NIMHD grant: Ancillary to HCHS/SOL: Gut Origins of Latino Diabetes Study
(R01 MD011389; R Kaplan, RD Burk, & RD Knight)

HCHS/SOL GOLD
Ancillary Study
Investigator Use Database Overview

February 2019
INV Version 2.0

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

Daniela Sotres-Alvarez
Yanping Teng

ACKNOWLEDGMENTS

HCHS/SOL GOLD Ancillary Study would not have been possible without the participation of HCHS/SOL participants and the support of study member staff. GOLD Ancillary Study (R01 MD011389) was funded by the National Institute of Diabetes and Digestive and Kidney Diseases.

Special thanks to (Alphabetically):

Alvarado, Cesar, BA
Clinic Manager
University of Illinois at Chicago

Burk, Robert D, MD
GOLD AS Principal Investigator
Albert Einstein College of Medicine

Crespo-Figueroa, Madeline, BS
Bronx Project Coordinator
Albert Einstein College of Medicine

Daviglus, Martha, MD, PhD
Chicago Principal Investigator
University of Illinois at Chicago

Gellman, Marc, PhD
Miami Principal Investigator
University of Miami

Gonzalez, Sara, PhD
Clinic Manager
Albert Einstein College of Medicine

Hernandez, Johanne
SD Project Coordinator
San Diego State University

Kaplan, Robert, PhD
GOLD AS Principal Investigator
Albert Einstein College of Medicine

Knight, Rob, PhD
Co-Investigator
University of California San Diego

Malo, Ana Rebeca
Clinic Manager
San Diego State University

Mirabal, Silvia
Clinic Manager
University of Miami

North, Kari, PhD
Co-Investigator
University of North Carolina at Chapel Hill

Sollecito, Christopher, BS
Research Technician
Albert Einstein College of Medicine

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

Talavera Ana, MPH
San Diego Project Coordinator
San Diego State University

Talavera Greg, MD, MPH
San Diego Principal Investigator
San Diego State University

Teng, Yanping, MSPH
Senior Biostatistician
University of North Carolina at Chapel Hill

HCHS/SOL GOLD Ancillary Study

Investigator Use Database INV2 (Feb. 2019)

Investigator Use Database Overview

Table of Contents

Updates to GOLD Data Release or Documentation

Version	Date	Description	Datasets	Documentation
1	11/30/2018	1st data release	_INV1	V1.0 (Nov 2018)
2	2/1/2019	**2nd data release. Updates:** - 56 participants are included which were excluded in INV1 release (N=3,433) because their consent form was not entered into HCHS/SOL Data Management System (CDART). All four sites verified having the hard copy of the inform consent and corrected this in CDART. As a result, the number of participants in GOLD is N=3,489. - GOLD sampling weights were updated to include all 3,489 participants. - Variables SURVEY_DATE and S1_FREQ are now included in GOLD_PART_DERV.	_INV2	V2.0 (Feb. 2019)

1. INTRODUCTION

This document describes the content and structure of the Investigator Use Database created for HCHS/SOL GOLD Ancillary Study. This database contains all the data collected for the **3,489 enrolled and consented GOLD participants,** subject to constraints (described within) to preserve participant confidentiality by de-identifying the data.

2. STUDY OBJECTIVES

The Hispanic/Latino population is the fasting growing segment of the US population. Diabetes disproportionately affects this group. National US 2007-09 data found that >20 yr old Hispanics (11.8%) have a 66% higher rate of diabetes compared to non-Hispanic whites (7.1%). In the population-based Hispanic Community Health Study (HCHS)/Study of Latinos (SOL), diabetes had a baseline prevalence of approximately 17%. Very recent data implicates the gut microbiome (GMB) as a key determinant of diabetes. Since different ancestral populations harbor different diabetes-associated sets of GMBs, it is necessary to study Hispanic/Latino populations with high rates of diabetes to determine the relationship between the GMB and diabetes. Understanding the relationship of the GMB to diabetes is anticipated to lead to a new era of prevention and treatment options, especially since therapeutic interventions are available that target the GMB. Nevertheless, there are major gaps in understanding the epidemiology of the GMB in the population and its role in the development of diabetes.

Specifically, GOLD: (i) collected stool samples from 3,489 subjects attending HCHS/SOL V2 at all four HCHS/SOL sites; (ii) analyzed stool samples for GMB; and (iii) will perform analysis of GMB integrated with extensive existing data in HCHS-SOL. Designed in the setting of a large, population based sample, HCHS-SOL is uniquely suited to discern ethnic differences in GMB makeup and to study the association of GMB with diabetes across birthplace/national background groups. GOLD will leverage the HCHS/SOL study by making use of existing HCHS-SOL data already in place including metabolomics, sociodemographics, health behaviors and genetics (2.5M SNP GWAS array and whole genome sequencing).

The aims of GOLD are:

Aim 1. Investigate factors affecting the gut microbiome (GMB) among Latinos. Investigators hypothesize that GMB composition differs with national background (e.g., Mexican, Puerto Rican, Cuban, etc.), birthplace (80% are foreign born), gender, age, adiposity, shared household, genetics and relatedness.

Aim 2. Evaluate the association of the gut microbiome (GMB) with the presence of diabetes and pre-diabetes. Investigators' hypothesis is that the GMB differs between across three groups defined at V2, including: *diabetes mellitus* (measured FPG≥126; HbA1c≥6.5%; 2hPG≥200 and/or diabetes medication use); *prediabetes* (FPG 100-125; HbA1c 5.7%-6.4%; 2hPG 140–199), and *normoglycemic* (normal FPG, HbA1c and 2hPG).

3. STUDY DESIGN

Original GOLD Study protocol established to collect and determine the genetic composition of the fecal microbiome from 2,000 HCHS/SOL cohort members. There were no exclusions, and all participants attending HCHS/SOL visit 2 were invited to participate. The study was fortunate to recruit more participants for a final sample size of 3,489 participants.

4. DATABASE STRUCTURE

4.1. Data Set Organization

There is one 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. Collection of direct measurements during examination procedures can also result in the creation of a data file.

Special **derived variable dataset (PART_DERV_GOLD_INV2)** has been created to with GOLD specific variables (e.g. age at GOLD clinic visit). The participant derived variable algorithms have been included in the GOLD Derived Variable Dictionary.

A codebook has been produced for each data set. A careful review of the codebook, 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. GOLD datasets

Administrative	Form Code (SAS dataset)	Number of observations
Pre-visit screening (eligibility, safety)	GOL	3,489
Inform Consent	GIC	3,489
Questionnaires		
Health and diet questions	GLE	3,488
Derived Variables		
Participant Derived	PART_DERV_GOLD	3,489

4.2. Form and Data Set Naming Conventions

Each GOLD data collection instrument (Form) has a unique three-letter mnemonic associated with it (e.g., GOL for the GOLD Enrollment and Tracking 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 GOLD Enrollment and Tracking data set is "GOL_INV2". The naming convention serves both to identify the originating form and provide version control when subsequent generations of 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.

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.

4.4. Common Variables Across Data Sets

Additional variables appear in every data set; OCCURRENCE, VERS and VISIT are meaningless in GOLD.

OCCURRENCE: It is 1 for all records except in GLE that has 2 participants for which a 2nd record was received and replaced 1st one (i.e. no multiple records per person)

VERS: It is 1 for all records (no change inversions in the FORMS)

VISIT: It is 1 for all records (1st data collection for GOLD).

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 GOLD Enrollment and Tracking form (GOL) is named GOL1 on the corresponding SAS file, GOL_INV2. Similarly, question 3 from the GOLD Enrollment and Tracking form is named GOL3.

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 GOLD clinic visit and appears in the derived variable data sets, GOLD_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

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. GOLD Enrollment and Tracking Form (GOL)

The GOLD Enrollment and Tracking form (GOL) includes the script for determining a participant's eligibility. The participant's eligibility and participation status was recorded in section A of the form. The Ancillary study sample Specimen kit distribution and return dates were recorded in section B. Section C is administrative use only section and was prefilled by HCHS/SOL data Management System.

5.2. GOLD Informed Consent Tracking Form (GIC)

This questionnaire is about the participants' informed consents on HCHS/SOL GOLD to collect questionnaire data and specimens, the use and the new collection of genetic data, allow the samples and data to be used for current and future research and agreement on future studies samples.

5.3. GOLD Questionnaire Form (GLE)

This brief questionnaire has 7 questions related to health, meat dietary preference, and special diet.

6. DERIVED FILES

6.1. GOLD_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 (3,489 observations). This file is a cross-section of "derived variables" whose values are defined based on combinations of data items (e.g. age at GOLD clinic visit). **GOLD sampling weights are included in this dataset.** See the separate document, *"GOLD 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 (WEIGHT_NORM_OVERALL_GOLD). 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.