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 Prevalence Queries Implementation Guide for Youth and Teen Data

Version 1.2

March 18, 2022

Record of Changes

Implementation Guide Changes

Version	Date	Author / Owner	Description of Change
DRAFT	July 9, 2020	Erin Tanenbaum / Health FFRDC	Initial Version
DRAFT	July 23, 2020	Erin Tanenbaum / Health FFRDC	Edited
DRAFT V10	August 11, 2021	Erin Tanenbaum/Health FFRDC	Added Social Determinants of Health, Glossary, and additional context
DRAFT	October 25, 2021	Erin Tanenbaum and Melissa Garcia/Health FFRDC	Edited
DRAFT v1.1	December 23, 2021	Erin Tanenbaum and the Health FFRDC	Edited draft
v1.2	March 18, 2022	Melissa Garcia	Edited draft Accepted by CDC on May 13, 2022

Methodology and SAS Programming Contributors

Name	Affiliation
Erin Tanenbaum	NORC at the University of Chicago
Scott Campbell	NORC at the University of Chicago
Devi Chelluri	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 prevalence queries (CODI-PQ) to generate BMI category prevalence estimates based on body mass index (BMI) percentiles[1] in youth and teens, aged 2 through 19, stratified by age, sex, and geography. Population estimates were obtained by applying statistical weights, imputation, and suppression criteria to electronic health record (EHR) non-probability-based samples. CODI-PQ were developed using a large ambulatory EHR dataset with coverage across the US. CODI-PQ were designed to use data from the CODI distributed health data network (DHDN) and other non-probability samples derived from EHR data.

1.1 Background

Public health surveillance of youth and teen 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. Self-reported or proxy-reported data can be subject to bias. 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

[1] About Child & Teen BMI | Healthy Weight, Nutrition, and Physical Activity | CDC

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. CODI-PQ 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-PQ algorithms can generate stable prevalence estimates at state, county, and zip code geographies from EHR data, depending on the data provided by the user, 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-PQ Implementation Guide for Youth and Teen Data is to provide a guide for CODI data partners[3] or end users to run the CODI-PQ. The Implementation Guide covers the following:

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

1.3 Scope

The CODI-PQ algorithms were created and tested with IQVIA's Ambulatory Electronic Medical Record (AEMR-US)[4] data and synthetic data generated for CODI using Synthea.™ [5] CODI-PQ analyze data for patient level records for children ages 2 through 19. Each record must include year of medical encounter, demographic information(age, sex, race, and some level of geographic location), and BMI percentile category. Patient-level records must include residential address information at the level of state, county, and zip code, or at the level of state and the ZIP Code Tabulation Area's first three digits (ZCTA-3). CODI-PQ leverage population counts from

[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] IQVIA's Ambulatory Electronic Medical Record (AEMR-US) database contains de-identified medical records and encounters from 44,000 physicians and 315 networks in the U.S. covering the period from January 2006 through May 2019. These data include provider medical specialty, patient variables such as examination date, year of birth, gender, race/ethnicity, and medical variables such as diagnoses, procedures, medication prescription orders, and patient and family history captured during a patient encounter. Contributing practices consist of medium to large physician offices, outpatient clinics, and physician groups. Because examination date and year of birth, but not age, were available, age was calculated from the examination date and the midpoint of the birth year (July 2).
[5] https://synthetichealth.github.io/synthea/

the American Community Survey (ACS). CODI-PQ 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-PQ can refer to Appendix A for more information. Technical staff preparing datasets for CODI-PQ can refer to Appendices C and D for detailed descriptions of the format required for input data. Explanation of CODI-PQ results can be found in Appendix F.

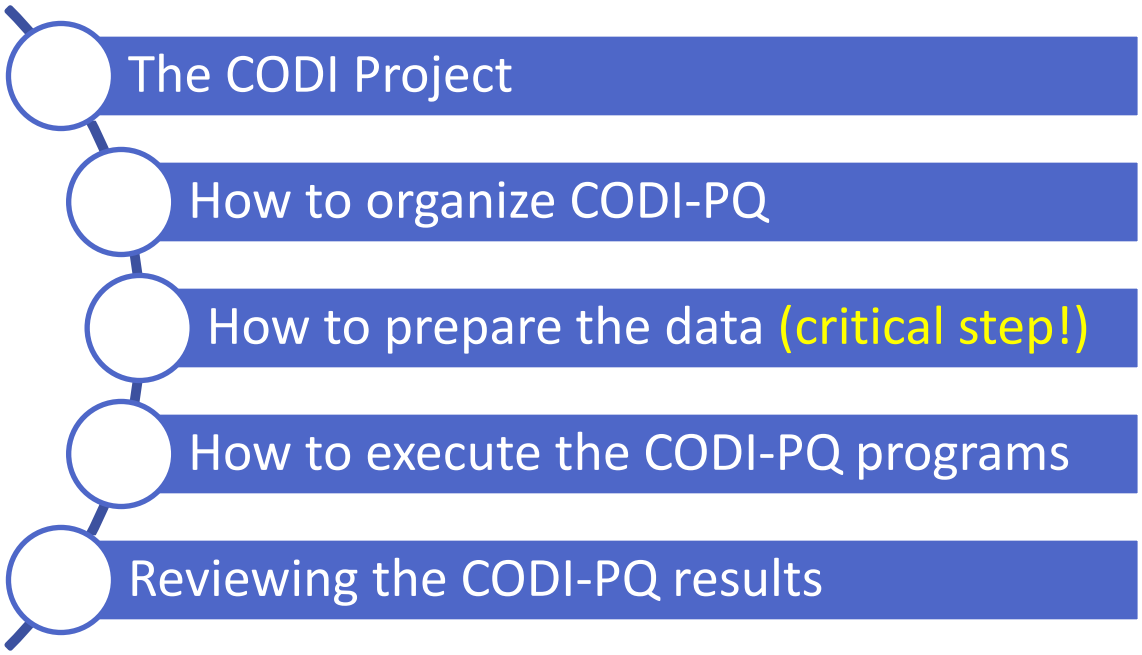
1.5 Document Organization

This document is organized as follows:

Section		Purpose
Section 1	Introduction	Provides a background for CODI-PQ
Section 2	User's Guide	Provides a general guide for users
Appendix A	Analysis Details	Provides detailed description of analysis
Appendix B	Social Determinants of Health	Provides methodology for prioritization of social determinants of health for statical weighting
Appendix C	ACS File Layouts	Table outlining the required ACS input file layouts
Appendix D	EHR File Layouts	Table outlining the required EHR input file layouts
Appendix E	CODI-PQ-GEO3 Example SAS Programs	Provides example SAS program for generating results by county or ZCTA-3
Appendix F	CODI-PQ Results Example	Provides CODI-PQ results data dictionary and example results
Appendix G	State FIPS codes	Provides list of state abbreviations
Appendix H	Glossary	Defines terms used in this document
Appendix I	Abbreviations and Acronyms	Defines acronyms used in this document
Appendix J	Bibliography	Lists sources used in preparing this document

2 User's Guide

The User's Guide section describes:



Organizing the CODI-PQ folders and properly preparing the data based on specifications are key steps to successful implementation.

2.1 CODI Concept

Figure 1 shows how CODI end 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-PQ. CODI-PQ 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.

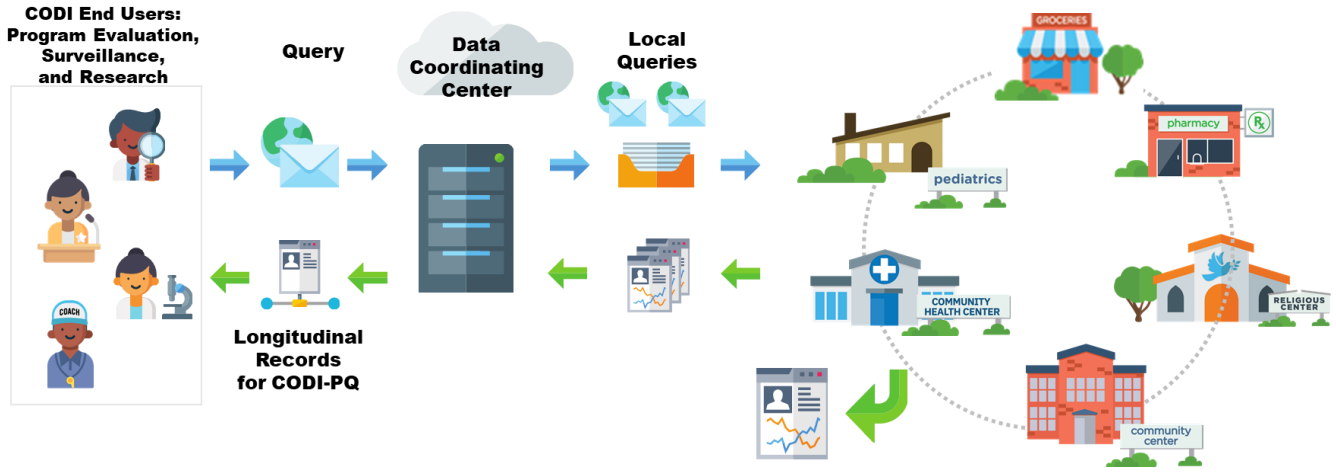


Figure 1.Data Partners with a Common Data Coordinating Center

2.2 About CODI-PQ

CODI-PQ are a set of programs that calculate youth and teen BMI percentile prevalence estimates from a non-probability sample[6] of EHR data. The CODI-PQ 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), and race imputation is conducted (PRE_PROCESSING_GEO3). In the prevalence query part, patients are selected based on user specifications, statistically weights are applied, variance estimates are calculated, results are suppressed (if needed), and prevalence results are output (CODI_PQ_GEO3).

For successful use of the CODI-PQ programs, end users are encouraged to carefully review the methodological details (described in appendices). Inputs for the CODI-PQ programs include EHR data supplied by the user and ACS data from 2019 supplied by the Health FFRDC[7]. Results can be calculated for a specific geography (e.g., state, state and county, state and ZCTA-3), subpopulation (e.g., age group, sex, race), or geography and subpopulation (e.g. age group by state and ZCTA-3).

Results are suppressed[8] if the user selects a geography or subpopulation with an insufficient number of patients for statistical weighting (see Appendix Section A.6) or if results do not meet NCHS suppression criteria (see Appendix Section A.10). The CODI-PQ programs user should have a working knowledge of SAS language and macros to select the population of interest, execute CODI-PQ, and review the SAS log.

[6] 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.
[7] ACS 2019 file for use with CODI-PQ is available for download from https://sft.mitre.org/#/folder/6281923. 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.
[8] 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.

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

- Impute race for youth and teens who are missing race information (optional)
- Calculate statistical weights with an EHR non-probability sample
- Calculate age-adjusted prevalence results (optional)
- Calculate youth and teen prevalence of BMI categories based on age- and sex-specific BMI percentiles[9], including:
  - **Underweight**: BMI less than the 5th percentile
  - **Healthy Weight**: BMI 5th percentile to less than the 85th percentile
  - **Overweight**: BMI 85th percentile to less than the 95th percentile
  - **Obesity**: BMI equal to or greater than the 95th percentile
    - **Severe Obesity**: 120 percent or greater of the BMI value for the 95th percentile
- Suppress prevalence estimates based on the 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-PQ require the following SAS features:

- 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-PQ

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

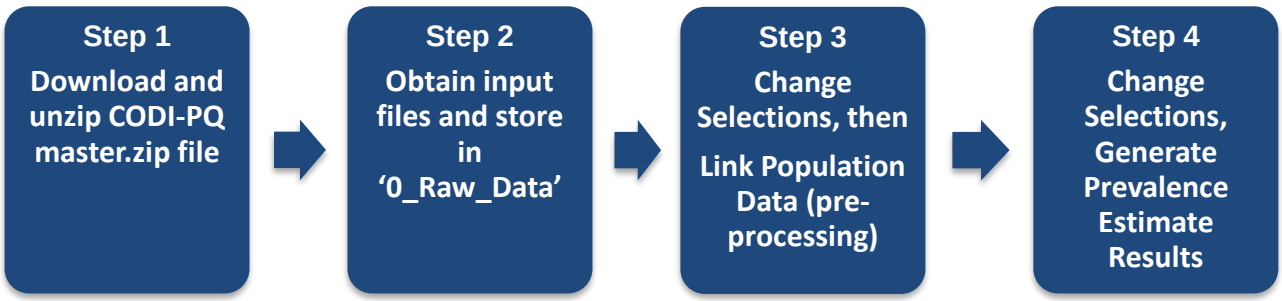


Figure 2. CODI-PQ Process

[9] https://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html#percentile

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

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

To begin, select the "Youth and Teens" folder and download "CODI-PQ-GEO3-master.zip." Note that "GEO3" refers to the program's options to estimate prevalence at the county or ZCTA3 level.

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 (Full Path Information set to on)**.

> **Unzip with Folders**
> After downloading CODI-PQ, unzip the SAS programs, and preserve the folder names by setting Full Path Information on. CODI-PQ includes "Quickstart" programs that automatically execute additional programs based on the folder structure in the zipped file.

CODI-PQ-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.

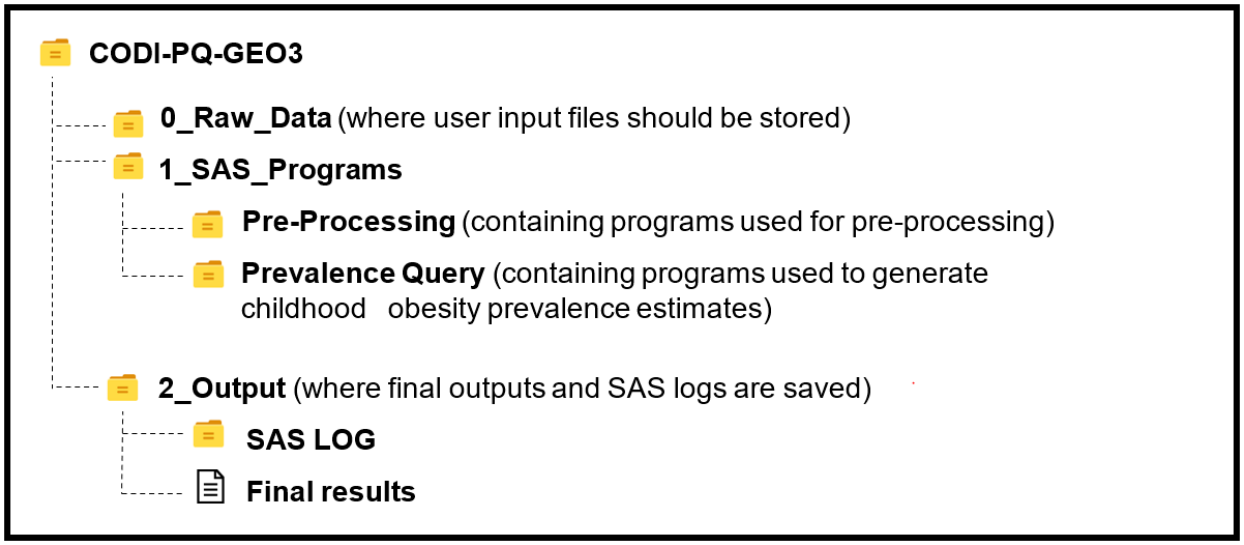


Figure 3. CODI-PQ-GEO3 Folder Structure[10]

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

[10] Note: Pre-Processing is labeled Pre_Processing_GEO3 within the folder. Prevalence query is labeled CODI_PQ_GEO3 within the folder.

models embedded into the SAS programs. For variable names, variable order, and a description of the file, see Appendix C.

2. **EHR data file** supplied by the end user in comma separated values (.csv). The EHR data file must:
    - Contain all variables in the order (sequence) expected. Variable names and order can be found in Appendix D.
    - Contain valid variable values as anticipated. Variable values can be found in Appendix D.
    - Have a unique identifier for all patients and the identifier is consistent between years.
    - Include 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 2 of each year was kept prior to executing pre-processing.
    - Include a valid height and weight value obtained on the same day to calculate the BMI and BMI percentile for all patients and categorize BMI status (e.g., underweight, healthy weight, etc.)
    - Have a geographic location of the patient's residency either as the state and ZCTA-3 or the state, county, and ZIP code.
    - Users may also wish to reconcile demographic characteristics across years for each patient, including:
        - Sex
        - Race

Additional variables on the file must be included even though their inclusion in the prevalence query is optional. If the file does not include these variables, researcher can add two columns to the end of their file with blank values. These variables include:

- A sickle cell disease[11] indicator
- A pregnancy indicator

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

2.4.3 STEP 3: Link Population (Pre-Processing)

Open the "Quickstart-Pre_Processing_CODI_PQ_GEO3" SAS program stored in "\1_SAS_Programs" and change the selection per the steps outlined in the tables below.

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 file. The programs also impute the race of those with unknown race and each time the program is submitted, new imputed race values are created and stored for each patient. For consistency, we encourage submitting the pre-processing programs only once for each EHR file. If additional data is later processed for the same patient, we encourage 1) replacing the race of all patients who were imputed before, but their race is now

[11] Note: The sickle cell disease indicator is not date sensitive since it is used for imputing race. Thus, the value can be calculated across all available data years and reconciled across years. Example: patient is identified with sickle cell disease in 2016. All records, regardless of year, for this patient would be identified as having sickle cell disease.

known, and 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 prevalence queries using the same pre-processed file 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"
2	Edit the SAS program within "SECTION 1: Input Folder and file names."	Follow the SAS programs and update the macro variable specifications, in particular (see Table 2)

Table 2. Change SAS Specifications, Section 1

SAS Macro Variable	Details	Example
ROOT_PRE	The core folder name where CODI-PQ-GEO3 is saved (see part 2.1.1). The SAS Programs folder and all other folders and files are stored in this directory.	%let ROOT_PRE = P:\CODI-PQ-master;
PRE_DEST	Following the example, the results from pre-processing will be generated and stored in folder P:\CODI-PQ-master \2_Output\Pre_Processed_***CODI_PQ_GEO3***	%let PRE_DEST = CODI_PQ_GEO3;
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. Important: the csv file must have all variables in the order specified, with the correct variable name, and with expected values. See C.1.	%let ACS_FILENAME = ACS_State_COUNTY;

EHR_FileNAME	The youth and teen level EHR file name (file described in 2.4.2). Do not include the extension. Important: the csv file must have all variables in the order specified, with the correct variable name, and with expected values. See D.1. For the example, the user will have stored their raw data in: P:\CODI-PQ-master\0_Raw_Data**EHR_csv_filename**.csv	%let EHR_FileNAME = EHR_csv_filename;
EHR_PRE_OUT	Optional, user can name the pre-processing output file (ACCEPTABLE VALUES: file name (no punctuations)). The example would be stored in: P:\CODI-PQ-master \2_Output\Pre_Processed_*CODI_PQ_GEO3***SAVES_EHR_HERE**	%let EHR_PRE_OUT = SAVES_EHR_HERE;
LOG_NAME_PRE	The name of the SAS log file. Quickstart_Pre_Processing_CODI_PQ. Users have the option to rename the log file name before it is created. The example SAS log would be stored in: P:\CODI-PQ-master \2_Output\SAS LOG***LogName***<Date and Time>.log. Note, the program automatically includes the date and time in all log file names*/	%let LOG_NAME_PRE = LogName;

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

Order	Description	Example or Detail
3	Section 2, edit the SAS program within "SECTION 2: Beginning and End Year of longitudinal EHR data." **In pre-processing, the start and end year includes all years within your patient level file (e.g., 2014 – 2019)**. In contrast, when generating prevalence estimate results, only include the specific year(s) requested for analysis (Part 4, e.g., 2019 if prevalence estimates with 2019 records are requested).	. /***/ %LET BEGIN_YEAR = **2014**; /***/ %LET END_YEAR = **2019**;
4	Edit the SAS program within "Section 4: County or ZCTA3 data (REQUIRED)" Edit with a Y or N. For example, Y for County level data, **N for ZCTA3 level data**. The ACS data file and EHR data file must have a variable GEO3 which includes three digits for either ZCTA3 or County codes. Both files must have equivalent geographic types. Thus, the files can have either have State+County or State+ZCTA3, but not both.	%LET COUNTY=**N**;
5	Save the Quickstart program.	SAS encourages saving all files before submitting the program.

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

Order	Description	Details
1	Submit the Quickstart program.	Submit the Quickstart program. The program completes all linkage and pre-processing tasks by looping through all needed SAS programs automatically.
2	Review the log.	Review the log for possible errors including words such as error, and uninitialized. Assuming no errors, 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_PQ_GEO3" SAS program stored in "\1_SAS_Programs" and change the selections outlined in the tables below.

The final results (CODI-PQ results) will be generated in .csv or Excel format and saved in "\2_Output." Appendix F provides an example of the results. Note that results are for the group of patients selected by the user. To calculate results for multiple geographic or demographic characteristics (e.g., first for youths ages 2 to 4 and then for teens aged 18 and 19), the user will need to update and execute the programs multiple times.

Note: the age ranges, sex, and races selected must match the data on the EHRs. For example, if all age ranges are selected by the user and the file has youths ages 2 to 12 but does not have teens aged 13 to 19, then the program will fail with an error message caused by insufficient sample size for patients aged 15 to 19.

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"
2	Edit the SAS program within "SECTION 1: Folder and file names" "SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE, STATE/ZCTA3, or STATE/COUNTY"	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_PQ	The core folder name, same as in pre-processing.	%let ROOT_PRE = P:\CODI-PQ-master;
PRE_DEST	The value from pre-processing quickstart variable pre_dest. See 2.4.3, Table 2. Following the example, the results from pre-processing were previously generated and stored in P:\CODI-PQ-master\2_Output\ Pre_Processed_**CODI_PQ_GEO3**	%let PRE_DEST = CODI_PQ_GEO3;
EHR_PRE_OUT	The youth and teen level EHR file from pre-processing. The example was stored in: P:\CODI-PQ-master \2_Output\Pre_Processed_*CODI_PQ_GEO3***SAVES_EHR_HERE**	%let EHR_PRE_OUT = SAVES_EHR_HERE;
LOG_NAME	The name of the resulting SAS log. Users have the option to rename the log file name before it is created. Following the example syntax, the SAS log will be stored in: P:\CODI-PQ-master\2_Output \SAS LOG**LogName**<Date and Time>.log. Note, the program automatically includes the date and time in all log file names.	%let LOG_NAME = LogName;

SAS Macro Variable	Details	Example
FileOUT_Name	The prefix for the resulting .csv or Excel file. Following the example syntax, the .csv or Excel file will be stored in: P:\CODI-PQ-master\ \2_Output**File_name**<Date and Time>.csv. Note, the program automatically includes the date and time in all results file names.	%LET FileOUT_Name = File_name;
SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE, STATE/ZCTA3, or STATE/COUNTY		
BEG_YEAR	Subsets the prevalence to medical encounters in this year for BMI prevalence. The prevalence will include adult EHR data from this year and after. Acceptable values must be a 4 digit year and the year must be present on the user's EHR file.	/***/ %LET BEG_YEAR = 2016;
END_YEAR	Subsets the prevalence to medical encounters through this year for BMI prevalence. The prevalence will include adult EHR data from this year and before. Acceptable values must be a 4 digit year and the year must be present on the user's EHR file. If the end year is not equal to the beginning year, then each patient's most recent record will be kept. Patients are not included multiple times within analytic results.	/***/ %LET END_YEAR = 2018;
ALL_STATES	Includes all states and the District of Columbia (national estimate) in the prevalence based on the geographic location of the youth or teen. For state or smaller geographies, set ALL_STATES = N; then by default the program will subset the prevalence based on the individual state or state+GEO3 values specified (in future step). Note that the EHR data file must have sufficient sample size in all states if set to yes (Y).	/*@Note: Include all geographical locations in file? (ACCEPTED VALUES: Y/N) ***/

SAS Macro Variable	Details	Example
ALL_AGES	Subsets the prevalence based on the age of the youth or teen. The user may either select to include all youth and teens aged 2 to 19 (Y) or alternatively may select age groups (N). Note: if ALL_AGES = Y; then by default the program will include all youth and teens aged 2 to 19. If ALL_AGES = N; then by default the program will subset the prevalence based on the individual age ranges selected (in a future step).	/***/ %LET ALL_AGES = Y;
ALL_SEXES	Subsets the prevalence based on the sex of the youth or teen. The user may either select to include all male and female youth and teens (Y) or alternatively may select either males or females (N). Note: if ALL_SEXES = Y; then by default the program will include both males and females. If ALL_SEXES = N; then by default the program will subset the prevalence based on the individual sex(es) selected (in a future step).	/***/ %LET ALL_SEXES = Y;
ALL_RACES	Subsets the prevalence based on the race of the youth or teen. The user may either select to include all races (Y) or alternatively may select race(s) (N). 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, Black, Asian, Other). If ALL_RACES = N; then by default the program will subset the prevalence based on the individual races selected (in a future step).	/***/ %LET ALL_RACES = Y;
SECTION 3: Additional Flags		
ACSCOUNTY	Specifies the geographic level of the ACS data.	/***/ %LET ACSCOUNTY = N; /*@Note: Is the ACS data at the county or ZCTA3 level? (ACCEPTABLE VALUES: Y for County level data, N for ZCTA3 level data) ***/

SAS Macro Variable	Details	Example
INCLUDE_PREGNANCY	Subsets EHR records based on pregnancy flag. **If set to yes (Y), then patients will be included regardless of pregnancy status.** If set to no (N,) then only patients with a pregnancy status set to no (0) will be included.	/***/ %LET INCLUDE_PREGNANCY = Y;
SAMPLE_CHECK	Executes an optional review of the EHR counts by age, race, and sex based on user defined criteria. All demographic categories selected by the user (e.g., sex male, etc.) will be displayed in the SAS output or results window. Each factor will include the factor, value (e.g., sex, male) and either "Sample Size Is Insufficient" if n <20 or "Sample Size Is Sufficient.	/***/ %LET SAMPLE_CHECK = Y;

Table 7. Change Specifications, Processing Steps, Continued

Order	Description	Details
3	Edit the SAS program within "SECTION 4: Only complete section 4 for any "N" values listed in section 2" "SECTION 5: Methodological option selections"	**Review specifications below.**

Table 8. Change Specifications, Processing Steps, Continued

SAS Macro Variable Category	Details	Example
SECTION 4: Only complete section 4 for any "N" values listed in section 2		
If ALL_STATES = N	GEO_GROUP informs the program the level of geography in the EHR and ACS data as well as in the GEO_LIST macro variable. The level of geography must match in all three locations. GEO_LIST subsets the EHR used in prevalence results based on the location of the patients. GEO_GROUP can take the value of a) STATE, b) ZCTA3, or c) COUNTY. Syntax for all three scenarios is described and shown in examples. Of note, values should be	/*Example 1:*/ /***/ %LET GEO_GROUP = STATE; /***/ %LET GEO_LIST = %STR('08', '10'); /*Example 2:*/ /***/ %LET GEO_GROUP = ZCTA3; /***/ %LET GEO_LIST = %STR('51221', '26486'); /*Example 3:*/

SAS Macro Variable Category	Details	Example
	surrounded by single quotes and comma delimited if more than one geography is to be included in the results. If ALL_STATES is set to yes (Y), then the SAS program does not review the user's responses to the GEO_LIST or GEO_GROUP. Example one selects patients in Colorado and Delaware. Example two uses ZCTA3 GEO_LIST identifiers that are 5 digit values composed of the 2-digit state FIPS code and the ZCTA3 or ZIP-3 code. '51221' selects patients living in Virginia (FIPS 51), within the ZIP-3 of 221 and '28486' selects patients living in Michigan (FIPS 24) with the ZIP-3 of 486. Example three selects patients living in Virginia, within Fauquier County and patients living in Virginia within Fairfax County.	/***/ %LET GEO_GROUP = COUNTY; /***/ %LET GEO_LIST = %STR('51061', '51059');
If ALL_AGES = N and CO_OCCURRING = N;	If ALL_AGES is set to no and CO_OCCURRING is set to no, the age macros (2-4, 5-9, 10-14, 15-17, 18-19) subset the prevalence based on the age of the youth or teen and the responses to each individual age macro. Note that if ALL_AGES is set to yes, then the SAS program does not review the age-specific macros.	%LET WGT_AGE_2_4 = N; %LET WGT_AGE_5_9 = N; %LET WGT_AGE_10_14 = Y; %LET WGT_AGE_15_17 = Y; %LET WGT_AGE_18_19 = Y;
If ALL_RACES = N;	If ALL_RACES is set to no, the race macros (White, Black, Asian, Other) subset the prevalence based on the race or imputed race of the youth or teen and the responses to each individual age macro. Note that if ALL_RACES is set to yes, 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;
If ALL_SEXES = N;	If ALL_SEXES is set to no, the sex macros (male, female) subset the prevalence based on the sex of the youth or teen and the responses to each individual sex macro. Note that if ALL_SEXES is set to yes, then the SAS program does not review the sex-specific macros	%LET SEX_MALE = N; %LET SEX_FEMALE = Y;

SAS Macro Variable Category	Details	Example
SECTION 5: Methodological option selections		
IMP_RACES	If IMP_RACES is set to yes (Y), then the program includes youth and teens with imputed race values. Otherwise, if IMP_RACES is set to no (N), then the patients with imputed races are excluded.	%LET IMP_RACES = Y;
AGE_ADJ	If AGE_ADJ is set to yes (Y), then the program generates age adjusted prevalence and standard errors. Otherwise, if AGE_ADJ is set to no (N), age adjusted prevalence is not generated. Crude and weighted estimates are generated by default.	%LET AGE_ADJ = Y;

Table 9. Change Specifications, Processing Steps, Continued

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

Table 10. CODI-PQ Execution Processing Steps

Order	Description	Details
1	Submit CODI-PQ Quickstart program.	Submit the Quickstart program. The program completes all tasks within the datasets 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, and uninitialized. Assuming no errors, continue to the next step. 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 described above 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 with a single year, consider using two or three years of data[12].

[12] 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.

2.4.5 Review BMI Category Prevalence Results

CODI-PQ generate prevalence outputs as a csv file. Table 11 provides an overview of the variables included for BMI category prevalence. Note, descriptive information about CODI-PQ user inputs, error codes, sources of technical documentation, caveats, and a possible citation begins with the rows labeled Order 3 and beyond. The exact notes displayed vary.

Table 11. CODI-PQ BMI Percentile Prevalence Results Data Dictionary

Column	Description
Order	Row order
BMI Category	The BMI category based on BMI percentile.
Sample	The observed (or unadjusted, or crude) count of youth and teens in the study population.
Population	The weighted (or adjusted) count of the study population.
Crude Prevalence	The observed (or unadjusted, or crude) prevalence in the study population.
Crude Prevalence Standard Error	The observed (or unadjusted, or crude) standard error in the study population.
Weighted Prevalence	Prevalence based on weighted counts. A sample weight is assigned to each sampled patient. It is a measure of the number of youth and teens in the population represented by that sample patient. See implementation guide, Appendix A. Sample Weights for more information.
Weighted Prevalence Standard Error	Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information.
Age-Adjusted Prevalence	Prevalence based on weighted, age-adjusted counts (optional). See implementation guide, Appendix A. Age Adjustment for more information.
Age-Adjusted Prevalence Standard Error	Standard error based on weighted, age-adjusted counts. See implementation guide, See implementation guide, Appendix A. Age Adjustment 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 C – ACS File Layouts
- Appendix D – EHR data File Layouts
- Appendix E – CODI-PQ-GEO3 Example SAS Programs
- Appendix F – CODI-PQ Results
- Appendix G – State FIPS Codes

Appendix A Analysis Details

A.1 Age Adjustment

Data are age-adjusted to eliminate differences in observed results that result from differences in the age distribution of the population among geographies. The projected 2000 U.S. population was used as the standard population.[13] The specific age groups used for age adjustment are 2 to 4 years, 5 to 14 years, and 15 to 19 years. Age-adjusted values may differ from weighted values even though age is used within the weighting program since the age distribution within a geography (GEO3) may differ from the nation.

Age adjustment, using the direct method, is the application of age-specific results in a population of interest to a standardized age distribution to eliminate differences in observed results that result from age differences in population composition. This adjustment is usually done when comparing two or more populations at one point in time or one population at two or more points in time.

Age-adjusted proportions are calculated by the direct method as follows:

$$\sum_{i=1}^{n} m_i \times (p_i/P)$$

where m_i = measure of the proportion in age group *i* in the population of interest, p_i = standard population in age group *i*, and n = total number of age groups over the age range of the age-adjusted prevalence.

$$\mathrm{P} = \sum_{i=1}^{n} p_i$$

Age adjustment by the direct method requires use of a standard age distribution. The standard for age adjusting proportions for data occurring after year 2000 is the year 2000 projected U.S. resident population.

Table 12. Projected Year 2000 U.S. Population Proportion Distribution by Age for Age Adjusting

Age	Proportion Distribution (weights)
All ages*	1
2 to 4	0.1605
5 to 9	0.2796
10 to 14	0.2815
15 to 17	0.1659
18 and 19	0.1123

*Figure is rounded up instead of down to force total to 1.0.

[13] Klein & Schoenborn, 2001.

Age-adjusted prevalence results and standard errors will typically be similar or identical to the weighted prevalence and standard errors. Age-adjusted results may differ from weighted results if one or more age group weighting cell was aggregated.

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 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.[14].

For youth and teens, BMI percentiles are based on age- and sex-specific growth charts and are often referred to as BMI-for-age percentiles. In youth and teens, a high amount of body fat can lead to weight-related diseases and other health issues. Having underweight can also put patients at risk for health issues.

Youth and teen BMI weight categories are described in section A.12

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-PQ on youth and teen data. Required input files are the following:

- EHR data file (data in csv format, provided by user) provided by the user
- American Community Survey (ACS) data file (provided by the Health FFRDC[15])

CODI-PQ are intended for use with all available EHR data for a geography or subpopulation. The programs were created and tested with IQVIA's Ambulatory Electronic Medical Record (AEMR)[16] data and synthetic data generated for CODI using Synthea.[17] The guide provided in this document is implemented through open-access programs.

The programs were tested using AEMR data and synthetic EHR data. Both provide a non-probability sample of longitudinally linked patients' medical records from within the U.S. CODI-PQ subset the file to youth and teens aged 2 to 19 years of age for BMI percentile prevalence. 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 either county or the first

[14] https://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html, Accessed March 9, 2022.
[15] ACS 2019 file for use with CODI-PQ is available for download from https://sft.mitre.org/#/folder/6281923. 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.
[16] CDC provided Ambulator Electronic Medical Record data under a Data Use Agreement with the Health FFRDC.
[17] The Synthea package is based on Walonoski, et al., 2017 and is available at: https://synthetichealth.github.io/synthea/.

three digits of the ZIP Code Tabulation Area (ZCTA-3)[18] associated with the patient's address. Patients are excluded from the analysis if their state and county or ZCTA-3 does not exist or if the ACS estimated population count within their county or ZCTA-3 equals 0.

Testing of CODI-PQ included 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 implausibility (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 or state and ZCTA-3.

A.4 Testing Data

A large sample of EHR data was used to build CODI-PQ. The sample was compared to Census estimates[19] to determine possible biases associated with the sample. Similar biases maybe in other EHR data, although users are encouraged to perform a similar analysis.

To obtain reliable BMI percentile results from EHRs, the data should be free of potential bias. BMI percentile prevalence results using incomplete EHR data could be biased if there are systematic differences in characteristics associated with BMI percentile between measured and non-measured patients. It is necessary to identify the factors leading to missingness (lack of weight and height measurements) and apply strategies, such as inverse-probability weighting (IPW) or imputation to obtain corrected results.

When using IPW, EHR data are weighted by the inverse of their probability of being sampled. The validity of IPW relies on including variables associated with missingness. In the same context, imputation generates unbiased results for the variables of interest through a program that replaces each missing value with a plausible value, thus creating complete datasets.

The purpose of reviewing the sample distribution to the Census population distribution was to review possible selection bias associated with missing weight and height data.

Table 13. American Community Survey Census Sample by Year for 2015-2019 for Youth and Teens Aged 2 to 19

2015	2016	2017	2018	2019
2,433,287	2,395,712	2,259,231	1,986,787	1,703,432

In 2015, there were 2,433,287 patients ages 2 to 19 in our sample EHR data file with plausible height and weight values from a total 74,123,130 patients living in the U.S. of that age, representing 3.3% of the population. This count of patients decreased with time to 1,703,432 patients in 2019 out of 74,012,765 patients living the U.S., representing 2.3% of the population.

[18] ZCTAs are areal representations of ZIP code service areas created by the Census Bureau. Approximately 99% of ZIP-3's with a population greater than zero are equal to ZCTA-3 and thus ZCTA-3 and ZIP-3 are used interchangeably within the analysis.

[19] Annual Estimates of the Resident Population by Sex, Age, Race, and Hispanic Origin for the United States: April 1, 2010, to July 1, 2019 (NC-EST2019-ASR6H); Source: U.S. Census Bureau, Population Division; Release Date: June 2020.

Weight and height measurements data stored in EHRs and pre-processed by growthcleanr were used to calculate the percentages by age with at least one measurement within a one-year period. The greatest percentages of measurements were seen for 2-year-olds (3.87%) and the lowest percentages of measurements were seen for 19-year-olds (2.61%). Counts were also compared by ACS age groups including aged 2 to 4 years, 5 to 9 years, 10 to 14 years, 15 to 17 years, and 18 to 19 years old. The lowest percentages of measures were seen for teens 18 to 19 years old (2.80%) and youth 5 to 9 (3.25%), and the greatest percentages of measures were seen for 2- to 4- year-olds (3.57%) and 10- to 14- year-olds (3.34%). This pattern remained consistent through 2019.

Table 14. Comparison of Patients with Plausible Weight and Height Data in the Electronic Health Records and the U.S. Population by Age, 2015

Age	Census 2015	EHR data 2015	
2 years	3,966,321	153,465	3.87%
3 years	3,974,351	135,299	3.40%
4 years	4,020,292	138,829	3.45%
5 years	4,017,589	143,688	3.58%
6 years	4,017,388	134,812	3.36%
7 years	4,145,872	129,926	3.13%
8 years	4,165,033	130,063	3.12%
9 years	4,130,887	127,399	3.08%
10 years	4,118,721	125,318	3.04%
11 years	4,126,769	136,027	3.30%
12 years	4,096,602	145,489	3.55%
13 years	4,080,092	139,636	3.42%
14 years	4,182,193	141,656	3.39%
15 years	4,247,757	143,304	3.37%
16 years	4,181,741	138,049	3.30%
17 years	4,189,329	133,084	3.18%
18 years	4,210,553	126,365	3.00%
19 years	4,251,640	110,878	2.61%
Total	74,123,130	2,433,287	3.28%
2 to 4 years	11,960,964	427,593	3.57%
5 to 9 years	20,476,769	665,888	3.25%
10 to 14 years	20,604,377	688,126	3.34%
15 to 17 years	12,618,827	414,437	3.28%
18 to 19 years	8,462,193	237,243	2.80%

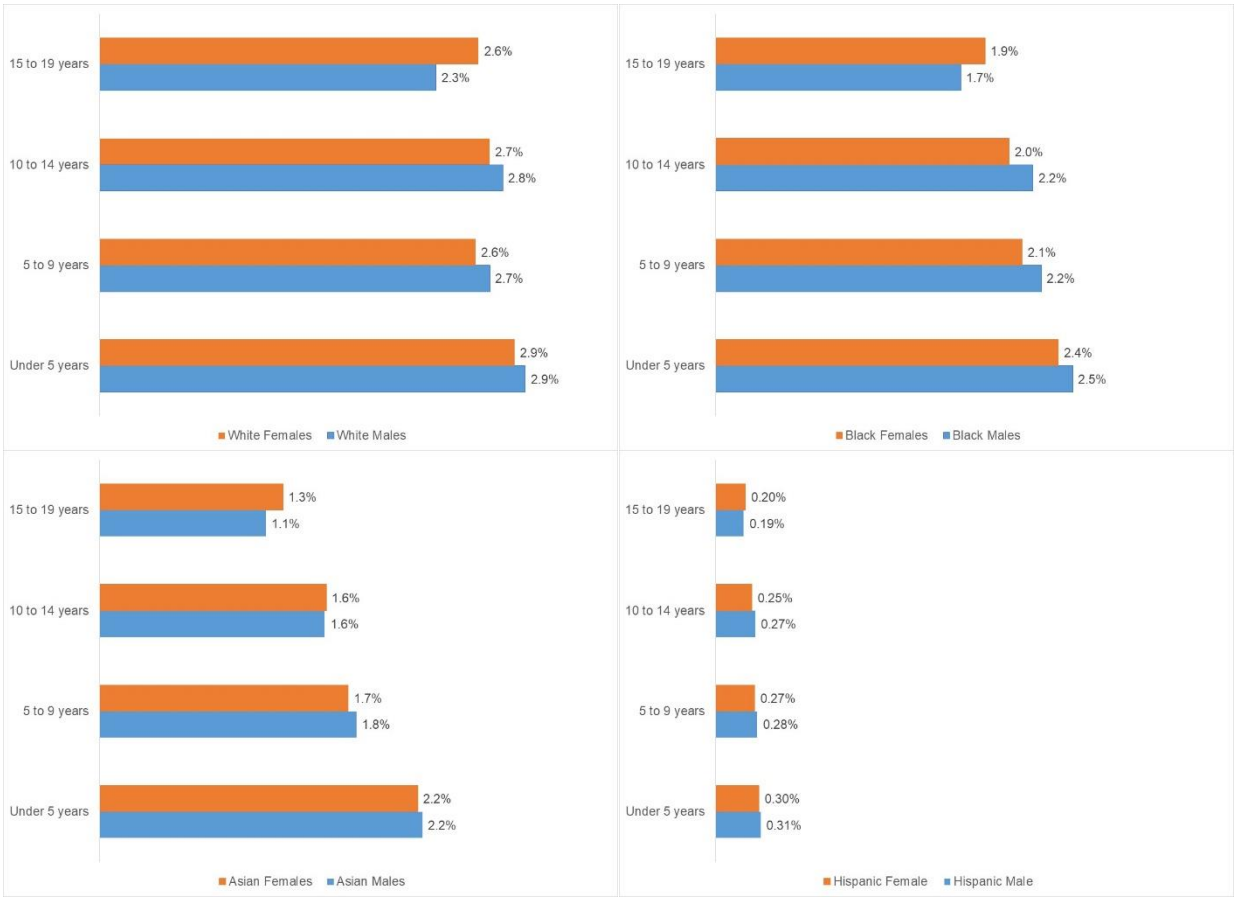


Figure 4. Comparison of Patients with Plausible Weight and Height Data in the Electronic Health Records and the U.S. Population by Race, Sex, and Age Group, 2015

For valid health prevalence results in youth and teens, we reviewed coverage of relevant characteristics related to BMI percentile, specifically sex and race/ethnicity.

Weight and height measurements data extracted from EHR systems and pre-processed by growthcleanr[20] were used to calculate the number of patients with plausible values per 1,000 persons in the population by age and sex (Figure 5. Comparison of Patients with Plausible Weight and Height Data in the Electronic Health Records and the U.S. Population by Age and Sex, 2015). Both males and females were included in the EHR, most often at the age of 2, with more males per 1,000 in the population being included than females. Males continued to outpace females for plausible values until puberty, when the rate of males with plausible height and weight data decreased with age, whereas females with plausible data remained relatively stable. By the age of 19, males were in the data at a rate of 21.2 per 1,000 persons, whereas females were in the data at a rate of 31.2 per 1,000 persons.

[20] Growthcleanr is available at https://github.com/carriedaymont/growthcleanr

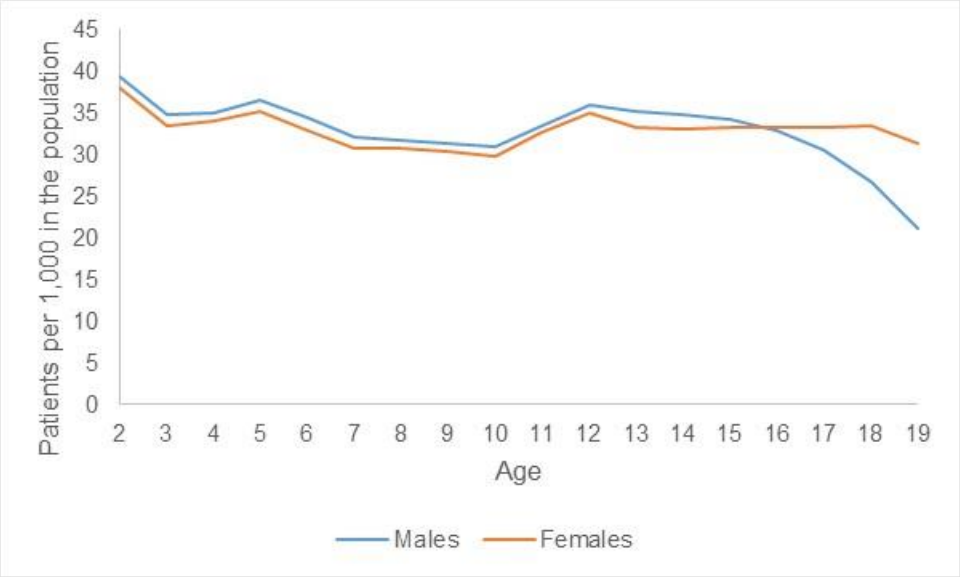


Figure 5. Comparison of Patients with Plausible Weight and Height Data in the Electronic Health Records and the U.S. Population by Age and Sex, 2015

A.4.1 Prevalence

A prevalence is either:

- **Crude**: the proportion of the sample that has a health condition (BMI percentile) at a point in time.
- **Weighted**: the proportion of the population that has a health condition at a point in time. See the Appendix A section "Statistical Weights" for more information.
- **Age-Adjusted**: the proportion of the population (adjusted by national distribution of age) that has a health condition at a point in time. See Appendix A section "Age Adjustment" for more information.

A.5 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 races).

These racial categories conform to previous work using a sample EHR data file. These categories are used because they are in the IQVIA AEMR dataset used for CODI-PQ development. Information on ethnicity is not captured in IQVIA AEMR and therefore not used in CODI-PQ development. We recognize that these categories may not accurately reflect the way that patients would self-identify and may conceal important differences within groups.

A.5.1 Race Exclusion

Statistical weighting programs require a large sample size (20 or more) in each stratum. If one or more racial groups has an insufficient sample size, the patients in the racial group impacted will automatically be excluded by the program.

A.5.2 Sickle Cell Disease

The race imputation uses presence of sickle cell disease (optional) to aid in imputing a patient's race due to its strong correlation with race. The sickle cell disease phenotype used in the development phase includes the following:

ICD-10 Codes: D57, D57.0, D57.00, D57.01, D57.02, D57.1, D57.2, D57.20, D57.21, D57.211, D57.212, D57.219, D57.3 (sickle cell trait), D57.4* (thalassemias), D57.40, D57.41, D57.411, D57.412, D57.419, D57.8, D57.80, D57.81, D57.811, D57.812, D57.819

ICD-9 Codes: 282.6, 282.60, 282.61, 282.62, 282.63, 282.64, 282.68, 282.69, 282.4* (thalassemias), 282.40, 282.41, 282.42, 282.43, 282.44, 282.45, 282.46, 282.47, 282.49, 282.5 (sickle cell trait)

SNOMED: Concept ID 22281 (CC 127040003), Concept ID 26942 (CC 417425009), Concept ID 40485018 (CC 444108000), Concept ID 4213628 (CC 417357006), Concept ID 4216915 (CC 417279003), Concept ID 30683 (CC 416180004), Concept ID 315523 (CC 36472007), Concept ID 443738 (CC 416826005), Concept ID 321263 (CC 417048006), Concept ID 25518 (CC- 16402000 sickle cell trait), Concept ID 24006 (CC 35434009), Concept ID 443721 (CC 417517009), Concept ID 443726 (CC 417683006)

The probability of each race, given presence of sickle cell disease was calculated from a combination of published incidence rates as well as verified with AEMR where race and sickle cell disease were available.

Table 15. Proportions of Sickle Cell Disease Used to Impute Race

Race	Sickle Cell Proportion
African American	94.49%
White	3.94%
Other	1.14%
Asian	0.42%

A.5.3 Race Imputation

Race is a required input for CODI-PQ. The data inputs and link population data (pre-processing) program inputs race for each youth or teen missing race information. The program operates sequentially in three phases, imputing race for youth and teens in one of the following three phases, those who:

1. Have sickle cell disease
2. Are identified as Hispanic and do not have sickle cell disease
3. Neither have sickle cell disease nor are identified as Hispanic

The race imputation relies on a combination of medical and ACS data.

Once complete, the results from each phase are aggregated with each youth or teen 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 four reasons:

1. The patient's geography is either invalid or did not have a population count in the 2019 ACS.
2. The patient's age is outside of the scope of the program or is unknown. Only persons ages 2 to 19 are in scope.
3. The sex of the patient is unknown.

CODI-PQ assign a value for race if a patient does not have a known racial value through statistical imputation. In testing, approximately 27% of the 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-PQ 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 exist in youth and teen BMI percentile 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-PQ impute race for those with unknown race using programs based on race and ethnicity of surrounding the community, ethnicity of the patient (where available if race is unavailable), sickle cell disease, age, and height. Statistical weights are calculated (based on each patient's age, sex, race, geography, and community characteristics) and used to adjust the EHR data non-probability sample to the population of interest. 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, weighted, and age-adjusted weighted results.

For records lacking race information, automated race imputation is employed in CODI-PQ data inputs and linked population data (pre-processing). Within the final program to calculate prevalence, the user specifies whether patients 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."

Race imputation occurs for each patient with an unknown race in three phases:

1. Patients with sickle cell disease
2. Patients identified as Hispanic but not identified with sickle cell disease
3. All other patients (neither have sickle cell disease nor are identified as Hispanic)

Table 16. Percentage of Patients Imputed for Each Phase in the Race Imputation Using AEMR Data

Phase	Percent of those Imputed
Phase 1: Imputed based on known chronic condition	6.8%
Phase 2: Imputed based on ethnicity	7.0%
Phase 3: All other patients with unknown race	86.3%

A.6 Statistical Weights

CODI and National AEMR data are derived from EHR data. As described in Appendix A, applying statistical weights is often used to reduce potential biases introduced by the EHR data sampling methodology. Ratio adjustments are applied to all sampled youth and teens. Ratio adjustment is a statistical weighting technique aimed to improve the accuracy of survey results by both reducing bias and increasing precision.[21] 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 dataset. 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-PQ raking includes social determinant of health categories – age, sex, race, 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 = age category less than or greater than current

Sex = do not aggregate

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

Education = community with a similar education category

Raking is completed by adjusting for one demographic variable (or dimension) at a time. For example, when weighting by age and sex, weights would first be adjusted for age groups, then those results would be adjusted by sex 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-PQ 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 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

[21] Little, 1993.

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 an age group of 2 to 4 years. Thus, CODI-PQ calculate the population of this age group by multiplying the count of persons aged 5 and under within each geography by the percentage of the national population that is aged 2 to 4, given that they are under the age of 5, based on Annual Estimates of the Resident Population by Single Year of Age and Sex for the U.S., July 1, 2018, U.S. Census Bureau. Population counts by race for those age less than 5 are also adjusted from Annual Estimates of the Resident Population using the same adjustment.

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 youth or teen'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.7 Prevalence Calculations

Crude prevalence is calculated as the count of the sample within each BMI percentile category.

To calculate the weighted prevalence of the population the sum of statistical weights within each BMI percentile 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 percentile, a Taylor-series approximation is used.[22]

Users are provided crude (unweighted) population, prevalence, and standard error, weighted population, prevalence, and standard error, and an optional age-adjusted prevalence and standard error. Age-adjusting aims to eliminate differences in results that result from differences in the age distribution of the population among geographies. The projected 2000 U.S. population was used as the standard population per current guide.

A.8 Sample Check

If SAMPLE_CHECK = Y, then the CODI-PQ execute an optional review of the sample size by age, race, and sex based on user defined criteria. All demographic categories selected by the user (e.g., sex male, etc.) will be displayed in the SAS output or results window. Each will include the factor, value (e.g., sex, male) and either "Sample Size Is Insufficient" if n <20 or "Sample Size Is Sufficient.Executes. This optional check is included to pinpoint potential sample size issues. For example, if the sample size is insufficient for males, the user may choose to execute CODI-PQ again after excluding males.

[22] Wolter, 2007.

A.9 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.

A.10 Suppression Criteria

Prevalence may be suppressed. CODI-PQ data suppression is adapted from the NCHS data presentation standards for reporting proportions in NCHS reports and data products,[23] developed by the Data Suppression Workgroup at NCHS.

The multistep NCHS Data Presentation Standards for Proportions are based on a minimum denominator sample size 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-PQ results. The Presentation Standards also provide guidance for identifying results for statistical review, CODI-PQ do not identify records for statistical review and leave this step for the user. The data presentation standards are described in Table 17 and Figure 6.

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, race, or sex categories. The suppression thresholds may also be altered by the user in the Quickstart program.

Table 17. NCHS Data Presentation Standards for Proportions

[23] Parker et al., 2017.

Statistic	Standard
Sample size	Proportions should be based on a minimum denominator sample size and effective denominator sample size (when applicable) of 30. 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[24]), 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.
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 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.

[24] 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).

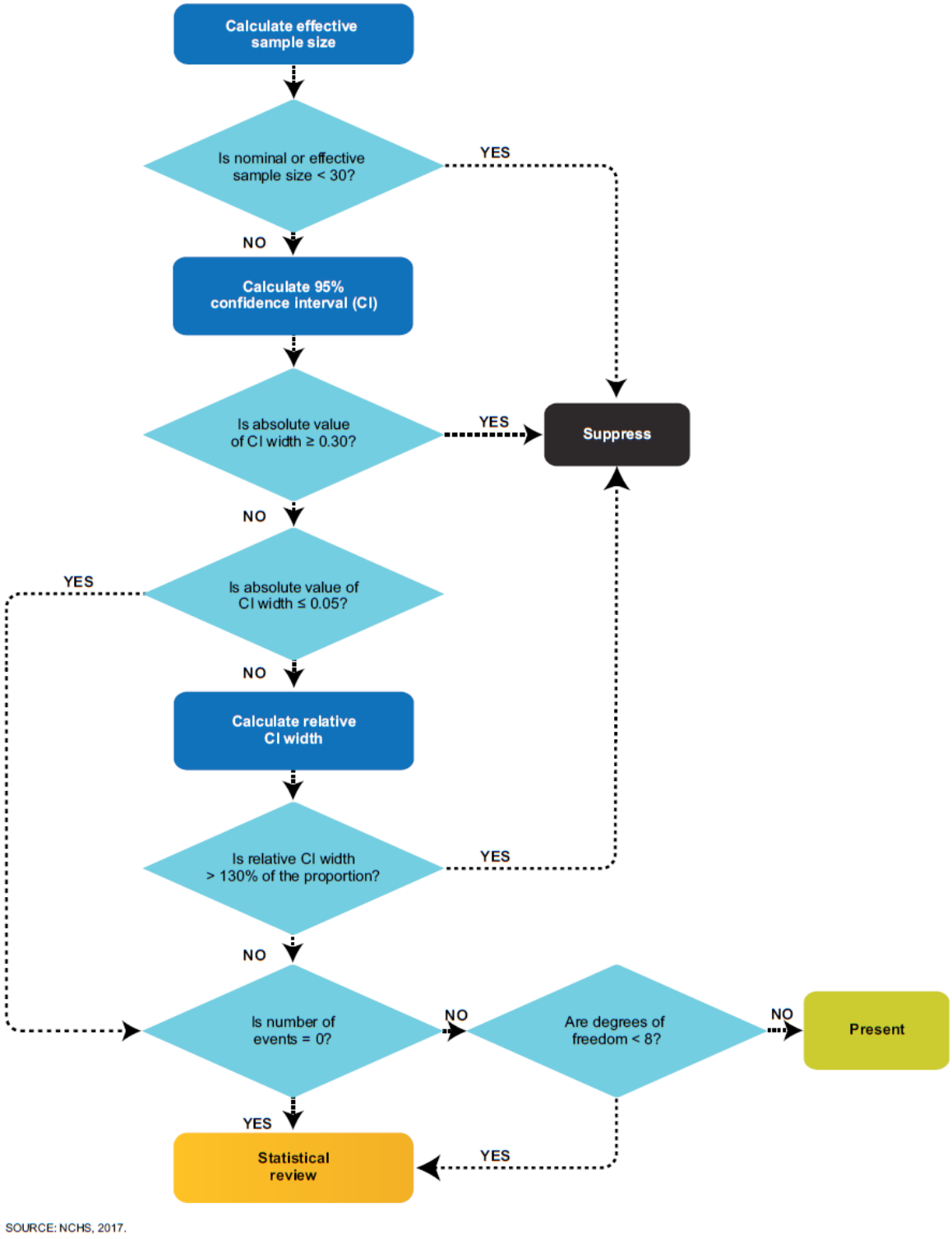


Figure 6. NCHS Suppression Standards[25]

[25] Parker et al., 2017

A.11 Variance

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

N_h = the number of youth and teens in stratum h ($h = 1, \dots, L$), where stratum refers to state-GEO3

Y_{hi} = the value of Y for youth or teen i of stratum h (often the possible values of Y are 0 and 1, as when Y indicates whether a youth or teen is or in a specified BMI percentile)

d_{hi} = 0 or 1, indicating whether youth or teen 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 youth and teens in stratum h

W_{hi} = the sampling weight for youth and teens i in stratum h

Y'_{hi} = the value of Y for youth and teen i in stratum h

d'_{hi} = the value of the domain indicator for youth and teen 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 youth and teens within stratum h, whereas for Y_{hi} and d_{hi} they refer to youth and teens 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.[26] 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.12 BMI Category

BMI category prevalence is calculated from a patient's BMI. EHR data included for analysis should have at most one BMI percentile assigned to each patient within a calendar year. If multiple BMIs are recorded in a single year, selecting a single BMI should be done at the user's discretion. Percentiles are based on a patient's age-sex BMI percentile value. Based on the 2000 CDC BMI-for-age growth charts, the BMI categories are defined as follows:

1. **Underweight**: BMI less than 5th percentile
2. **Healthy Weight**: BMI 5th percentile to less than the 85th percentile
3. **Overweight**: BMI 85th to less than the 95th percentile
4. **Obesity**[27]: BMI equal to or greater than 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.13 ZCTA-3

A ZCTA is a statistical geographic entity that approximates the delivery area for a U.S. Postal Service five-digit (ZCTA) ZIP code. ZCTAs are aggregations of census blocks that have the same predominant ZIP code associated with the residential mailing addresses in the U.S. Census Bureau's Master Address File. ZCTAs do not precisely depict ZIP code delivery areas, and do not include all ZIP codes used for mail delivery. The U.S. Census Bureau has established ZCTAs as a new geographic entity similar to, but replacing, data tabulations for ZIP codes undertaken in conjunction with the 1990 and earlier censuses. For more information, refer to census.gov.[28]

A ZCTA-3 includes the first three digits of a five-digit ZCTA. Three-digit ZCTAs (ZCTA-3), representing the first three digits of a ZIP code, were generated from the AEMR and ACS data.

[26] Wolter, 2007.
[27] Note: prevalence of obesity will include two categories: those that are category 4 and 4a.
[28] https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

Once ZCTA's are aggregated as ZCTA-3's, then the first three digits of a residential ZIP code is equivalent to a ZCTA-3 in over 99% of the population.

A.14 Limitations

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

- Representativeness of CODI-PQ results – CODI-PQ 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.
- Random missingness of plausible height or weight – CODI-PQ 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-PQ 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 patient. Multiple-imputation of race is not employed in CODI-PQ 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 respondents 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.
- Probabilistic 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).
- Measurement error – Height and weight measurement protocols may differ between medical providers, even with clear protocols aimed to increase consistency between medical professionals,[29] leading to potential measurement error. Additionally, height

[29] Best & Shepherd, 2020.

and weight values in EHR data are subject to data entry errors or software glitches. All CODI-PQ 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 patient-level records (encounters) could result in unstable results and reflect poor EHR coverage, a small underlying population, and/or a rare encounter. CODI-PQ suppresses results based on published small sample guidelines using the National Center for Health Statistics Data Presentation Standards for Proportions[30].

[30] Parker JD, Talih M, Malec DJ, et al, 2017.

Appendix B Social Determinants of Health

Data from CODI and AEMR are not a random sample, making adjusting the non-probability sample on observed differences between the sample and the target population beneficial. Therefore, SDOH will be used to statistically weight those non-probability samples. To identify the best SDOH for weighting samples of U.S. youth and teens aged 2-19 years, the Health FFRDC identified and prioritized SDOH associated with youth and teen BMI percentile in the U.S. and by subpopulations within the U.S., including age, race, sex, and sub-geography.

B.1 Prioritizing Social Determinants of Health

Thirty-six SDOH concepts were considered for analysis. The list of concepts was a result of subject matter expert brainstorm sessions in 2019 between Health FFRDC and CDC staff. All concepts readily available at the geography of interest were included in the analysis. Table 18 provides a list of concepts included in the SDOH analysis. Table 19 outlines additional concepts considered for inclusion that did not meet our final criteria for inclusion.

Table 18. Included Concepts: Social Determinants of Health

Approved Concepts	Source(s) with Concept Available
Age	CODI, AEMR, ACS
Region	Census, U.S. Department of Health and Human Services
Education	ACS
Employment/Unemployment	ACS
Ethnicity	AEMR, ACS
Head-of-household	ACS
Housing	ACS
Housing Stability	ACS
Income	ACS
Income Inequality	ACS
Poverty	ACS
Primary Language	ACS
Medical Insurance	CODI, ACS
Race	CODI, AEMR, ACS
Sex	CODI, AEMR, ACS
Transportation	ACS
Adult BMI Percentile (underweight, healthy weight, overweight, obesity, severely obesity)	AEMR

Table 19. Additional Concepts Considered: Social Determinants of Health

Additional Concepts Considered
Pediatric Medical Complexity Algorithm
Primary language
Urbanicity
Walkability
Attention Deficit Hyperactivity Disorder (ADHD)
Air quality
Asthma
Deprivation index
Domestic violence
Food desert/swamp
Home value
Incarceration history
Material security
Migrant/seasonal work
Park access
Refugee status
Safety
Social integration and support
Stress/social vulnerability index
Veteran status

B.2 From Concepts to Measures

After SDOH concepts were identified, each data source was reviewed for measures relating to the concept. A comprehensive list of measures was thus identified for analysis. Table 19 provides a full list of measures included in the analysis. In this section we share the methodology and results for both the bivariate and multivariable analyses performed for the SDOH prioritization. Within each methodology section, we share the models, data source used for the analysis, universe (population included in sample), covariate(s) and dependent variable(s) included in the analysis, goodness of fit requirements, and how the results are reported.

B.3 Bivariate Analysis

There are multiple ways to characterize some SDOH concepts, leading to multiple candidate measures[31] of a SDOH concept. Up to six possible measures were compared to one another within each SDOH concept. The best measure within each concept was determined using a bivariate analysis. The measure within each concept that best explains the outcome (youth and

[31]Within this report, the term measurement means the process by which we describe and ascribe meaning to the key concepts we are investigating. At its core, measurement is about defining one's terms in as clear and precise a way as possible.

teen BMI percentile) was considered to outrank other measures. Once the best measure was determined, the concepts were also ranked against one another. These results generate a ranking of measures and concepts without taking other factors into consideration.

B.3.1 SDOH Prioritization Bivariate Methodology

We performed logistic regressions for each approved concept with multiple plausible specifications (measures). The specifications included:

Models: Cumulative Logit and Multinomial Logit

When there is a natural ordering to the response categories, a cumulative logit model is appropriate. This logit models the cumulative probability that a response (Y) can be classified at or below a given category. In its simplest form, it assumes that the same proportionality constant applies for each parameter to each cumulative level.

The cumulative logit model has the following form:

$$logit[P(Y \leq j)] = \alpha_j + \beta x, where\ j = 0\ to\ J-1$$

$$logit[P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = \log\left[\frac{\pi_0 + \cdots \pi_j}{\pi_{j+1} + \cdots \pi_j}\right]\ where\ j = 0\ to\ J-1$$

Although cumulative logit models rank the dependent variable, a multinomial logit model was also fit. Multinomial logit models are an extension of logistic regression models in which the response can be classified into three or more unordered categories. In these analyses, the response variable is classified into five categories. The multinomial model was fit in addition to or in place of the cumulative logit model for three reasons:

- One or more assumptions were breached thus making a multinomial model the next most appropriate model form.
- The multinomial model took considerably more computing time.
- The rank order of the models was identical in all instances where both models were fit.

The multinomial logit model has the following form. For K classes multinomial problem where labels ranged from [0, K-1], we can generalize it via:

$$log\frac{P(y=1|\beta x, \overline{w})}{P(y=0|\beta x, \overline{w})} = \beta x, \overline{w}_1 \qquad P(y=0|\beta x, \overline{w}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\beta x, \overline{w}_1)}$$

$$log\frac{P(y=2|\overrightarrow{x}, \overline{w})}{P(y=0|\overrightarrow{x}, \overline{w})} = \beta x, \overline{w}_2 \quad \Rightarrow \quad P(y=1|\beta x, \overline{w}) = \frac{\exp(\beta x, \overline{w}_2)}{1 + \sum_{i=1}^{K-1} \exp(\beta x, \overline{w}_2)}, where\ k$$

$$\ldots \qquad \ldots$$

$$log\frac{P(y=k|\beta x, \overline{w})}{P(y=0|\beta x, \overline{w})} = \beta x, \overline{w}_k \qquad P(y=k|\beta x, \overline{w}) = \frac{\exp(\beta x, \overline{w}_k)}{1 + \sum_{i=1}^{k} \exp(\beta x, \overline{w}_i)}$$

$$= 0\ to\ K-1\ and\ \overline{w} = (\overrightarrow{w}_1, \overrightarrow{w}_2, \ldots, \overrightarrow{w}_k)^T$$

$$becomes\ a\ (K-1)(N+1)\ matrix, where\ N\ is\ the\ number\ of\ features$$

Data Source: 1.8 million records from the 2016 AEMR linked to the 2018 ACS by ZIP Code Tabulation Area (3 digit) as defined by the Census Bureau (i.e., ZCTA-3), plus confirmation file from the 2018 AEMR linked to the 2018 ACS by ZCTA-3.

Universe: youth and teens, aged 2 to 19.

Covariate: one approved measure within a concept.

Dependent variables: weight classification (underweight, healthy weight, overweight, obesity, severe obesity).

Goodness of fit: McFadden pseudo R-squared. We used pseudo R-squared measures for evaluating "goodness of fit" in regression models with categorical dependent variables. Log-likelihood-based pseudo R-squared represent the improvement in model likelihood over a null model.

Reporting: Within each concept, we ranked measures from those with the highest pseudo R-squared to those with the lowest pseudo R-squared. Across concepts, we selected the measure with the highest pseudo R-squared.

B.3.2 Results

Table 20 provides a ranked list of concepts. Age of the patient from AEMR ranked the most predictive followed by adult BMI percentile, educational attainment of population aged 45 to 64 years of age, income of a single parent, race, and medical insurance.

The statistical weights cell formulation is based on the Childhood Obesity Data Initiative – Social Determinants of Health (SDOH) Prioritization report[32] as well as other artifacts. The SDOH report was developed in partnership between the Health FFRDC and CDC, with feedback from leading health researchers. Table 20 includes the finalized list of BMI percentile related social determinants of health used in CODI-PQ.

Table 20. Final List of Prioritized Social Determinants of Health

Concept	Measure	Source
Age	Age of patient, continuous	EHR
Race	Patient race, categorical	EHR
Sex	Patient sex, categorical	EHR
Education	Educational attainment of population aged 25 to 64 years of age	ACS

B.4 Multivariate Analysis

Seventeen concepts were calculated and ranked for inclusion in the CODI weighting program through bivariate analyses. The top twelve concepts from those bivariate analyses were further calculated in combination using a multivariable method to determine a possible mix of concepts to include in the weighting programs.

[32] The CODI SDOH Prioritization report is available at https://github.com/NORC-UChicago/CODI-PQ.

B.4.1 SDOH Prioritization Multivariate Methodology

The bivariate analyses ranked concepts against one another but not in combination. A multivariable analysis was needed to determine the best set of SDOH as they relate to youth and teen BMI percentile. Regressions were performed as follows:

Model: Stepwise Cumulative Logit and Multicategory Logit. With stepwise regression, effects are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates if no further effect can be added to the model or if the current model is identical to a previously calculated model.

Data Source: 1.8 million records from the 2016 AEMR linked to the 2018 ACS by ZCTA-3, plus confirmation file from the 2018 AEMR linked to the 2018 ACS by ZCTA-3.

Universe: youth or teen, aged 2 to 19.

Covariates: the top 12 ranked concepts from the bivariate analysis.

Dependent variables: weight classification (underweight, healthy weight, overweight, obesity, severely obesity).

Reporting: Across models, we selected the concepts based on the order in which they are included in the final model. We removed concepts or updated the measure within each concept as appropriate and reported on the updated order.

B.4.2 Results

Three multivariable models were run. In model 1, all concepts were ranked. Results indicate age remained most predictive, followed by race, adult BMI percentile, sex, medical insurance, head-of-household, income, and primary language. Adult BMI percentile ranked third amongst all concepts, but the adult's data source is not nationally representative because it is also calculated from the non-probability sample of AEMR. As such, inclusion of those data is not possible without applying appropriate statistical weights.

Model 2 excluded the adult BMI percentile. Results indicate age remained most predictive, followed by race, education, sex, primary language, transportation, and income.

Model 3 assumed age, race, and sex were required for subpopulation results and excluded the adult BMI percentile. Results indicate education was next most predictive followed by primary language, transportation, head-of-household, and income.

B.5 Summary

Model 3 in Table 23 represents the final ranking of concepts. Within model 3, the educational attainment of the population aged 45 to 64 was found to be most predictive of youth and teen BMI percentile. Although this is the most predictive measure within the concept, this education measure may have limited impact in areas with a small number of persons in this age range. For this reason, a final model was calculated with educational attainment of the population aged 25 to 64, without the inclusion of the adult BMI percentile.

Once the new education measure was found to be equally relevant, a final regression was performed using 2018 AEMR data. The order in which measures were added to the model remained consistent, confirming our findings. The final prioritization is in Table 21.

Table 21. Social Determinants of Health: Summary of Findings

Rank	Concept	Measure	Source
1	Age	Age of patient, continuous	AEMR
1	Race	Patient race, categorical	AEMR
1	Sex	Patient sex, categorical	AEMR
2	Education	Educational attainment of population aged 25 to 64 years of age	ACS
3	Primary Language	Language spoken at home, population aged 5 plus	ACS
4	Transportation	Means of transportation to work	ACS
5	Head-of-household	Percent of households in geography by head of household (female with children, male with children, married with children, grandparent with children)	ACS
6	Income	Family income, with kids less than 18 years of age (categorical)	ACS
7	Poverty	Percent of population who receive supplemental security income (SSI), cash public assistance income, or food stamps/snap in the past 12 months	ACS
8	Employment/Unemployment	Percent of population that is unemployed, Categorical	ACS
9	Medical Insurance	Health insurance coverage status of population	ACS

B.6 Conclusion

Bivariate and multivariable analyses provided a systematic method to assess SDOH against youth and teen BMI percentile in the U.S. AEMR's nationwide data files generate a large number of real-world values of BMI percentile after adjusting for observed differences in demographics relative to the American Community Survey by state and ZCTA-3. The systematic analysis was limited in its knowledge of the data sources available for weighting. Should additional nationally representative files become available, the results could change with the inclusion of different concepts, measures, or sources. For example, the AEMR adult BMI percentile data is a non-probability-based sample and thus requires statistical weights be applied before its use.

B.7 Limitations, SDOH Prioritization

The limitations of this analysis include the following identified items as well as others. First, our comprehensive list of SDOH was compiled from a team of subject matter experts in 2019. Future work may consider comparing other concepts to further strengthen the findings or may consider innovative concepts (e.g., use of a phenotype or household BMI percentile).

Second, inclusion of possible SDOH was limited based on readily available data sources, concepts, and measures within those sources. If other SDOH had been available or included in the analyses, the results may have differed. Future work may consider comparing other data sources, concepts, and measures.

Third, our analysis relied on logistic regression, a statistical model commonly used to rank measures against a dependent variable. Other analytic tools are available including use of data science tools such as a random forest model. The results may be consistent with the multinomial logit analysis or may show additional nuances not possible to see with the current analysis. Random Forests are another way to extract information from a set of data. The appeals of this type of model are:

- It does not assume that the model has a linear relationship like regression models do.
- It utilizes ensemble learning. If we were to use just 1 decision tree, we would not be using ensemble learning. A random forest takes random samples, forms many decision trees, and then averages out the leaf nodes to get a clearer model.
- When fine-tuned through careful statistical analysis, it is highly likely to perform better than logistic regression.

As such, random forest models are more time consuming. Logistic regressions tend to be as or more effective than random forests if time to tune is not available so long as the number of covariates is small in comparison to the number of records.

Fourth, analysis of adult data has revealed a strong correlation between the number of medical encounters and the percentage of the population with BMI percentile. A similar trend can also be seen in youth and teen data, although additional research is needed.

B.8 Future Research Opportunities

Additional research is warranted, with the following suggestions:

- Include a nationally representative adult BMI percentile file
- Explore how SDOH change as the patient ages
- Explore SDOH of the adult population and the household
- Explore additional data sources when available
- Explore a random forest model or other functional form

B.9 Supplemental Tables

Table 22. Additional Concepts and Measures Considered: Social Determinants of Health

Concept	Measure	Source (if available)
Environmental	Air quality	
Food desert/swamp	Food access research atlas	https://www.ers.usda.gov/data-products/food-access-research-atlas/ https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data/
Food desert/swamp	Food security	https://www.ers.usda.gov/webdocs/DataFiles/50764/techdoc2018.pdf?v=7286.8 https://thedataweb.rm.census.gov/ftp/cps_ftp.html#cpssupps
Food desert/swamp	Food environment atlas	https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/
Health	ADHD	
Health	Asthma	
Income inequality	U.S. Census Bureau	https://www.census.gov/topics/income-poverty/income-inequality/about/metrics.html
Income inequality	Deprivation index	https://www.neighborhoodatlas.medicine.wisc.edu/
Income inequality	Theil index	
Income inequality	Atkinson index	
Income inequality	Mean Log Deviation (MLD)	
Safety	Uniform Crime Reporting Program	https://www.fbi.gov/services/cjis/ucr/
Safety	Motor vehicle crashes	https://one.nhtsa.gov/Data/Fatality-Analysis-Reporting-System-(FARS)
Safety	Domestic violence	
Walkability	Walk Score	https://www.walkscore.com/cities-and-neighborhoods/ https://www.walkscore.com/methodology.shtml
Walkability	National walkability index	https://catalog.data.gov/dataset/walkability-index https://www.epa.gov/sites/production/files/2014-03/documents/sld_userguide.pdf https://catalog.data.gov/harvest/object/4f569c8f-4f9a-4d00-bf2b-e242296be0ce/html
Walkability	Park access	National Environmental Public Health Tracking Network (NEPHTN) ephtracking.cdc.gov
Wealth	Home value	

Table 23. Social Determinants of Health Measures

Concept	Measure	Source
Age	Age of patient, categorical	AEMR
Age	Age of population in geography, categorical	ACS
Age	Age of patient, continuous	AEMR
Region	Census division (9)	Census

Concept	Measure	Source
Region	Census region (4)	Census
Region	Health and Human Services region (10)	U.S. Department of Health and Human Services
Education	Educational attainment of population aged 18 to 24 years of age	ACS
Education	Educational attainment of population aged 25 to 34 years of age	ACS
Education	Educational attainment of population aged 35 to 44 years of age	ACS
Education	Educational attainment of population aged 45 to 64 years of age	ACS
Education	Educational attainment of population aged 18 years of age or older	ACS
Education	Educational attainment of adult females	ACS
Education	Educational attainment of adult males	ACS
Employment/Unemployment	Percent of population that is unemployed, categorical	ACS
Ethnicity	Percentage of youth or teen in geography who are of Hispanic Origin	ACS
Head-of-household	Percent of households in geography by head of household (female with children, male with children, married with children, grandparent with children)	ACS
Housing	Tenure of occupied housing (rent, own)	ACS
Housing Stability	Percent of households not living in the same house a year ago	ACS
Income	Family income, with kids less than 18 years of age (categorical)	ACS
Income	Median family income	ACS
Income	Median household income	ACS
Income	Married family income, with kids less than 18 years of age (categorical)	ACS
Income	Income of a single parent (categorical)	ACS
Income Inequality	Gini Index	ACS
Poverty	Percent of population who receive supplemental security income (SSI), cash public assistance income, or food stamps/snap in the past 12 months	ACS
Poverty	Percent of family households living in poverty	ACS
Poverty	Percent of population living in poverty	ACS
Poverty	Percent of children living in poverty, categorized by age of children	ACS
Poverty	Percent of children living in poverty	ACS
Poverty	Percent of population living in poverty, by sex	ACS
Primary Language	Language spoken at home, population aged 5+	ACS
Medical Insurance	Health Insurance coverage status of population	ACS
Race	ACS race distribution	ACS
Race	Race, geographic distribution	AEMR
Race	Patient race, categorical	AEMR
Sex	Patient sex, categorical	AEMR
Sex	Sex, geographic distribution	AEMR
Sex	Sex of youth or teens, geographic distribution	ACS
Transportation	Means of transportation to work	ACS
Adult BMI percentile	Adult BMI percentile, categorical	AEMR

Table 24. Social Determinants of Health: Bivariate Ranked Concepts and Measures

Rank	Concept	Measure	Source	Measure Rank
1	Age	Age of patient, categorical	AEMR	1
1	Age	Age of patient, continuous	AEMR	2
1	Age	Age of population in geography, categorical	ACS	3
2	Adult BMI percentile	Adult BMI percentile, categorical	AEMR	1
3	Education	Educational attainment of population aged 45 to 64 years of age	ACS	1
3	Education	Educational attainment of adult females	ACS	2
3	Education	Educational attainment of adult males	ACS	3
3	Education	Educational attainment of population aged 35 to 44 years of age	ACS	4
3	Education	Educational attainment of population aged 18 years of age or older	ACS	5
3	Education	Educational attainment of population aged 25 to 34 years of age	ACS	6
3	Education	Educational attainment of population aged 18 to 24 years of age	ACS	7
4	Income	Income of a single parent (categorical)	ACS	1
4	Income	Family income, with kids less than 18 years of age (categorical)	ACS	2
4	Income	Married family income, with kids less than 18 years of age (categorical)	ACS	3
4	Income	Median family income	ACS	4
4	Income	Median household income	ACS	5
5	Race	Patient race, categorical	AEMR	1
5	Race	ACS race distribution	ACS	2
5	Race	Race, geographic distribution	AEMR	3
6	Medical Insurance	Health insurance coverage status of population	ACS	1
7	Poverty	Percent of children living in poverty, categorized by age of children	ACS	1
7	Poverty	Percent of population living in poverty, by sex	ACS	2
7	Poverty	Percent of children living in poverty	ACS	3
7	Poverty	Percent of population living in poverty	ACS	4
7	Poverty	Percent of family households living in poverty	ACS	5
7	Poverty	Percent of population who receive supplemental security income (SSI), cash public assistance income, or food stamps/snap in the past 12 months	ACS	6
8	Transportation	Means of transportation to work	ACS	1
9	Head-of-household	Percent of households in geography by head of household (female with children, male with children, married with children, grandparent with children)	ACS	1
10	Primary Language	Language spoken at home, population aged 5+	ACS	1
11	Employment/Unemployment	Percent of population that is unemployed, categorical	ACS	1
12	Sex	Patient sex, categorical	AEMR	1
12	Sex	Sex, geographic distribution	AEMR	2
12	Sex	Sex of youth or teen, geographic distribution	ACS	3

Rank	Concept	Measure	Source	Measure Rank
13	Ethnicity	Percentage of youth or teen in geography who are of Hispanic Origin	ACS	1
14	Region	Health and Human Services region (10)	U.S. Department of Health and Human Services	1
14	Region	Census region (4)	Census	2
14	Region	Census division (9)	Census	3
15	Housing Stability	Percent of households not living in the same house a year ago	ACS	1
16	Housing	Tenure of occupied housing (rent, own)	ACS	1
17	Income Inequality	Gini Index	ACS	1

Table 25. Social Determinants of Health: Multivariable Ranked Measures

Model	Rank	Concept	Comments
1	1	Age	
1	2	Race	
1	3	Adult BMI percentile	
1	4	Sex	
1	5	Medical Insurance	
1	6	Head-of-household	
1	7	Income	
1	8	Primary Language	
1	9	Transportation	
1	10	Poverty	
1	11	Employment/Unemployment	
1	12	Education	
2	1	Age	
2	2	Race	
2	3	Education	
2	4	Sex	
2	5	Primary Language	
2	6	Transportation	
2	7	Income	
2	8	Head-of-household	
2	9	Medical Insurance	
2	10	Poverty	
2	NA	Adult BMI percentile	Excluded
2	NA	Employment/Unemployment	Excluded
3	1	Age	
3	1	Race	
3	1	Sex	
3	2	Education	
3	3	Primary Language	
3	4	Transportation	
3	5	Head-of-household	
3	6	Income	
3	7	Poverty	
3	8	Employment/Unemployment	

Model	Rank	Concept	Comments
3	9	Medical Insurance	
3	NA	Adult BMI percentile	Excluded

Appendix C ACS File Layouts

C.1 ACS Input File Layout

The following variables are included in both the ZCTA-3 file as well as the County file. ACS data is imported in the CODI-PQ and require a csv file with the following variable names, possible variable values, and in the order listed below. Variable geoid (option 1) is for ZCTA-3 files only and geoid (option 2) is for the County file only.

Table 26. ACS Input File Layout, CSV File

Variable Name	Label	Description	Format	Example
Geoid	Geoid (option 1)	3 digits ZIP Code Tabulation Areas (ZCTAs) followed by two-letter State Abbreviations	Character	221
Geoid	Geoid (option 2)	3-digit County FIPS Code	Character	059
State_code	State FIPS code	2-digit State Abbreviation	Character	VA
b01001f_001	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Total: Total:	Population count	Number	8199
b01001f_003	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: Under 5 years	Population count	Number	256
b01001f_004	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 5 to 9 years	Population count	Number	246
b01001f_005	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 10 to 14 years	Population count	Number	495
b01001f_006	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 15 to 17 years	Population count	Number	297
b01001f_007	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 18 and 19 years	Population count	Number	145
b01001f_018	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: Under 5 years	Population count	Number	271
b01001f_019	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 5 to 9 years	Population count	Number	188
b01001f_020	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 10 to 14 years	Population count	Number	267

Variable Name	Label	Description	Format	Example
b01001f_021	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 15 to 17 years	Population count	Number	139
b01001f_022	SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 18 and 19 years	Population count	Number	134
b01001e_001	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Total: Total:	Population count	Number	278
b01001e_003	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: Under 5 years	Population count	Number	0
b01001e_004	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 5 to 9 years	Population count	Number	28
b01001e_005	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 10 to 14 years	Population count	Number	28
b01001e_006	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 15 to 17 years	Population count	Number	0
b01001e_007	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 18 and 19 years	Population count	Number	0
b01001e_018	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: Under 5 years	Population count	Number	0
b01001e_019	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 5 to 9 years	Population count	Number	0
b01001e_020	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 10 to 14 years	Population count	Number	0
b01001e_021	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 15 to 17 years	Population count	Number	0
b01001e_022	SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 18 and 19 years	Population count	Number	10
b01001d_001	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Total: Total:	Population count	Number	14824

Variable Name	Label	Description	Format	Example
b01001d_003	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: Under 5 years	Population count	Number	277
b01001d_004	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 5 to 9 years	Population count	Number	222
b01001d_005	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 10 to 14 years	Population count	Number	276
b01001d_006	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 15 to 17 years	Population count	Number	263
b01001d_007	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 18 and 19 years	Population count	Number	774
b01001d_018	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: Under 5 years	Population count	Number	101
b01001d_019	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 5 to 9 years	Population count	Number	237
b01001d_020	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 10 to 14 years	Population count	Number	355
b01001d_021	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 15 to 17 years	Population count	Number	242
b01001d_022	SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 18 and 19 years	Population count	Number	1404
b01001c_001	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Total: Total:	Population count	Number	755
b01001c_003	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: Under 5 years	Population count	Number	36
b01001c_004	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 5 to 9 years	Population count	Number	44
b01001c_005	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 10 to 14 years	Population count	Number	3
b01001c_006	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 15 to 17 years	Population count	Number	22
b01001c_007	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 18 and 19 years	Population count	Number	37
b01001c_018	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: Under 5 years	Population count	Number	0

Variable Name	Label	Description	Format	Example
b01001c_019	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 5 to 9 years	Population count	Number	14
b01001c_020	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 10 to 14 years	Population count	Number	11
b01001c_021	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 15 to 17 years	Population count	Number	9
b01001c_022	SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 18 and 19 years	Population count	Number	26
b01001b_001	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Total: Total:	Population count	Number	13407
b01001b_003	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: Under 5 years	Population count	Number	283
b01001b_004	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 5 to 9 years	Population count	Number	222
b01001b_005	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 10 to 14 years	Population count	Number	425
b01001b_006	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 15 to 17 years	Population count	Number	439
b01001b_007	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 18 and 19 years	Population count	Number	429
b01001b_018	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: Under 5 years	Population count	Number	485
b01001b_019	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 5 to 9 years	Population count	Number	254
b01001b_020	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 10 to 14 years	Population count	Number	359
b01001b_021	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 15 to 17 years	Population count	Number	189
b01001b_022	SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 18 and 19 years	Population count	Number	561
b01001a_001	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Total: Total:	Population count	Number	426175

Variable Name	Label	Description	Format	Example
b01001a_003	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: Under 5 years	Population count	Number	9461
b01001a_004	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 5 to 9 years	Population count	Number	9446
b01001a_005	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 10 to 14 years	Population count	Number	11409
b01001a_006	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 15 to 17 years	Population count	Number	7231
b01001a_007	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 18 and 19 years	Population count	Number	8352
b01001a_018	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: Under 5 years	Population count	Number	8754
b01001a_019	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 5 to 9 years	Population count	Number	10226
b01001a_020	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 10 to 14 years	Population count	Number	10640
b01001a_021	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 15 to 17 years	Population count	Number	7731
b01001a_022	SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 18 and 19 years	Population count	Number	9368
b01001g_001	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Total: Total:	Population count	Number	12014
b01001g_003	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: Under 5 years	Population count	Number	732
b01001g_004	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 5 to 9 years	Population count	Number	659
b01001g_005	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 10 to 14 years	Population count	Number	823
b01001g_006	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 15 to 17 years	Population count	Number	491
b01001g_007	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 18 and 19 years	Population count	Number	501
b01001g_018	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: Under 5 years	Population count	Number	683
b01001g_019	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 5 to 9 years	Population count	Number	650
b01001g_020	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 10 to 14 years	Population count	Number	652
b01001g_021	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 15 to 17 years	Population count	Number	410

Variable Name	Label	Description	Format	Example
b01001g_022	SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 18 and 19 years	Population count	Number	651
b03002_012	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Hispanic or Latino:	Population count	Number	56886
b03002_013	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: White alone	Population count	Number	43689
b03002_014	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Black or African American alone	Population count	Number	1753
b03002_015	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: American Indian and Alaska Native alone	Population count	Number	196
b03002_016	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Asian alone	Population count	Number	253
b03002_017	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone	Population count	Number	150
b03002_018	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Some other race alone	Population count	Number	7913
b03002_019	HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Two or more races: Two or more races:	Population count	Number	2932
b15001_011	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: 25 to 34 years:	Population count	Number	28989
b15001_017	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Bachelor's degree	Population count	Number	7476
b15001_018	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Graduate or professional degree	Population count	Number	2604
b15001_019	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: 35 to 44 years:	Population count	Number	24797
b15001_025	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Bachelor's degree	Population count	Number	4891
b15001_026	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Graduate or professional degree	Population count	Number	3258

Variable Name	Label	Description	Format	Example
b15001_027	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: 45 to 64 years:	Population count	Number	62253
b15001_033	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Bachelor's degree	Population count	Number	11482
b15001_034	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Graduate or professional degree	Population count	Number	8394
b15001_058	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Bachelor's degree	Population count	Number	8124
b15001_059	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Graduate or professional degree	Population count	Number	5237
b15001_060	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: 35 to 44 years:	Population count	Number	26186
b15001_066	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Bachelor's degree	Population count	Number	6646
b15001_067	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Graduate or professional degree	Population count	Number	5657
b15001_068	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: 45 to 64 years:	Population count	Number	67764
b15001_074	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Bachelor's degree	Population count	Number	14062
b15001_075	SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Graduate or professional degree	Population count	Number	11986

C.2 ACS for Use with GEO3 Data

Table 27. ACS Pre-Processing Results File Layout – GEO3[33]

Variable Name	Format	Length
Geography	Character	5
GEO3	Character	3
State_FIPS	Character	2
TOTAL_ACS_POPULATION	Number	8
AGE_L5_MALE_WHITE	Number	8
AGE_5_9_MALE_WHITE	Number	8
AGE_10_14_MALE_WHITE	Number	8
AGE_15_17_MALE_WHITE	Number	8
AGE_18_19_MALE_WHITE	Number	8
AGE_L5_FEMALE_WHITE	Number	8
AGE_5_9_FEMALE_WHITE	Number	8
AGE_10_14_FEMALE_WHITE	Number	8
AGE_15_17_FEMALE_WHITE	Number	8
AGE_18_19_FEMALE_WHITE	Number	8
AGE_L5_MALE_BLACK	Number	8
AGE_5_9_MALE_BLACK	Number	8
AGE_10_14_MALE_BLACK	Number	8
AGE_15_17_MALE_BLACK	Number	8
AGE_18_19_MALE_BLACK	Number	8
AGE_L5_FEMALE_BLACK	Number	8
AGE_5_9_FEMALE_BLACK	Number	8
AGE_10_14_FEMALE_BLACK	Number	8
AGE_15_17_FEMALE_BLACK	Number	8
AGE_18_19_FEMALE_BLACK	Number	8
AGE_L5_MALE_ASIAN	Number	8
AGE_5_9_MALE_ASIAN	Number	8
AGE_10_14_MALE_ASIAN	Number	8
AGE_15_17_MALE_ASIAN	Number	8
AGE_18_19_MALE_ASIAN	Number	8
AGE_L5_FEMALE_ASIAN	Number	8
AGE_5_9_FEMALE_ASIAN	Number	8
AGE_10_14_FEMALE_ASIAN	Number	8
AGE_15_17_FEMALE_ASIAN	Number	8

[33] The variables are needed at a minimum for the CODI-PQ results. Actual ACS file may include additional variables.

Variable Name	Format	Length
AGE_18_19_FEMALE_ASIAN	Number	8
AGE_L5_MALE_OTHER	Number	8
AGE_5_9_MALE_OTHER	Number	8
AGE_10_14_MALE_OTHER	Number	8
AGE_15_17_MALE_OTHER	Number	8
AGE_18_19_MALE_OTHER	Number	8
AGE_L5_FEMALE_OTHER	Number	8
AGE_5_9_FEMALE_OTHER	Number	8
AGE_10_14_FEMALE_OTHER	Number	8
AGE_15_17_FEMALE_OTHER	Number	8
AGE_18_19_FEMALE_OTHER	Number	8
AGE_25_64_BACH_GRAD	Number	8
AGE_25_64_BACH_GRAD_GTR20PERC	Number	8
TOTAL_LATIN	Number	8
LATIN_WHITE	Number	8
LATIN_BLACK	Number	8
LATIN_ASIAN	Number	8
LATIN_OTHER	Number	8

Appendix D EHR File Layouts

D.1 EHR Input File Layout

D.1.1 EHR GEO3 Data

User-provided EHR data are imported in CODI-PQ and require a csv file with the following variable names, variable value options (case sensitive), and in the order listed below.

Table 28. EHR Input File Layout for GEO3-Level Programs, CSV File[34]

[34] Only one record per patient per year is allowed. A patient may be included multiple times if multiple years are included in the input data file.

Variable Name	Description	Format	Valid values	Example
SUBJID	Patient Identifier	Character	Character value of maximum length 25	123456789
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 data approximations)	11
RACE_ETH	Patient's race if known or ethnicity when race is not known	Character	"AFRICAN AMERICAN" "Black" "ASIAN" "Asian" "CAUCASIAN" "White" "HISPANIC" "Hispanic" "OTHER" "Other" "UNKNOWN" "Unknown"	CAUCASIAN
STATE_ABR	Patient's residential state, two-letter state abbreviations	Character	any postal abbreviation of state (see appendix G for possible values)	VA
GEO3	Either: Patient's residential a) county code (5 digits) or b) ZIP-3 (3 digits)	Number	Any numeric value	221
WEIGHT_CATEGORY	Patient's BMI percentile. See section A.2 and A.12 for more information.	Character	Normal or Healthy Weight Obese Overweight Severe Obesity Underweight	Overweight
YEAR	Year of the medical encounter	Number	Yyyy	2018
SCDCNT	Sickle-Cell indicator	Number	Program treats patients with a count of 1 or higher as having sickle cell disease. If sickle cell disease information is not available, set this value to blank or zero.	2
PREGNANCY_FLAG	Pregnancy flag	Number	0/1 If pregnancy information is not available, set this value to zero or blank.	1
ZIP	5-digit ZIP code	Character	5-digit ZIP code. Required for county level records. Optional for ZCTA-3 level records.	20814

D.2 EHR Results File Layout for GEO3

Table 29. EHR Pre-Processing Results File Layout – GEO3[35]

Variable Name	Format	Example
SUBJID	Character	12345626
Ageyr	Number	16
AGE_CATEGORIES	Character	15 – 17
WTCAT	Character	(2) Healthy Weight (5th to <85th percentile)
STATE_ALPHA	Character	MI
STATE_FIPS	Character	48
ZIP	Character	20184
GEO3	Character	100
Geography	Character	48100
SCDCNT	Number	0
PREGNANCY_FLAG	Number	1
Race	Character	Black
Sex	Character	Female
Year	Number	2018
SCD	Number	0
Imputed_Race	Character	Black
Race_Imputed	Number	0

[35] The variables are needed at a minimum for the CODI-PQ results. Actual file may include additional variables.

Appendix E CODI-PQ-GEO3 Example SAS Programs

E.1 Data Inputs and Link Population Data (Pre-processing) Quickstart with GEO3 Data

Appendix E.2 includes a program to generate results. This example is for the pre-processing data inputs and link population data step.

Text highlighted in yellow has been reviewed and approved or reviewed and edited from its original values. The program uses the data inputs: ACS_State_ZCTA3 and EHR_filename located in the P:\Example\0_Raw_Data folder. The file processes EHR data between 2016 and 2019 and creates a SAS file named CODI_EHR_READY stored in the folder P:\Example\2_Output\Pre_Processed_CODI_PQ. Both the raw data for ACS and EHR file include ZCTA3, not county codes (COUNTY = N). The SAS log is stored in P:\Example\2_Output\SAS LOG\This_is_ the_Log <plus date and time information>.log. Text between /* and */ are comments in SAS. Comments may vary slightly in CODI-PQ from the example below.

```
/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
/*************************************************************************************************************************************/
/************************************** -- USER SELECTION CRITERA SECTIONS 1 through 4 -- **************************************/
/******************* -- PLEASE UPDATE THE BLACK TEXT AFTER THE EQUAL SIGN (ACCEPTED VALUES LISTED IN SAS NOTE) -- *******************/
/*SECTION 1: Input Folder and file names***/
/***/ %LET ROOT_PRE = P:\Example; /*@Note: base directory (ACCEPTABLE VALUES: computer directory name) ***/
***/
/***/ %LET PRE_DEST = CODI_PQ; /*@Note: Suffix name for EHR Output folder (ACCEPTABLE VALUES: folder name (no punctuation) ***/
/***/ %LET ACS_FILENAME = ACS_State_ZCTA3;  /*@Note: ACS file name (ACCEPTABLE VALUES: file name, do not include ".csv") ***/
/***/ %LET EHR_FILENAME = EHR_filename; /*@Note: EHR file name (ACCEPTABLE VALUES: file name, do not include ".csv") ***/
/***/ %LET LOG_NAME_PRE = This_is_the_Log; /*@Note: SAS log file name prefix ACCEPTABLE VALUES: SAS file name (no punctuation) ***/

/*SECTION 2: Beginning and End Year of longitudinal EHR data
                                                                    ***/
/***/ %LET BEGIN_YEAR = 2016; /*@Note: LONGITUDINAL Start year (ACCEPTABLE VALUES: 4-digit numeric year) ***/
/***/ %LET END_YEAR   = 2019; /*@Note: LONGITUDINAL End year (ACCEPTABLE VALUES: 4-digit numeric year) ***/

/*SECTION 3: OPTIONAL Output File Name Suffix
                                                                                ***/
/***/ %LET EHR_PRE_Out = CODI_EHR_READY; /*@Note: EHR output file name (ACCEPTABLE VALUES: SAS file name (no punctuation) ***/

/*SECTION 4: County or ZCTA3 data (REQUIRED) ***/
/***/ %LET COUNTY=N; /*@Note: County/ZCTA3 indicator (ACCEPTABLE VALUES: Y for County level data, N for ZCTA3 level data ***/
/*************************************************************************************************************************************/
/***Note: Root directory includes subfolders: "..\0_Raw_Data"
                                             "..\1_SAS_Programs"
                                             "..\1_SAS_Programs\Pre_Processing_GEO3"
                                             "..\1_SAS_Programs\CODI_PQ_GEO3"
                                             "..\2_Output" and
                                             "..\2_Output\SAS LOGS" ***/
/***NOTE: SAS programs must be stored in the PROGS directory including: Macro1-CODI_PQ.sas,
```

```
                                                                    Macro2-CODI_PQ.sas,
                                                                    Macro3-CODI_PQ.sas,
                                                                    Macro4-CODI_PQ.sas,
                                                                    Module1-CODI_PQ.sas,
                                                                    Module2-CODI_PQ.sas,
                                                                    Macro1-CODI_PQ-Co_occurring.sas,
                                                                    Macro2-CODI_PQ-Co_occurring.sas,
                                                                    Macro3-CODI_PQ-Co_occurring.sas,
                                                                    Macro4-CODI_PQ-Co_occurring.sas,
                                                                    Module1-CODI_PQ-Co_occurring.sas,
                                                                    Module2-CODI_PQ-Co_occurring.sas, ***/
/*NOTE: query output is stored as a csv file in "..\2_Output" named after a time/date stamp and CODI_Prevalence_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 */
/*****************************************************************************************************************************/
/*****************************************************************************************************************************/
/*****************************************************************************************************************************/
/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
```

E.2 Generate Results Example with GEO3 Data

Appendix E.1 includes a program to generate pre-processed results that are then used in this program.

Text highlighted in yellow has been reviewed and approved or reviewed and edited from its original values. The program uses the data inputs: ACS_State_ZCTA3 and CODI_EHR_READY located in the P:\Example\2_Output\Pre_Processed_CODI_PQ folder.

Subpopulation selected by the user includes: EHR records from 2017 including patients 2 to 14 years of age who are either white or Asian, living in Jefferson County (059) Colorado (FIPS code = 08) or Yuma County (125) 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). Note that Sex_Male and Sex_Female have not been edited since the ALL_SEXES is turned on to yes.

The file processes a SAS file named CODI_EHR_READY stored in the folder P:\Example\2_Output\Pre_Processed_CODI_PQ. The SAS log is stored in P:\Example\2_Output\SAS LOG\CODI_PQ_ZCTA3_PEDS<plus date and time information>.log. Text between /* and */ are comments in SAS. Comments may vary slightly in CODI-PQ from the example below.

Methods: Include imputed race information and calculate the age-adjusted prevalence.

```
/*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_PQ            = P:\Example;  /*@Note: same as in pre-processing quickstart program, base directory (ACCEPTABLE VALUES:
computer directory name)                                  ***/
/***/ %LET PRE_DEST           = CODI_PQ; /*@Note: same as in pre-processing quickstart program, Suffix name for EHR Output folder
(ACCEPTABLE VALUES: folder name (no puctuation)                     ***/
/***/ %LET EHR_PRE_OUT = CODI_EHR_Ready; /*@Note: same as in pre-processing quickstart program, name the pre-processing output file
(ACCEPTABLE VALUES: file name (no punctuations)) ***/
/***/ %LET LOG_NAME            = CODI_PQ_ZCTA3_PEDS; /*@Note: Name for SAS log storage location
                                ***/
/***/ %LET FileOUT_Name     = Results_Here; /*@Note: Output file name; MUST BE FEWER THAN 28 CHARACTERS
                           ***/

/*SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE, STATE/ZCTA3, or STATE/COUNTY
                  ***/
/***/ %LET BEG_YEAR             = 2017; /*@Note: Beginning year of analysis                        (ACCEPTED VALUES: 4-Digit numeric,
2015-2019) ***/
/***/ %LET END_YEAR             = 2017; /*@Note: End year of analysis                                                (ACCEPTED VALUES: 4-
Digit numeric, 2015-2019) ***/
/***/ %LET ALL_STATES   = N;    /*@Note: Include all geographical locations in file?(ACCEPTED VALUES: Y/N)
                  ***/
/***/ %LET ALL_AGES    = N;     /*@Note: Include all age ranges?                                  (ACCEPTED VALUES: Y/N)
                           ***/
/***/ %LET ALL_SEXES    = Y;    /*@Note: Include all sex values?                                  (ACCEPTED VALUES: Y/N)
                           ***/
/***/ %LET ALL_RACES    = N;    /*@Note: Include all race categories?                             (ACCEPTED VALUES: Y/N)
                             ***/

/*SECTION 3: Additional flags
                                                                    ***/
/***/ %LET ACSCOUNTY   = N; /*@Note: Is the ACS data at the county or ZCTA3 level? (ACCEPTABLE VALUES: Y for County level data, N for
ZCTA3 level data)      ***/
/***/ %LET INCLUDE_PREGNANCY = Y; /*@Note: Include pregnant individuals and non-pregnant individuals (Y) or non-pregnant individuals
only (N)                                (ACCEPTED VALUES: Y/N)                                                  ***/
/***/ %LET SAMPLE_CHECK = Y; /*@Note: Run check for sufficient sample size of query      (ACCEPTED VALUES: Y/N)
                  ***/

/*SECTION 4: Only complete section 4 for any "N" values listed in section 2
                                                ***/
/*IF ALLGEOGRAPHIES= N THEN SELECT STATE CODES OR STATE AND COUNTY CODES BELOW:
                                           ***/
       /***  (ACCEPTED VALUES: SINGLE QUOTES SURROUNDING 2 OR 5-Digit CODES w/ "," BETWEEN MULTIPLE SELECTIONS, )
                  ***/
/*IF ALLSTATES = N THEN SELECT ONE OR MORE AGE CATEGORIES BELOW:
                                          ***/
       /***/ %LET GEO_GROUP    = ZCTA3;         /*@Note: Level of geography                                  (ACCEPTED VALUES:
STATE, ZCTA3, COUNTY)             ***/
       /***/ %LET GEO_LIST     = %STR('08059', '08125'); /*@Note: IF GEO_GROUP="STATE" then populate with State FIPS code(s) for
example, 36 is New York, If GEO_GROUP="ZCTA3" then populate with FIPS State+ZCTA3 code(s), ***/
```

```
                                                                                                                         /*If
Geo_Group="COUNTY" then populate with FIPS State+FIPS County code(s) (ACCEPTED VALUES: 2-digit state FIPS or 5-digit state FIPS+ZCTA3
***/
                                                                                                                         /*or
5-digit state FIPS+county FIPS (Must be surrounded by single quotation and comma delimited))***/
/*IF ALL_AGES = N THEN SELECT ONE OR MORE AGE CATEGORIES BELOW:
                                                        ***/
       /***/ %LET WGT_AGE_2_4   = Y;   /*@Note: Age Range: 2 to 4                              (ACCEPTED VALUES: Y/N)
                                  ***/
       /***/ %LET WGT_AGE_5_9   = Y;   /*@Note: Age Range: 5 to 9                              (ACCEPTED VALUES: Y/N)
                               ***/
       /***/ %LET WGT_AGE_10_14  = Y;        /*@Note: Age Range: 10 to 14                                    (ACCEPTED
VALUES: Y/N)                                  ***/
       /***/ %LET WGT_AGE_15_17  = N;        /*@Note: Age Range: 15 to 17                                    (ACCEPTED
VALUES: Y/N)                               ***/
       /***/ %LET WGT_AGE_18_19  = N;        /*@Note: Age Range: 18 to 19                                    (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)                         ***/
/*IF ALL_SEXES = N THEN SELECT MALE OR FEMALE BELOW:
                                                        ***/
       /***/ %LET SEX_MALE      = Y;  /*@Note: Sex: Male                                           (ACCEPTED VALUES: Y/N)
                                    ***/
       /***/ %LET SEX_FEMALE  = Y;    /*@Note: Sex: Female                                         (ACCEPTED VALUES: Y/N)
                                    ***/

/*SECTION 5: Methodological option selections
                                                                            ***/
       /***/ %LET IMP_RACES = Y;      /*@Note: Include imputed race values?                    (ACCEPTED VALUES: Y/N)
                            ***/
       /***/ %LET AGE_ADJ     = Y;    /*@Note: Produce age-adjusted estimates?             (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_PQ.sas,
                                                            Macro2-CODI_PQ.sas,
```

```
                                                    Macro3-CODI_PQ.sas,
                                                    Macro4-CODI_PQ.sas,
                                                    Module1-CODI_PQ.sas,
                                                    Module2-CODI_PQ.sas***/
/***NOTE: query output is stored as a csv file in "..\2_Output" named after a time/date stamp and the designated FileOUT_Name
***/
/***NOTE: to save a SAS file of the weighted observations remove comment from lines 273-276    ***/
/*****************************************************************************************************************************************
/
/***STOP STOP STOP 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    */
/*****************************************************************************************************************************************
/
/*****************************************************************************************************************************************
/
/*****************************************************************************************************************************************
/
```

Appendix F CODI-PQ Results

F.1 Example BMI Percentile Prevalence

Once complete, CODI-PQ generate prevalence results as a csv file. Table 30 provides an overview of the variables included for BMI percentile prevalence, and Table 31 and 34 provide example results. User inputs, error codes, sources of technical documentation, caveats, and a possible citation begins with the row labeled order 3 and continues thereafter. The number of rows and exact wording of text will vary based on the criteria selected.

Table 30. CODI-PQ Results Data Dictionary

Column	Description
Order	Row order
Weight Category	A categorical value based on BMI percentile.
Sample	The observed (or unadjusted, or crude) count of youth and teens in the study population.
Population	The weighted (or adjusted) count of the study population.
Crude Prevalence	The observed (or unadjusted, or crude) prevalence in the study population.
Crude Prevalence Standard Error	The observed (or unadjusted, or crude) standard error in the study population.
Weighted Prevalence	Prevalence based on weighted counts. A sample weight is assigned to each sample patient. It is a measure of the number of youth and teens in the population represented by that sample patient. See implementation guide, Appendix A. Sample Weights for more information.
Weighted Prevalence Standard Error	Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information.
Age-Adjusted Prevalence	Prevalence based on weighted, age-adjusted counts. See implementation guide, Appendix A. Age Adjustment for more information.

Table 31. Results Example from Synthetic Data[36]

Order	Weight Category	Sample	Population	Crude Prevalence	Crude Prevalence Standard Error	Weighted Prevalence	Weighted Prevalence Standard Error	Age-Adjusted Prevalence	Age-Adjusted Prevalence Standard Error
1	(1) Underweight (<5th percentile)	61	5,801	4.16	0.25	4.80	0.82	4.80	0.82

[36] Note: borders and shading are for demonstration purposes only. CSV exports columns separated with a comma. The results can be imported into Excel. Results based on synthetic data. Exact wording in Order 3 through 19 may vary from those shown in the example.

Order	Weight Category	Sample	Population	Crude Prevalence	Crude Prevalence Standard Error	Weighted Prevalence	Weighted Prevalence Standard Error	Age-Adjusted Prevalence	Age-Adjusted Prevalence Standard Error
1	(2) Healthy Weight (5th to <85th percentile)	824	82,020	56.17	1.3	67.88	3.58	67.88	3.58
1	(3) Overweight (85th to <95th percentile)	251	14,752	17.11	0.98	12.21	2.42	12.21	2.42
1	(4) Obesity (>=95th percentile)	331	18,255	22.56	1.09	15.11	2.89	15.11	2.89
1	(4b) Severe Obesity (>=120% of the 95th percentile)	100	6,085	7.52	1.01	5.25	2.22	5.25	2.22
2	Totals:	1,467	120,828						
3	Dataset: AEMR, 2015-2019								
4	Query Parameters: AGE RACE SEX GEOGRAPHY YEAR								
5	AGE: (10 - 14, 15 - 17, 18 - 19)								
6	SEX: (Female)								
7	RACE: (White, Black, Asian, Other)								
8	RACE Suppressed: No.								
9	RACE Imputed: People with unknown race were excluded.								
10	Geography: (08810)								
11	Year: 2019								
12	Weighting cells were collapsed for: (Geography)								
13	AGE adjusted?: (Yes)								
13	Error Codes: (None)								
14	Technical Documentation: See https://github.com/NORC-UChicago/CODI-PQ for more information and full details on data sources and methodologies.								
15	Query Date: Friday, 1 July 2021 4:09:46 PM								
15	Suggested Citation: Tanenbaum, E., Campbell, S., Chelluri, D., Zalsha, S.,								

Order	Weight Category	Sample	Population	Crude Prevalence	Crude Prevalence Standard Error	Weighted Prevalence	Weighted Prevalence Standard Error	Age-Adjusted Prevalence	Age-Adjusted Prevalence Standard Error
	Boim, J., Paddock, S., Copeland, K. (2021). Clinical and Community Data Initiative Prevalence Queries (CODI-PQ) SAS programs (version 2015-2019).								
16	Caveats								
17	Patients with either missing or invalid age, sex, height, weight, or geography are not included in counts and prevalence results.								
18	The method used to calculate the standard errors are documented in the technical documentation.								
19	The population are based on age-race-sex-location specific counts from the 2014-2018 American Community Survey Five-year Estimates released by the Census Bureau on December 19, 2019.								

Table 32. Example Results with Errors (Insufficient Sample Size), Error Messages Are Shown in Order Row 13[37]

[37] Note: borders and shading are for demonstration purposes only. CSV exports columns separated with a comma. The results can be imported into Excel. Results based on synthetic data. Exact wording in Order 3 through 22 may vary from those shown in the example.

Order	Weight Category	Sample	Population	Crude Prevalence	Crude Prevalence Standard Error	Weighted Prevalence	Weighted Prevalence Standard Error	Age-adjusted Prevalence	Age-adjusted Prevalence Standard Error
1	(1) Underweight (<5th percentile)								
1	(2) Healthy Weight (5th to <85th percentile)								
1	(3) Overweight (85th to <95th percentile)								
1	(4) Obesity (>=95th percentile)								
1	(4b) Severe Obesity (>=120% of the 95th percentile)								
2	Totals:								
3	Dataset: IQVIA, 2015-2019								
4	Query Parameters: AGE RACE SEX GEOGRAPHY YEAR								
5	AGE: (02 - 04)								
6	SEX: (Female)								
7	RACE: (Other)								
8	RACE Suppressed: (Error)								
9	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%.								
10	Geography: (11) District of Columbia								
11	Years: 2017 - 2018								
12	Weighting cells were collapsed for: (Error)								
13	AGE adjusted: (Yes)								
14	Error Codes: (Current selection criteria return an insufficient number of youth and teens and do not meet minimum threshold to create sample weights. Ensure that selections are correct (e.g., correct list of state codes or ZCTA-3values) or include additional geographic or demographic categories (e.g., add additional communities or include additional or all races, age groups, sex, etc.).)								

Order	Weight Category	Sample	Population	Crude Prevalence	Crude Prevalence Standard Error	Weighted Prevalence	Weighted Prevalence Standard Error	Age-adjusted Prevalence	Age-adjusted Prevalence Standard Error
15	Technical Documentation: See https://github.com/NORC-UChicago/CODI-PQ for more information and full details on data sources and methodologies.								
16	Query Date: Friday, August 27, 2021 1:49:38 PM								
17	Suggested Citation: Tanenbaum, E., Campbell, S., Chelluri, D., Zalsha, S., Boim, J., Paddock, S., Copeland, K. (2021). Clinical and Community Data Initiative Prevalence Query (CODI-PQ) SAS programs (version 2015-2019)								
18	Caveats								
19	Patients with either missing or invalid age, sex, height, weight, or geography are not included in counts and prevalence results.								
20	The method used to calculate the standard errors are documented in the technical documentation.								
21	The population is based on age-race-sex-location specific counts from the 2014-2018 American Community Survey Five-year Estimates released by the Census Bureau on December 19, 2019.								
22	The method used to calculate age-adjusted prevalence is documented in the technical documentation.								

F.2 Possible Result Errors

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

Table 33: CODI-PQ Results Error Codes

Error	Description
One or more demographic or geographic category has no groups selected. One or more group must be selected in each category. Please ensure each demographic and geographic category has one or more groups selected (e.g., age group, select an age range for inclusion).	One or more categories are not selected. For example, a minimum of one year, sex, age group, geography, and racial group must be selected (Y). See Step 2.4.4 for more details.
Years are out of scope. CODI-PQ was developed between 2019 and 2021, see Implementation Guide for more details.	Starting year cannot be before 2000, ending year cannot be after 2030. CODI-PQ may be inappropriate to implement on medical encounters outside of 2015 through 2021. Please review the methodology in full to determine whether CODI-PQ is appropriate for your needs.
Geographic level (GEO_GROUP) has been left blank or has been set to an unacceptable value. To remedy issue, please update the GEO_GROUP variable to either STATE, ZCTA-3, or county.	The observed (or unadjusted, or crude) count of youth and teens in the study population.
State and/or GEO3 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.).	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; /*@Note: Level of geography (ACCEPTED VALUES: STATE/ZCTA3, or STATE/COUNTY) ***/ /***/ %LET GEO_LIST = %STR('08', ‘10’); If GEO_GROUP = STATE/ZCTA3; /*@Note: Level of geography (ACCEPTED VALUES: STATE/ZCTA3, or STATE/COUNTY) ***/ /***/ %LET GEO_LIST = %STR('51221', ‘51224’); If GEO_GROUP = STATE/COUNTY; /*@Note: Level of geography (ACCEPTED VALUES: STATE/ZCTA3, or STATE/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, sex), geographies, or years.	To determine the demographic group(s) with insufficient sample size, consider running the optional sample check (Sample_Check = Y).

Error	Description
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 [4] 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."
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 34: CODI-PQ 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 age group, race, sex, and geography. If the sample size is insufficient in an age group or geography, weighting cells may be collapsed (combined). Race and sex do not allow for consolidation of weighting cells.
CODI PQ 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-PQ methodology is appropriate for your use case when used outside of these date ranges.	Users may choose to employ CODI-PQ outside of the testing period. It is recommended that the user carefully review all methods prior to doing so.

F.3 Example Sample Checker Results

CODI-PQ generates optional sample size checker results in the SAS output or results window. Table 38 provides an example results table. Factors include age categories, race, and sex. The values include all demographics selected by the user.

For example, the results below were generated based on the user's selections of only including persons ages 2 – 4, race either of Black or Asian, and both male and female. The user may choose to exclude males and rerun the analysis if partial information is of interest. Why? Results show see that male has an insufficient number of patients as well as Asian. CODI-PQ automatically removes racial categories if the sample size is insufficient (see A.6 Statistical Weights) but does not remove male or female automatically.

Table 35: CODI-PQ Sample Size Checker Results

Checker_Results	Factor	Value
Sample Size Is Sufficient	Age Categories	2 - 4
Sample Size Is Sufficient	Race	Black
Sample Size Is Insufficient	Race	Asian
Sample Size Is Sufficient	Sex	Female
Sample Size Is Insufficient	Sex	Male

Appendix G 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 36. 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

Name	Postal Code	FIPS
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
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

Name	Postal Code	FIPS
Washington	WA	53
West Virginia	WV	54
Wisconsin	WI	55
Wyoming	WY	56

Appendix H Glossary

ACS – American Community Survey. CODI-PQ rely on ACS population counts for statistical weighting.

AEMR – Ambulatory Electronic Medical Record. Used to test CODI-PQ.

AEMR-US – Ambulatory United States Electronic Medical Record data

Source: AEMR-US version 5 OMOP 5 (Aug 2019 release) accessed through the E360TM Software-as-a-Service (SaaS) Platform.

Age-Adjusted Prevalence – Is a prevalence that controls for the effects of differences in population age distributions. When comparing across geographic areas, age adjusting is typically used to control for the influence that different population age distributions might have on health encounter prevalences. Age-adjustment (or age standardization) is the same as calculating a weighted average. It weights the age-specific prevalence observed in a population of interest by the proportion of each age group in a standard population. The standard population are published by the CDC and represent the U.S. 2000 population in each age group.

Age Groups – Age groups include ages 2 to 4, 5 to 9, 10 to 14, 15 and 17, and 18 to 19 years of age.

BMI – Body Mass Index. Used to categorize a person's height and weight into various categories (e.g., underweight, overweight, etc.).

BMI Percentile – Categorization of a youth or teen's height, weight, age, and sex into one of five categories: underweight, healthy weight, overweight, obesity, and/or severe obesity.

CDC – Centers for Disease Control and Prevention

CDM – Common Data Model

Census Tract – Small, relatively permanent statistical subdivisions of a county or equivalent entity that are updated prior to each decennial census. The primary purpose of census tracts is to provide a stable set of geographic units for the presentation of statistical data. Census tracts generally have a population size between 1,200 and 8,000 people.

CODI – Previously the "Childhood Obesity Data Initiative" currently the "The Clinical and Community Data Initiative." CODI 1.0 and 2.0 are projects 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-PQ – CODI prevalence queries (CODI_PQ in SAS programs)

CODI-PQ-GEO3 – CODI PQ 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 percentile – is the total number of people within a particular BMI percentile (e.g., severe obesity) in a specified geographic area (state, county, ZCTA-3, 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 either 1) the state and county or 2) the state and ZCTA-3.

GEO3 – Geographic area identified by three numbers. GEO3 is defined based on either the state and 1) county or 2) ZCTA-3.

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 value 5th percentile to less than the 85th percentile

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 – 95th percentile to less than 120 percent of the body mass index (BMI) value for the 95th percentile.

Overweight – Body Mass Index value 85th to less than the 95th percentile

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-PQ – a set of SAS programs that are executed once and only once per AEMR 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 PQ.

Race Imputation – Imputing missing race data, see also imputation. Setting race imputation to yes allows the programs to include all available EHR data for youth and teens even if the 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 youth and teens in the study population.

SDOH – Social Determinants of Health

Severe Obesity – 120 percent or greater of the BMI value for the 95th percentile

Statistical Weights – A statistical weight is an amount given to increase or decrease the importance of an item. Weights are commonly given for people when a sample and not a census is taken. The value of the weight can be thought of as denoting the number of youth and teens in the population represented by that sample person 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.

Teen – A person aged 13 to 19

Underweight – Body Mass Index value less than 5th percentile

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

Weight Category – Categorization of a person's height, weight, age, and sex (BMI percentile) into one of five categories: underweight, healthy weight, overweight, obesity, and/or severe 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.

Youth – A person aged 2 to 12

ZCTA-3 – The first three digits of a ZIP code tabulation area (ZCTA3 in SAS)

ZCTA-3 data – Refers to a data file of ER linked to a patient's ZIP-3 and thus ZCTA-3

Appendix I 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-PQ	Clinical and Community Data Initiative 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
ZCTA	ZIP Code Tabulation Area

Appendix J 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.

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

Izrael, D., Hoaglin, D. C., & Battaglia, M. P. (2004, May). "To rake or not to rake is not the question anymore with the enhanced raking macro," In Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference.

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.

DATA RIGHTS NOTICE

This tool was produced for the U. S. Government under Contract Number 75FCMC18D0047, and is subject to Federal Acquisition Regulation Clause 52.227-14, Rights in Data-General.

No other use other than that granted to the U. S. Government, or to those acting on behalf of the U. S. Government under that Clause is authorized without the express written permission of The MITRE Corporation.

To the extent necessary MITRE hereby grants express written permission to use, reproduce, distribute, and otherwise leverage this implementation guide.

For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000.