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 Adult Prevalence Queries Implementation Guide
Version 1.2

March 18, 2022

Record of Changes

Version	Date	Author / Owner	Description of Change
1.0	November 1, 2021	Erin Tanenbaum / Health FFRDC	Initial Version
1.1	December 23, 2021	Erin Tanenbaum / Health FFRDC	Updated Draft
1.2	March 18, 2022	Melissa Garcia	Updated Draft

Methodology and SAS Programming Contributors

Name	Affiliation
Erin Tanenbaum	NORC at the University of Chicago
Devi Chelluri	NORC at the University of Chicago
Shalima Zalsha	NORC at the University of Chicago
Jason Boim	NORC at the University of Chicago
Kennon Copeland	NORC at the University of Chicago
Susan Paddock	NORC at the University of Chicago
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-APQ, 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 programs, referenced here as the CODI adult prevalence queries (CODI-APQ) to generate body mass index (BMI)[1] category prevalence estimates based on BMI and diabetes prevalence in adults, aged 20 through 64 stratified by age, sex, and geography. Population estimates were obtained by applying statistical weights, imputation, and suppression criteria to EHR non-probability-based samples. CODI-PQ were developed using a large ambulatory EHR dataset with coverage across the US. CODI-APQ 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 adult obesity often relies on self-report surveys such as the Behavioral Risk Factor Surveillance System. This self-reported data can be expensive to collect, limited in geographic specificity, subject to self-reporting bias, and may struggle with low 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, diagnosis, observations, prescriptins, 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 on large numbers of patients within related geographies. Despite these advantages, aggregate EHR data at the population level are subject to bias.

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

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

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-APQ 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-APQ 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-APQ Implementation Guide is to provide a guide for CODI Data Partners[3] or end users to run the CODI-APQ. The Implementation Guide covers the following:

- CODI-APQ data inputs and linking to population data (pre-processing)
- Generating results in CODI-APQ
- Understanding the CODI-APQ results
- Methodological details

1.3 Scope

The CODI-APQ 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-APQ require patient level records for patients ages 20 through 64. Each record must include year of medical encounter, BMI category, and demographic information (age, sex, race,ethnicity, and some level of geographic location).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-APQ leverage population counts from the American Community Survey (ACS) for statistical weighting. CODI-APQ

[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 encounter 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 data, year of birth, gender, race and ethnicity, and medical variables such as diagnoses, procedures, medication prescriptions records, 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/

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 programming language and macros. Those interested in statistical analysis details used in CODI-APQ can refer to Appendix A for more information. Technical staff preparing datasets for CODI-APQ can refer to Appendices B and C for detailed descriptions of the format required for input data. Explanation of CODI-APQ results can be found in Appendix E.

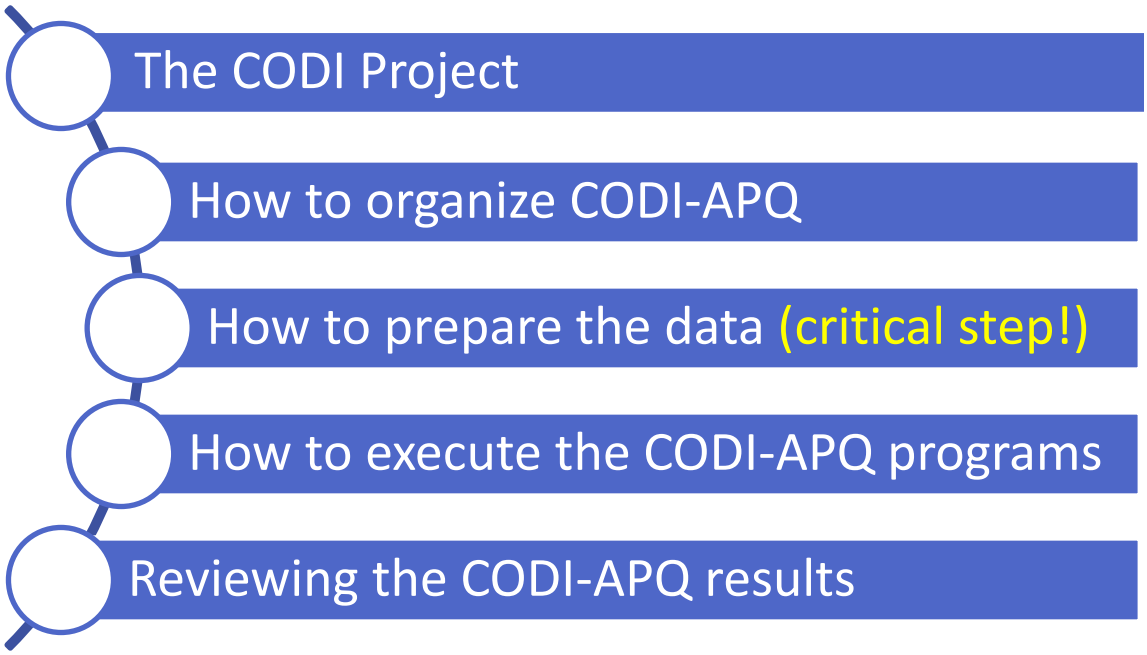
1.5 Document Organization

This document is organized as follows:

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

2 User's Guide

The User's Guide section describes:



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

2.1 CODI Concept

Figure

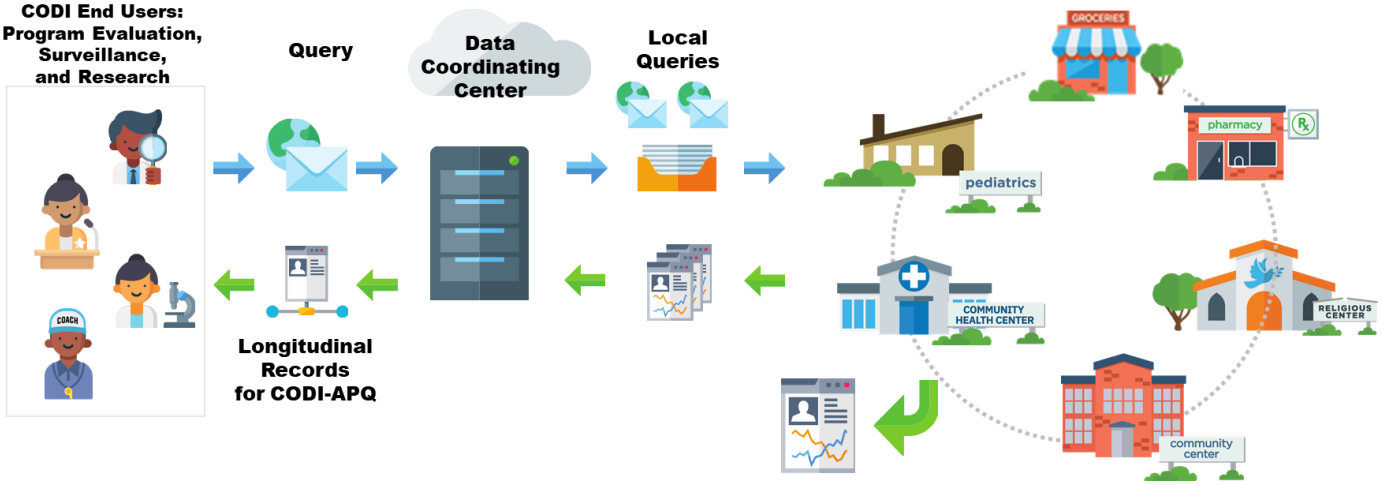


Figure 1. Data Partners with a Common Data Coordinating Center

1 shows how CODI users (e.g., researchers, community-based program evaluators) interact with the data coordinating center, which distributes their research queries to data partners. The Data Doordinating Center assembles the results into longitudinal records, which are sent to the CODI end usersW. CODI end users use the patient-level longitudinal records to create prevalence

estimates with CODI-APQ. CODI-APQ 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.

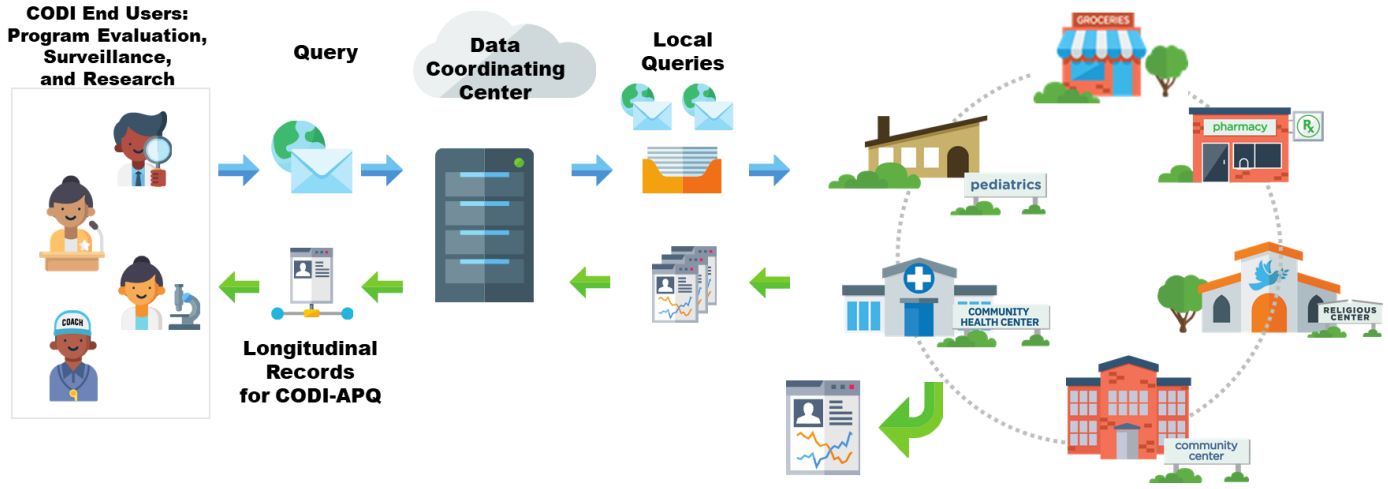


Figure 1. Data Partners with a Common Data Coordinating Center

2.2 About CODI-APQ

CODI-APQ algorithms are contained in a set of programs that calculate adult BMI and diabetes prevalence estimates from a non-probability sample[6] of EHR data. The CODI-APQ programs are are divided into two parts: 1) adult pre-processing, and 2) the adult prevalence query. In pre-processing, patient data are imported into SAS, formatted and linked to the American Community Survey (ACS), and race imputation is conducted in the pre-processing steps (CODI_APQ_PRE_PROCESSING_GEO3). In the prevalence query step, patients are selected based on user specifications, statistically weighted, variance estimates are calculated, results are suppressed (if needed), and prevalence results are output (CODI_APQ_GEO3).

For successful use of the CODI-APQ programs, end users are encouraged to carefully review the methodological details (described in appendices).Inputs for the CODI-APQ 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).

[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-APQ 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.

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-APQ programs user should have a working knowledge of SAS programming language and macros to select the population of interest, execute CODI-APQ, and review the SAS log.

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

- Impute race for patients who are missing race information (optional)
- Calculate statistical weights with an EHR non-probability sample
- Calculate age-adjusted prevalence results (optional)
- Calculate adult BMI category[9] prevalence by BMI, including:
  - **Underweight**: BMI less than 18.5 kg/m^2
  - **Healthy Weight**: BMI greater than or equal to 18.5 and less than 25 kg/m^2
  - **Overweight**: BMI greater than or equal to 25 and less than 30 kg/m^2
  - **Obesity**: BMI greater than or equal to 30 kg/m^2
    - **Obesity Class 1**: BMI greater than or equal to 30 and less than 35 kg/m^2
    - **Obesity Class 2**: BMI greater than or equal to 35 and less than 40 kg/m^2
    - **Obesity Class 3**: BMI greater than or equal to 40 kg/m^2
- Suppress prevalence estimates based on the National Center for Health Statistics (NCHS) Data Presentation Standards for Proportions

2.3 SAS Setup

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

The four-step process to run the CODI-APQ is outlined in Figure 2. Note, the pre-processing program is labeled Pre_Processing_CODI_APQ_GEO3 within the folder and the prevalence query program is labeled CODI_APQ_GEO3.

[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.
[9] https://www.cdc.gov/obesity/adult/defining.html

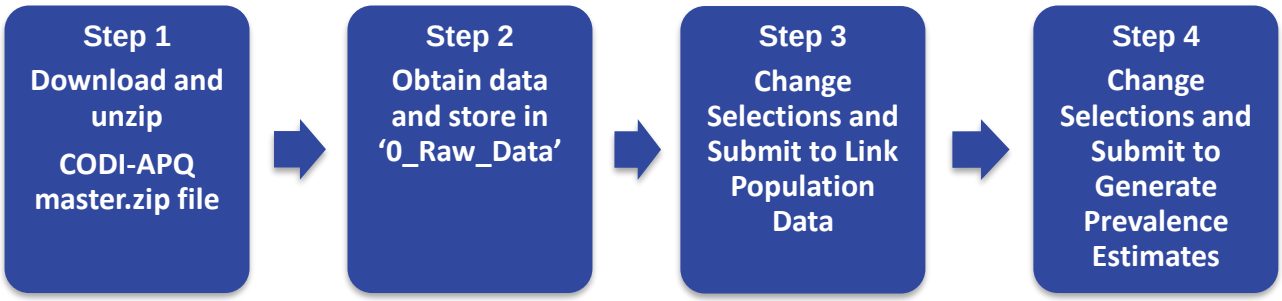


Figure 2. CODI-APQ Process

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

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

To begin, select the "Adults Age 20 to 64" folder and download "CODI-APQ-GEO3-master.zip." Note that "GEO3" refers to the program's options to estimate prevalence at the county or ZCTA3 level.

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

> **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-APQ-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.

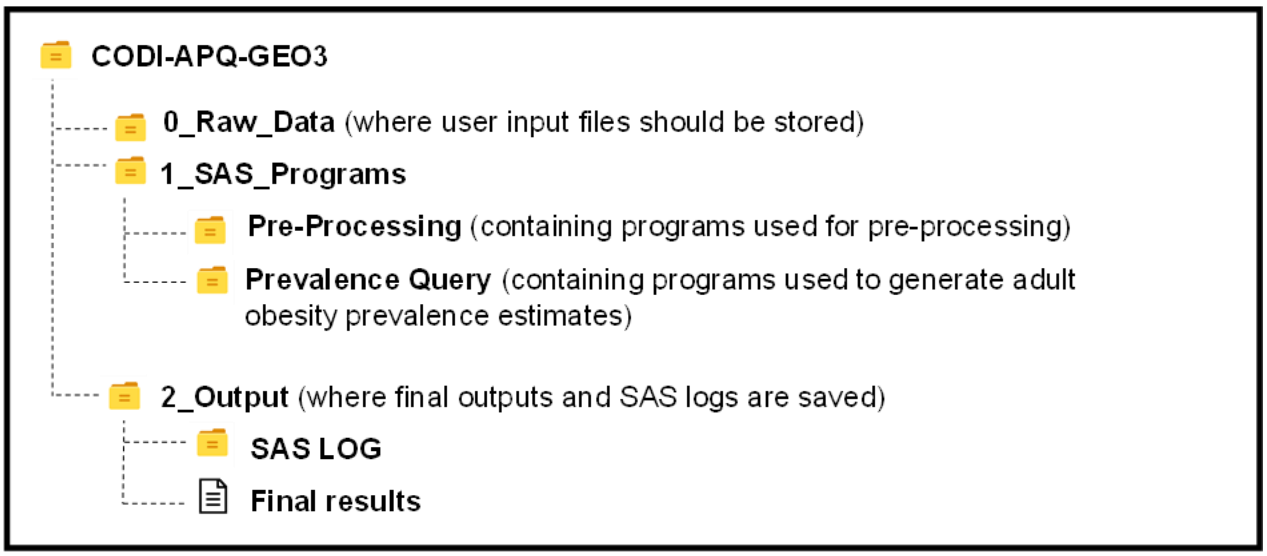


Figure 3 CODI-APQ-GEO3 Folder Structure[10]

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 program is labeled Pre_Processing_CODI_APQ_GEO3 within the folder. Prevalence query program is labeled CODI_APQ_GEO3 within the folder.

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

2. **EHR data file** supplied by the end user in comma separated values (.csv). The EHR file must:
    - Contain all variables in the order (sequence) expected. Variable names and order can be found in Appendix C.
    - Contain valid variable values as anticipated. Variable values can be found in Appendix C.
    - 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 for all patients (underweight, healthy weight, etc.) and categorized 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[11].
    - Users may also wish to reconcile demographic characteristics for each patient across years, 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, the researcher can add three columns to the end of their file with blank values. These variables include:

- A sickle cell disease[12] indicator
- A pregnancy indicator
- A diabetes spectrum indicator (prediabetes, diabetes, no evidence of diabetes)

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

2.4.3 STEP 3: Link Population (Pre-Processing)

Open the "Quickstart-Pre_Processing_CODI_APQ_GEO3" SAS program stored in "\1_SAS_Programs" and change the selection per the steps outlined in the tables below.

[11] ZCTA-3 and COUNTY are defined as: ZCTA-3 – ZIP Code Tabulation Area (ZCTA) is a 5-digit code assigned by the Census Bureau and the ZCTA-3 is the first three digits of the ZCTA. Information on ZCTAs and ZCTA-3s can be found through the Census Bureau (https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html) and COUNTY – FIPS county code.

[12] 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.

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 once and only once for each EHR data 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 known, 2) keeping the imputed race value consistent for patients who were imputed before and their race value is still unknown.

A new folder ("\2_Output\Pre-Processed_…") will be created upon completion of the programs. In this folder, two SAS7bdat files (user input ACS file and pre-processed CODI file) will be generated. **Once pre-processing is complete, the user can submit an unlimited number of adult 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 the SAS programs folder, data files, etc. are stored.	%let ROOT_PRE = P:\CODI-APQ-master;
PRE_DEST	Following the example, the results from pre-processing will be generated and stored in folder P:\CODI-APQ-master\2_Output\ Pre_Processed_**CODI_APQ**	%let PRE_DEST = **CODI_APQ**;
ACS_FILENAME	The American Community Survey file name. 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 B.1.	%let ACS_FILENAME = ACS_State_COUNTY;
EHR_FileNAME	The adult-level EHR data file. The file must be in csv format. 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 C.1. For the example, the user will have stored their raw data in: P:\CODI-APQ-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-APQ-master\2_Output\ Pre_Processed_CODI_APQ**CODI_APQ_GEO3**	%LET EHR_PRE_Out = **CODI_APQ_GEO3**;
LOG_NAME_PRE	The SAS log is automatically stored and the user can specify the name of the file. The example SAS log would be stored in: P:\CODI-APQ-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 Details
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	Section 3, edit the SAS program within "SECTION 3: Optional Results File Name Suffix." This is optional and changes the name of the output patient-level file. The example would be stored in: P:\CODI-APQ-master\2_Output\ Pre_Processed_CODI_APQ**CODI_APQ_GEO 3.**	%LET EHR_PRE_Out = **CODI_APQ_GEO3**;
5	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**;
6	Save the Quickstart program.	SAS encourages saving all files before submitting the program.

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

Order	Description	Details
1	Submit the Quickstart program.	Submit the Quickstart program. The program completes all tasks within the data sets and proc statements in the Quickstart program and moves to the next SAS program automatically through an include statement.
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_APQ_GEO3" SAS program stored in "\1_SAS_Programs" and change the selections within the program per the steps outlined in the table below.

The final results (CODI-APQ results) will be generated in Excel format and saved in "\2_Output." Appendix E provides examples of the results. Note that results are for the group of patients selected by the user. To calculate results for multiple geographic or demographic characteristics (e.g., first for females and then for males), 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 patients aged 20 to 29 but does not have patients aged 30 to 64, then the program will fail with an error message caused by insufficient sample size for patients aged 30 to 64.

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: Input 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 by Sections

SAS Macro Variable	Details	Example
SECTION 1: Folder and file names		
ROOT_PQ	The core folder name is same as in pre-processing.	%let ROOT_PQ = P:\ CODI-APQ-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-APQ-master\2_Output\ Pre_Processed_**CODI_APQ**	%let PRE_DEST = **CODI_APQ**;
EHR_PRE_OUT	The adult level EHR data file described in pre-processing P:\CODI-APQ-master\2_Output\ Pre_Processed_CODI_APQ**CODI_APQ_GEO3**	%LET EHR_PRE_Out = **CODI_APQ_GEO3**;
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-APQ-master\CODI-APQ-GEO3\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;
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-APQ-master\ \2_Output**File_name**<Date and Time>.xls. 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		

SAS Macro Variable	Details	Example
BEG_YEAR	Subsets the prevalence to medical encounters in this year for diabetes or 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 diabetes or 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 in the prevalence based on the geographic location of the adult. If 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). If set to yes (Y) then the EHR must have sufficient sample size in all states in the U.S.	/***/ %LET ALL_STATES = N;
ALL_AGES	Subsets the prevalence based on the age of the adult. The user may either select to include all adults aged 20 to 64 or alternatively may select age groups. Note: if ALL_AGES = Y; then by default the program will include all adults aged 20 to 64. If ALL_AGES = N; then by default the program will subset the prevalence based on the individual age ranges selected (in future step).	/***/ %LET ALL_AGES = Y;

SAS Macro Variable	Details	Example
ALL_SEXES	Subsets the prevalence based on the sex of the adult. The user may either select to include all male and female adults or alternatively may select either males or females. 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 future step).	/***/ %LET ALL_SEXES = Y;
ALL_RACES	Subsets the prevalence based on the race of the adult. The user may either select to include all races or alternatively may select specific race(s). Inclusion or exclusion of imputed race is not impacted by the choice made in this step. Note: if ALL_RACES = Y; then by default the program will include all races (White, Black, Asian, Other). If ALL_RACES = N; then by default the program will subset the prevalence based on the individual races selected (in future step).	/***/ %LET ALL_RACES = Y;
SECTION 3: Additional Flags		
ACSCOUNTY	Specifies the geographic level of the ACS data. If equal to no (N) then the EHR and ACS data files includes state and ZCTA3 codes. If equal to yes (Y) then the files include state and county codes.	/***/ %LET ACSCOUNTY = N; /
INCLUDE_PREGNANCY	Subsets EHR records based on pregnancy flag. If set to Y, then patients will be included regardless of pregnancy status. If set to N, then only patients with a pregnancy status set to no (0) will be included.	/***/ %LET INCLUDE_PREGNANCY = Y;

SAS Macro Variable	Details	Example
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. age 50-64, etc.) will be displayed in the SAS output or results window. Each factor will include the factor, value (e.g. age, 50-64) and either “Sample Size Is Insufficient” if n <20 or “Sample Size Is Sufficient”.	/***/ %LET SAMPLE_CHECK = Y;
CO_OCCURRING	Set to yes (Y) if diabetes prevalence is requested.	/***/ %LET CO_OCCURRING = Y;
CO_OCCURING_COND_VAR	Required only when co_occurring is set to yes. Provide the variable name (also known as field or column) where the diabetes spectrum variable is stored in the raw data file (from pre-processing step). This value will stay the same as in the pre-processing step.	/***/ %LET CO_OCCURING_COND_VAR =DIABETES_SPECTRUM;

Table 7. Change Specifications, Processing Steps by Sections, Continued

Order	Description	Details
3	Edit the SAS program within “SECTION 4: Only complete section 3 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(s)
SECTION 4: Only complete section 3 for any "N" values listed in section 2		

SAS Macro Variable Category	Details	Example(s)
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 senarios is described and show. Of note, values should be 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.	/*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:*/ /***/ %LET GEO_GROUP = COUNTY; /***/ %LET GEO_LIST = %STR('51061', '51059');
If ALL_AGES = N;	If ALL_AGES is set to no (N), the age macros (20-24, 25-29, 30-34, 35-44, 45-54, 55-64) subset the prevalence based on the age of the adult and the responses to each individual age macro. Note that if ALL_AGES is set to yes (Y), then the SAS program does not review the user's responses to the age-specific macros.	%LET WGT_AGE_20_24 = N; %LET WGT_AGE_25_29 = N; %LET WGT_AGE_30_34 = Y; %LET WGT_AGE_35_44 = Y; %LET WGT_AGE_45_54 = Y; %LET WGT_AGE_55_64 = Y;

SAS Macro Variable Category	Details	Example(s)
If ALL_RACES = N;	If ALL_RACES is set to no (N), the race macros (White, Black, Asian, Other) subset the prevalence based on the race (or imputed race) of the adult and the responses to each individual age macro. Note that if ALL_RACES is set to yes (Y), then the SAS program does not review the race-specific macros.	%LET RACE_WHITE = N; %LET RACE_BLACK = Y; %LET RACE_ASIAN = Y; %LET RACE_OTHER = Y;
If ALL_SEXES = N;	If ALL_SEXES is set to no (N), the sex macros (male, female) subset the prevalence based on the sex of the adult and the responses to each individual sex macro. Note that if ALL_SEXES is set to yes (Y), then the SAS program does not review the sex-specific macros.	%LET SEX_MALE = N; %LET SEX_FEMALE = Y;
SECTION 5: Methodological option selections		
IMP_RACES	If IMP_RACES is set to yes (Y), then the program includes adults with imputed race values as well as adults with race values provided by EHR. Otherwise, if IMP_RACES is set to no (N), then the patients with imputed races are excluded and only patients with race values provided by EHR are included.	%LET IMP_RACES = Y;
AGE_ADJ	If AGE_ADJ is set to yes (Y), then the program generates crude, weighted, and age adjusted prevalence and standard errors. Otherwise, if AGE_ADJ is set to no (N), age adjusted prevalence is not generated. Note that crude and weighted results are always provided.	%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-APQ Execution Processing Steps

Order	Description	Details
1	Submit CODI-APQ Quickstart program.	Submit the Quickstart program. The program completes all tasks within the data sets and proc statements in the Quickstart program and moves to the next SAS program automatically through an include statement.
2	Review the log.	Review the log for possible errors including words such as error 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 for a single year, consider using two or three years of data[13].

2.4.5 Review BMI Category Prevalence Results

CODI-APQ generate BMI category prevalence (or BMI prevalence) outputs as an Excel file. Table 11 provides an overview of the variables included when diabetes prevalence is set to no. Note, descriptive information about CODI-APQ 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 vary.

Table 11. CODI-APQ BMI Prevalence Results Data Dictionary

Column	Description
Order	Row order
BMI Category	The BMI category based on BMI.
Sample	The observed (or unadjusted, or crude) count of adults 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 adults in the population represented by that sample patient. See Appendix A.6 Statistical Weights for more information.
Weighted Prevalence Standard Error	Standard error based on weighted counts. See Appendix A.11 Variance for more information.
Age-Adjusted Prevalence	Prevalence based on weighted, age-adjusted counts (optional). See Appendix A.1 Age Adjustment for more information.
Age-Adjusted Prevalence Standard Error	Standard error based on weighted, age-adjusted counts. See Appendix A.1 Age Adjustment for more information.

[13] 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.

2.4.6 Review BMI and Co-Occurring Conditions Prevalence Results

CODI-APQ generate optional diabetes prevalence results as an Excel file. Table 12 provides an overview of the Excel Worksheets and Table 13 provides results by Worksheet class and variables (Excel cells) generated when the co-occurring condition is set to yes. Variable names repeat between Excel Worksheets. Inclusion criteria determines the sample represented within the Excel Worksheet. Note, descriptive information about CODI-APQ user inputs, error codes, sources of technical documentation, caveats, and a possible citation are found in rows labeled Order 3-24.

Table 12. CODI-APQ Diabetes Prevalence Results, Description of Excel Worksheets

Excel Worksheet	Worksheet Class	Description and Inclusion Criteria
Counts	Counts	Provides crude and weighted counts. Does not include prevalence estimates. Inclusion: all patients.
BMI Category	BMI prevalence	Provides results as described in Section 2.4.5 Review BMI prevalence Results Inclusion: all patients.
Diabetes Spectrum	Diabetes prevalence	Provides results for diabetes counts and prevalence. Inclusion: all patients.
Diabetes Underweight	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = underweight. Inclusion: patients with a BMI equal to underweight.
Diabetes Healthy Weight	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = healthy weight. Inclusion: patients with a BMI equal to healthy weight.
Diabetes Overweight	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = overweight. Inclusion: patients with a BMI equal to overweight.
Diabetes Obese	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = obese. Inclusion: patients with a BMI equal to obesity.
Diabetes Obese Class 1	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = obese class 1. Inclusion: patients with a BMI equal to obese class 1.
Diabetes Obese Class 2	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = obese class 2. Inclusion: patients with a BMI equal to obese class 2.
Diabetes Obese Class 3	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = obese class 3. Inclusion: patients with a BMI equal to obese class 3.

Table 13. CODI-APQ Diabetes Prevalence Results Data Dictionary

Worksheet Class	Column	Description
Counts	Order	Row order
Counts	Condition	Co-Occurring Condition based on the Diabetes Spectrum (no evidence of diabetes, pre-diabetes, diabetes).
Counts	BMI Category	The BMI category based on BMI.
Counts	Sample	The observed (also known as unadjusted, or crude) count of adults in the study population.

Worksheet Class	Column	Description
Counts	Population	The weighted (or adjusted) count of the study population.
Counts	Population (Age Adjusted)	The weighted, age-adjusted count of the study population (optional).
BMI prevalence	Order	Row order
BMI prevalence	BMI Category	The BMI category based on BMI.
BMI prevalence	Sample	The observed (also known as unadjusted, or crude) count of adults in the study population.
BMI prevalence	Population	The weighted (or adjusted) count of the study population.
BMI prevalence	Population (Age Adjusted)	The weighted, age-adjusted count of the study population (optional).
BMI prevalence	Crude Prevalence	The observed (or unadjusted, or crude) prevalence in the study population.
BMI prevalence	Crude Prevalence Standard Error	The observed (or unadjusted, or crude) standard error in the study population.
BMI prevalence	Weighted Prevalence	Prevalence based on weighted counts. A sample weight is assigned to each sampled patient. It is a measure of the number of adults in the population represented by that sample patient. See Appendix A.6 Statistica Weights for more information.
BMI prevalence	Weighted Prevalence Standard Error	Standard error based on weighted counts. See Appendix A.11 Variance for more information.
BMI prevalence	Age-Adjusted Prevalence	Prevalence based on weighted, age-adjusted counts (optional). See Appendix A.1 Age Adjustment for more information.
BMI prevalence	Age-Adjusted Prevalence Standard Error	Standard error based on weighted, age-adjusted counts (optional). See Appendix A.1 Age Adjustment for more information.
Diabetes prevalence, BMI by diabetes prevalence	Order	Row order
Diabetes prevalence, BMI by diabetes prevalence	Condition	Co-Occurring Condition based on the Diabetes Spectrum.
Diabetes prevalence, BMI by diabetes prevalence	Sample	The observed (or unadjusted, or crude) count of adults in the study population.
Diabetes prevalence, BMI by diabetes prevalence	Population	The weighted (or adjusted) count of the study population.
Diabetes prevalence, BMI by diabetes prevalence	Population (Age Adjusted)	The weighted, age-adjusted count of the study population (optional).
Diabetes prevalence, BMI by diabetes prevalence	Crude Prevalence	The observed (or unadjusted, or crude) prevalence in the study population.
Diabetes prevalence, BMI by diabetes prevalence	Crude Prevalence Standard Error	The observed (or unadjusted, or crude) standard error in the study population.

Worksheet Class	Column	Description
Diabetes prevalence, BMI by diabetes prevalence	Weighted Prevalence	Prevalence based on weighted counts. A sample weight is assigned to each sampled patient. It is a measure of the number of adults in the population represented by that sample patient. See Appendix A.6 Statistical Weights for more information.
Diabetes prevalence, BMI by diabetes prevalence	Weighted Prevalence Standard Error	Standard error based on weighted counts. See Appendix A.11 Variance for more information.
Diabetes prevalence, BMI by diabetes prevalence	Age-adjusted Prevalence	Prevalence based on weighted, age-adjusted counts (optional). See Appendix A.1 Age Adjustment for more information.
Diabetes prevalence, BMI by diabetes prevalence	Age-Adjusted Prevalence Standard Error	Standard error based on weighted, age-adjusted counts (optional). See Appendix A.1 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 B – ACS File Layouts
- Appendix C – EHRs File Layouts
- Appendix D – CODI-APQ-GEO4 Example SAS Programs
- Appendix E – CODI-APQ Results Example
- Appendix F – State FIPS Codes
- Appendix G – Glossary
- Appendix H – Abbreviations and Acronyms
- Appendix I – Bibliography

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.[14] The specific age groups used for age adjustment are 20 to 24 years, 25 to 34 years, 35 to 44 years, 45 to 54 years, and 55 to 64 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.

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 can be used to screen for potential weight and health-related issues.

[14] Klein & Schoenborn, 2001.

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

BMI categories are described in section A.11.

For more information, see:

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

A.3 Data Sources (Inputs)

This document provides an implementation guide for CODI-APQ on adult data. Required input files are the following:

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

CODI-APQ 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 AEMRdata and synthetic EHR data. Both provide a non-probability sample of longitudinally linked patients' medical records from within the U.S. CODI-APQ subset the file to adults aged 20 to 64 years of age. 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 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-APQ 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 plausibility. [19]

[15] ACS 2019 file for use with CODI-APQ 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/.
[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] 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 Diabetes Spectrum

Prevalence is calculated from an optional indicator for diabetes spectrum. Diabetes indicators are based on the user's defined phenotype and each patient can have one diabetes status assigned within a calendar year. Each adult may be classified as one of the following categories:

1. **Diabetes**: evidence of Type I or Type II diabetes
2. **Prediabetes**: evidence of prediabetes
3. **No Evidence**: no evidence of prediabetes or diabetes

CODI-APQ include algorithms for prevalence of diabetes but can be modified by the end user for other conditions. The algorithm was developed using a categorical variable for the diabetes spectrum where 1=no evidence of diabetes, 2=pre-diabetes, and 3=diabetes. Laboratory values, medications, and diagnosis codes in the AEMR data were used to define the variable for each individual. A reference (or complete description) of the diabetes spectrum phenotype used for testing CODI-APQ will be provided in future releases of this user's guide. Note, definitions of diabetes or other chronic conditions may vary by organization.

A.4.1 Prevalence

A prevalence is either:

- **Crude**: the proportion of the sample that has a health condition (BMI category or diabetes) 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.6 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.11 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 data set used for CODI-APQ development. Information on ethnicity is not captured in IQVIA AEMR and therefore not used in CODI-APQ 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 14. 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-APQ. The data inputs and link population data (pre-processing) program inputs race for each adult missing race information. The program operates sequentially in three phases, imputing race for adults 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 adult 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 three 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 age 20 to 64 are in scope.
3. The sex of the patient is unknown.

CODI-APQ 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-APQ 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 adult BMI and diabetes 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-APQ 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-APQ 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 15. 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. Applying statistical weights is often used to reduce potential biases introduced by the EHR data sampling methodology. Ratio adjustments are applied to all sampled adults. Ratio adjustment is a statistical weighting technique aimed to improve the accuracy of survey results by both reducing bias and increasing precision.[20] One way to accomplish this goal is known as iterative proportional fitting or raking. Raking adjusts the data so that groups that are underrepresented in the sample can be accurately represented in the final data set. Raking accurately matches sample distributions to known demographic characteristics of populations. The use of raking reduces nonresponse bias and has been shown to reduce error within sample results.

Implementing raking programs requires 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-APQ 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-APQ 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:

$$MSE = Variance + (Bias)^2$$

[20] Little, 1993.

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 MSE as low as possible. There is no exact rule for this trade-off because the amount of bias is unknown.

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

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

A.7 Prevalence Calculations

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

To calculate the weighted prevalence of the population the sum of statistical weights within each BMI or diabetes category is divided by the sum of statistical weights within the EHR. To control extreme weights which may increase the variance, extreme weights are trimmed. To calculate the variance of BMI, a Taylor-series approximation is used.[21]

CODI-APQ provide users with 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-APQ 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". 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-APQ again after excluding males.

[21] 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-APQ data suppression is adapted from the NCHS data presentation standards for reporting proportions in NCHS reports and data products,[22] 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 NCHS Data Presentation Standards for Proportions are applied to all CODI-APQ results. The Presentation Standards also provide guidance for identifying results for statistical review, CODI-APQ do not identify records for statistical review and leave this step for the user. The data presentation standards are described in

Table 1616 and **Error! Reference source not found.**4.

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.

[22] Parker et al., 2017.

Table 16. NCHS Data Presentation Standards for Proportions

Statistic	Standard
Sample size	Proportions should be based on a minimum denominator sample size and effective denominator sample size (when applicable) of 30. 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[23]), 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.

[23] 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).

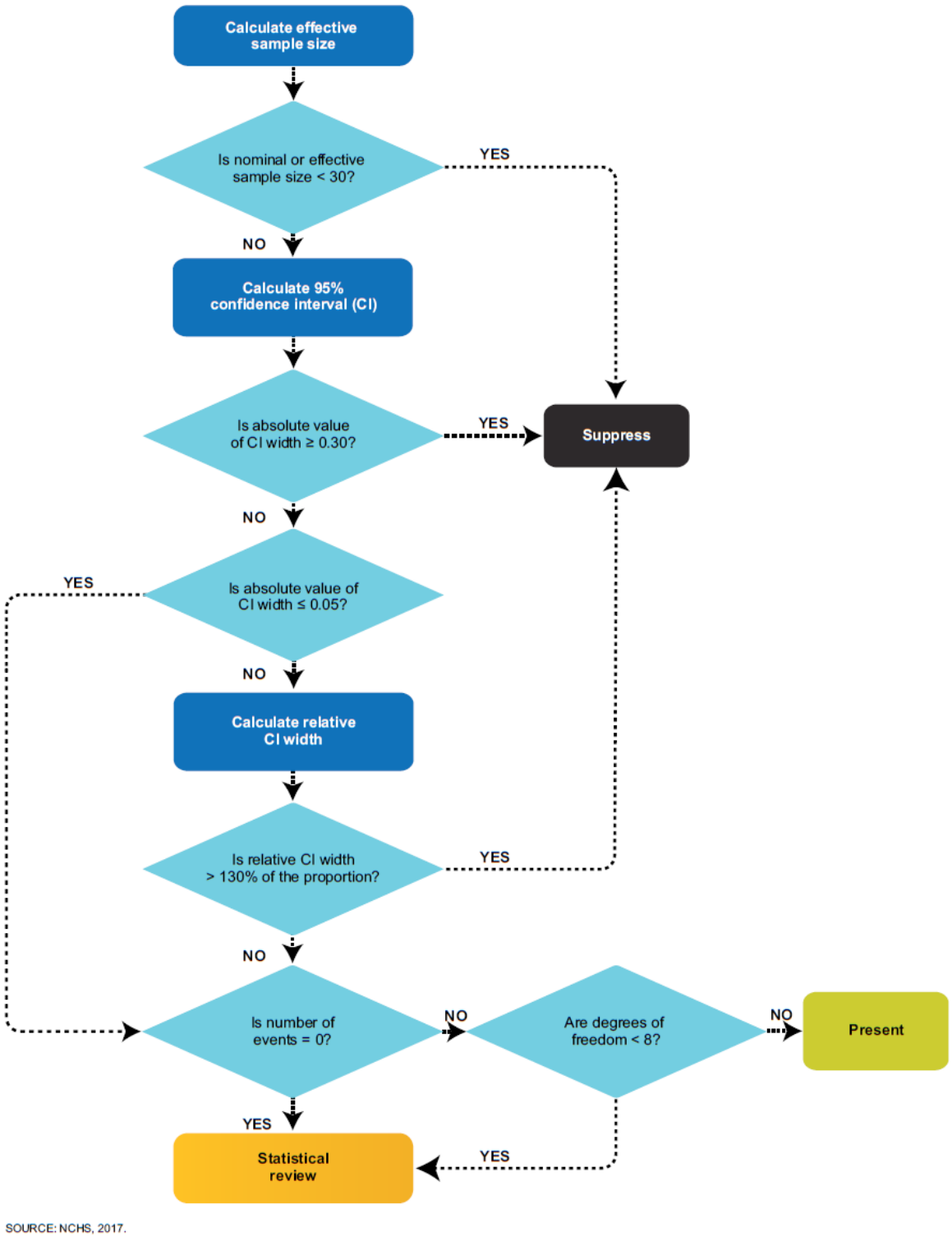


Figure 4. NCHS Suppression Standards[24]

A.11Variance

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

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

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

d_{hi} = 0 or 1, indicating whether adult 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 adults in stratum h

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

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

d'_{hi} = the value of the domain indicator for adult 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 adults within stratum h, whereas for Y_{hi} and d_{hi} they refer to adults in the population in stratum h. Then, the ratio estimator for θ_d is:

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

[24] Parker et al., 2017.

To calculate the variance of $\widehat{\theta}_d$, a Taylor-series approximation is used.[25] 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

Prevalence is calculated from a patient's BMI category. EHR data included for analysis should have at most one BMI category assigned to each patient within a calendar year. If multiple BMIs are recorded in a single year, selecting a single BMI should be done at the user's discretion. The adult BMI categories for prevalence are defined as follows:

- **Underweight**: BMI less than 18.5 kg/m^2
- **Healthy Weight**: BMI greater than or equal to 18.5 and less than 25 kg/m^2
- **Overweight**: BMI greater than or equal to 25 and less than 30 kg/m^2
- **Obesity**: BMI greater than or equal to 30 kg/m^2
    - **Obesity Class 1**: BMI greater than or equal to 30 and less than 35 kg/m^2
    - **Obesity Class 2**: BMI greater than or equal to 35 and less than 40 kg/m^2
    - **Obesity Class 3**: BMI greater than or equal to 40 kg/m^2

For more information, visit:

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

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

[25] Wolter, 2007.

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.[26]

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

- Representativeness of CODI-APQ results - CODI-APQ 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-APQ 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-APQ 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-APQ 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.
- Diabetes spectrum phenotype and sample population– CODI-APQ require the use of a phenotype to estimate the prevalence of prediabetes and diabetes within a population. The sample population and phenotype require careful consideration of the pros and cons. For example, if the sample population will be subset to patients with only wellness visits, it can greatly impact the denominator and the prevalence estimate. Similarly, a single, standard phenotype is not available for diabetes spectrum. Thus, the user may consider different versions of phenotypes used in public health.

[26] https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

- Measurement error - Height and weight measurement protocols may differ between medical providers, even with clear protocols aimed to increase consistency between medical professionals,[27] leading to potential measurement error. Additionally, height and weight values in EHR data are subject to data entry errors or software glitches. All CODI-APQ 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-APQ suppress results based on published small sample guidelines using the National Center for Health Statistics Data Presentation Standards for Proportions[28].

[27] Best & Shepherd, 2020.
[28] Parker JD, Talih M, Malec DJ, et al, 2017.

Appendix B ACS File Layouts

B.1 ACS Input File Layout

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

Table 17. ACS Input File Layout, CSV File

Variable Name	Description	Format	Example
ZCTA3	Geoid (option 1) 3 digits ZIP Code Tabulation Areas (ZCTAs) followed by two-letter State Abbreviations	Character	221VA
County_Code	Geoid (option 2) 3-digit County FIPS Code	Character	059
State_code	Geoid (option 2) State FIPS code, 2-digit State Code	Character	08
B01001A_008	AGE_20_24_MALE_WHITE	Number	8199
B01001A_009	AGE_25_29_MALE_WHITE	Number	256
B01001A_010	AGE_30_34_MALE_WHITE	Number	246
B01001A_011	AGE_35_44_MALE_WHITE	Number	495
B01001A_012	AGE_45_54_MALE_WHITE	Number	297
B01001A_013	AGE_55_64_MALE_WHITE	Number	145
B01001A_023	AGE_20_24_FEMALE_WHITE	Number	271
B01001A_024	AGE_25_29_FEMALE_WHITE	Number	188
B01001A_025	AGE_30_34_FEMALE_WHITE	Number	267
B01001A_026	AGE_35_44_FEMALE_WHITE	Number	139
B01001A_027	AGE_45_54_FEMALE_WHITE	Number	134
B01001A_028	AGE_55_64_FEMALE_WHITE	Number	278
B01001B_008	AGE_20_24_MALE_BLACK	Number	0
B01001B_009	AGE_25_29_MALE_BLACK	Number	28
B01001B_010	AGE_30_34_MALE_BLACK	Number	28

Variable Name	Description	Format	Example
B01001B_011	AGE_35_44_MALE_BLACK	Number	0
B01001B_012	AGE_45_54_MALE_BLACK	Number	0
B01001B_013	AGE_55_64_MALE_BLACK	Number	0
B01001B_023	AGE_20_24_FEMALE_BLACK	Number	0
B01001B_024	AGE_25_29_FEMALE_BLACK	Number	0
B01001B_025	AGE_30_34_FEMALE_BLACK	Number	0
B01001B_026	AGE_35_44_FEMALE_BLACK	Number	10
B01001B_027	AGE_45_54_FEMALE_BLACK	Number	14824
B01001B_028	AGE_55_64_FEMALE_BLACK	Number	277
B01001C_008	AGE_20_24_MALE_AIAN	Number	222
B01001C_009	AGE_25_29_MALE_AIAN	Number	276
B01001C_010	AGE_30_34_MALE_AIAN	Number	263
B01001C_011	AGE_35_44_MALE_AIAN	Number	774
B01001C_012	AGE_45_54_MALE_AIAN	Number	101
B01001C_013	AGE_55_64_MALE_AIAN	Number	237
B01001C_023	AGE_20_24_FEMALE_AIAN	Number	355
B01001C_024	AGE_25_29_FEMALE_AIAN	Number	242
B01001C_025	AGE_30_34_FEMALE_AIAN	Number	1404
B01001C_026	AGE_35_44_FEMALE_AIAN	Number	755
B01001C_027	AGE_45_54_FEMALE_AIAN	Number	36
B01001C_028	AGE_55_64_FEMALE_AIAN	Number	44
B01001D_008	AGE_20_24_MALE_ASIAN	Number	3
B01001D_009	AGE_25_29_MALE_ASIAN	Number	22
B01001D_010	AGE_30_34_MALE_ASIAN	Number	37

Variable Name	Description	Format	Example
B01001D_011	AGE_35_44_MALE_ASIAN	Number	0
B01001D_012	AGE_45_54_MALE_ASIAN	Number	14
B01001D_013	AGE_55_64_MALE_ASIAN	Number	11
B01001D_023	AGE_20_24_FEMALE_ASIAN	Number	9
B01001D_024	AGE_25_29_FEMALE_ASIAN	Number	26
B01001D_025	AGE_30_34_FEMALE_ASIAN	Number	13407
B01001D_026	AGE_35_44_FEMALE_ASIAN	Number	283
B01001D_027	AGE_45_54_FEMALE_ASIAN	Number	222
B01001D_028	AGE_55_64_FEMALE_ASIAN	Number	425
B01001E_008	AGE_20_24_MALE_NHPI	Number	439
B01001E_009	AGE_25_29_MALE_NHPI	Number	429
B01001E_010	AGE_30_34_MALE_NHPI	Number	485
B01001E_011	AGE_35_44_MALE_NHPI	Number	254
B01001E_012	AGE_45_54_MALE_NHPI	Number	359
B01001E_013	AGE_55_64_MALE_NHPI	Number	189
B01001E_023	AGE_20_24_FEMALE_NHPI	Number	561
B01001E_024	AGE_25_29_FEMALE_NHPI	Number	426175
B01001E_025	AGE_30_34_FEMALE_NHPI	Number	9461
B01001E_026	AGE_35_44_FEMALE_NHPI	Number	9446
B01001E_027	AGE_45_54_FEMALE_NHPI	Number	11409
B01001E_028	AGE_55_64_FEMALE_NHPI	Number	7231
B01001F_008	AGE_20_24_MALE_OTHER	Number	8352
B01001F_009	AGE_25_29_MALE_OTHER	Number	8754
B01001F_010	AGE_30_34_MALE_OTHER	Number	10226

Variable Name	Description	Format	Example
B01001F_011	AGE_35_44_MALE_OTHER	Number	10640
B01001F_012	AGE_45_54_MALE_OTHER	Number	7731
B01001F_013	AGE_55_64_MALE_OTHER	Number	9368
B01001F_023	AGE_20_24_FEMALE_OTHER	Number	12014
B01001F_024	AGE_25_29_FEMALE_OTHER	Number	732
B01001F_025	AGE_30_34_FEMALE_OTHER	Number	659
B01001F_026	AGE_35_44_FEMALE_OTHER	Number	823
B01001F_027	AGE_45_54_FEMALE_OTHER	Number	491
B01001F_028	AGE_55_64_FEMALE_OTHER	Number	501
B01001G_008	AGE_20_24_MALE_GE2R	Number	683
B01001G_009	AGE_25_29_MALE_GE2R	Number	650
B01001G_010	AGE_30_34_MALE_GE2R	Number	652
B01001G_011	AGE_35_44_MALE_GE2R	Number	410
B01001G_012	AGE_45_54_MALE_GE2R	Number	651
B01001G_013	AGE_55_64_MALE_GE2R	Number	56886
B01001G_023	AGE_20_24_FEMALE_GE2R	Number	43689
B01001G_024	AGE_25_29_FEMALE_GE2R	Number	1753
B01001G_025	AGE_30_34_FEMALE_GE2R	Number	196
B01001G_026	AGE_35_44_FEMALE_GE2R	Number	253
B01001G_027	AGE_45_54_FEMALE_GE2R	Number	150
B01001G_028	AGE_55_64_FEMALE_GE2R	Number	7913
B03002_012	TOTAL_LATIN	Number	2932
B03002_013	LATIN_WHITE	Number	28989
B03002_014	LATIN_BLACK	Number	7476

Variable Name	Description	Format	Example
B03002_015	LAT_AIAN	Number	2604
B03002_016	LAT_ASIAN	Number	24797
B03002_017	LAT_NHPI	Number	4891
B03002_018	LAT_OTHER	Number	3258
B03002_019	LAT_GE2R	Number	62253
B15001_011	AGE_25_34_MALE_EDUC	Number	11482
B15001_017	AGE_25_34_MALE_BACHELOR	Number	8394
B15001_018	AGE_25_34_MALE_GRAD_PROF	Number	8124
B15001_019	AGE_35_44_MALE_EDUC	Number	5237
B15001_025	AGE_35_44_MALE_BACHELOR	Number	26186
B15001_026	AGE_35_44_MALE_GRAD_PROF	Number	6646
B15001_027	AGE_45_64_MALE_EDUC	Number	5657
B15001_033	AGE_45_64_MALE_BACHELOR	Number	67764
B15001_034	AGE_45_64_MALE_GRAD_PROF	Number	14062
B15001_052	AGE_25_34_FEMALE_EDUC	Number	11986
B15001_058	AGE_25_34_FEMALE_BACHELOR	Number	1000
B15001_059	AGE_25_34_FEMALE_GRAD_PROF	Number	2000
B15001_060	AGE_35_44_FEMALE_EDUC	Number	100
B15001_066	AGE_35_44_FEMALE_BACHELOR	Number	450
B15001_067	AGE_35_44_FEMALE_GRAD_PROF	Number	120
B15001_068	AGE_45_64_FEMALE_EDUC	Number	43
B15001_074	AGE_45_64_FEMALE_BACHELOR	Number	2356345
B15001_075	AGE_45_64_FEMALE_GRAD_PROF	Number	2235

B.2 ACS for Use with GEO3 Data

Table 18. ACS Pre-Processing Results File Layout – GEO3

Variable Name	Format	Length
Geography	Number	8
State_Alpha	Number	8
State_FIPS	Number	8
GEO3	Number	8
TOTAL_ACS_POPULATION	Number	8
AGE_20_24_MALE_WHITE	Number	8
AGE_25_29_MALE_WHITE	Number	8
AGE_30_34_MALE_WHITE	Number	8
AGE_35_44_MALE_WHITE	Number	8
AGE_45_54_MALE_WHITE	Number	8
AGE_55_64_MALE_WHITE	Number	8
AGE_20_24_FEMALE_WHITE	Number	8
AGE_25_29_FEMALE_WHITE	Number	8
AGE_30_34_FEMALE_WHITE	Number	8
AGE_35_44_FEMALE_WHITE	Number	8
AGE_45_54_FEMALE_WHITE	Number	8
AGE_55_64_FEMALE_WHITE	Number	8
AGE_20_24_MALE_BLACK	Number	8
AGE_25_29_MALE_BLACK	Number	8
AGE_30_34_MALE_BLACK	Number	8
AGE_35_44_MALE_BLACK	Number	8
AGE_45_54_MALE_BLACK	Number	8
AGE_55_64_MALE_BLACK	Number	8
AGE_20_24_FEMALE_BLACK	Number	8
AGE_25_29_FEMALE_BLACK	Number	8
AGE_30_34_FEMALE_BLACK	Number	8
AGE_35_44_FEMALE_BLACK	Number	8
AGE_45_54_FEMALE_BLACK	Number	8
AGE_55_64_FEMALE_BLACK	Number	8
AGE_20_24_MALE_ASIAN	Number	8
AGE_25_29_MALE_ASIAN	Number	8
AGE_30_34_MALE_ASIAN	Number	8
AGE_35_44_MALE_ASIAN	Number	8

Variable Name	Format	Length
AGE_45_54_MALE_ASIAN	Number	8
AGE_55_64_MALE_ASIAN	Number	8
AGE_20_24_FEMALE_ASIAN	Number	8
AGE_25_29_FEMALE_ASIAN	Number	8
AGE_30_34_FEMALE_ASIAN	Number	8
AGE_35_44_FEMALE_ASIAN	Number	8
AGE_45_54_FEMALE_ASIAN	Number	8
AGE_55_64_FEMALE_ASIAN	Number	8
AGE_20_24_MALE_OTHER	Number	8
AGE_25_29_MALE_OTHER	Number	8
AGE_30_34_MALE_OTHER	Number	8
AGE_35_44_MALE_OTHER	Number	8
AGE_45_54_MALE_OTHER	Number	8
AGE_55_64_MALE_OTHER	Number	8
AGE_20_24_FEMALE_OTHER	Number	8
AGE_25_29_FEMALE_OTHER	Number	8
AGE_30_34_FEMALE_OTHER	Number	8
AGE_35_44_FEMALE_OTHER	Number	8
AGE_45_54_FEMALE_OTHER	Number	8
AGE_55_64_FEMALE_OTHER	Number	8
AGE_25_64_BACH_GRAD	Number	8
AGE_25_64_BACH_GRAD_GTR10PERC	Number	8
TOTAL_LATIN	Number	8
LATIN_WHITE	Number	8
LATIN_BLACK	Number	8
LATIN_ASIAN	Number	8
LATIN_OTHER	Number	8

Appendix C EHR File Layouts

C.1 EHR Input File Layout

C.1.1 EHR GEO3 Data

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

Table 19. EHR Iut File Layout for GEO3-Level Programs, CSV 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 this may be an approximate value due to birth ata approximation)**	**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"**	**WHITE**
STATE_ABR	**Patient's residential state, two-letter state abbreviations**	**Character**	**Any postal abbreviation of state. See Appendix F for a list of possible values.**	**MI**
GEO3	**Either: Patient's residential a) county code (5 digits) or b) ZIP-3 (3 digits)**	**Number**	**Any numeric value**	**059**
WEIGHT_CATEGORY	**Patient's BMI Category. See section A.2 and A.12 for more information.**	**Character**	**Normal or Healthy Weight Obese (Class 1) Obese (Class 2) Obese (Class 3) Overweight Underweight**	**Overweight**
YEAR	**Year of the medical encounter**	**Number**	**Yyyy**	**2018**
DIABETES_SPECTRUM	**Patient's diabetes status. See section A.4 for more information.**	**Number**	**<blank> = No Evidence of Diabetes 1 = Prediabetes, 2 = Diabetes, or 3 = No Evidence of Diabetes**	**1**
SCD	**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**

Variable Name	Description	Format	Valid values	Example
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**

C.2 EHR Results File Layout for GEO3

Table 20. GEO3

Variable Name	Format	Example
SUBJID or PATID	Character	12345626
Ageyr	Number	16
AGE_CATEGORIES	Character	15 – 17
WTCAT	Character	(2) Healthy Weight (BMI greater than or equal to 18.5 and less than 25 kg/m2)
STATE_ALPHA	Character	MI
STATE_FIPS	Character	48
ZIP	Character	20184
GEO3	Character	100
Geography	Character	48100
DIABETES SPECTRUM	Number	1
SCD	Number	0
PREGNANCY_FLAG	Number	1
Race	Character	Black
Sex	Character	Female
Year	Number	2018
Imputed_Race	Character	Black
Race_Imputed	Number	0

Appendix D CODI-APQ-GEO3 Example SAS Programs

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

Appendix D.1 includes a program to generate a pre-processed file using the Quickstart pre-processing program. This example uses COUNTY data.

Text highlighted in yellow has been reviewed and approved or reviewed and edited from its original values. The program uses the data inputs: ACS_COUNTY, EHR_COUNTY. The file processes EHR data between 2016 and 2019 and creates a SAS file named CODI_APQ_COUNTY stored in the folder P:\Example\2_Output\Pre_Processed_CODI_APQ. The SAS log is stored in P:\Example\2_Output\SAS LOG\QS_Pre_Processing <plus date and time information>.log. Text between /* and */ are comments in SAS. Comments may vary slightly in CODI-APQ from the example below.

```
/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
/*************************************************************************************************************************************/
/******************* -- PREPROCESSING ALGORITHM USER INPUT SECTION (PLEASE COMPLETE SECTIONS 1-3 BELOW) -- **********************/
/***************** -- 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_APQ; /*@Note: Suffix name for EHR Output folder (ACCEPTABLE VALUES: folder name (no puctuation)***/
/***/ %LET ACS_FILENAME = ACS_COUNTY; /*@Note: ACS file name (ACCEPTABLE VALUES: file name, do not include ".csv")***/
/***/ %LET EHR_FILENAME = EHR_COUNTY; /*@Note: EHR file name (ACCEPTABLE VALUES: file name, do not include ".csv")***/
/***/ %LET LOG_NAME_PRE = QS_Pre_Processing; /*@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_APQ_COUNTY; /*@Note: EHR output file name (ACCEPTABLE VALUES: SAS file name (no punctuation)***/

/*SECTION 4: County or ZCTA3 data (REQUIRED)***/
/***/ %LET COUNTY = Y; /*@Note: County/ZCTA3 indicator (ACCEPTABLE VALUES: Y for County level data, N for ZCTA3 level data***/

/***Note: ROOT_PRE directory includes subfolders:
                                                        "..\0_Raw_Data"
                                                        "..\1_SAS_Programs"
                                                        "..\02_Output" and
                                                        "..\02_Output\SAS LOGS"***/
/***NOTE: SAS programs must be stored in the PROGS_PRE directory including:
                                                        Module0-Pre_Processing_CODI_APQ.sas
                                                        Module1-Pre_Processing_CODI_APQ.sas
                                                        Module2-Pre_Processing_CODI_APQ.sas
```

```
                                                                    Module3-Pre_Processing_CODI_APQ.sas***/
/*****************************************************************************************************************************************/
/***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. */
```

D.2 Generate Results Example with GEO3 Data

Appendix D.2 includes a program excerpt to generate prevalence results using the Quickstart program and data inputs: ACS_COUNTY and CODI_Ready in the folder P:\Example\2_Output\Pre_processed_CODI_APQ. This example uses COUNTY data. Words highlighted in yellow correspond with changes in the QuickStart to generate user defined results.

The file processes EHR data for a subpopulation and a given analysis and creates a csv file with output named "CODI_COUNTY_Prevalence_Query_Report_[DateTime].csv" stored in the folder P:\Example\2_Output\. The SAS log is stored in P:\Example\ 2_Output\SAS LOGS\CODI_APQ_LOG.log.

Subpopulation: EHR records from 2017 including patients 20 to 64 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 equals yes.

Methods: Includes diabetes prevalence estimates in the results, patients regardless of pregnancy status, includes actual and imputed race, and calculates the age-adjusted prevalence. Text between /* and */ are comments in SAS. Comments may vary slightly in CODI-APQ 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 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: base directory (ACCEPTABLE VALUES: computer directory name)  ***/
/***/ %LET PRE_DEST           = CODI_APQ;      /*@Note: Suffix name of pre-processing output folder (ACCEPTABLE VALUES: folder name
(no punctuations)) ***/
/***/ %LET EHR_PRE_OUT = CODI_APQ_COUNTY;      /*@Note: Suffix name of pre-processing output file (ACCEPTABLE VALUES: file name (no
punctuations)) ***/
/***/ %LET LOG_NAME           = CODI_APQ_LOG; /*@Note: Name for SAS log storage location                                    ***/
```

```
/***/ %LET FileOUT_Name    = CODI_COUNTY;             /*@Note: Output file name; MUST BE FEWER THAN 28 CHARACTERS              ***/

/*SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE OR STATE/COUNTY CODE          ***/
/***/ %LET BEG_YEAR            = 2017;  /*@Note: Beginning year of analysis       (ACCEPTED VALUES: 4-Digit year from EHR) ***/
/***/ %LET END_YEAR            = 2017;  /*@Note: End year of analysis      (ACCEPTED VALUES: 4-Digit year from EHR) ***/
/***/ %LET ALL_STATES   = N;     /*@Note: Include all geographical locations in file?       (ACCEPTED VALUES: Y/N) ***/
/***/ %LET ALL_AGES     = Y;     /*@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   = Y;          /*@Note: Is the ACS data at the county (Y) or ZCTA3 (N) level? (ACCEPTABLE VALUES: Y for County
level data, N for ZCTA3 level data) ***/
/***/ %LET INCLUDE_PREGNANCY = Y;     /*@Note: Include non-pregnant and 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)             ***/
/***/ %LET CO_OCCURRING = Y;          /*@Note: Include co-occurring conditions            (ACCEPTED VALUES: Y/N) ***/
/***/ %LET CO_OCCURRING_COND_VAR =DIABETES_SPECTRUM; /*@Note: Include co-occurring condition variable/field/column name***/

/*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    = COUNTY;       /*@Note: Level of geography (ACCEPTED VALUES: STATE, ZCTA3, or COUNTY)***/
       /***/ %LET GEO_LIST     = %STR('08059', '08125'); /*@Note: IF GEO_GROUP="STATE" then populate with State FIPS code(s), 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 AGE_20_24 = Y;      /*@Note: Age Range: 20 to 24 (ACCEPTED VALUES: Y/N)                                    ***/
       /***/ %LET AGE_25_29 = Y;      /*@Note: Age Range: 25 to 29 (ACCEPTED VALUES: Y/N)                                 ***/
       /***/ %LET AGE_30_34 = Y;      /*@Note: Age Range: 30 to 34 (ACCEPTED VALUES: Y/N)                                    ***/
       /***/ %LET AGE_35_44 = Y;      /*@Note: Age Range: 35 to 44 (ACCEPTED VALUES: Y/N)                                 ***/
       /***/ %LET AGE_45_54 = Y;      /*@Note: Age Range: 45 to 54 (ACCEPTED VALUES: Y/N)                                 ***/
       /***/ %LET AGE_55_64 = Y;      /*@Note: Age Range: 55 to 64 (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    = N;   /*@Note: Include patients with imputed race values? Y = imputed and EHR provided race, N = EHR
race only (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_APQ.sas,
                                                       Macro2-CODI_APQ.sas,
                                                       Macro3-CODI_APQ.sas,
                                                       Macro4-CODI_APQ.sas,
                                                       Module1-CODI_APQ.sas,
                                                       Module2-CODI_APQ.sas,
                                                       Macro1-CODI_APQ-Co_occurring.sas,
                                                       Macro2-CODI_APQ-Co_occurring.sas,
                                                       Macro3-CODI_APQ-Co_occurring.sas,
                                                       Macro4-CODI_APQ-Co_occurring.sas,
                                                       Module1-CODI_APQ-Co_occurring.sas,
                                                       Module2-CODI_APQ-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 STOP
STOP*/
/*** DO NOT EDIT BEYOND THIS POINT    DO NOT EDIT BEYOND THIS POINT        DO NOT EDIT BEYOND THIS POINT        DO NOT EDIT BEYOND
THIS POINT    */
/*************************************************************************************************************************************/
/*Note: subsection of the full program. Be sure to only edit this section but submit the full program. */
```

Appendix E CODI-APQ Results

E.1 Example BMI Category Prevalence

Once complete, CODI-APQ generate prevalence results as Excel files. Table 2121 provides an overview of the variables included when diabetes prevalence is set to no, and

Table 2222 provides example results based on synthetic data. Note in the sample provided in

Table 2323 descriptive information about CODI-APQ 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 output will vary based on the criteria selected.

Table 21. CODI-APQ Results Data Dictionary

Column	Description
Order	Row order
Weight Category	A categorical value based on BMI.
Sample	The observed (or unadjusted, or crude) count of adults 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 adults in the population represented by that sample patient. See Appendix A.6 Statistical Weights for more information.
Weighted Prevalence Standard Error	Standard error based on weighted counts. See Appendix A.11 Variance for more information.
Age-adjusted Prevalence	Prevalence based on weighted, age-adjusted counts. See Appendix A.1 Age Adjustment for more information.

Table 22. Results Example from Synthetic Data[29]

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 (BMI<18.5)	3,246	33,384	10.27	0.03	7.21	0.04	7.21	0.04
1	(2) Healthy Weight (18.5<=BMI<25)	8,190	159,717	25.90	0.13	34.50	0.16	34.50	0.16
1	(3) Overweight (25<=BMI<30)	10,374	143,974	32.81	0.14	31.10	0.17	31.10	0.17
1	(4) Obesity (Classes 1, 2, and 3) (BMI 30+)	9,811	125,831	31.03	0.15	27.18	0.08	27.18	0.09
1	(4a) Obesity (Class 1) (30<=BMI<35)	5,468	45,681	17.29	0.12	9.87	0.04	9.87	0.05
1	(4b) Obesity (Class 2) (35<=BMI<40)	1,354	67,213	4.28	0.1	14.52	0.02	14.52	0.02
1	(4c) Obesity (Class 3) - Severe Obesity (BMI 40+)	2,989	12,937	9.45	0.09	2.79	0.02	2.79	0.02
2	Totals:	31,621	462,906						
3	Query Version: CODI-APQ GEO3 2015-2019								
4	Query Parameters: AGE RACE SEX GEOGRAPHY YEAR								
5	AGE: (20 - 24, 25 - 29, 30 - 34, 35 - 44, 45 - 54)								
6	SEX: (Female)								
7	Pregnancy: included in analysis								
8	RACE: (White, Black, Asian)								
9	RACE Suppressed: (None)								
10	RACE Imputed: People with unknown race were excluded.								

[29] Note: borders and shading are for demonstration purposes only. Result column headers and content below the results may vary from that presented.

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
11	Geography: (99999)								
12	Year: 2019								
13	Weighting cells were collapsed for: (None)								
14	Age-Adjusted: (Yes)								
15	Error Codes: (One or more rows has suppressed results. Percentages are not available for all results due to suppression constraints.)								
16	Implementation Guide: See https://github.com/NORC-UChicago/CODI-PQ for more information and full details on data sources and calculations.								
17	Query Date: Wednesday, December 8, 2021 3:49:27 PM								
18	Suggested Citation: Tanenbaum, E., Campbell, S., Chelluri, D., Zalsha, S., Boim, J., Paddock, S., Copeland, K. (2021). Clinical and Community Data Initiative Adult Prevalence Queries (CODI-APQ) SAS programs (version 2015-2019).								
18	The Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare federally funded research and development center, Health FFRDC. Retrieved from https://github.com/NORC-UChicago/CODI-PQ on Wednesday, December 8, 2021 3:49:27 PM								
19	Caveats								
20	Patients with either missing or invalid age, sex, height, weight, or geography are not included in results.								
21	The standard error calculations are documented in the Implementation Guide.								

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
22	The population is calculated from age-race-sex-geography specific counts from the user provided American Community Survey Five-year Estimates.								
23	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-APQ methodology is appropriate for your use case when used outside of these date ranges.								
24	The age-adjusted prevalence calculations are documented in the Implementation Guide.								

Table 23. Example Results with Errors (insufficient sample size), Error Message Is Shown in Order = 15[30]

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 (BMI<18.5)								
1	(2) Healthy Weight (18.5<=BMI<25)								
1	(3) Overweight (25<=BMI<30)								
1	(4) Obesity (Classes 1, 2, and 3) (BMI 30+)								
1	(4a) Obesity (Class 1) (30<=BMI<35)								
1	(4b) Obesity (Class 2) (35<=BMI<40)								
1	(4c) Obesity (Class 3) - Severe Obesity (BMI 40+)								
2	Totals:								
3	Query Version: CODI-APQ GEO3 2015-2019								
4	Query Parameters: AGE RACE SEX GEOGRAPHY YEAR								
5	AGE: (20 - 24)								
6	SEX: (Female)								
7	Pregnancy: excluded from analysis								
8	RACE: (Black, Asian)								
9	RACE Suppressed: (Error)								
10	Imputed race: People with unknown race were excluded.								

[30] Note: exact text and row orders vary based on user specifications.

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
11	Geography: (51221)								
12	Year: 2019								
13	Weighting cells were collapsed for: (Error)								
14	Age-Adjusted: (No)								
15	Error Codes: (Current selection returns an insufficient number of patients and does not meet the minimum threshold to calculate sample weights. Ensure that selections are correct (e.g., correct list of state codes or GEO3 values) or include additional geographic or demographic categories (e.g., add additional communities or include additional or all races, age groups, sex, etc.).)								
16	Implementation Guide: See https://github.com/NORC-UChicago/CODI-PQ for more information and full details on data sources and calculations.								
17	Query Date: Wednesday, December 8, 2021 5:18:02 PM								

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
18	Suggested Citation: Tanenbaum, E., Campbell, S., Chelluri, D., Zalsha, S., Boim, J., Paddock, S., Copeland, K. (2021). Clinical and Community Data Initiative Adult Prevalence Query (CODI-APQ) SAS programs (version 2015-2019).								
18	The Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare federally funded research and development center, Health FFRDC. Retrieved from https://github.com/NORC-UChicago/CODI-PQ on Wednesday, December 8, 2021 5:18:02 PM								
19	Caveats								
20	Patients with either missing or invalid age, sex, height, weight, or geography are not included in results.								
21	The standard error calculations are documented in the Implementation Guide.								

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
22	The population estimates are based on age-race-sex-location specific counts from the 2015-2019 American Community Survey Five-year Estimates released by the Census Bureau on December 9, 2020.								
23	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-APQ methodology is appropriate for your use case when used outside of these date ranges.								

E.2 Diabetes Prevalence

CODI-APQ generate optional diabetes prevalence results as an Excel file. Table 24 provides an overview of the Excel Worksheets and

Table 25 provides results by Worksheet class and variables (Excel cells) generated when the co-occurring condition is set to yes. Variable names repeat between Excel Worksheets. Inclusion criteria determines the sample represented within the Excel Worksheet. Note, descriptive information about CODI-APQ user inputs, error codes, sources of technical documentation, caveats, and a possible citation begins with the row labeled order 3 and continues through order 24. The number of rows output will vary based on the criteria selected.

Table 24. CODI-APQ Diabetes Prevalence Results, Description of Excel Worksheets

Excel Worksheet	Worksheet Class	Description and Inclusion Criteria
Counts	Counts	Provides crude and weighted counts. Does not include prevalence estimates. Inclusion: all patients.
Weight Category	BMI prevalence	Provides results as described in Section 2.5.5 Review BMI prevalence Results Inclusion: all patients.
Diabetes Spectrum	Diabetes prevalence	Provides results for diabetes counts and prevalence. Inclusion: all patients.
Diabetes Underweight	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = underweight. Inclusion: patients with a BMI equal to underweight.
Diabetes Healthy Weight	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = healthy weight. Inclusion: patients with a BMI equal to healthy weight.
Diabetes Overweight	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = overweight. Inclusion: patients with a BMI equal to overweight.
Diabetes Obese	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = obese. Inclusion: patients with a BMI equal to obesity.
Diabetes Severe Obese	BMI by Diabetes prevalence	Provides results for diabetes counts and prevalence for all patients identified as BMI = obese class I. Inclusion: patients with a BMI equal to severe obesity.

Table 25. CODI-APQ Diabetes Prevalence Results Data Dictionary

Worksheet Class	Column	Description
Counts	Order	Row order
Counts	Condition	Co-Occurring Condition based on the Diabetes Spectrum (no evidence of diabetes, pre-diabetes, diabetes).
Counts	Weight Category	The weight category based on BMI.
Counts	Sample	The observed (also known as unadjusted, or crude) count of adults in the study population.
Counts	Population	The weighted (or adjusted) count of the study population.

Worksheet Class	Column	Description
Counts	Population (Age Adjusted)	The weighted, age-adjusted count of the study population (optional).
BMI prevalence	Order	Row order
BMI prevalence	Weight Category	The weight category based on BMI.
BMI prevalence	Sample	The observed (also known as unadjusted, or crude) count of adults in the study population.
BMI prevalence	Population	The weighted (or adjusted) count of the study population.
BMI prevalence	Population (Age-Adjusted)	The weighted, age-adjusted count of the study population (optional).
BMI prevalence	Crude Prevalence	The observed (or unadjusted, or crude) prevalence in the study population.
BMI prevalence	Crude Prevalence Standard Error	The observed (or unadjusted, or crude) standard error in the study population.
BMI prevalence	Weighted Prevalence	Prevalence based on weighted counts. A sample weight is assigned to each sampled patient. It is a measure of the number of adults in the population represented by that sample patient. See Appendix A.6 Statistical Weights for more information.
BMI prevalence	Weighted Prevalence Standard Error	Standard error based on weighted counts. See Appendix A.11 Variance for more information.
BMI prevalence	Age-Adjusted Prevalence	Prevalence based on weighted, age-adjusted counts (optional). See Appendix A.1 Age Adjustment for more information.
BMI prevalence	Age-Adjusted Prevalence Standard Error	Standard error based on weighted, age-adjusted counts (optional). See implementation guide, See Appendix A.1 Age Adjustment for more information.
Diabetes prevalence, BMI by diabetes prevalence	Order	Row order
Diabetes prevalence, BMI by diabetes prevalence	Condition	Co-Occurring Condition based on the Diabetes Spectrum.
Diabetes prevalence, BMI by diabetes prevalence	Sample	The observed (or unadjusted, or crude) count of adults in the study population.
Diabetes prevalence, BMI by diabetes prevalence	Population	The weighted (or adjusted) count of the study population.
Diabetes prevalence, BMI by diabetes prevalence	Population (Age-Adjusted)	The weighted, age-adjusted count of the study population (optional).
Diabetes prevalence, BMI by diabetes prevalence	Crude Prevalence	The observed (or unadjusted, or crude) prevalence in the study population.
Diabetes prevalence, BMI by diabetes prevalence	Crude Prevalence Standard Error	The observed (or unadjusted, or crude) standard error in the study population.
Diabetes prevalence, BMI by diabetes prevalence	Weighted Prevalence	Prevalence based on weighted counts. A sample weight is assigned to each sampled

Worksheet Class	Column	Description
		patient. It is a measure of the number of adults in the population represented by that sample patient. See Appendix A.6 Statistical Weights for more information.
Diabetes prevalence, BMI by diabetes prevalence	Weighted Prevalence Standard Error	Standard error based on weighted counts. See Appendix A.11 Variance for more information.
Diabetes prevalence, BMI by diabetes prevalence	Age-Adjusted Prevalence	Prevalence based on weighted, age-adjusted counts (optional). See Appendix A.1 Age Adjustment for more information.
Diabetes prevalence, BMI by diabetes prevalence	Age-Adjusted Prevalence Standard Error	Standard error based on weighted, age-adjusted counts (optional). See Appendix A.1 Age Adjustment for more information.

E.3 Possible Result Errors

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

Table 26. CODI-APQ 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).
Years are out of scope. CODI-APQ were developed between 2019 and 2021, see Implementation Guide for more details.	Starting year cannot be before 2000, ending year cannot be after 2030. CODI-APQ may be inappropriate to implement on medical encounters outside of 2015 through 2021. Please review the methodology in full to determine whether CODI-APQ are 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; /***/ %LET GEO_LIST = %STR('08', ‘10’); If GEO_GROUP = ZCTA3; /***/ %LET GEO_LIST = %STR('51221', ‘51224’); If GEO_GROUP = COUNTY; /***/ %LET GEO_LIST = %STR('08001', ‘08002’);

Error	Description
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).
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 27: CODI-APQ 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-APQ were developed between 2019 and 2021 and tested with EHR from 2015 through 2019. Please review the Implementation Guide in full to determine whether CODI-APQ methodology is appropriate for your use case when used outside of these date ranges.	Users may choose to employ CODI-APQ outside of the testing period. It is recommended that the user carefully review all methods prior to doing so.

E.4 Example Sample Checker Results

CODI-APQ generate optional sample size checker results in the SAS output or results window. Table 28 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 age 20 – 24, 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 because results show that male category has an insufficient number of patients, as well as Asian. CODI-APQ automatically remove racial categories if the sample size is insufficient (see A.6 Statistical Weights) but does not remove male or female automatically.

Table 28: CODI-APQ Sample Size Checker Results

Checker Results	Factor	Value
Sample Size Is Sufficient	Age Categories	20 - 24
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 F State FIPS codes

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

Table 29: State FIPS Codes

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

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

Appendix G Glossary

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

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

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 20 to 24, 25 to 29, 30 to 34, 35 to 44, 45 to 54, and 55 to 64 years of age.

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

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-APQ – CODI prevalence queries (CODI_APQ in SAS programs)

CODI-APQ-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 – is the total number of people within a particular BMI (e.g., underweight) 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

Diabetes Spectrum - Categorization of a person's diabetes and prediabetes status into one of three categories: no evidence of diabetes, prediabetes, or diabetes.

Diabetes – A category in the diabetes spectrum. Both Type I and Type II diabetes are included.

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 greater than or equal to 18.5 and less than 25

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.

No Evidence of Diabetes - Part of the diabetes spectrum, lacking sufficient evidence that the patient has prediabetes or diabetes.

Obesity – Body Mass Index greater than or equal to 30

Obesity Class 1 – Body Mass Index greater than or equal to 30 and less than 35

Obesity Class 2 – Body Mass Index greater than or equal to 35 and less than 40

Obesity Class 3 – Body Mass Index greater than or equal to 40

Overweight - Body Mass Index greater than or equal to 25 and less than 30

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 Networks

Prediabetes – A category in the diabetes spectrum based on a phenotype

Pre-Processing CODI-APQ – 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 adults 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 adults in the study population.

SDOH – Social Determinants of Health

Statistical Weights - A statistical weight is an amount given to increase or decrease the importance of an item. Weights are commonly given for people when a sample and not a census is taken. The value of the weight can be thought of as denoting the number of adults 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.

Underweight - Body Mass Index value less than 18.5

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) into one of five categories: underweight, healthy weight, overweight, obesity (class 1, 2, or 3).

Weights – See Statistical Weights or Weight Category

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

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 H Abbreviations and Acronyms

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

Appendix I Bibliography

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

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

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

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

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

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

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.

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

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

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 document 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.