MITRE

The Clinical and Community Data Initiative

Sponsor: Centers for Disease Control and Prevention
Dept. No.: P351
Contract No.: 75FCMC18D0047
Project No.: 37208164

Clinical and Community Data Initiative Household Prevalence Queries Implementation Guide
Version 1.1

March 18, 2021

Record of Changes

Version	Date	Author / Owner	Description of Change
DRAFT	November 30, 2021	Erin Tanenbaum / Health FFRDC	Initial Draft
DRAFT 1.0	December 23, 2021	Erin Tanenbaum / Health FFRDC	Updated Draft
DRAFT 1.1	March 18, 2022	Melissa Garcia	Update Draft

Methodology and SAS Programming Contributors

Name	Affiliation
Erin Tanenbaum	NORC at the University of Chicago
Shalima Zalsha	NORC at the University of Chicago
Scott Campbell	NORC at the University of Chicago
Devi Chelluri	NORC at the University of Chicago
Jason Boim	NORC at the University of Chicago
Kennon Copeland	NORC at the University of Chicago
Susan Paddock	NORC at the University of Chicago
Dawn Heisey-Grove	MITRE
Melissa Garcia	MITRE
Andrew Gregorowicz	MITRE
Kris Mork	MITRE
Daniel Chudnov	MITRE
Melissa Bruno	MITRE
Samantha Lange	CDC
Raymond King	CDC

Contact Information

For answers to questions about CODI-PQ, contact:

Erin Tanenbaum
Senior Statistician
NORC at the University of Chicago
4350 East-West Highway, 8th Floor, Bethesda MD 20814
Email: Tanenbaum-Erin@norc.org
NORC.org

Table of Contents

List of Figures

List of Tables

1 Introduction

As part of the Centers for Disease Control and Prevention's (CDC) efforts to promote health, prevent disease, injury, and disability, and prepare for emerging health threats, the Division of Nutrition, Physical Activity, and Obesity partnered with the Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare Federally Funded Research and Development Center (Health FFRDC) on the Clinical and Community Data Initiative (CODI). CODI brings together data stored across different sectors and organizations to create individual-level, linked longitudinal records that include SDOH, clinical and community interventions, and health outcomes. The CODI infrastructure expands the ability to standardize, integrate, query, share, and analyze these data in a manner that preserves privacy and supports community efforts to improve health using data-driven approaches. This includes the development of statistical methods and tools to extrapolate information captured in an electronic health record, which is a convenience sample or non-probability sample, to the general population.

The Health FFRDC developed open-access statistical programs, referenced here as the CODI household prevalence queries (CODI-HPQ) to generate BMI category prevalence estimates based on body mass index (BMI)[1] in adults, youth, and teens living within the same household. CODI-HPQ were designed to use data from the CODI distributed health data network (DHDN) and other non-probability samples derived from Electronic Health Records (EHRs).

1.1 Background

Public health surveillance of household obesity often relies on self-report surveys such as the Youth Risk Behavior Surveillance System surveys in which data for children is provided by a parent or the Behavioral Risk Factor Surveillance System for adults. Self-reported or proxy-reported data can be subject to bias. Additionally, these surveys can be expensive to administer, limited in geographic specificity, and may struggle with response rates and timeliness. Data from EHRs have the potential to play a significant role in obesity population health surveillance, programs, interventions, and evaluations. EHR data – measurements, diagnoses, observations, prescriptions, and procedures – provide non-probability samples of health outcomes among the care-seeking population and the opportunity to provide decision makers with detailed, timely, and accurate information of large numbers of patients within proximal geographies. Despite these advantages, aggregate EHR data at the population level are subject to bias.

Several factors influence the relevance of EHR data for population health. First, the representativeness of the EHR cohort to the population of interest within a geographic or other unit of investigation (e.g., similarity in distribution of sex, race, and age). Second, the proportion of the population captured by a health system's EHR. Third, the number of events captured in the EHR cohort. A small number of events could result in unstable estimates and reflect poor EHR coverage, a small underlying population (e.g., rural community) and/or a rare event. Finally, the data generating process in an EHR depends on when and why a patient visits a healthcare provider, resulting in missing values that may be attributed to a lack of occurrence of that event, a lack of documentation of that event, or lack of data collection. Statical methods and data standards can be used to address these limitations.

[1] About Adult BMI.

CODI-HPQ provide a suite of tools to address some of these limitations and to calculate population obesity prevalence estimates from EHR data using statistical weights, imputation, and suppression criteria. Statistical weighting is used to reduce non-probability sample bias and produce representative distributions of the populations of interest. Imputation is used to infer missing race/ethnicity and enable estimation across subpopulations. The National Center for Health Statistics (NCHS) Data Suppression Criteria for Proportion[2] is adopted as standard to suppress statistically unreliable estimates and ensure limited disclosure of information when samples are small. The CODI-HPQ algorithms can generate stable prevalence estimates at state and county geographies from EHR data, with the aim to improve access to timely data on local disease burden to inform prevention and other public health activities.

1.2 Purpose

The purpose of the CODI-HPQ Implementation Guide is to provide a guide for CODI data partners[3] or end users to run the CODI-HPQ. The Implementation Guide covers the following:

- CODI-HPQ data inputs and link population data (pre-processing)
- Generating results in CODI-HPQ
- Understanding the CODI-HPQ results
- Methodological details

Contact Information

For answers to questions about CODI-HPQ, contact:

Erin Tanenbaum

Senior Statistician

NORC at the University of Chicago

4350 East-West Highway, Suite 800

Bethesda, MD 20814

Email: Tanenbaum-Erin@norc.org

1.3 Scope

The CODI-HPQ algorithms were created and tested with synthetic data generated for CODI using Synthea.™[4] CODI-HPQ analyze data for patient level records for patients ages 2 through 64. Each record must include year of medical encounter, demographic information (age, sex,

[2] Parker et al., 2017.
[3] CODI data partners are organizations and institutions which facilitate CODI data exchange by contributing and hosting data that can be accessed through the CODI infrastructure for queries and other research or programmatic uses of the data.
[4] https://synthetichealth.github.io/synthea/

race, and some level of geographic location), and BMI category. Patient-level records must include a household identifier (see Appendix A) along with residential address information at the level of state and county codes. CODI- leverages population counts from the American Community Survey (ACS). CODI-HPQ assume that end users include all EHR data for a geography and/or subpopulation that they have available.

All statistical programs described in this document were created and tested using SAS 9.4 software (SAS Institute, Inc., Cary, North Carolina). The guidance provided in this document is implemented through open-access programs.

1.4 Audience

The audience for this IG is CODI data partners and end users. The user should have a working knowledge of SAS language and macros. Those interested in statistical analysis details used in CODI-HPQ can refer to Appendix A for more information. Technical staff preparing datasets for CODI-HPQ can refer to Appendices B and C for detailed descriptions of the format required for input data. Explanation of CODI-HPQ results can be found in Appendix E.

1.5 Document Organization

This document is organized as follows:

Section		Purpose
Section 1	Introduction	Provides a background for CODI-HPQ
Section 2	User's Guide	Provides a general guide for users
Appendix A	Analysis Details	Provides detailed description of analysis
Appendix B	ACS File Layouts	Table outlining the required ACS input file layouts
Appendix C	EHR File Layouts	Table outlining the required EHR input file layouts
Appendix D	CODI-HPQ GEO3 Example SAS Programs	Provides example SAS program
Appendix E	CODI-HPQ Results	Provides CODI-HPQ results data dictionary and example results
Appendix F	State FIPS codes	Provides list of state abbreviations
Appendix G	Glossary	Defines terms used in this document
Appendix H	Abbreviations and Acronyms	Defines acronyms used in this document
Appendix I	Bibliography	Lists sources used in preparing this document

2 User's Guide

The User's Guide section describes:

1. The CODI Project
2. How to prepare your data for the programs
3. How to run the CODI-HPQ programs (Important step! Carefully review specifications.)
4. The CODI-HPQ output

2.1 CODI Concept

Error! Reference source not found.1 shows how CODI users (e.g., researchers, community-based program evaluators) interact with the data coordinating center, which distributes their research queries to data partners. The data coordinating center assembles the results into longitudinal records, which are sent to the CODI end users. CODI end users use the patient-level longitudinal records to create prevalence estimates with CODI-HPQ. CODI-HPQ can also be used on cross-sectional data. Additional CODI details can be found in the documentation available through GitHub at https://github.com/mitre/codi.

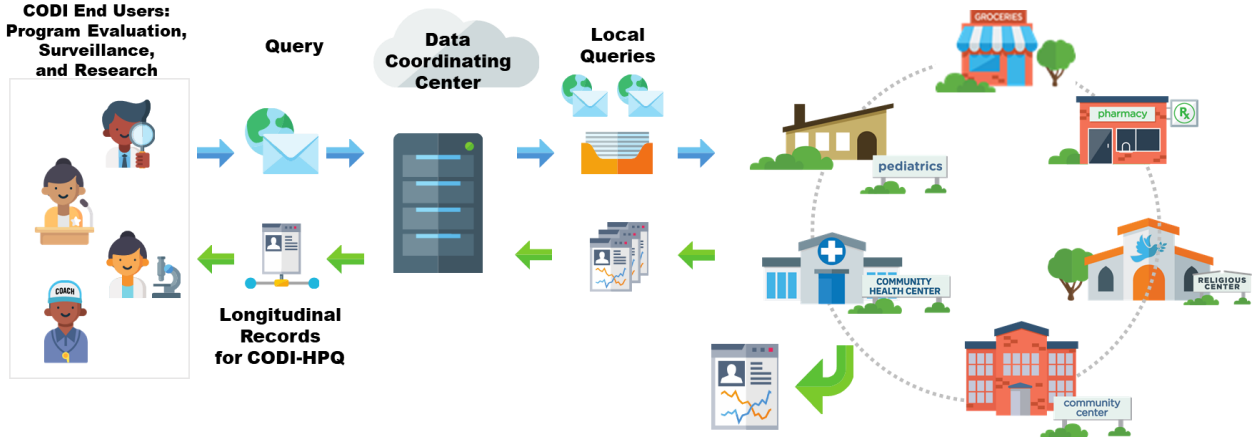

Figure 1. Data Partners with a Common Data Coordinating Center

2.2 About CODI-HPQ

CODI-HPQ are a set of programs that calculates household BMI prevalence estimates from a non-probability sample[5] of EHR patients linked to the same household. CODI-HPQ programs are divided into two parts: 1) pre-processing, and 2) the prevalence query. In pre-processing, patient data are imported into SAS and linked to the American Community Survey (ACS), household characteristics are summarized, and race imputation is conducted in the pre-processing steps (CODI_HPQ_PRE_PROCESSING_GEO3). In the prevalence query part, households are selected based on user specifications, statistically weighted, variance estimates

[5] Non-probability sample is a group of individuals based on a sampling method in which not all members of the population have an equal chance of being a part of the sample. In probability sampling, each member of the population has a known chance of being selected. Thus, probability sampling is more stringent than non-probability sampling.

are calculated, results are suppressed (if needed), and prevalence results are output (CODI_HPQ_GEO3).

For successful use of the CODI-HPQ programs, end users are encouraged to carefully review the methodological details (described in appendices). Inputs for the CODI-HPQ programs include EHR data supplied by the user and ACS data from 2019 supplied by the Health FFRDC[6]. Results can be calculated for a specific geography (e.g., state, state and county), subpopulation (e.g., youth and teen age group, number of adults in the household, race), or geography and subpopulation (e.g., age group by state and ZCTA-3).

Results are suppressed[7] if the user selects a geography or subpopulation with an insufficient number of households for statistical weighting (see Appendix Section A.5) or if results do not meet NCHS suppression criteria (see Appendix Section A.8). The CODI-HPQ programs user should have a working knowledge of SAS language and macros to select the population of interest, execute CODI-HPQ, and review the SAS log.

The programs described in the User's Guide are designed to:

- Link patients based on a user specified household identifier
- Designate one adult as the householder
- Impute race for householders who are missing race information (optional)
- Calculate statistical weights with an EHR non-probability sample
- Calculate household BMI prevalence by BMI, including:
  - **Youth and teen**
    - **No youth or teen has obesity**; with BMI percentile less than 95th percentile
    - **One or more youth or teen has obesity**; with BMI percentile greater than or equal to the 95th percentile
  - **Adults**
    - **No adult has obesity**: BMI less than 30 kg/m^2
    - **One or more adult has obesity**: BMI greater than or equal to 30 kg/m^2
- Suppress prevalence estimates based on the National Center for Health Statistics (NCHS) Data Presentation Standards for Proportions

2.3 SAS Setup

All statistical programs described in the User's Guide were created and tested using SAS 9.4 software (SAS Institute, Inc., Cary, North Carolina) in a Windows environment. CODI-HPQ require the following SAS features:

[6] ACS 2019 file for use with CODI-HPQ is available for download by request; contact CODI@cdc.gov. The 2019 ACS data was used for model calibration. Use of other years of ACS data requires recalibration of the model due to changes in population counts.
[7] SAS outputs a dot (.) instead of a numeric value when results are suppressed. Suppression occurs by row and may include one or more than one row of results.

- BASE SAS
- SAS STAT
- The ability to import a file from csv into SAS
- The ability to export a file from SAS into csv

2.4 Step-By-Step Process to Run CODI-HPQ

The four-step process to run the CODI-HPQ is outlined below:

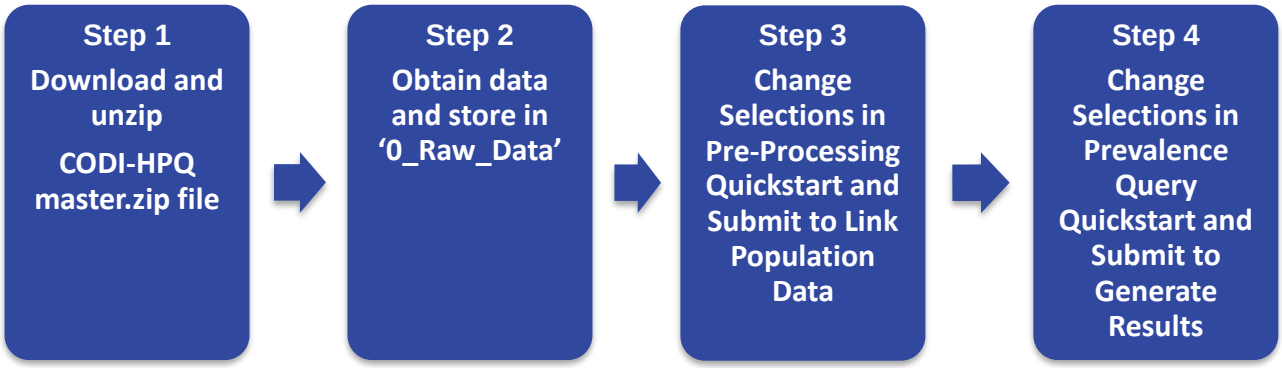

Figure 2. CODI-HPQ Process

2.4.1 STEP 1: Download and Unzip CODI-HPQ-master.zip File

Access CODI-HPQ programs on GitHub: https://github.com/NORC-UChicago/CODI-PQ.

To begin, select the "Households" folder and download "CODI-HPQ-GEO3-master.zip". Note that "GEO3" refers to the county code which is three digits long.

Use file compression software to unzip the files. Be sure the option is selected to unzip both files and folders and preserve the folder names.

CODI-HPQ-GEO3's folder structure is shown in the figure below. Note that folders and subfolders have been created and structured in a way to make it easier for the user to organize the input and results files.

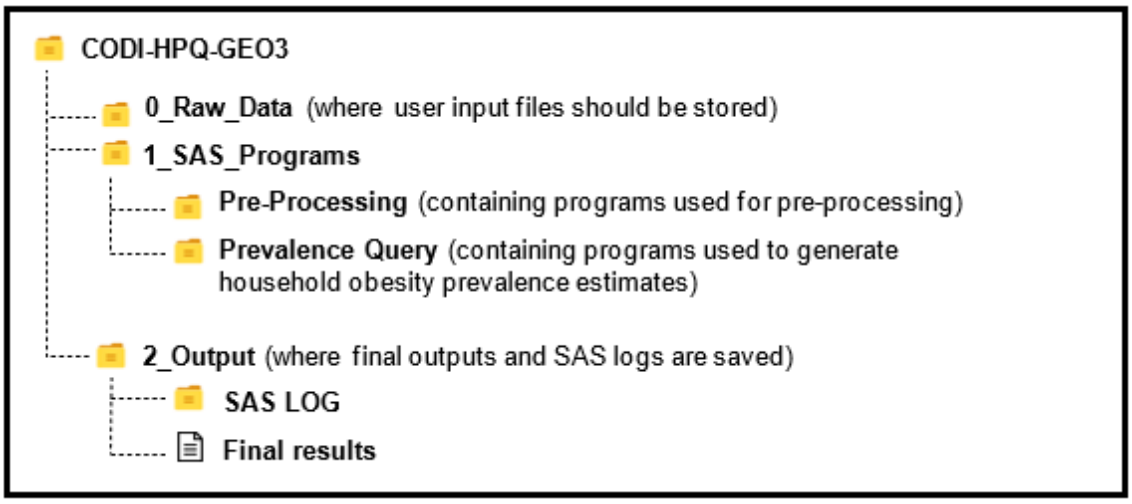

Figure 3. CODI-HPQ-GEO3 Folder Structure

2.4.2 STEP 2: Obtain Input Files and Store Them in the '0_Raw_Data' Folder

Required input files include:

1. **ACS data file** of specific variables from the 2019 ACS can be downloaded from the Health FFRDC via Secure File Transfer Protocol (SFTP). (Contact CODI@cdc.gov for permission to access this file via SFTP.) This file is cited to ensure consistency with the models embedded into the SAS programs. For variable names, variable order, and a description of the file, see Appendix .B
2.
3. **EHR data file** supplied by the end user in comma separated values (.csv). The EHR data file is assumed to contain:
    - All **variables in the order (sequence) required for accurate results**. Variable names and order can be found in Appendix C.
    - **Valid variable values** as anticipated. Variable values can be found in Appendix C.
    - A **unique identifier for all patients** and the identifier is consistent between years.
    - A **unique identifier for all households** and the identifier is consistent between years unless the household moves to a new housing unit. See A.1 for more details.
    - A **maximum** of one record per patient per year. The user can choose the record kept so it aligns with analysis goals. For testing purposes, the event date closest to July 1 of each year was kept prior to executing pre-processing.
    - A valid height and weight value obtained on the same day which was used to calculate BMI and assign the **BMI category** for all patients (underweight, healthy weight, etc.)
    - Have a geographic location of the **patient's residency (state and county[8])**.
    - Have the **same residency location for persons with the same household identifier**. Note: CODI-HPQ randomly select one adult as the householder and will use this information to determine geographic location.
    - **Users may also wish to reconcile racial** characteristics (optional) for each patient across years[9].

A full description of the EHR data file format is available in Appendix C.

2.4.3 STEP 3: Link Population (Pre-Processing)

Open the "Quickstart-Pre_Processing_CODI_HPQ_GEO3" SAS program stored in "\1_SAS_Programs\Pre-Processing," change selection per the steps outlined in the tables below.

[8] COUNTY is based on FIPS state and county code.
[9] If creating racial estimates for more than one time point, allowing race to change over time will create increased volatility in the estimates.

Note that the pre-prevalence program should be submitted once and only once per file. As such, include the start and end years for the full EHR file. The programs also impute the race of householders with unknown race. Thus, each time the program is submitted, new imputed race values are created and stored. For consistency, we encourage submitting the pre-processing programs once and only once for each EHR file. If additional data is later processed for the same households, we encourage 1) replacing the race of all patients who were imputed before, but their race is now known, 2) keeping the imputed race value consistent for patients who were imputed before and their race value is still unknown.

A new folder ("\2_Output\Pre-Processed_…") will be created upon completion of the programs. In this folder, two SAS7bdat files (user input ACS file and pre-processed CODI file) will be generated. Once pre-processing is complete, the user can submit an unlimited number of household prevalence queries using the same pre-processed files each time.

Table 1. Change Specifications, Pre-Processing Steps

Order	Description	Details
1	Open the Pre-processing Quickstart program.	The Quickstart program is stored in the folder: "..\1_SAS_Programs\Pre-Processing\Quickstart-Pre_Processing_CODI_HPQ_GEO3.sas"
2	Edit the SAS program within "SECTION 1: Input Folder and file names."	Follow the SAS programs and update the macro variable specifications (see Table 2).

Table 2. Change SAS Specifications

SAS Macro Variable	Details	Example
ROOT_HPQ	The core folder name where CODI-HPQ-GEO3 are saved. The SAS Programs folder and all other folders and files are stored in this directory.	%let ROOT_HPQ = C:\Example\CODI_HPQ_1130;
PRE_HDEST	The folder name for results inside the "\2_Output"	%let PRE_HDEST = CODI_HPQ_PRE; /* output would be stored in C:\Example\CODI-HPQ_1130\2_Output\Pre_Processed_**CODI_HPQ_PRE***/
EHR_H_PRE_OUT	User can name the suffix of the pre-processing output file (ACCEPTABLE VALUES: file name (no punctuations)).	%let EHR_H_PRE_OUT = CODI_HPQ_Preprocessed_Filename;
EHR_FILENAME	The comma delimited (csv) person level EHR file from part 2.4.2. This file includes patients age 2 to 64. Do not include the extension (e.g. .csv).	%let EHR_FileNAME = EHR_example_Household;
ACS_FILENAME	The American Community Survey file name from part 2.4.2. The file is in csv format. Do not include the extension in the file name.	%let ACS_FILENAME = ACS_COUNTY;
LOG_NAME_PRE	The name of the SAS log file. Users have the option to rename the log file name before it is created.	%let LOG_NAME_PRE = LogNameHERE; /*the SAS log will be stored in: C:\Example\CODI-HPQ_1130\2_Output\SAS LOG**LogNameHERE**<Date and Time>.log. Note, the program automatically includes the date and time in all log file names*/

Table 3. Change Specifications, Pre-Processing Steps, Continued

Order	Description	Details
3	Save the Quickstart program.	SAS encourages saving all files before submitting the program.

Table 4. Pre-Processing CODI-HPQ Program Execution Steps

Order	Description	Details
1	Submit the Quickstart program.	Submit the Quickstart program. The program completes all tasks within the data sets and proc statements in the Quickstart program and moves to the next SAS program automatically through an include statement. It is important to submit the full SAS program.
2	Review the log.	Review the log for possible errors including words such as error, repeat, and uninitialized. Assuming no errors[10], continue to Part 4. In the event of errors, reassess the location of the files and the file formats.

2.4.4 STEP 4: Generate Prevalence Estimate Results

Open the "Quickstart-CODI_HPQ_GEO3" SAS program stored in "\1_SAS_Programs\ Prevalence Query" and change the selections within the program per the steps outlined in the tables below.

The final results (CODI-HPQ results) will be generated in Excel format and saved in "\2_Output." Appendix E provides examples of the results. Note that results are for the group of households selected by the user. To calculate results for multiple geographic or demographic characteristics (e.g., by race), the user will need to update and execute the programs multiple times.

Note: the age ranges and races selected must match the data on the EHRs. For example, if estimates for Asian households (only) is selected by the user and the file does not have Asian householders then the program will fail with an error message caused by insufficient sample size.

[10] The iterative proportional fitting macro does create uninitialized comments.

Table 5. Change Specifications, Processing Steps

Order	Description	Details
1	Open the Quickstart program.	The Quickstart program is stored in the folder: "..\1_SAS_Programs\Prevalence Query" and is named "Quickstart-CODI_HPQ_GEO3"
2	Edit the SAS program within "SECTION 1: Folder and file names"; "SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE OR STATE/COUNTY CODE"	Follow the SAS programs and update the macro variable specifications.

Table 6. Change Specifications, Processing Steps

SAS Macro Variable	Details	Example
SECTION 1: Folder and file names		
ROOT_HPQ	The core folder name same as in pre-processing.	%let ROOT_HPQ = C:\Example\CODI_HPQ_1130;
PRE_HDEST	The pre-processing quickstart variable pre_hdest.	%let PRE_HDEST = CODI_HPQ_PRE;
EHR_H_PRE_OUT	The patient level EHR file name same as in pre-processing.	%LET EHR_H_PRE_OUT = CODI_HPQ_Preprocessed_Filename; /*following the same example as above, the results from pre-processing will be stored as a SAS data file in C:\Example\CODI-HPQ_1130\2_Output\ Pre_Processed_**SAVE_PRE_FILE_HERE** */
LOG_NAME	The name of the resulting SAS log. Users have the option to rename the log file name before it is created.	%let LOG_NAME = THISisTHElog; /*the SAS log will be stored in: C:\Example\CODI-HPQ_1130\2_Output\SAS LOG**THISisTHElog**<Date and Time>.log. Note, the program automatically includes the date and time in all log file names*/
FileOUT_Name	The prefix for the resulting Excel file.	%LET FileOUT_Name = File_name; /*the .csv or Excel file will be stored in: C:\Example\CODI-HPQ_1130\2_Output**File_name**<Date and Time>.xls. Note, the program automatically includes the date and time in all results file names*/
SECTION 2: Subset data based on specifications INCLUDING YEAR,		

SAS Macro Variable	Details	Example
GEOGRAPHY, STATE, or STATE/COUNTY CODE		
ALL_H_STATES	Includes all states (including D.C.) in the prevalence based on the geographic location of the household. If ALL_H_STATES = N; then by default the program will subset the prevalence based on the individual state or state and county specified (in future step).	%LET ALL_H_STATES = N; /*@Note: EHRs file includes all of the US? (ACCEPTED VALUES: Y/N) ***/
H_YEAR	Subsets the prevalence to medical encounters in this year. The prevalence will include EHR data from this year only.	/***/ %LET H_YEAR = 2016; /*@Note: Year of analysis (ACCEPTED VALUES: 4-Digit numeric, e.g. 2019) ***/
ALL_H_AGES	Subsets the prevalence based on the age of the children in the household. If ALL_H_AGES = N; then by default the program will subset the prevalence based on the individual age groups specified (in future step).	%LET ALL_H_AGES = Y; /***(ACCEPTED VALUES: Y/N) ***/
ALL_RACES	Subsets the prevalence based on the race of the householder. The user may either select to include all races or alternatively may select race(s). Inclusion or exclusion of imputed race is not impacted by the choice made in this step. Note: if ALL_RACES = Y; then by default the program will include all races (White,	/***/ %LET ALL_RACES = Y; /*@Note: Include all race categories? (ACCEPTED VALUES: Y/N) ***/

SAS Macro Variable	Details	Example
	Black, Asian, Other). If ALL_RACES = N; then by default the program will subset the prevalence based on the individual races selected (in future step).	

Table 7. Change Specifications, Processing Steps, Continued

SAS Macro Variable Category	Details	Example
SECTION 3: Only complete section 3 for any "N" values listed in section 2		
If ALL_H_STATES = N	GEO_H_GROUP informs the program what level of geography is to be used in the GEO_H_LIST macro variable. GEO_H_LIST subsets the prevalence based on the location of the household. GEO_H_GROUP can take the value of a) STATE, or b) STATE/COUNTY. Below are possible values for two scenarios. Of note, values should be surrounded by single quotes and comma delimited if more than one geography is to be included in the results. a) If GEO_H_GROUP=STATE; then the program defaults to using state FIPS codes. For example, %STR('08', '10') would select Colorado and Delaware. b) If GEO_H_GROUP=STATE/COUNTY ; then the program defaults to using a concatenated state FIPS and County code(s). For example, %STR('51061', '51059') would select patients living in Virginia, within Fauquier County and Fairfax County.	/***/ %LET GEO_H_GROUP = STATE; /*@Note: Level of geography (ACCEPTED VALUES: STATE, or STATE/COUNTY) ***/ /***/ %LET GEO_H_LIST = %STR('08', '10'); Or for state and county: /***/ %LET GEO_H_GROUP = STATE/COUNTY; /*@Note: Level of geography (ACCEPTED VALUES: STATE, or STATE/COUNTY) ***/ /***/ %LET GEO_H_LIST = %STR('51061', '51059');
If ALL_RACES = N;	If ALL_RACES is set to no, the race macros (White, Black, Asian, Other) subset the household prevalence based on the race or imputed race of the householder. Note that if ALL_RACES is set to Y, then the SAS program does not review the race-specific macros.	%LET RACE_WHITE = N; %LET RACE_BLACK = Y; %LET RACE_ASIAN = Y; %LET RACE_OTHER = Y;

SAS Macro Variable Category	Details	Example
SECTION 4: Methodological option selections		
IMP_RACES	If IMP_RACES is set to Y (yes), then the program includes households with imputed householder race values. Otherwise, if IMP_RACES is set to N (no), then the households with imputed householder race are excluded.	%LET IMP_RACES = Y;

Table 8: Change Specifications, Processing Steps, Continued

Description	Details
Save the Quickstart program.	It is encouraged to save the Quickstart **program before submitting in SAS.**

Table 9. CODI-HPQ Execution Processing Steps

Order	Description	Details
1	Submit CODI-HPQ Quickstart program.	Submit the Quickstart program. The program completes all tasks within the data sets and proc statements in the Quickstart program and moves to the next SAS program automatically through an include statement.
2	Review the log.	Review the log for possible errors including words such as error, , repeat, and uninitialized. Assuming no errors, continue to step 3. In the event of errors, reassess the location of the files and the file formats.
3	Review the results.	Review the results for possible data suppression or errors. Consider a statistical review based on the NCHS data presentation standards. In the event of errors reassess the choices and re-submit. In the event of data suppression, consider expanding your selection criteria and re-submit. For example, if prevalence results cannot be created for a single county, consider using two or more counties of data[11].

2.4.5 Review BMI Category Prevalence Results

CODI-HPQ generate prevalence outputs as an Excel file. Table 10 provides an overview of the variables included. Note, descriptive information about CODI-HPQ user inputs, error codes, sources of technical documentation, caveats, and a possible citation begins with the rows labeled Order 5.

[11] Note: If more than one year is selected, the first record of each SUBJID is kept with all subsequent records excluded from prevalence results to meet statistical weighting assumptions.

Table 10. CODI-HPQ BMI Prevalence Results Data Dictionary

Column	Description
Order	Row order
Youth and Teens Weight Category	A categorical value based on BMI percentile of youth and teen(s) in households.
Adults Weight Category	A categorical value based on BMI of adult(s) in households.
Sample	The observed (or unadjusted, or crude) count of households in the study population.
Population	The weighted (or adjusted) count of households.
Crude Prevalence	The observed (or unadjusted, or crude) household prevalence in the study population.
Crude Prevalence Standard Error	The observed (or unadjusted, or crude) household standard error in the study population.
Weighted Prevalence	Household prevalence based on weighted counts. A sample weight is assigned to each sampled household. It is a measure of the number of households in the population represented by that sample household. See implementation guide, Appendix A. Statistical Weights for more information.
Weighted Prevalence Standard Error	Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information.

2.5 Additional Details for Users

Further detail on file layouts for input and results is provided in the following appendices:

- Appendix B – ACS File Layouts
- Appendix C – EHR Data File Layouts
- Appendix D – CODI-HPQ-GEO4 Example SAS Programs
- Appendix E – CODI-HPQ Results Example
- Appendix F – State FIPS Codes
- Appendix G – Glossary
- Appendix H – Abbreviations and Acronyms
- Appendix I – Bibliography

Appendix A Analysis Details

A.1 Household, Household Identifier, and Householder

A.1.1 Household

A household according to the United States Census Bureau (U.S. Census) consists of all the people who occupy a housing unit. A house, an apartment or other group of rooms, or a single room, is regarded as a housing unit when it is occupied or intended for occupancy as separate living quarters; that is, when the occupants do not live with any other persons in the structure and there is direct access from the outside or through a common hall.

A household includes the related family members and all the unrelated people, if any, such as lodgers, foster children, wards, or employees who share the housing unit. A person living alone in a housing unit, or a group of unrelated people sharing a housing unit such as partners or roomers, is also counted as a household.

The Household Prevalence Query includes households that have a minimum of one adult and one child (age 2 through 18) assigned to the same household identifier. All counts of households within CODI-HPQ are based on the Census definition of households with children.

A.1.2 Household Identifier

CODI-HPQ require all persons have a household identifier. Based on the Census definition of a household, the household identifier should be identical to all persons living in the same housing unit. If an individual moves to a different housing unit, then the household identifier should change. Since each person should only be on the file once per year, a person cannot have two or more household identifiers per year. A proxy for housing unit is often a person's address, although more than one housing unit could live at the same address.

In the synthetic dataset used to develope CODI-HPQ, individuals and households were linked across organizations using a process called Privacy Preserving Record Linkage (PPRL). The same process will be used in CODI. PPRL involves each data owner garbling select personally identifiable information (PII) for individuals and households, and then sending that garbled data to the linkage agent for matching. This process ensures privacy is protected because it is essentially impossible to reverse the garbling function to determine the original PII, however the encoding format has properties that enable matching of similar values. The linkage agent then performs a matching process that compares records and households across all data owners. Matched individuals are assigned a Link ID, and matched households are assigned a Household Link ID. The linkage agent returns to each data owner a mapping of that data owner's record numbers to Link IDs and Household IDs. Those mappings are stored by the data owner as part of the CODI Record Linkage Data Model (RLDM).

More detail on the PPRL process is available in the CODI PPRL Implementation Guide (IG): https://raw.githubusercontent.com/mitre/codi/main/CODI%20PPRL%20Implementation%20Guide.pdf (Sections 2-3 within the pdf).

More detail on the RLDM is available in the CODI Data Model IG: https://raw.githubusercontent.com/mitre/codi/main/CODI%20Data%20Model%20Implementation%20Guide.pdf (Appendix C within the pdf).

A.1.3 Householder

According to the U.S. Census, a householder refers to the person (or one of the people) in whose name the housing unit is owned or rented or, if there is no such person, any adult member, excluding roomers, boarders, or paid employees. If the house is owned or rented jointly by a married couple, the householder may be either the husband or the wife. The person designated as the householder is the "reference person" to whom the relationship of all other household members, if any, is recorded. The number of householders is equal to the number of households.

The Household Prevalence Query randomly assigns householder status to one adult assigned to each household identifier. If more than one adult is assigned the same household identifier, then the random selection is performed based on the householder age distribution within the state and county using ACS's distribution of age of householder.

Since the number of adults may vary in EHR data from year to year, the householder is randomly assigned independently from previous assignments.

A.2 Body Mass Index

Body mass index (BMI) is a patient's weight in kilograms divided by the square of height in meters. A high BMI can be an indicator of high body fatness. BMI can be used to screen for weight categories that may lead to health problems, but it is not diagnosis of a patient's body fatness or health.

For adults age 20 through 64, BMI is a person's weight in kilograms divided by the square of height in meters. A high amount of body fat can lead to weight-related diseases and other health issues. Being underweight can also put patients at risk for health issues.

BMI categories are described in section A.10.

For more information, see:

https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html.

For youth and teens, BMI is age- and sex-specific and is often referred to as BMI-for-age. BMI is not used as a diagnostic tool for youth (or children) and teens; however, it is used to screen for potential weight and health-related issues.[12]

Youth and teen BMI Percentile categories are described in section A.10.

For more information, see:
https://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html.

A.3 Data Sources (Inputs)

This document provides an implementation guide for CODI-HPQ on patient level data. Required input files are the following:

[12] https://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html, Accessed March 9, 2022.

- EHR data files (data in csv format, provided by user) provided by the user
- American Community Survey (ACS) data file (provided by the Health FFRDC[13])

CODI-HPQ are intended for use with all available EHR data for a geography or subpopulation. The programs were created and tested with the Ambulatory Electronic Medical Record (AEMR)[14] data and synthetic data generated for CODI using Synthea.[15] The guide provided in this document is implemented through open-access programs.

The programs were tested using synthetic EHR data which provides a non-probability sample of longitudinally linked patients' medical records from within the United States. CODI-HPQ subset the file to households with one or more adults aged 20 to 64 years of age and one or more youth or teens aged 2 to 18. The programs assume a maximum of one record per year per patient. Data should include patient identifiers that link medical encounters to demographic and geographic characteristics including year of birth, race, ethnicity (when race is not available), sex, state, and county associated with the patient's address. Patients are excluded from the analysis if their state and county does not exist or if the ACS estimated population count within their county equals 0.

Testing of CODI-HPQ included patient-level EHR data pre-processed using 'growthcleanr.' The 'growthcleanr' package is a publicly available program for identifying biological implausible height and weight measurements in longitudinal files at https://github.com/mitre/growthcleanr-web. The program evaluates data against published growth trajectory charts for youth, teens and adults and flags measurements for plausibility (Daymont et al., 2017).

To statistically weight EHR data to the general population, the 2015-2019 American Community Survey (ACS) 5-year, population estimates by age, race, sex, and community educational attainment are used. Population counts are available by state and county.

A.3.1 Prevalence

A prevalence is either:

- **Crude**: the proportion of the sample that has a health condition (BMI) at a point in time.
- **Weighted**: the proportion of the population within the BMI group at a point in time. See the Appendix A section "Statistical Weights" for more information.

A.4 Race

Race is defined by one of the following categories: White, Black, Asian (including Native Hawaiian and other Pacific Islanders), and Other (including American Indian and Alaskan Native, some other race, two or more race).

[13] ACS 2019 file for use with CODI-HPQ is available for download by request; contact CODI@cdc.gov. The 2019 ACS data was used for model calibration. Use of other years of ACS data requires recalibration of the model due to changes in population counts.
[14] CDC provided Ambulator Electronic Medical Record data under a Data Use Agreement with the Health FFRDC.
[15] The Synthea package is based on Walonoski, et al., 2017 and is available at https://synthetichealth.github.io/synthea/.

These racial categories conform to previous work using a sample EHRs file. These categories are used because they are the race breakdowns available when CODI-HPQ were created, though we recognize that these categories may not accurately reflect the way that patients would self-identify and may conceal important differences within groups.

A.4.1 Race Imputation

Race is a required input for CODI-HPQ. The data inputs and link population data (pre-processing) program inputs race for each householder missing race information. The program operates sequentially in three phases, householders who:

1. Have household members with known race
2. Householder identified as Hispanic
3. Householder identified as non-Hispanic

The race imputation relies on ACS data.

Once complete, the results from each phase are aggregated with each householder with an EHR-provided race, an imputed race, or categorized as "unknown."

A patient's race may be missing after race imputation for one of two reasons:

1. The patient's geography is either invalid or did not have a population count in the 2019 ACS.
2. The sex of the patient is unknown.

CODI-HPQ assign a value for race if a patient does not have a known racial value through statistical imputation. In testing, approximately 27% of IQVIA's AEMR records were missing race (values of "unknown"), yet biases by race were found when compared to the national distribution. Specifically, from a national file, white was overrepresented, and all non-white races were underrepresented. In addition, some electronic records do not store both race and ethnicity separately, thus CODI-HPQ reassign all records that are assigned a "race" of Hispanic (note: Hispanic is an ethnicity, not a race).

As of 2019, racial and ethnic disparities were detected in obesity prevalence in the U.S. To reduce these disparities, high-quality data on race are needed. However, these data are often missing in some portion of EHR data. CODI-HPQ impute race for householders with unknown race using programs based on race and ethnicity of persons in the same household, surrounding the community, ethnicity of the patient (where available if race is unavailable), and age. Statistical weights are calculated and used to adjust the EHR data non-probability sample to the population of interest (see A.5 Statistical Weights). Weights are derived from individual-level demographic and social determinant of health (SDOH) data available in the EHR, as well as population-level SDOH proxies derived from the ACS data. Calculated prevalence is included as crude and weighted results.

For records lacking race information, automated race imputation is employed in CODI-HPQ data inputs and linked population data (pre-processing). Within the final program to calculate prevalence, the user specifies whether householders with imputed race should be included in the results. Records with a race value are included in the prevalence independent of whether imputed race is assigned as "yes" or "no."

A.5 Statistical Weights

CODI and National AEMR data are derived from EHRs. Applying statistical weights is often used to reduce potential biases introduced by the EHR data sampling methodology. Ratio adjustments are applied to all sampled households. Ratio adjustment is a statistical weighting technique aimed to improve the accuracy of survey results by both reducing bias and increasing precision.[16] One way to accomplish this goal is known as iterative proportional fitting or raking. Raking adjusts the data so that groups that are underrepresented in the sample can be accurately represented in the final data set. Raking accurately matches sample distributions to known demographic characteristics of populations. The use of raking reduces nonresponse bias and has been shown to reduce error within sample results.

Implementing raking programs require the specification of appropriate weighting classes or cells. Data used to form classes for adjustments must be available for both sample and the population. CODI-HPQ raking includes social determinant of health categories – age of children, number of adults in the household, race of householder, and education categories in the surrounding area (based on percentage of adults in the community with a bachelor's degree or higher). Once formed, the weighting classes are assessed, and cells with small sample counts are aggregated with their nearest neighbor to reduce prevalence variability. The collapsing follows these guide points:

Age of children = do not aggregate, instead exclude small cell categories from prevalence results

Race = do not aggregate, instead exclude small cell categories from prevalence results

Education = community with a similar education category

Number of adults = do not aggregate, instead exclude small cell category from prevalence results (one adult, or two or more adults)

Raking is completed by adjusting for one demographic variable (or dimension) at a time. For example, when weighting by age of children and race, weights would first be adjusted for age of children, then those results would be adjusted by racial groups. The calculations continue in an iterative process until all group proportions in the sample approach those of the population, or after a set number of iterations. Once raked, weight trimming is used to reduce errors in the outcome caused by unusually high or low weights in some categories.

The fundamental objective of CODI-HPQ is to generate statistics that reduce bias and are sufficiently precise to satisfy the goals of the expected analyses of the data. In general, the goal is to keep the mean squared error (MSE) of the primary statistics of interest as low as possible. The MSE of a survey result is:

$$\text{MSE} = \text{Variance} + (\text{Bias})^2$$

The purpose of weighting adjustments is to reduce bias. Thus, the application of weighting adjustments usually results in lower bias in the associated survey statistics, but at the same time

[16] Little, 1993.

adjustments may result in some increases in variances of the survey results when compared with crude variances.

The increases in variance result from the added variability in the sampling weights due to the adjustments. Thus, the user who uses the weights should review the variability in the sampling weights caused by these adjustments. A trade-off is made between variance and bias to keep the MSE as low as possible. There is no exact rule for this trade-off because the amount of bias is unknown.

The five-year estimates of ACS do not include households with an age group of 0 to 1 years. Thus, CODI-HPQ overestimate households with children age 6 years or younger.

ACS race is categorized to match the EHR data file and grouped as White, African American, Asian (including Native Hawaiian and other Pacific Islanders), and other (including American Indian and Alaskan Native, some other race, two or more races).

ACS educational attainment (bachelor's degree or more) is linked by geography (state and GEO3) based on the patient's residential address. Once linked, education is calculated as the percent of the population aged 25 to 64 who have earned a bachelor's degree or more within the adult's geography. Educational attainment is then dichotomized based on the value: 20% of the population with a bachelor's degree or more. Approximately 52% of counties in the U.S. fall above 20%, and 48% fall below.

A.6 Prevalence Calculations

Crude prevalence is calculated as the count of the sampled households within each BMI category.

To calculate the weighted prevalence of the population the sum of statistical weights (households) within each BMI category is divided by the sum of statistical weights within the EHR. To control extreme weights which may increase the variance, extreme weights are trimmed. To calculate the variance of BMI, a Taylor-series approximation is used.[17]

Users are provided crude (unweighted) population, prevalence, and standard error, weighted population, prevalence, and standard error.

A.7 Standard Error

The precision of a sample can be measured using a variety of calculations, including the standard error, confidence interval, and the margin of error. The standard error is the most commonly used measure of the precision of a value and provides a gauge of how close a value is likely to be to the true population value in the absence of any bias. See Appendix A.11 Variance for more information.

[17] Wolter, 2007.

A.8 Suppression Criteria

Prevalence may be suppressed. CODI-HPQ data suppression is adapted from the NCHS data presentation standards for reporting proportions in NCHS reports and data products,[18] developed by the Data Suppression Workgroup at NCHS.

The multistep NCHS Data Presentation Standards for Proportions are based on a minimum denominator sample size of households and on the absolute and relative widths of a confidence interval calculated using the Clopper-Pearson method. The National Center for Health Statistics (NCHS) Data Presentation Standards for Proportions are applied to all CODI-HPQ results. The Presentation Standards also provide guidance for identifying results for statistical review, CODI-HPQ do not identify records for statistical review and leaves this step for the user. The data presentation standards are described in Table 11 and **Error! Reference source not found.**.

If one or more rows are suppressed, the user may select to increase their research criteria by including additional years of data, increasing the geography, or including more age or race categories. The suppression thresholds may also be altered by the user in the Quickstart program.

Table 11. NCHS Data Presentation Standards for Proportions

Statistic	Standard
Sample size	Proportions should be based on a minimum denominator sample size and effective denominator sample size (when applicable) of 30 households. Results with either a denominator sample size or an effective denominator sample size (when applicable) less than 30 should be suppressed. If the number of encounters is 0 (or its complement[19]), then the denominator sample size should be used to obtain confidence intervals. If all other criteria are met for presentation, a result based on 0 encounters (or its complement) should be flagged for statistical review by the clearance official. The review could result in either the presentation or the suppression of the proportion.
Confidence interval	If the sample size criterion is met, calculate a 95% two-sided confidence interval using the Clopper-Pearson method, or the Korn-Graubard method for complex surveys, and obtain its width.
Small absolute confidence interval width	If the absolute confidence interval width is greater than 0.00 and less than or equal to 0.05, then the proportion can be presented if the number of encounters is greater than 0 and the degrees of freedom criterion (below) is met. If the number of encounters is 0 (or its complement) or the degrees of freedom criterion is not met, then the result should be flagged for statistical review by the clearance official. The review could result in either the presentation or the suppression of the proportion.
Large absolute confidence interval width	If the absolute confidence interval width is greater than or equal to 0.30, then the proportion should be suppressed.
Relative confidence interval width	If the absolute confidence interval width is between 0.05 and 0.30 and the relative confidence interval width is more than 130%, then the proportion should be suppressed.

[18] Parker et al., 2017.

[19] The complement of a proportion p is (1 – p). The complement of the number of encounters in the numerator for p is the number of encounters in the numerator for (1 – p).

Statistic	Standard
Relative confidence interval width	If the absolute confidence interval width is between 0.05 and 0.30 and the relative confidence interval width is less than or equal to 130%, then the proportion can be presented if the degrees of freedom criterion below is met. If the degrees of freedom criterion is not met, then the result should be flagged for statistical review by the clearance official. The review could result in either the presentation or the suppression of the proportion.
Degrees of freedom	When applicable for complex surveys, if the sample size (households) and confidence interval criteria are met for presentation and the degrees of freedom are fewer than 8, then the proportion should be flagged for statistical review. This review may result in either the presentation or the suppression of the proportion.
Complementary proportions	If all criteria are met for presenting the proportion but not for its complement, then the proportion should be shown. A footnote indicating that the complement of the proportion may be unreliable should be provided by the end user and is not provided by CODI-HPQ.

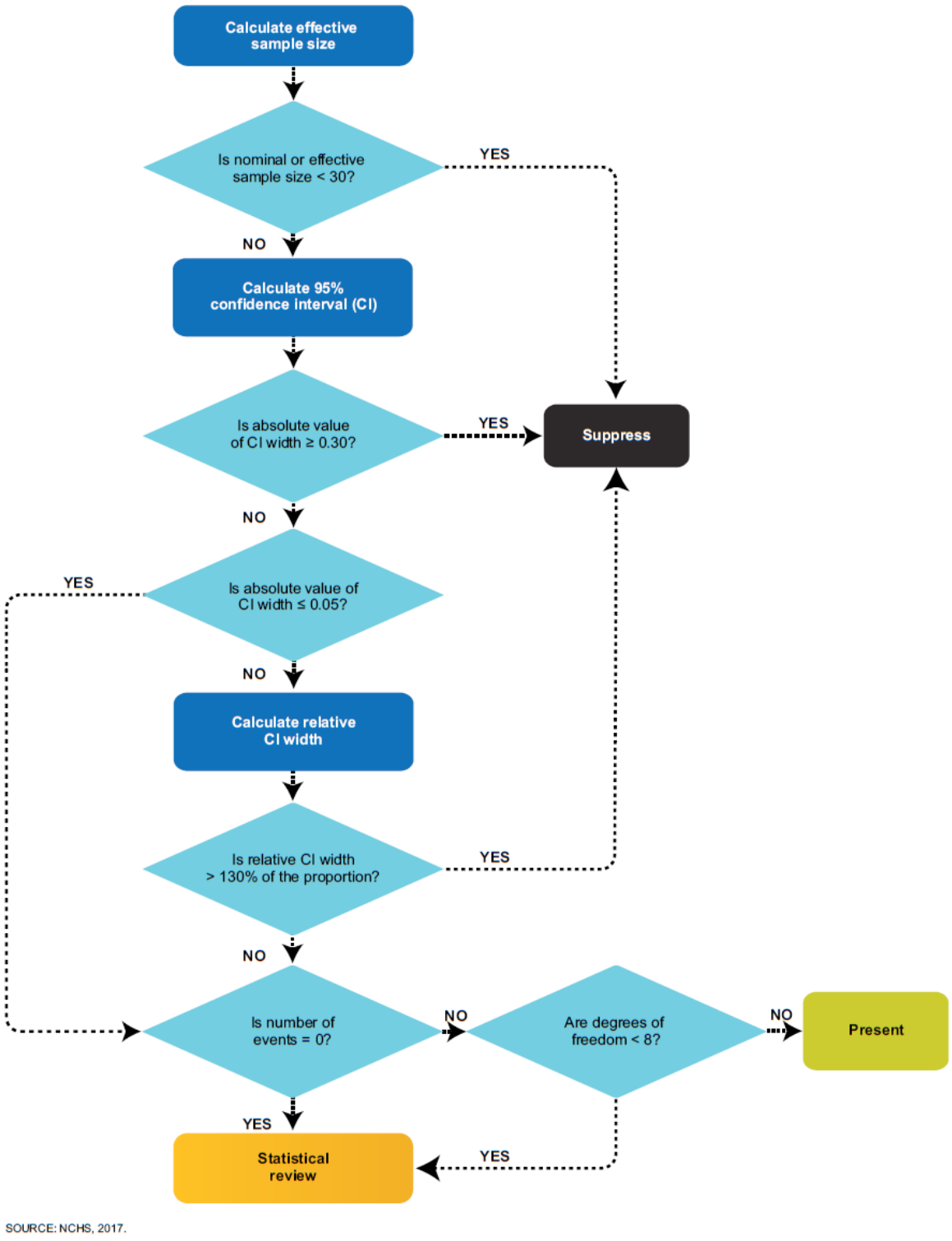

Figure 4. NCHS Suppression Standards

A.9 Variance

BMI prevalence is derived using the household sample weights and data on BMI categories. BMI prevalence is a ratio, and the ratio estimator, $\hat{\theta}$, corresponds to a population parameter, θ, such as the true but unknown BMI prevalence. To define the population parameter, let:

N_h = the number of households in stratum h ($h = 1, \ldots, L$), where stratum refers to state-GEO3

Y_{hi} = the value of Y for households i of stratum h (often the possible values of Y are 0 and 1, as when Y indicates whether a household has a specified BMI value)

d_{hi} = 0 or 1, indicating whether household i of stratum h belongs to a particular domain (such as a specified race)

$$Y_{dh} = \sum_{i=1}^{N_h} d_{hi} Y_{hi}$$

$$T_{dh} = \sum_{i=1}^{N_h} d_{hi}$$

Then, adding the subscript d to indicate the role of the domain, the ratio is the parameter of interest.

$$\theta_d = \frac{\sum_{h=1}^{L} Y_{dh}}{\sum_{h=1}^{L} T_{dh}}$$

In the sample, let:

n_h = the number of sample households in stratum h

W_{hi} = the sampling weight for households i in stratum h

Y'_{hi} = the value of Y for household i in stratum h

d'_{hi} = the value of the domain indicator for household i in stratum h

$$\hat{Y}_{dh} = \sum_{i=1}^{n_h} d'_{hi} W_{hi} Y'_{hi}$$

$$\hat{T}_{dh} = \sum_{i=1}^{n_h} d'_{hi} W_{hi}$$

The distinction between Y'_{hi} and Y_{hi} and between d'_{hi} and d_{hi} is merely that for Y'_{hi} and d'_{hi} the subscript i refers to sampled households within stratum h, whereas for Y_{hi} and d_{hi} they refer to households in the population in stratum h. Then, the ratio estimator for θ_d is:

$$\hat{\theta}_d = \frac{\sum_{h=1}^{L} \hat{Y}_{dh}}{\sum_{h=1}^{L} \hat{T}_{dh}}$$

To calculate the variance of $\widehat{\theta}_d$, a Taylor-series approximation is used.[20] Within stratum h, linearization yields the new variable.

$$Z_{hi} = \frac{d'_{hi} W_{hi} (Y'_{hi} - \hat{\theta}_d)}{\sum_{h=1}^{L} \hat{T}_{dh}}$$

Then, letting

$$\bar{Z}_h = \frac{\sum_{i=1}^{n_{th}} Z_{hi}}{n_h}$$

the Taylor-series approximation to the variance of $\hat{\theta}_d$ is

$$v(\hat{\theta}_d) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (Z_{hi} - \bar{Z}_h)^2$$

A.10 BMI Categories

Prevalence is calculated from a patient's BMI category. EHR data included for analysis should have at most one BMI category assigned to each patient within a calendar year. BMI is a person's weight in kilograms divided by the square of height in meters. Based on the 2000 CDC Growth Chart, the adult BMI categories for prevalence are as follows:

- **Does not have obesity:** BMI less than 30 kg/m^2
- **Has obesity**: BMI greater than or equal to 30 kg/m^2

The SAS program categorizes records into the above categories based on the following input values:

Underweight: BMI less than 18.5 kg/m^2

Healthy Weight: BMI greater than or equal to 18.5 and less than 25 kg/m^2

Overweight: BMI greater than or equal to 25 and less than 30 kg/m^2

Obesity: BMI greater than or equal to 30 kg/m^2

Obesity Class 1: BMI greater than or equal to 30 and less than 35 kg/m^2

Obesity Class 2: BMI greater than or equal to 35 and less than 40 kg/m^2

Obesity Class 3: BMI greater than or equal to 40 kg/m^2

For more information, visit:

https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html.

Youth and teen BMI categories are based on a patient's age-sex BMI percentile value. Based on the 2000 CDC Growth Chart, the percentile for prevalence are as follows:

1. **Does not have obesity**: BMI less than 95th percentile
2. **Has obesity**: BMI greater than or equal to 95th percentile

[20] Wolter, 2007.

The SAS program categorizes records into the above categories based on the following input values:

Underweight: BMI less than 5th percentile
Healthy Weight: BMI 5th percentile to less than the 85th percentile
Overweight: BMI 85th to less than the 95th percentile
Obesity[21]: BMI 95th percentile to less than 120 percent of the BMI value for the 95th percentile

a. **Severe Obesity**: 120 percent or greater of the BMI value for the 95th percentile

For more information, visit:
https://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html.

A.11 Limitations

CODI-HPQ users should consider the following limitations related to the program development, the data inputs required, and the results:

- Representativeness of CODI-HPQ results – CODI-HPQ results may differ from those based on a probability-based survey that could be more representative of the general population.
- Inclusion in EHRs – EHR data represent the care-seeking population for all medical providers included within a sample.
- Inclusion of household – CODI-HPQ require at a minimum an adult age 20 to 64 and a youth or teen age 2 to 17. It is assumed that all persons living within the same household are included in the EHR data.
- Linkage of persons by household identifier – CODI-HPQ assume all persons with the same household identifier lived together in the same household during the medical encounter year.
- Record linkage strategies include false links and missed matches.
    - It is recommended that the user become familiar with any record linkage strategy and its limitations.
    - If the linkage errors are not properly taken into account, biased estimates and mis-relationships between variables recorded in different sources (i.e. household linkage, person 1 in source A and person 2 in source B) may result (Di Consiglio and Tuoto, 2018).
    - If the user has information about how linkage error affects the distribution of household obesity, consider using techniques for quantitative bias analysis, to adjust for these errors (Lash, 2011, Schneeweiss, 2006).

[21] Note: prevalence of obesity will include two categories: those that are category 4 and 4a.

- Random missingness of plausible height or weight - CODI-HPQ patient inclusion requires a plausible height and weight value. It is assumed that if patients are missing height and weight from EHR data, it is missing at random.
- Random missingness of demographic and geographic characteristics- CODI-HPQ patient inclusion requires a valid and known age, sex, and geographic location to be reported. The race of each patient is also needed, although the program imputes race for patients missing race. It is assumed that if patients are missing age, sex, and/or geographic location from EHR data, it is missing at random.
- Race imputation - Race imputation assigns one value of race per householder. Multiple-imputation of race is not employed in CODI-HPQ to allow for a) analysis of large EHR files without the need for increasing the length of the original file and b) ease in counting number of households in the crude results. Variance for those with imputed race is likely smaller than those with known race. Also, race imputation does not analyze a patient's first and last name. Other EHR race imputation methodologies have utilized the patient's first and last name with positive results.[22]
- Measurement error - Height and weight measurement protocols may differ between medical providers, even with clear protocols aimed to increase consistency between medical professionals,[23] leading to potential measurement error. Additionally, height and weight values in EHR data are subject to data entry errors or software glitches. All CODI-HPQ EHR data were cleaned using growthcleanr. Growthcleanr scans all available height and weight values and flags values that are implausible; however, users must decide to exclude the implausible values, recognizing that biologically acceptable values may still have errors. See Methods for more information about growthcleanr.
- Small sample sizes - A small number of households could result in unstable results and reflect poor EHR coverage, a small underlying population, and/or a rare encounter. CODI-HPQ suppress results based on published small sample guidelines using the National Center for Health Statistics Data Presentation Standards for Proportions[24].

[22] Fiscella & Fremont, 2006.
[23] Best & Shepherd, 2020.
[24] Parker JD, Talih M, Malec DJ, et al, 2017.

Appendix B ACS File Layouts

B.1 ACS Input File Layout

The following variables are included in the County file. ACS data is imported in the CODI-HPQ and require a csv file with the following variable names, possible variable values, and in the order listed below.

Table 12. ACS Input File Layout, CSV File

Variable Name	Label	Description	Format	Example
State_Code	FIPS State Code	2-digit State Code	Numeric	8
County_Code	FIPS County Code	3-digit County FIPS Code	Numeric	59
female_rel_child_6_17	Female householder, no spouse present: with related children 6 to 17 years only	Count of female householders, no spouse present: with related children 6 to 17 years only	Numeric	7816
female_rel_child_l18	Female householder, no spouse present: with related children under 18 years	Count of female householders, no spouse present: with related children under 18 years	Numeric	11389
female_rel_child_l6	Female householder, no spouse present: with related children under 6 years only	Count of female householders, no spouse present: with related children under 6 years only	Numeric	2031
hh_fam	Family households	Count of family households	Numeric	149418
hh_fam_asian	Family households: Asian	Count of family households with a householder who is Asian alone	Numeric	3813
hh_fam_black	Family households: Black	Count of family households with a householder who is Black or African American alone	Numeric	1065
hh_fam_ppl_l18	Family households: with one or more people under 18 years	Count of family households with one or more people under 18 years	Numeric	63619
hh_fam_white	Family households: White	Count of family households with a householder who is White alone	Numeric	139207

Variable Name	Label	Description	Format	Example
hh_income_25_44	Householder 25 to 44 years	Count of households with a householder between the age of 25 to 44 years	Numeric	75954
hh_income_45_64	Householder 45 to 64 years	Count of households with a householder between the age of 45 to 64 years	Numeric	91604
hh_income_65pl	Householder 65 years and over	Count of households with with a 65 years of age or over householder	Numeric	57727
hh_income_l25	Householder under 25 years	Count of households with a householder under 25 years of age	Numeric	6999
hh_owner_bachelors_plus	Householder's education attaintment, owner-occupied: Bachelor's degree or higher	Count of owner-occupied households with a Bachelor's degree or higher householder's education attainment.	Numeric	89248
hh_renter_bachelors_plus	Householder's education attaintment, renter-occupied: Bachelor's degree or higher	Count of renter-occupied households with a Bachelor's degree or higher householder's education attainment.	Numeric	22260
hh_tenure_educ_total	Tenure by educational attainment of householder	Total count of households	Numeric	232284
male_rel_child_6_17	Male householder, no spouse present: with related children 6 to 17 years only	Count of male householders, no spouse present: with related children 6 to 17 years only	Numeric	4146
male_rel_child_l18	Male householder, no spouse present: with related children under 18 years	Count of male householders, no spouse present: with related children under 18 years	Numeric	5702
male_rel_child_l6	Male householder, no spouse present: with related children under 6 years only	Count of male householders, no spouse present: with related children under 6 years only	Numeric	1037
married_rel_child_6_17	Married-couple family household: with related children 6 to 17 years only	Count of married-couple family households with related children 6 to 17 years only	Numeric	25656

Variable Name	Label	Description	Format	Example
married_rel_child_l18	Married-couple family household: with related children under 18 years	Count of married-couple family households with related children under 18 years	Numeric	46232
married_rel_child_l6	Married-couple family household: with related children under 6 years only	Count of married-couple family households with related children under 6 years only	Numeric	11951
LAT_AIAN	American Indian and Alaska Native alone, Hispanic or Latino	Total population of American Indian and Alaska Native alone with Hispanic or Latino origin	Numeric	1447
LAT_ASIAN	Asian alone, Hispanic or Latino	Total population of Asian alone with Hispanic or Latino origin	Numeric	261
LAT_BLACK	Black or African American alone, Hispanic or Latino	Total population of Black or African American alone with Hispanic or Latino origin	Numeric	707
LAT_GE2R	Two or more races, Hispanic or Latino	Total population of people with two or more races with Hispanic or Latino origin	Numeric	4589
LAT_NHPI	Native Hawaiian and Other Pacific Islander alone, Hispanic or Latino	Total population of Native Hawaiian and Other Pacific Islander alone with Hispanic or Latino origin	Numeric	24
LAT_OTHER	Some other race alone, Hispanic or Latino	Total population of some other race alone with Hispanic or Latino origin	Numeric	9728
LAT_WHITE	White alone, Hispanic or Latino	Total population of White alone with Hispanic or Latino origin	Numeric	71574
NON_LATX_AIAN	American Indian and Alaska Native alone, not Hispanic or Latino	Total population American Indian and Alaska Native alone with no Hispanic or Latino origin	Numeric	2527
NON_LATX_ASIAN	Asian alone, not Hispanic or Latino	Total population of Asian alone with no Hispanic or Latino origin	Numeric	15995

Variable Name	Label	Description	Format	Example
NON_LATX_BLACK	Black or African American alone, not Hispanic or Latino	Total population of Black or African American alone with no Hispanic or Latino origin	Numeric	6132
NON_LATX_GE2R	Two or more races, not Hispanic or Latino	Total population of people with two or more races with no Hispanic or Latino origin	Numeric	11547
NON_LATX_NHPI	Native Hawaiian and Other Pacific Islander alone, not Hispanic or Latino	Total population of Native Hawaiian and Other Pacific Islander alone with no Hispanic or Latino origin	Numeric	297
NON_LATX_OTHER	Some other race alone, not Hispanic or Latino	Total population of some other race alone with no Hispanic or Latino origin	Numeric	925
NON_LATX_WHITE	White alone, not Hispanic or Latino	Total population of White alone with no Hispanic or Latino origin	Numeric	449045
TOTAL_LAT	Hispanic or Latino	Total population of people with Hispanic or Latino origin	Numeric	88330
TOTAL_NON_LATX	Not Hispanic or Latino	Total population of people with no Hispanic or Latino origin	Numeric	486468

B.2 ACS for Use with GEO3 Data

Table 13. ACS Pre-Processing Results File Layout – GEO3

Variable Name	Label	Description	Format	Example
State_Code	FIPS State Code	2-digit State Code	Numeric	8
County_Code	FIPS County Code	3-digit County FIPS Code	Numeric	59
female_rel_child_6_17	Female householder, no spouse present: with related children 6 to 17 years only	Count of female householders, no spouse present: with related children 6 to 17 years only	Numeric	7816
female_rel_child_l18	Female householder, no spouse present: with related children under 18 years	Count of female householders, no spouse present: with related children under 18 years	Numeric	11389

Variable Name	Label	Description	Format	Example
female_rel_child_l6	Female householder, no spouse present: with related children under 6 years only	Count of female householders, no spouse present: with related children under 6 years only	Numeric	2031
hh_fam	Family households	Count of family households	Numeric	149418
hh_fam_asian	Family households: Asian	Count of family households with a householder who is Asian alone	Numeric	3813
hh_fam_black	Family households: Black	Count of family households with a householder who is Black or African American alone	Numeric	1065
hh_fam_ppl_l18	Family households: with one or more people under 18 years	Count of family households with one or more people under 18 years	Numeric	63619
hh_fam_white	Family households: White	Count of family households with a householder who is White alone	Numeric	139207
hh_income_25_44	Householder 25 to 44 years	Count of households with a householder between the age of 25 to 44 years	Numeric	75954
hh_income_45_64	Householder 45 to 64 years	Count of households with a householder between the age of 45 to 64 years	Numeric	91604
hh_income_65pl	Householder 65 years and over	Count of households with with a 65 years of age or over householder	Numeric	57727
hh_income_l25	Householder under 25 years	Count of households with a householder under 25 years of age	Numeric	6999
hh_owner_bachelors_plus	Householder's education attaintment, owner-occupied: Bachelor's degree or higher	Count of owner-occupied households with a Bachelor's degree or higher householder's education attainment.	Numeric	89248

Variable Name	Label	Description	Format	Example
hh_renter_bachelors_plus	Householder's education attaintment, renter-occupied: Bachelor's degree or higher	Count of renter-occupied households with a Bachelor's degree or higher householder's education attainment.	Numeric	22260
hh_tenure_educ_total	Tenure by educational attainment of householder	Total count of households	Numeric	232284
male_rel_child_6_17	Male householder, no spouse present: with related children 6 to 17 years only	Count of male householders, no spouse present: with related children 6 to 17 years only	Numeric	4146
male_rel_child_l18	Male householder, no spouse present: with related children under 18 years	Count of male householders, no spouse present: with related children under 18 years	Numeric	5702
male_rel_child_l6	Male householder, no spouse present: with related children under 6 years only	Count of male householders, no spouse present: with related children under 6 years only	Numeric	1037
married_rel_child_6_17	Married-couple family household: with related children 6 to 17 years only	Count of married-couple family households with related children 6 to 17 years only	Numeric	25656
married_rel_child_l18	Married-couple family household: with related children under 18 years	Count of married-couple family households with related children under 18 years	Numeric	46232
married_rel_child_l6	Married-couple family household: with related children under 6 years only	Count of married-couple family households with related children under 6 years only	Numeric	11951
LAT_AIAN	American Indian and Alaska Native alone, Hispanic or Latino	Total population of American Indian and Alaska Native alone with Hispanic or Latino origin	Numeric	1447
LAT_ASIAN	Asian alone, Hispanic or Latino	Total population of Asian alone with Hispanic or Latino origin	Numeric	261

Variable Name	Label	Description	Format	Example
LAT_BLACK	Black or African American alone, Hispanic or Latino	Total population of Black or African American alone with Hispanic or Latino origin	Numeric	707
LAT_GE2R	Two or more races, Hispanic or Latino	Total population of people with two or more races with Hispanic or Latino origin	Numeric	4589
LAT_NHPI	Native Hawaiian and Other Pacific Islander alone, Hispanic or Latino	Total population of Native Hawaiian and Other Pacific Islander alone with Hispanic or Latino origin	Numeric	24
LAT_OTHER	Some other race alone, Hispanic or Latino	Total population of some other race alone with Hispanic or Latino origin	Numeric	9728
LAT_WHITE	White alone, Hispanic or Latino	Total population of White alone with Hispanic or Latino origin	Numeric	71574
NON_LATX_AIAN	American Indian and Alaska Native alone, not Hispanic or Latino	Total population American Indian and Alaska Native alone with no Hispanic or Latino origin	Numeric	2527
NON_LATX_ASIAN	Asian alone, not Hispanic or Latino	Total population of Asian alone with no Hispanic or Latino origin	Numeric	15995
NON_LATX_BLACK	Black or African American alone, not Hispanic or Latino	Total population of Black or African American alone with no Hispanic or Latino origin	Numeric	6132
NON_LATX_GE2R	Two or more races, not Hispanic or Latino	Total population of people with two or more races with no Hispanic or Latino origin	Numeric	11547
NON_LATX_NHPI	Native Hawaiian and Other Pacific Islander alone, not Hispanic or Latino	Total population of Native Hawaiian and Other Pacific Islander alone with no Hispanic or Latino origin	Numeric	297

Variable Name	Label	Description	Format	Example
NON_LATX_OTHER	Some other race alone, not Hispanic or Latino	Total population of some other race alone with no Hispanic or Latino origin	Numeric	925
NON_LATX_WHITE	White alone, not Hispanic or Latino	Total population of White alone with no Hispanic or Latino origin	Numeric	449045
TOTAL_LAT	Hispanic or Latino	Total population of people with Hispanic or Latino origin	Numeric	88330
TOTAL_NON_LATX	Not Hispanic or Latino	Total population of people with no Hispanic or Latino origin	Numeric	486468
GEOGRAPHY	State and County FIPS code	5-digit State and county code combination	Character	08059
P_BA	Percent of householders with a BA degree or higher	Percentage of householders in the county with a Bachelor's degree or higher	Numeric	0.48
BA_G20	Education level indicator	Indicator if more than 20% of households in the county has a Bachelor's degree or higher	Numeric	1

Appendix C EHR File Layouts

C.1 EHR Input File Layout

C.1.1 EHR GEO3 Data

EHR data are imported in the CODI-HPQ and require a csv file with the following variable names, possible variable values, and in the order listed below.

Table 14. EHR Input File Layout for GEO3-Level Programs, CSV File[25]

[25] One record per patient per year, thus a patient may be included multiple times in the EHR.

Variable Name	Description	Format	Valid values	Example
SUBJID or PATID	**Patient Identifier**	**Character**	**Character value of maximum length 50.**	**S123456789**
HOUSEHOLD_ID	**Household Identifier**	**Character**	**Character value of maximum length 50.**	**H22182123412**
SEX_NUM	**Sex of patient where 0 is male, 1 is female**	**Number**	**0 1**	**0**
AGEYEARS	**Age of patient in years at the time of the medical encounter**	**Number**	**Count of years as whole numbers (NOTE that this may be approximate value due to birth date approximation)**	**11**
RACE_ETH	**Patient's race if known or ethnicity when race is not known**	**Character**	**"Black" "AFRICAN AMERICAN" "Asian" "White" "CAUCASIAN" "Hispanic" "HISPANIC" "Other" "OTHER" "Unknown" "UNKNOWN" ""**	**WHITE**
WEIGHT_CATEGORY	**Patient's BMI Percentile. See section A.2 and A.11 for more information.**	**Character**	**"Normal or Healthy Weight" "Obese" "Obesity" "Does Not Have Obesity" "Severe Obesity" "Obese Class 1" "Obese Class 2" "Obese Class 3" "Obesity Class 1" "Obesity Class 2" "Obesity Class 3" "Overweight" "Underweight"**	**Overweight**

Variable Name	Description	Format	Valid values	Example
YEAR	**Year of the medical encounter**	**Number**	**Yyyy**	**2018**
COUNTY_FIPS_CODE	**Patient's residential county code**	**Number**	**Any county FIPS numeric value (up to 3 digits)**	**59**
STATE_FIPS_CODE	**Patient's residential state code**	**Number**	**Any state FIPS code (up to two digits). See Appendix F for a list of possible values.**	**08**

C.2 EHR Results File Layout for GEO3

Table 15. GEO3 Results

Variable Name	Description	Format	Example
PATID	Patient Identifier	Character	S123456789
HOUSEHOLD_ID	Household Identifier	Character	H22182123412
AGEYEARS	Age of patient in years at the time of the medical encounter	Number	11
RACE_ETH	Patient's race if known or ethnicity when race is not known	Character	White
WEIGHT_CATEGORY	Patient's BMI Percentile. See section A.2 and A.11 for more information.	Character	Obese
YEAR	Year of the medical encounter	Number	2018
COUNTY_FIPS_CODE	Patient's residential county code	Number	5
STATE_FIPS_CODE	Patient's residential state code	Number	8
GEOGRAPHY	Patient's residential state-county FIPS code (5-digits)	Character	08005
SEX	Sex of patient (Male or Female)	Character	Female
ADULTS	Weight category based on the weight of the adults in the household	Character	One or more adults with obesity
YOUTH_AND_TEENS	Weight category based on the weight of the youth and teens in the household	Character	No youth or teens with obesity
RACE_IMPUTED	Race imputation indicator ("Y" for imputed, "N" for not imputed)	Character	Y
HOUSEHOLD_CHILDAGE_CAT	Age category of youth and teens in the household	Character	6 to 17 years only
BA_G20	Flag indicating if the county where the patient resides has 20% or more of its householders have a Bachelor's degree or higher.	Numeric	1
NB_ADULTS	Number of adults in the household	Numeric	2
IMPUTE_RACE	Imputed race value	Character	White

Appendix D CODI-HPQ-GEO3 Example SAS Programs

D.1 Data Inputs and Link Population Data (Pre-Processing) Quickstart with GEO3 Data

Appendix D.1 includes a program to generate a pre-processed file using the Quickstart pre-processing program. This example uses COUNTY data. Text highlighted in yellow has been reviewed and approved or reviewed and edited from its original values.

The program uses the data inputs: ACS_COUNTY and EHR_Synthetic_Household. The file processes EHR data and creates SAS files named CODI_HPQ_Example_PRE stored in pre-processing output folder C:\Example\CODI_HPQ_1130 \2_Output\CODI_HPQ_PRE. The SAS log is stored in C:\Example\CODI_HPQ_1130 \2_Output\SAS LOG\ HPQ_LOG_Pre_processing_<plus date and time information>.log.

```
/*****************************************************************************************************************************/
/************************************** -- USER INPUT FOLDER -- ***************************************************************/
/****************** -- PLEASE UPDATE THE BLACK TEXT AFTER THE EQUAL SIGN (ACCEPTED VALUES LISTED IN SAS NOTES) -- ******************/

/*SECTION 1: Folder and file names                                                    ***/
/***/ %LET ROOT_HPQ    = C:\Example\CODI_HPQ_1130;   /*@Note: base directory (ACCEPTABLE VALUES: computer directory name) ***/
/***/ %LET PRE_HDEST   = CODI_HPQ_PRE;                      /*@Note: Suffix name of pre-processing output folder (ACCEPTABLE VALUES:
folder name (no punctuations)) ***/
/***/ %LET EHR_H_PRE_OUT= CODI_HPQ_Preprocessed_Filename ;          /*@Note: Suffix name of pre-processing output file(ACCEPTABLE
VALUES: file name (no punctuations)) ***/
                                                                          ***/

/***/ %LET EHR_FILENAME = EHR_Synthetic_Household; /*@Note: EHR file name (ACCEPTABLE VALUES: file name, do not include ".csv") ***/
/***/ %LET ACS_FILENAME = ACS_COUNTY;  *@Note: ACS person-level file name (ACCEPTABLE VALUES: file name, do not include ".csv") ***/

/***/ %LET LOG_NAME_PRE       = HPQ_LOG_Pre_processing;     /*@Note: SAS log file name prefix ACCEPTABLE VALUES: SAS file name (no
punctuation)

/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
/***Note: ROOT_PRE directory includes subfolders:
                                                          "..\0_Raw_Data" sss
                                                          "..\1_SAS_Programs"
                                                          "..\02_Output" and
                                                          "..\02_Output\SAS LOG"***/
/***NOTE: SAS programs must be stored in the PROGS_PRE directory including:
                                                          Module1-Pre_Processing_CODI_HPQ.sas***/
/*****************************************************************************************************************************/
/***STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP*********/
/*** DO NOT EDIT BEYOND THIS POINT    DO NOT EDIT BEYOND THIS POINT        DO NOT EDIT BEYOND THIS POINT    *********/
/*****************************************************************************************************************************/
/*****************************************************************************************************************************/
/*****************************************************************************************************************************/
/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
```

D.2 Generate Results Example with GEO3 Data

Appendix D2. includes a program excerpt to generate prevalence results using the Quickstart program and needed data inputs. This example uses COUNTY data. Text highlighted in yellow has been reviewed and approved or reviewed and edited from its original values.

The file processes EHR data for a subpopulation and a given analysis and creates a csv file with output named "HPQ_EXAMPLE_OUTPUT_WIMPUTE.csv" stored in the folder C:\Example\CODI_HPQ_1130 \2_Output. The SAS log is stored in C:\Example\CODI_HPQ_1130\2_Output\SAS LOGS\HPQ_EXAMPLE_LOG_<plus date and time information>.log.

Subpopulation: EHR records from 2019 including households with children 6 years of age or younger (but not households with children who are both under 6 as well as 7 to 17 years of age) who are either white or Asian, living in Jefferson County (059) Colorado (FIPS code = 08) or Baca (005) Colorado (FIPS code = 08) see: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697 to determine the correct value (08059 is combined state and County).

Methods: Include imputed race information.

```
/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
/*************************************************************************************************************************************************/
/************************************************ -- USER SELECTION CRITERA SECTIONS 1 through 5 -- ************************************************/
/****************** -- PLEASE UPDATE THE BLACK TEXT AFTER THE EQUAL SIGN (ACCEPTED VALUES LISTED IN SAS NOTE) -- ******************/
/*************************************************************************************************************************************************/

/*SECTION 1: Folder and file names                                                    ***/
/***/ %LET ROOT_HPQ           = C:\Example\CODI_HPQ_1130;   /*@Note: base directory, same as in pre-processing SAS programs
(ACCEPTABLE VALUES: computer directory name) ***/
/***/ %LET PRE_HDEST   = CODI_HPQ_PRE;                     /*@Note: Suffix name of pre-processing output folder, same as in pre-
processing SAS programs (ACCEPTABLE VALUES: folder name (no punctuations)) ***/
/***/ %LET EHR_H_PRE_OUT= CODI_HPQ_Preprocessed_Filename ;          /*@Note: Suffix name of pre-processing output file, same as in pre-
processing SAS programs (ACCEPTABLE VALUES: file name (no punctuations)) ***/
/***/ %LET LOG_NAME    = HPQ_EXAMPLE_LOG;                   /*@Note: Name for SAS log storage location            ***/
/***/ %LET FileOUT_Name = HPQ_EXAMPLE_OUTPUT_WIMPUTE;       /*@Note: Output file name                                        ***/

/*SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE OR STATE/COUNTY CODE           ***/
/***/ %LET ALL_H_STATES        = N;     /*@Note: EHRs file includes all of the US?  (ACCEPTED VALUES: Y/N)                ***/
/***/ %LET H_YEAR       = 2019;  /*@Note: year of analysis                                                ***/
/***/ %LET ALL_H_AGES   = N;     /*@Note: Include all youth age ranges? (ACCEPTED VALUES: Y/N)                      ***/
/***/ %LET ALL_RACES    = N;     /*@Note: Include all householder race categories? (ACCEPTED VALUES: Y/N)  ***/

/*SECTION 3: Only complete section 3 for any "N" values listed in section 2                  ***/
/*IF ALL_H_STATES= N THEN SELECT STATE CODES OR STATE AND COUNTY CODES BELOW: ***/
/***/ %LET GEO_H_GROUP  = STATE;             /*@Note: Level of geography (ACCEPTED VALUES: STATE or COUNTY)***/
/***/ %LET GEO_H_LIST  = %STR('08','37');     /*@Note: IF GEO_GROUP="STATE" then populate with State FIPS code(s), If
GEO_GROUP="COUNTY" then populate with FIPS State+FIPS County code(s) (ACCEPTED VALUES: 2-digit state FIPS for STATE or  5-digit state
FIPS+county FIPS for COUNTY (Must be surrounded by single quotation and comma delimited))***/
```

```
/*IF ALL_AGES = N THEN SELECT ONE OR MORE AGE CATEGORIES BELOW:     ***/
       /***/ %LET AGE_UNDER_6 = Y;     /*@Note: Children's Age Range: include households that have kids under the age of 6, but do not
have kids 6 to 17 years of age (ACCEPTED VALUES: Y/N)                                                    ***/
       /***/ %LET AGE_6_17 = N;        /*@Note: Children's Age Range: include households that have kids of age 6 or older, but do not
have kids under the age of 6 (ACCEPTED VALUES: Y/N) ***/
       /***/ %LET AGE_BOTH = N;        /*@Note: Children's Age Range: include households that have 1) kids of age 6 or older, and also
have 2) kids under the age of 6 (ACCEPTED VALUES: Y/N) ***/

/*IF ALL_RACES = N THEN SELECT ONE OR MORE RACE BELOW                          ***/
       /***/ %LET RACE_WHITE  = Y;     /*@Note: White (ACCEPTED VALUES: Y/N)                                        ***/
       /***/ %LET RACE_BLACK  = N;     /*@Note: Black/African American ACCEPTED VALUES: Y/N)          ***/
       /***/ %LET RACE_ASIAN  = Y;     /*@Note: Asian                 (ACCEPTED VALUES: Y/N)                  ***/
       /***/ %LET RACE_OTHER  = N;     /*@Note: Other                 (ACCEPTED VALUES: Y/N)                                  ***/


/*SECTION 4: Methodological option selections
                                                        ***/
       /***/ %LET IMP_RACES = Y;       /*@Note: Include imputed householder race values? (ACCEPTED VALUES: Y/N)      ***/
/*************************************************************************************************************************************/
/***Note: Root directory includes subfolders: "..\0_Raw_Data"
                                             "..\1_SAS_Programs"
                                             "..\2_Output" and
                                             "..\2_Output\SAS LOGS" ***/
/***NOTE: SAS programs must be stored in the PROGS directory including: Macro1-CODI_HPQ_GEO3.sas,
                                                                       Macro2-CODI_HPQ_GEO3.sas,
                                                                       Module1-CODI_HPQ_GEO3.sas,
                                                                       Module2-CODI_HPQ_GEO3.sas,
                                                                       Module3-CODI_HPQ_GEO3.sas,***/
/*NOTE: query output is stored as a csv file in "..\2_Output" named after a time/date stamp and CODI_HPrevalence_Query_Report     ***/
/*************************************************************************************************************************************/
/*************************************************************************************************************************************/
/***STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP STOP */
/*** DO NOT EDIT BEYOND THIS POINT    DO NOT EDIT BEYOND THIS POINT DO NOT EDIT BEYOND THIS POINT DO NOT EDIT BEYOND THIS POINT    */
/*************************************************************************************************************************************/
/*************************************************************************************************************************************/
/*************************************************************************************************************************************/
/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
```

Appendix E CODI-HPQ Results

E.1 Example BMI Prevalence

Once complete, CODI-HPQ generate prevalence results as an Excel file. Table 16 provides an overview of the variables included and example results based on synthetic data.

Table 16. CODI-HPQ Results Data Dictionary

Column	Description
Order	Row order
Youth and Teens Weight Category	A categorical value based on BMI percentile of youth and teen(s) in households.
Adults Weight Category	A categorical value based on BMI of adult(s) in households.
Sample	The observed (or unadjusted, or crude) count of households in the study population.
Population	The weighted (or adjusted) count of households.
Crude Prevalence	The observed (or unadjusted, or crude) household prevalence in the study population.
Crude Prevalence Standard Error	The observed (or unadjusted, or crude) household standard error in the study population.
Weighted Prevalence	Household prevalence based on weighted counts. A sample weight is assigned to each sampled household. It is a measure of the number of households in the population represented by that sample household. See implementation guide, Appendix A. Statistical Weights for more information.
Weighted Prevalence Standard Error	Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information.

Table 17. Results Example from Synthetic Data[26]

Order	Youth and Teens Weight Category	Adults Weight Category	Sample	Population	Crude Prevalence	Crude Prevalence Standard Error	Weighted Prevalence	Weighted Prevalence Standard Error
1	No youth or teens with obesity	No adults with obesity	356	166678	54.94	1.95	51.97	4.3
2	No youth or teens with obesity	One or more adults with obesity	212	101354	32.72	1.84	31.6	3.54
3	One or more youth or teens with obesity	No adults with obesity	36	8761	5.56	0.9	2.73	0.66
4	One or more youth or teens with obesity	One or more adults with obesity	44	43923	6.79	0.99	13.7	4.89
5	Query Version: CODI-HPQ GEO3 2021							
6	Race: (White, Black, Asian)							
7	Race Suppressed: (None)							
8	Imputed race: 32.05% of race values were imputed. Please be advised, prevalence estimates may incur additional bias with imputed race values. Extreme caution is recommended when the proportion of imputed race values exceeds 40%.							
9	Number of Adults in Household Suppressed: (None)							
10	Age of Children: (Under 6 years only, 6 to 17 years only)							
11	Age of Children Suppressed: Childage Suppressed: (None)							
12	Geography: (08) Colorado, (37) North Carolina							
13	Year: 2019							
14	Weighting cells were collapsed for: (Geography)							
15	Error codes: (None)							
16	Implementation Guide: See https://github.com/NORC-UChicago/CODI-PQ for more information and full details on calculations.							
17	Query Date: Monday, December 20, 2021, 3:20:40 PM							
18	Suggested Citation: Tanenbaum, E., Zalsha, S. (2021). Clinical and Community Data Initiative Household Prevalence Query (CODI-HPQ) SAS programs.							
19	The Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare federally funded research and development center, Health FFRDC. Retrieved from https://github.com/NORC-UChicago/CODI-PQ after December 31, 2021. The standard error calculations are documented in the Implementation Guide.							
20	Patients with either missing or invalid age, sex, height, weight, or geography are not included in results. The household estimates are based on American Community Survey Five-year Estimates.							
21	CODI-HPQ was developed between 2019 and 2021 and tested with EHR from 2015 through 2019. Please review the Implementation Guide in full to determine whether CODI-HPQ methodology is appropriate for your use case when used outside of these date ranges.							

[26] Note: borders and shading are for demonstration purposes only. CSV exports columns separated with a comma. The results can be imported into Excel.

Table 18. Example Results with Errors (Insufficient Sample Size), Error Messages Are Shown in Row Order 15

Order	Youth and Teens Weight Category	Adults Weight Category	Sample	Population	Crude Prevalence	Crude Prevalence Standard Error	Weighted Prevalence	Weighted Prevalence Standard Error
1	No youth or teens with obesity	No adults with obesity						
2	No youth or teens with obesity	One or more adults with obesity						
3	One or more youth or teens with obesity	No adults with obesity						
4	One or more youth or teens with obesity	One or more adults with obesity						
5	Query Version: CODI-HPQ GEO3 2021							
6	Race: (Asian, Other)							
7	Race Suppressed: (Asian, Other)							
8	Imputed race: (Error)% of race values were imputed. Please be advised, prevalence estimates may incur additional bias with imputed race values. Extreme caution is recommended when the proportion of imputed race values exceeds 40%.							
9	Number of Adults in Household Suppressed: (Error)							
10	Age of Children: (Under 6 years only)							
11	Age of Children Suppressed: (Error)							
12	Geography: (08059)							
13	Year: 2019							
14	Weighting cells were collapsed for: (Error)							
15	Error codes: Current selections return an insufficient number of patients and do not meet minimum threshold to estimate sample weights. Consider including additional demographic categories (e.g., races, age groups) or geographies.							
16	Implementation Guide: See https://github.com/NORC-UChicago/CODI-PQ for more information and full details on calculations.							
17	Query Date: Monday, December 20, 2021, 1:46:01 PM							
18	Suggested Citation: Tanenbaum, E., Zalsha, S. (2021). Clinical and Community Data Initiative Household Prevalence Query (CODI-HPQ) SAS programs.							
19	The Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare federally funded research and development center, Health FFRDC. Retrieved from https://github.com/NORC-UChicago/CODI-PQ after December 31, 2021. The standard error calculations are documented in the Implementation Guide.							
20	Patients with either missing or invalid age, sex, height, weight, or geography are not included in results. Household estimates are based on American Community Survey Five-year Estimates.							
21	CODI-HPQ was developed between 2019 and 2021 and tested with EHR from 2015 through 2019. Please review the Implementation Guide in full to determine whether CODI-HPQ methodology is appropriate for your use case when used outside of these date ranges.							

E.2 Possible Result Errors

There are several reasons that CODI-HPQ may not produce some or all results as described in the table that follows.

Table 19: CODI-HPQ Results Error Codes

Error Provided in Output/Results	Description
One or more demographic or geographic category has no groups selected. One or more groups must be selected in each category. Please ensure each demographic and geographic category has one or more groups selected (e.g. children's age, race).	One or more categories are not selected. For example, a minimum of one age group, and racial group must be selected (Y).
Year is out of scope or no year selected. CODI-HPQ were developed between 2019 and 2021, see Implementation Guide for more details.	Starting year cannot be before 2000, ending year cannot be after 2030. CODI-HPQ may be inappropriate to implement on medical encounters outside of 2015 through 2021. Please review the methodology in full to determine whether CODI-HPQ are appropriate for your needs.
Geographic level (GEO_H_GROUP) has been left blank or has been set to an unacceptable value. To remedy issue, please update the GEO_H_GROUP variable to either STATE or COUNTY.	CODI-HPQ may create estimates based on either a state identifier or a state and county identifier.
STATE and/or COUNTY is incorrectly specified. Review the lists and ensure each value is: Surrounded by quotations, Comma delimited, and/or The correct length (e.g., '08001', '08002', '08003, etc. for COUNTY and '08', '37', etc. for STATE).	Ensure the GEO_LIST is set to the correct format. 1. State is a FIPS number, not a state abbreviation, 2. All numbers must be in single quotes, 3. there is a space and a comma whenever selecting multiple locations, and 4. the text is within the function %STR( ); Examples: If GEO_GROUP = STATE; /***/ %LET GEO_LIST = %STR('08', '10'); If GEO_GROUP = COUNTY; /***/ %LET GEO_LIST = %STR('08001', '08002');
Current selections return an insufficient number of patients and do not meet minimum threshold to estimate sample weights. Consider including additional demographic categories (e.g., races, age groups) or geographies.	Select a larger sample.
Iterative proportional fitting weighting routine has failed to converge. Please revise selection criteria and rerun algorithm.	Weighting is not possible using iterative proportionate fitting under certain circumstances. For example, according to a SAS SUGI paper, (Izrael, 2004) "Oh and Scheuren (1978) note that the available convergence proofs make strong assumptions about the cell counts in the cross-classification of the raking variables – that no cells are empty or that some particular combination of nonempty cells is present. They recommend setting up the raking problem in a "sensible" manner to avoid: 1) imposing too many marginal constraints on the sample, 2) defining marginal categories that contain a small percentage of the sample, and 3) imposing contradictory constraints on the sample. … Convergence may be slow if 1) any categories contain fewer than 5% of the sample cases, 2) the size of the difference between each control total and the weighted sample margin prior to raking. If some differences are large, the number of iterations will typically be higher."

Error Provided in Output/Results	Description
A SAS error has occurred within the algorithm. Review the SAS log or contact a system administrator for further assistance.	SAS errors occur when syntax is not properly specified. Common reasons for SAS errors include missing semi-colons, single or double quotes, mismatched quotes, deleting the "/*" that is before a comment or "*/" after a comment, etc., etc. In addition to reviewing your SAS code and log, consider contacting SAS technical support, and/or make a new copy of the software from Github.

Additional messages may be displayed but are not indicative of an error. For example, the percentage of persons with imputed race that are included in the prevalence estimates.

Table 20: CODI-HPQ Results Error Codes

Comment	Description
RACE Imputed: (Error) of race values were imputed. Please be advised, prevalence may incur additional bias with imputed race values. Extreme caution is encouraged when the proportion of imputed race values exceeds 40%.	If the user allows records with imputed race to be included in the analysis, then the percentage of records (crude) with imputed race is reported in the results.
Weighting cells were consolidated for:	Statistical weighting is conducted by number of adults, age of children, race, and geography. If the sample size is insufficient in an age group or geography, weighting cells may be collapsed (combined). Race and number of adults do not allow for consolidation of weighting cells.
CODI-HPQ were developed between 2019 and 2021 and tested with EHR from 2015 through 2019. Please review the Implementation Guide in full to determine whether CODI-HPQ methodology is appropriate for your use case when used outside of these date ranges.	Users may choose to employ CODI-HPQ outside of the testing period. It is recommended that the user carefully review all methods prior to doing so.

Appendix F State FIPS Codes

Note: for a list of all state and county codes, visit USDA's website https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697

Table 21: State FIPS Codes

Name	Postal Code	FIPS
Alabama	AL	01
Alaska	AK	02
Arizona	AZ	04
Arkansas	AR	05
California	CA	06
Colorado	CO	08
Connecticut	CT	09
Delaware	DE	10
District of Columbia	DC	11
Florida	FL	12
Georgia	GA	13
Hawaii	HI	15
Idaho	ID	16
Illinois	IL	17
Indiana	IN	18
Iowa	IA	19
Kansas	KS	20
Kentucky	KY	21
Louisiana	LA	22
Maine	ME	23
Maryland	MD	24
Massachusetts	MA	25
Michigan	MI	26
Minnesota	MN	27
Mississippi	MS	28
Missouri	MO	29
Montana	MT	30
Nebraska	NE	31
Nevada	NV	32
New Hampshire	NH	33
New Jersey	NJ	34
New Mexico	NM	35
New York	NY	36

Name	Postal Code	FIPS
North Carolina	NC	37
North Dakota	ND	38
Ohio	OH	39
Oklahoma	OK	40
Oregon	OR	41
Pennsylvania	PA	42
Rhode Island	RI	44
South Carolina	SC	45
South Dakota	SD	46
Tennessee	TN	47
Texas	TX	48
Utah	UT	49
Vermont	VT	50
Virginia	VA	51
Washington	WA	53
West Virginia	WV	54
Wisconsin	WI	55
Wyoming	WY	56

Appendix G Glossary

ACS – American Community Survey. CODI-HPQ relies on ACS household counts for statistical weighting.

Age Groups – Age groups are for youth and teens only and include households with patients 2 to 6 years of age only, 7 to 17 years of age only, or households with both age ranges (would include households with 2 or more children only).

BMI – Body Mass Index. Used to categorize a person's height and weight into various categories (e.g., with obesity, does not have obesity, etc.)

CDC – Centers for Disease Control and Prevention

CDM – Common Data Model

CODI – Previously the "Childhood Obesity Data Initiative" currently the "The Clinical and Community Data Initiative." CODI is a project led by the Centers for Disease Control and Prevention originally designed to enhance data capacity for users interested in exploring the efficacy of weight-related intervention and prevention strategies.

CODI-HPQ – CODI household prevalence queries (CODI_HPQ in SAS programs)

CODI-HPQ-GEO3 – CODI HPQ applied on EHR with state and a three digit geographic identifier

Converge – Property (exhibited by the statistical weighting function) of approaching a limit more and more closely as an argument (variable) of the function increases or decreases or as the number of terms of the series increases. Crude Prevalence of BMI – is the total number of people within a particular BMI (e.g., underweight) in a specified geographic area (state, county, etc.) for a specified group of people (age, race, or all people) divided by the total population for the same geographic area and same specified group for a specific time period (e.g., 2016) and multiplied by 100.

COUNTY Data – When referenced in all capital letters, it refers to EHR data linked to a patient's state and county FIPS code.

CSV – Comma Separated Value. All input files should be in CSV.

DHDN - Distributed Health Data Network

EHR – Electronic Health Records. Digital records of patient health information. An EHR contains the patient's records from multiple providers and provides a more holistic, long-term view of a patient's health.

EMR – Electronic Medical Records. Digital records of patient health information. A digital version of a patient's chart.

Execute - In SAS software is the process by which a computer or virtual machine executes the instructions of a computer program. The term run is used synonymously in SAS. A related definition refers to the specific action of a user starting, launching, or invoking a program.

FFRDC – Federally Funded Research and Development Center

FIPS Codes – Numbers which uniquely identify geographic areas. The number of digits in FIPS codes vary depending on the level of geography. State-level FIPS codes have two digits, county-level FIPS codes have five digits of which the first two digits are the FIPS code of the state to which the county belongs followed by three digits which represent a county within the state.

Geographic Area – Geographic area is defined based on the patient's residential state and county.

GEO3 – Geographic area identified by three numbers. GEO3 is defined based on the state and county.

Growthcleanr - An open-source R package for assessing height and weight record data from EHR systems, focused on categorizing the plausibility of individual record based on longitudinal analysis of each patient subject.

Health FFRDC- Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare federally funded research and development center

Healthy Weight – Body Mass Index greater than or equal to 18.5 and less than 25

Household – According to the U.S. Census, a household consists of all the people who occupy a housing unit. See census.gov for more information. CODI-HPQ assign patients to households based on the household identifier.

Householder – According to the U.S. Census, the householder refers to the person (or one of the people) in whose name the housing unit is owned or rented (maintained) or, if there is no such person, any adult member, excluding roomers, boarders, or paid employees. If the house is owned or rented jointly by a married couple, the householder may be either the husband or the wife. The person designated as the householder is the "reference person" to whom the relationship of all other household members, if any, is recorded. The number of householders is equal to the number of households. See census.gov for more information. CODI-HPQ randomly designate one patient age 20 to 64 as the householder.

Informed Presence – The belief that patients do not randomly go to the provider's office and thus are not randomly included in EHR data.

Imputation – Estimating a value for a specific data item (e.g., race) where the response is missing or unusable.

Iterative Proportional Fitting – (IPF or raking) is an iterative algorithm for proportionally adjusting a matrix or contingency table of non-negative elements to produce a new 'similar' table with specified positive marginal totals in at least two dimensions.

MSE – Mean Squared Error

NCHS – National Center for Health Statistics

NHANES – National Health and Nutrition Examination Survey, a probability-based survey that might be more representative of the general population.

Obesity – Body Mass Index greater than or equal to 30 kg/m^2 for adults or greater than or equal to 95^{th} percentile for youth and teens.

Open-Access Program – A program made freely available to libraries and end users.

Open-Source Program – A program made freely available to libraries and end users, written in software that is free of charge.

PCORnet – Patient Centered Outcomes Research Network

Pre-Processing CODI-HPQ – a set of SAS programs that are executed once and only once per EHR data file. It is also known as the data inputs and link population data.

Prevalence – Proportion of a particular population found to be affected by a medical condition at a specific time.

PUF – Public Use File

Quickstart – A SAS program which requires user input. Only the Quickstart programs are needed along with user specifications to run the pre-processing and/or the HPQ.

Race Imputation – Imputing missing race data, see also imputation. Setting race imputation to yes allows the programs to include all available EHR data for households even if the householder's medical record did not include a known race. See Imputation for further clarification.

Random Sample - A method of selecting a sample from a population in such a way that every possible sample that could be selected has a predetermined probability of being selected.

RDM – CODI Research Data Model

RLDM – CODI Record Linkage Data Model

Run – In SAS software is the process by which a computer or virtual machine executes the instructions of a computer program. The term execute is used synonymously. A related definition refers to the specific action of a user starting, launching, or invoking a program.

SAS – SAS is a statistical software suite

Sample – The observed (or unadjusted, or crude) count of households in the study population.

SDOH – Social Determinants of Health

Statistical Weights - A statistical weight is an amount given to increase or decrease the importance of an item. Weights are commonly given for people or households when a sample and not a census is taken. The value of the weight can be thought of as denoting the number of households in the population represented by that sample household in EHR, accounting for differences between the distribution of the sample and total populations.

Note: the use of statistical weights is encouraged for all analyses because the data comes from a nonprobability sample with no known probabilities of selection. Failure to use statistical weights may yield biased results and overstated significance levels.

Suppression/Presentation Guidelines for Proportions – Guidelines used by all of HHS which provide criteria for presenting or suppressing proportions. The multistep NCHS Data Presentation Standards for Proportions are based on a minimum denominator sample size and on criteria based on the absolute and relative widths of a CI calculated using the Clopper-Pearson method.

Synthea – An open-source, synthetic patient generator that models the medical history of synthetic patients.

Variance – A measure of how far a set of numbers is spread out from their average value.

Weight Category – Categorization of a household's members height, weight, age, and sex (BMI) into one of four categories: household with no youth, teens, or adults with obesity, household with no youth or teens with obesity and one or more adults with obesity, household with one or more youth or teens with obesity and no adults with obesity, or household with one or more youth or teens with obesity and one or more adults with obesity.

Weights – See Statistical Weights or Weight Category

Weighted Prevalence – Prevalence based on weighted counts where are equal to crude prevalence with statistical weights applied.

Appendix H Abbreviations and Acronyms

ACRONYM	DEFINITION
ACS	American Community Survey
ADHD	Attention Deficit Hyperactivity Disorder
AEMR	Ambulatory Electronic Medical Record
BMI	Body Mass Index
CDC	Centers for Disease Control and Prevention
CI	Confidence Interval
CODI	Clinical and Community Data Initiative
CODI-HPQ	Clinical and Community Data Initiative Household Prevalence Queries
CSV	Comma Separated Value
DHDN	Distributed Health Data Network
EHR	Electronic Health Record
EMR	Electronic Medical Record
FFRDC	Federally Funded Research and Development Center
HHS	U.S. Department of Health and Human Services
IG	Implementation Guide
IPW	Inverse-Probability Weighting
MSE	Mean Square Error
NCHS	National Center for Health Statistics
NHANES	National Health and Nutrition Examination Survey
PUF	Public Use File
SAS	A Statistical Software Suite
SDOH	Social Determinants of Health
SFTP	Secured File Transfer Protocol

Appendix I Bibliography

Anderson, R.N., & Rosenberg, H.M. (1998). Report of the second workshop on age adjustment. National Center for Health Statistics. *Vital Health Stat* 4(30).

Best, C., & Shepherd, E. (2020). Accurate measurement of weight and height 2: Calculating height and BMI. *Nursing Times* [online]; 116: 5, 42-44.

Bower, J.K., Patel, S., Rudy, J.E., & Felix, A.S. (2017). Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: Finding the signal through the noise. *Current Epidemiology Reports*, 4(4), 346-352. doi:10.1007/s40471-017-0130-z.

Christopher, A. S., McCormick, D., Woolhandler, S., Himmelstein, D. U., Bor, D. H., & Wilper, A. P. (2016). Access to Care and Chronic Disease Outcomes Among Medicaid-Insured Persons Versus the Uninsured. *American Journal of Public Health, 106*(1), 63-69.

Daymont, C., Ross, M.E., Localio, A.R., Fiks, A.G., Wasserman, R.C., & Grundmeier, R.W. (2017). Automated identification of implausible values in growth data from pediatric electronic health records, *Journal of the American Medical Informatics Association*, 24(6) 1080–1087, https://doi.org/10.1093/jamia/ocx037

Di Consiglio, L., & Tuoto, T. (2018). When adjusting for the bias due to linkage errors: a sensitivity analysis. *Statistical Journal of the IAOS*, 34(4), 589-597.

Fiscella, K., & Fremont, A. M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research*, 41(4p1), 1482-1500.

Flood, T.L., Zhao, Y.-Q., Tomayko, E.J., Tandias, A., Carrel, A.L., & Hanrahan, L.P. (2015). Electronic health records and community health surveillance of childhood obesity. *American Journal of Preventive Medicine*, 48(2), 234-240. doi:10.1016/j.amepre.2014.10.020

Goldstein, B. A., Bhavsar, N. A., Phelan, M., & Pencina, M. J. (2016). Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *American Journal of Epidemiology, 184*(11), 847-855. doi:10.1093/aje/kww112

Hilliard, Paul J., (2017). "Using New SAS 9.4 Features for Cumulative Logit Models with Partial Proportional Odds." Paper Accompaniment for E-Poster 406-2017 Available: https://support.sas.com/resources/papers/proceedings17/0406-2017.pdf

Klein, R. J., & Schoenborn, C. A. (2001). Age adjustment using the 2000 projected U.S. population. *Healthy People 2000 statistical notes*, (20), 1–9.

Kuczmarski RJ, Ogden CL, Guo SS, et al. 2000 CDC growth charts for the United States: methods and development. *Vital Health Stat 11*. 2002;(246):1-190

Lash, T. L., Fox, M. P., & Fink, A. K. (2011). *Applying quantitative bias analysis to epidemiologic data*. Springer Science & Business Media.

Little, R. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88(423), 1001-1012. doi:10.2307/2290792

Oh, H. Lock and Scheuren, Fritz (1978), "Some Unresolved Application Issues in Raking Ratio Estimation." 1978 Proceedings of the Section on Survey Research Methods, Washington, DC: American Statistical Association, pp. 723-728.

Parker, J.D., Talih, M., Malec, D.J., et al. (2017) National Center for Health Statistics data presentation standards for proportions. National Center for Health Statistics. *Vital Health Stat* 2(175).

Romo, M. L., Chan, P. Y., Lurie-Moroni, E., Perlman, S. E., Newton-Dame, R., Thorpe, L. E., & McVeigh, K. H. (2016). Characterizing Adults Receiving Primary Medical Care in New

York City: Implications for Using Electronic Health Records for Chronic Disease Surveillance. *Preventing Chronic Disease, 13*, E56-E56. doi:10.5888/pcd13.150500

Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety*, 15(5), 291-303.

The SAS Institute. "The Logistic Procedure." Using the statistical software SAS® software (SAS Institute. 2011). SAS Institute Inc., SAS 9.4 Help and Documentation, Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_logistic_toc.htm

U.S. Census Bureau. (2020). Annual estimates of population by sex, age, race, and Hispanic origin for the United States: April 1, 2010, to July 1, 2019 (NC-EST2019-ASR6H). Washington, DC: U.S. Census Bureau, Population Division; Release Date: June 2020.

Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018 Mar 1;25(3):230-238. doi: 10.1093/jamia/ocx079. Erratum in: J Am Med Inform Assoc. 2018 Jul 1;25(7):921. PMID: 29025144; PMCID: PMC7651916.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. Springer.

NOTICE