2) that identifies participants who moved within San Diego County area between baseline examination and Visit 2.
CASAS_MICRO_ PARTICIPANT_V2	ID	- ONE record per participant (n=1,680). - Dataset includes CASAS participants with geocoded addresses from HCHS/SOL baseline and Visit 2 (Visit 2 time point data only, includes those who did/did not moved during the same time period). - There is an indicator variable (SHARED) that identifies participants chosen for household-level analysis and (ADDRESS_ID) that uniquely identifies participants who live in the same household.

4.2. Form and Data Set Naming Conventions

Each SOL CASAS data collection instrument (Form) has a unique three-letter mnemonic associated with it (e.g., PAE for the SOL CASAS Physical Activity English form). Corresponding data sets begin with the same first three letters of the mnemonic, followed by the character string "_INV2" for Investigator Use, Version 1. For example, the Physical Activity data set is "PAE_INV2". The naming convention serves both to identify the originating form 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 the ID. This ID is the same as the one in HCHS/SOL baseline and all its ancillary studies. It is a random 8-digit identification code, unique to each HCHS/SOL participant.

Only for ACTICAL data some of the datasets have additional KEY identifiers. See section 6.4 in this Overview document and SOL CASAS Derived Variable Dictionary.

4.4. Common Variables Across Data Sets

Additional variables appear in every data set; VERS and VISIT are meaningless in SOL CASAS:

1) VERS: Version of the data collection form. A one-character variable indicating which version of the paper form was used to collect the data. However, in SOL CASAS there were no changes in the FORMS.
2) VISIT: It is 1 for all records (1st data collection for SOL CASAS).

4.5. Variable Naming Conventions

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 and then the question number as indicated on the form. For example, question 1 on the Physical Activity form (PAE) is named PAE1 on the corresponding SAS file, PAE_INV2. Similarly, question 3 from the Physical Activity form is named PAE3.

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 NIDDKD/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. STAFF ID codes were deleted across all forms and not substituted. DATE OF BIRTH at

HCHS/SOL baseline was not distributed; instead it was converted to age at CASAS clinic visit and appears in the derived variable data sets, CASAS_PART_DERV_INV2.

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
.C	Only used in dataset CASAS_PA_EPOCH to identify non-wear time using Choi's algorithm.

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 "**≤ .Z**" which will detect these special missing values rather than "**= .**", which will not. Alternatively, the SAS missing function can be used.

5. DESCRIPTION OF DATA COLLECTION FORMS

5.1. Eligibility/Participation Checklist (CEL). NOT DISTRIBUTED.

The Individual Eligibility/Participation Checklist (CEL) includes the script for determining a participant's eligibility. The participant's eligibility and participation status are recorded in sections 1 and 2 of the form. Sections 3 and 4 are completed the day of the SOL CASAS study visit. Recruiters are responsible for verifying that all individuals meet the eligibility criteria for inclusion in the study.

5.2. Neighborhood Selection (NSE)

This questionnaire is a nine-item scale about the participants' reasons for moving to their current neighborhood. On a scale of 1 to 5, with 1 being "not at all important" and 5 being "very important", participants rate the importance of the various factors in their decision to move to their current neighborhood. This measure has been used in prior studies to account for neighborhood preference (Sallis JF *et al*, Soc Sci Med 2009; Norman GJ, *Am J Health Behav.* 2013).

5.3. Physical Activity (PAE)

The physical activity questionnaire also used at HCHS/SOL baseline was based on the commonly used Global Physical Activity Questionnaire (GPAQ) which captures self-report physical activity during a typical week. A modified version of the World Health

Organization (WHO) GPAQ was used to obtain information about participants' habitual activities in three domains: work-related (both moderate and vigorous levels), transportation (moderate level), and leisure or recreational (both moderate and vigorous levels). It also included a question on sitting. The following changes were made from the WHO original GPAQ for use in HCHS/SOL and SOL CASAS Ancillary Study:

Work:
(original #3, PAE3) essentially the same question but stated in a longer format
(original #6, PAE6) essentially the same question but stated in a longer format

Transport:
(original #8, PAE8) "for at least 10 minutes continuously" was omitted
(original #9, PAE9) question was worded slightly differently

Recreation/Leisure:
(PAE11) new item that was not asked originally that quantifies the types of vigorous activities that are done
(PAE15) new item that was not asked originally that quantifies the types of moderate activities that are done
(original #15, PAE17) essentially the same question but stated in a longer format

Sitting:
No changes

A trained interviewer conducted the interview and obtained the physical activity information using clarification and prompts within the form. Interviewers consistently reminded participants during the interview that activities should only be reported if the duration was at least ten minutes. Interviewers helped participants recall their activities during a typical week by asking them to list their activities during the previous week and then clarifying whether that was a typical week.

6. DERIVED FILES

6.1. CASAS_PART_DERV (Participant Derived Variables)

The Participant Derived Variable dataset is not associated solely with any particular form because it contains variables from many forms and files. There is one record per enrolled participant (1,776 observations). This file is a cross-section of "derived variables" whose values are defined based on combinations of data items (e.g. age at CASAS clinic visit or GPAQ variables). CASAS sample weight are included in this dataset. See the separate document, *"SOL CASAS Derived Variable Dictionary"* for the definitions of the variables included in this special purpose file. Statistical analysis using HCHS/SOL data must account for the complex sampling design by specifying strata (STRAT), primary sampling unit (PSU_ID) and sample weights (CASAS_WEIGHT_NORM). 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.

6.2. CASAS_PA_EPOCH (Counts at the 15sec EPOCH level)

Dataset with counts for each EPOCH (15 sec). See table below for details.

6.3. CASAS_PA_DAYS (Actical variables at the day level from 15sec EPOCH)

WARNING: IT IS RECOMMENDED TO USE CASAS_PA_MIN_DAYS WHICH IS HOW BASELINE DATA WAS COLLECTED.

This dataset has summarized counts and minutes at the calendar day level. NOTE that first the counts were classified into intensities (sedentary, light, moderate and vigorous) at the 15sec level using the baseline cutpoints divided by four. Second, the indicators were summed to obtain the number of minutes in the intensity for that specific day. This causes less minutes in higher intensities because it is possible that accumulated counts per minute are below the cutpoint but individual epochs are not (see example below). This is a concern specially if we want to compare activity variables to baseline where an EPOCH of 1 min was specified and used to classify the minute level. Hence, we rescored the data by first collapsing the EPOCH (15sec) to the MINUTE level and then classified the counts into the intensities using exactly the same cutpoints used at baseline.

MIN	EPOCH	Counts/epoch	Sedentary using ≤25 as cutpoint to classify a an epoch of 15 sec	Sedentary using ≤100 as cutpoint to classify an epoch of 1 minute
1	1	25	Yes	
1	2	26	0	
1	3	26	0	
1	4	26	0	
	Total	**103**	**.25 min**	**0 min**
2	5	10	Yes	
2	6	25	Yes	
2	7	25	Yes	
2	8	25	Yes	
	Total	**85**	**1 min**	**1 min**

6.4. CASAS_PA_DERV (Actical variables at the participant level from 15sec EPOCH)

WARNING: IT IS RECOMMENDED TO USE CASAS_PA_MIN_DERV INSTEAD. See explanation in CASAS_PA_DAYS which was used to create this dataset.

6.5. CASAS_PA_MIN (Counts at the MINUTE level)

This dataset has counts for each MINUTE (60 sec). It was created by collapsing four records of CASAS_PA_EPOCH to combine the 15sec epoch into one minute. See table 4.1. for details.

6.6. CASAS_MIN_SEDBOUTS (Sedentary bout length at the bout level)

This dataset has the sedentary bout length for each bout. See table 4.1. for details.

6.7. CASAS_PA_MIN_DAYS (Actical variables at the day level)

This dataset has summarized variables (counts, minutes, bouts) at the calendar day level. See table 4.1. for details.

6.8. CASAS_PA_MIN_DERV (Actical variables at the participant level)

This dataset has summarized variables (counts, minutes in four different intensities, sedentary bouts) for participants with at least 3 adherent days (i.e. at least 10hrs of wear time). See table 4.1. for details.

6.9. CASAS_MICRO_PARTICIPANT_V2

This dataset has the micro-environment variables for visit 2 geocode address available (N=1,680). Please refer to the “SOL CASAS Micro Data MAPS Scoring Dictionary” for more detailed information including data sources.

6.11. HCHS_GIS_MACRO_BUFFER_V1

This dataset has, for all SD participants, the macro-environment variables using two different buffers (600m and 1,600m) and baseline geocoded adress (N=3,854), not only Casas participants. Please refer to the “HCHSCasasGISMacro_Codebook” and “HCHSCasasGISMacro_Codebook_Supplement” for detailed information including data sources and timepoints.

6.12. HCHS_GIS_MACRO_BUFFER_V2

This dataset has, for all SD participants, the macro-environment variables using two different buffers (600m and 1,600m) and visit 2 geocoded address (N=3,854). Please refer to the “HCHSCasasGISMacro_Codebook” and “HCHSCasasGISMacro_Codebook_Supplement” for detailed information including data sources and timepoints.

7. REFERENCES

Choi L., Liu, Z., Matthews, C.E., Buchowski M.S. Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exerc* 43 (2011), 357-364.

Colley R., Garriguet D., Janssen I., Craig C., Clarke J., Tremblay M.,. Physical activity of Canadian adults: Accelerometer results from the 2007 to 2009 Canadian Health Measures Survey. *Statistics Canada no. 82-003-XPE, Health Reports* 22 (2011), 7-14.

Norman GJ1, Carlson JA, O'Mara S, Sallis JF, Patrick K, Frank LD, Godbole SV. Neighborhood preference, walkability and walking in overweight/obese men. *Am J Health Behav.* 2013 Mar;37(2):277-82. doi: 10.5993/AJHB.37.2.15.

Sallis JF, Saelens BE, Frank LD, Conway TL, Slymen DJ, Cain KL, Chapman JE, Kerr J. Neighborhood built environment and income: examining multiple health outcomes. *Soc Sci Med*, 68 (7) (2009), pp. 1285-1293.

2008 US Physical Activity Guidelines:
http://www.health.gov/paguidelines/guidelines/default.aspx

World Health Organization (WHO) Global Physical Activity Questionnaire (GPAQ)
http://www.who.int/chp/steps/GPAQ/en/index.html