NHLBI grant: Ancillary to HCHS/SOL: Cardiometabolic Outcomes in Multi-ethnic Physical Activity & Sedentary Behavior Study (R01 HL136266; Y Mossavar-Rahmani, RC Kaplan, & V Ramachandran)

COMPASS Ancillary Study
Investigator Use Database Overview

October 2020
(Documentation V1.1)
INV1 data

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

Daniela Sotres-Alvarez
Franklyn Gonzalez II

ACKNOWLEDGMENTS

COMPASS Study would not have been possible without the participation of HCHS/SOL participants and the support of study member staff. COMPASS Study (R01 HL136266) was funded by the National Heart, Lung, and Blood Institute.

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:

HCHS/SOL Bronx Site
Albert Einstein College of Medicine

Madeleine Crespo-Figueroa
Project Director/Clinic Manager

Robert Kaplan, PhD
COMPASS Multiple Principal Investigator

Yasmin Mossavar-Rahmani, PhD, RD, CDN
COMPASS Multiple Principal Investigator

Marlen Murillo-Pimentel
Study Coordinator

Xiaonan Xue, PhD
Co-Investigator

HCHS/SOL Chicago Site
University of Illinois at Chicago

Cesar Alvarado
Project Manager

Marta Daviglus, MD, PhD
Co-Investigator

Veronica Herzog
Clinic Manager/Project Coordinator

Kimberly Manzanares
Coordinator

HCHS/SOL Miami Site
University of Miami

Eduardo Alvarez
Clinic Manager/Project Coordinator

Claudia Chambers
Clinic Operations Manager

Marc Gellman, PhD
Co-Investigator

Maria Pattany
Project Manager

Neil Schneiderman, Ph.D.
Co-Investigaor

HCHS/SOL Coordinating Center
University of North Carolina at Chapel Hill

Franklyn Gonzalez II, MS
Project Manager

Kelly Evenson, PhD
Co-Investigator

Daniela Sotres-Alvarez, DrPH
Co-Investigator

Framingham Cohort
Boston University

Joanne Murabito, MD, ScM
Co-Investigator

Vasan Ramachandran, MD
COMPASS Multiple Principal Investigator

Nicole Spartano, PhD
Co-Investigator

HCHS/SOL San Diego Site (CASAS ANCILLARY)

Mathew Allison, MD, MPH
CASAS Multiple Principal Investigator
University of California San Diego

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

Linda Gallo, PhD
CASAS Multiple Principal Investigator
San Diego State University

Johanne Hernandez
HCHS/SOL Project Manager
San Diego State University

Tasi Rodriguez
CASAS Project Manager

Ana Talavera, MPH
HCHS/SOL Project Director
San Diego State University

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

COMPASS Investigator Use Database Overview INV1

Table of Contents

Updates to COMPASS Data Release or Documentation

DATA Version	Date	Description	Datasets	Documentation Version
INV1	4/30/20	1st data release It includes COMPASS data (Bronx, Chicago and Miami) plus CASAS data (San Diego)	Suffix: _INV1 PAE, CFE COMPASS_PART_DERV COMPASS_PA_EPOCH COMPASS_PA_MIN COMPASS_MIN_SEDBOUTS COMPASS_PA_MIN_DAYS COMPASS_PA_MIN_DERV	V1 (Apr. 2020)
INV1	10/12/20	2nd data release - **CFE (Activity Feedback form)**. An identifier (WEARORDER) was added to distinguish among records from same participant but from different Actical devices. Twenty records were deleted because of being all empty or no Actical data available. - **COMPASS_PART_DERV**. Sampling weights slightly changed due to a typo that has been corrected in the variable that determines eligibility and participation criteria. - **Accelerometer datasets.** A handful of corrections to mismatched IDs from COMPASS participants (Bronx, Chicago or Miami) resulted in 4 more adherent participants.	Suffix: _INV1 **Replacement (all files):** PAE (identical), CFE COMPASS_PART_DERV COMPASS_PA_EPOCH COMPASS_PA_MIN COMPASS_MIN_SEDBOUTS COMPASS_PA_MIN_DAYS COMPASS_PA_MIN_DERV	V1.1 (Oct. 2020)

1. INTRODUCTION

This document describes the content and structure of the Investigator Use Database created for HCHS/SOL COMPASS Ancillary Study. This database includes data from COMPASS Ancillary which recruited participants in the Bronx, Chicago and Miami (N=2,880), and from CASAS Ancillary study which recruited participants in San Diego (N=1,776). This database contains all the data collected for these 4,656 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

Aim 1. Among nondiabetics, to identify physical activity (PA) and sedentary behavior (SB) patterns associated with conversion to diabetes up to 12 years by adding a second accelerometry measure to HCHS/SOL. We will examine influence of bout length and intensity of PA to define the dose-response relationships affecting diabetes risk.

Aim 2. Among nondiabetics, to identify the relationship between moderate to vigorous physical activity (MVPA), light physical activity (LPA) and SB with incident cardiovascular disease events and mortality, in order to define the magnitude of risks and dose-response for duration, intensity and bout length.

Aim 3. To investigate demographic and psychosocial correlates associated with 6+ year changes in patterns of PA and SB in Hispanics/Latinos and non-Hispanics/Latinos without a diagnosis of diabetes.

The hypothesized predictors of PA/SB include age, sex and race/ethnic and national background – comparing a mostly (80+%) immigrant HCHS-SOL Hispanic/Latino population versus US born Hispanics/Latinos and non-Hispanics/Latinos. We hypothesize that younger, non-Hispanic/Latino white adults and women will have high risk of decreasing MVPA and increasing SB over time. Using the rich characterization of the cohorts, additional analyses will examine mental and physical well-being, employment and economic indicators, acculturation, diet, comorbidities and health care use in relation to changes in MVPA, LPA and SB. Thus we will not only identify the patterns of PA and SB that predict disease (Aims 1+2), but also the barriers and facilitators of these patterns as well as the groups in greatest need of intervention (Aim 3).

3. STUDY DESIGN

COMPASS is an HCHS/SOL ancillary study which collected data from three sites: Bronx (NY), Chicago (IL), and Miami (FL). It complements with accelerometer data collected by CASAS Ancillary which collected data at San Diego (CA) site.

3.1. Participants

COMPASS Ancillary recruited participants in the Bronx, Chicago and Miami (N=2,880) from 2/28/2017 to 11/15/2019. CASAS Ancillary recruited participants in San Diego

(N=1,776) from 12/17/2015 to 9/30/2017. Participants were considered eligible to participate in COMPASS or CASAS if they have completed both HCHS/SOL baseline and Visit 2, could walk one block without help, and were not identified as diabetic from HCHS/SOL V1, HCHS/SOL V2, or COMPASS screening data (CEP1).

Center	Frequency	%
Bronx	1,000	21.48
Chicago	1,023	21.97
Miami	857	18.41
San Diego	1,776	38.14
Total	**4,656**	**100.00**

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 COMPASS 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 COMPASS Ancillary Study datasets

SOL COMPASS Dataset (_INV1 extension)	Key fields (identify each record uniquely)	Dataset description
PAE	ID	Self-reported physical activity (GPAQ) (N=4,644)
CFE	ID + WEARORDER	Activity Feedback Form (N=2,596)
COMPASS_ PART_DERV	ID	Participant derived variables and sampling weights (N=4,656)
COMPASS_ PA_EPOCH	ID + STARTDATE+ EPOCH	- Multiple records per participant; one record per each epoch (15 seconds) with counts. There are 261,447,617 records. - Multiple devices (identified by STARTDATE) per participant if wore it more than once. Almost all worn the device once. However, in COMPASS 26 participants worn it twice. In CASAS, 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.
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).		
COMPASS_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. Almost all worn the device once. However, in COMPASS 23 participants worn it twice. In CASAS, 82 participants worn it two times and three participants worn it three times. - There are 65,534,561 records. - Nonwear time is identified with SAS special missing (.C) determined using Choi's algorithm.
COMPASS _ MIN _SEDBOUTS	ID + STARTDATE + NDAY + BOUT NDAY is a consecutive day of Actical wear, redefined to tract bouts across two days (midnight)	- Multiple records per participant; one record for each sedentary bout per day and per actical worn; - There are 2,325,971 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. - 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.
COMPASS _ 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 48,018 records. - Each ID has as many days as the device was worn (range from 2 to 12 days). DAY 1 is the next day after the clinic visit day (protocol).
COMPASS_MIN_ PA_DERV	ID	- ONE record per participant (n=4,491). - 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.

4.2. Form and Data Set Naming Conventions

Each COMPASS data collection instrument (Form) has a unique three-letter mnemonic associated with it (e.g., PAE for the COMPASS Physical Activity English form). Corresponding data sets begin with the same first three letters of the mnemonic, followed by the character string "_INV1" for Investigator Use, Version 1. For example, the Physical Activity data set is "PAE_INV1". 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 COMPASS Derived Variable Dictionary.

4.4. Common Variables Across Data Sets

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

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 COMPASS there were no changes in the FORMS.
2) VISIT: It is 1 for all records (1st data collection for SOL COMPASS).

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_INV1. 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 here we distribute age at the clinic visit

for COMPASS or CASAS (for participants in San Diego), and it is included in the derived variable data set COMPASS_PART_DERV_INV1.

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 COMPASS_PA_EPOCH to identify non-wear time using Choi's algorithm.
.L	Only used in dataset COMPASS_PA_EPOCH to identify hours before start time when midnight was not specified as protocol stated.
.M	Only used in dataset COMPASS_PA_EPOCH to identify character "M" read from the AWC file

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 (CEP). NOT DISTRIBUTED.

The Individual Eligibility/Participation Checklist (CEP) includes the script for determining a participant's eligibility. Recruiters are responsible for verifying that all individuals meet the eligibility criteria for inclusion in the study.

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, CASAS and COMPASS Ancillary Studies.

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.

5.4. Accelerometer Feedback Form (CFE)

This form has 3 questions that were answered in paper and returned with the accelerometer. It asked whether the amount of activity performed during the last week (while woaring the activity monitor) was similar to a typical week. It also askwed whether the participant had biked, swam or left weights. This is a multiple-record dataset because the form was completed each time the Actical device was worn.

6. DERIVED FILES

6.1. COMPASS_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 (4,656 observations). This file is a cross-section of "derived variables" whose values are defined based on combinations of data items (e.g. age at COMPASS clinic visit or GPAQ variables). COMPASS sample weight are included in this dataset. See the separate document, *"SOL COMPASS 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 (COMPASS_WEIGHT_NORM) which includes CASAS participants. 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. COMPASS_PA_EPOCH (Counts at the 15sec EPOCH level)

Dataset with counts for each EPOCH (15 sec).

6.5. COMPASS_PA_MIN (Counts at the MINUTE level)

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

6.6. COMPASS_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. COMPASS_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. COMPASS_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.

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.

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