|  |  |
| --- | --- |
| MITRE Logo |  |
|  | **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 1.0** |
| The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.  **For Internal MITRE Use.** This document was prepared for authorized distribution only. It has not been approved for public release.  ©2021 The MITRE Corporation.  All rights reserved. | **November 1, 2021** |

Record of Changes

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author / Owner | Description of Change |
| DRAFT | November 1, 2021 | Erin Tanenbaum / Health FFRDC | Initial Version |

**Table of Contents**

[1 Introduction 6](#_Toc86246311)

[1.1 Background 6](#_Toc86246312)

[1.2 Purpose 7](#_Toc86246313)

[1.3 Scope 7](#_Toc86246314)

[1.4 Audience 7](#_Toc86246315)

[1.5 Document Organization 7](#_Toc86246316)

[2 User’s Guide 8](#_Toc86246317)

[2.1 CODI Concept 8](#_Toc86246318)

[2.2 About CODI-APQ 9](#_Toc86246319)

[2.3 SAS Setup 10](#_Toc86246320)

[2.4 Step-by-step Process to run CODI-APQ 10](#_Toc86246321)

[2.4.1 STEP 1: Download and unzip CODI-APQ-master.zip file 10](#_Toc86246322)

[2.4.2 STEP 2: Obtain input files and store them in the ‘0\_Raw\_Data’ folder 11](#_Toc86246323)

[2.4.3 STEP 3: Link Population (Pre-processing) 12](#_Toc86246324)

[2.4.4 STEP 4: Generate Results 14](#_Toc86246325)

[2.4.5 Review BMI prevalence Results 21](#_Toc86246326)

[2.5 Additional Details for Users 24](#_Toc86246327)

[Appendix A Analysis Details 25](#_Toc86246328)

[Appendix B ACS File Layouts 39](#_Toc86246329)

[Appendix C EHR File Layouts 49](#_Toc86246330)

[Appendix D CODI-APQ-GEO3 Example SAS Programs 52](#_Toc86246331)

[Appendix E CODI-APQ Results 56](#_Toc86246332)

[Appendix F State FIPS codes 66](#_Toc86246333)

[Appendix G Glossary 68](#_Toc86246334)

[Appendix H Abbreviations and Acronyms 72](#_Toc86246335)

[Appendix I Bibliography 73](#_Toc86246336)

List of Figures

[Figure 1. Data Partners with a Common Data Coordinating Center 4](#_Toc85790550)

[Figure 2. CODI-APQ Process 5](#_Toc85790551)

[Figure 3. CODI-APQ-GEO3 Folder Structure 6](#_Toc85790552)

[Figure 4. Comparison of Patients with Plausible Weight and Height Data in the Electronic Health Records and the US Population by Race, Sex, and Age Group, 2015 25](#_Toc85790553)

[Figure 5. Comparison of Patients with Plausible Weight and Height Data in the Electronic Health Records and the US Population by Age and Sex, 2015 26](#_Toc85790554)

[Figure 6. NCHS Suppression Standards 33](#_Toc85790555)

List of Tables

[Table 1. Change Specifications, Processing Steps. 7](#_Toc85789247)

[Table 2. Change Specifications, SAS Macro Variables 8](#_Toc85789248)

[Table 3. Change Specifications, Processing Steps, Continued 8](#_Toc85789249)

[Table 4. Pre-processing CODI-APQ Program Execution Processing Steps 9](#_Toc85789250)

[Table 5. Change Specifications, Processing Steps 10](#_Toc85789251)

[Table 6. Change Specifications, Macro Variables 10](#_Toc85789252)

[Table 7. Change Specifications, Processing Steps, Continued 14](#_Toc85789253)

[Table 8. Change Specifications, SAS Macro Variables 14](#_Toc85789254)

[Table 9. Change Specifications, Processing Steps, Continued 16](#_Toc85789255)

[Table 10. CODI-APQ Execution Processing Steps 16](#_Toc85789256)

[Table 11. CODI-APQ BMI prevalence Results Data Dictionary 16](#_Toc85789257)

[Table 12. CODI-APQ Diabetes Prevalence Results, Description of Excel Worksheets 17](#_Toc85789258)

[Table 13. CODI-APQ Diabetes Prevalence Results Data Dictionary 18](#_Toc85789259)

[Table 14. Projected Year 2000 U.S. Population Proportion Distribution by Age for Age Adjusting 21](#_Toc85789260)

[Table 15. American Community Survey Census Sample by Year for 2015-2019 for Adults Aged 2 to 19 24](#_Toc85789261)

[Table 16. Comparison of Patients with Plausible Weight and Height Data in The Electronic Health Records and the US Population by Age, 2015 24](#_Toc85789262)

[Table 17. Proportions of Sickle Cell Disease Used to Impute Race 27](#_Toc85789263)

[Table 18. Percentage of Patients Imputed for Each Phase in the Race Imputation Using AEMR Data 29](#_Toc85789264)

[Table 19. NCHS Data Presentation Standards for Proportions 31](#_Toc85789265)

[Table 20. Included Concepts: Social Determinants of Health 38](#_Toc85789266)

[Table 21. Additional Concepts Considered: Social Determinants of Health 39](#_Toc85789267)

[Table 22. Final List of Prioritized Social Determinants of Health 41](#_Toc85789268)

[Table 23. Social Determinants of Health: Summary of Findings 43](#_Toc85789269)

[Table 24. Additional Concepts and Measures Considered: Social Determinants of Health 45](#_Toc85789270)

[Table 25. Social Determinants of Health Measures 45](#_Toc85789271)

[Table 26. Social Determinants of Health: Bivariate Ranked Concepts and Measures 47](#_Toc85789272)

[Table 27. Social Determinants of Health: Multivariable Ranked Measures 48](#_Toc85789273)

[Table 28. ACS Input File Layout, CSV File 50](#_Toc85789274)

[Table 29. ACS Pre-processing Results File Layout – GEO3 58](#_Toc85789275)

[Table 30. EHR Input File Layout for GEO3-Level Programs, CSV File 60](#_Toc85789276)

[Table 31. GEO3 63](#_Toc85789277)

[Table 32. CODI-APQ Results Data Dictionary 68](#_Toc85789278)

[Table 33. Results Example from Synthetic Data 69](#_Toc85789279)

[Table 34. Example Results with Errors (insufficient sample size), error messages are shown in row order 71](#_Toc85789280)

[Table 35. Example Results with Errors (out of scope years), error messages are shown in row order 74](#_Toc85789281)

[Table 36. CODI-APQ Diabetes Prevalence Results, Description of Excel Worksheets 76](#_Toc85789282)

[Table 37. CODI-APQ Diabetes Prevalence Results Data Dictionary 77](#_Toc85789283)

[Table 38. State FIPS Codes 80](#_Toc85789284)

# 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, and the Center for Surveillance, Epidemiology, and Laboratory Services 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)](https://www.cdc.gov/obesity/initiatives/codi/community-and-clinical-data-initiative.html). CODI will expand the ability to capture, standardize, integrate, and query existing patient-level electronic health record (EHR) and non-medical community data. To accomplish this, open-access programs were created to support research and program evaluation on prevalence of adult weight-status categories and diabetes, using a distributed health data network (DHDN) infrastructure. CODI will expand the traditional DHDN infrastructure[[1]](#footnote-2) to include data from both clinical data partners and non-clinical, community-based data partners.

The Health FFRDC developed open-access programs, referenced here as the CODI adult prevalence queries (CODI-APQ) to generate weight-status-category prevalence based on body mass index (BMI)[[2]](#footnote-3) and diabetes prevalence in adults, aged 20 through 64 stratified by age, sex, and geography. CODI-APQ were designed to use data from the CODI DHDN and other non-probability samples derived from Electronic Health Records (EHRs).

## Background

Public health surveillance of adult obesity often relies on self-report surveys or by proxy, which are subject to bias. These surveys can be expensive to administer, limited in geographic specificity, and may struggle with response rates and timeliness. Data from EHRs have the potential to play a significant role in obesity population health surveillance, programs, interventions, and evaluations. While EHR data are non-probability samples of health outcomes among the care-seeking population, the data – measurements, diagnoses, observations, prescriptions, and procedures – offer an opportunity to provide decision makers with detailed, timely, and accurate information at the local, state, and regional levels for surveillance, research, and evaluation.

Population estimates for surveillance can be obtained by applying statistical weights, imputation, and suppression criteria to EHR data. Applying statistical weights can reduce potential biases introduced by the EHR sampling. CODI-APQs provide a tool for calculating prevalence estimates by weighting user provided EHR or DHDN data for a target population. 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 use, with the aim to improve access to timely data on local disease burden to inform prevention and other public health activities.

## Purpose

The purpose of the CODI-APQ Implementation Guide is to provide a guide for CODI data partners[[3]](#footnote-4) or data users to run the CODI-APQ. The Implementation Guide covers the following:

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

## Scope

CODI-APQ were created and tested with IQVIA's Ambulatory Electronic Medical Record (AEMR-US)[[4]](#footnote-5) data and synthetic data generated for CODI using Synthea.[[5]](#footnote-6) CODI-APQ require patient level records for patients ages 20 through 64. Each record must include year of medical encounter, and demographic information, BMI category. Patient-level records must include residential address information at the level of state, county, and zip code, or at the level of state and the ZIP Code Tabulation Area’s first three digits (ZCTA-3). CODI-APQ leverages population counts from the American Community Survey. CODI-APQ assumes end users include all available EHR data for a geography and/or subpopulation.

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.

## Audience

The audience for this IG is CODI data partners and data users. The user should have a working knowledge of SAS 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 D.

## 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 |

# User’s Guide

The User’s Guide section describes

1. The CODI Project
2. How to prepare your data for the programs
3. How to run the CODI-APQ programs
4. The CODI-APQ output

## CODI Concept

Figure 1 shows how CODI users (e.g., researchers, community-based program evaluators) interact with the data coordinating center, which distributes their research queries to data partners. The data coordinating center assembles the results into longitudinal records, which are sent to the CODI end-users. CODI end-users use the patient-level longitudinal records to create prevalence estimates with CODI-APQ. Additional CODI details can be found in the documentation available through GitHub at https://github.com/mitre/codi.

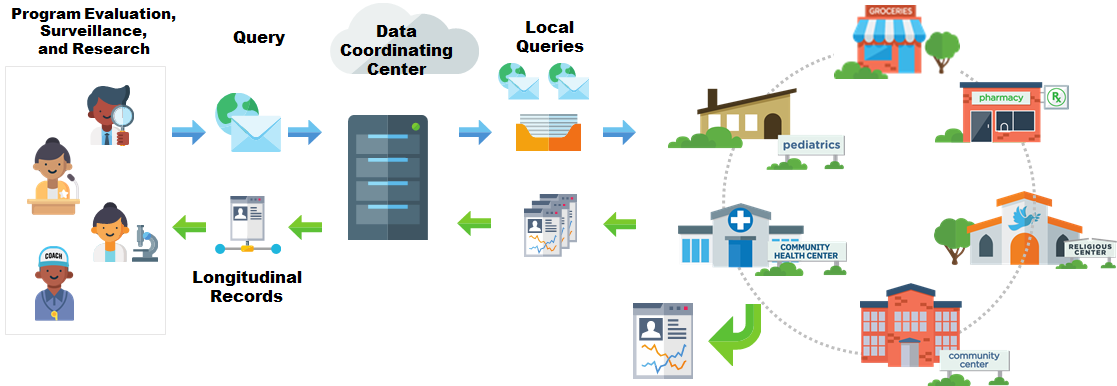


Figure . Data Partners with a Common Data Coordinating Center

## About CODI-APQ

CODI-APQ is a set of programs that calculate adult BMI and diabetes prevalence estimates from a non-probability sample[[6]](#footnote-7) of EHRs.

CODI-APQ programs require careful review of the methodological details (described in appendices), EHRs supplied by the user, and ACS data from 2019 supplied by the Health FFRDC[[7]](#footnote-8). 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.

Results are suppressed if the user selects a geography or subpopulation with an insufficient number of patients for statistical weighting or if results do not meet NCHS suppression criteria. The CODI-APQ programs user should have a working knowledge of SAS 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 missing race information (optional),
* Calculate statistical weights with an EHR non-probability sample,
* Calculate age-adjusted prevalence results (optional),
* Calculate adult BMI prevalence by BMI, including:
* **Underweight**: BMI less than 18.5
* **Healthy Weight**: BMI greater than or equal to 18.5 and less than 25
* **Overweight**: BMI greater than or equal to 25 and less than 30
* **Obesity**: BMI greater than or equal to 30
  + **Obesity Class 1**: BMI greater than or equal to 30 and less than 35
  + **Obesity Class 2**: BMI greater than or equal to 35 and less than 40
  + **Obesity Class 3**: BMI greater than or equal to 40
* Suppress prevalence estimates based on the National Center for Health Statistics (NCHS) Data Presentation Standards for Proportions.

## 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 requires 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

## Step-by-step Process to run CODI-APQ

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

Figure . CODI-APQ Process

### STEP 1: Download and unzip CODI-APQ-master.zip file

Access CODI-APQ programs on GitHub: [https://github.com/NORC-UChicago/CODI-APQ](https://github.com/NORC-UChicago/CODI-PQ).

Select 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 Winzip or other available software to unzip the files. Be sure the option is selected to unzip both files and folders and preserve the folder names.

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.

Text

Description automatically generated

Figure . CODI-APQ-GEO3 Folder Structure

### STEP 2: Obtain input files and store them in the ‘0\_Raw\_Data’ folder

Required input files include:

1. **ACS data file** (downloaded from <https://sft.mitre.org/#/folder/6281923>) The 2019 ACS data can be downloaded from the Health FFRDC Secured File Transfer Protocol (SFTP) site to ensure consistency with the models embedded into the SAS programs.
2. **EHR file** supplied by the end user in comma separated values (.csv). The EHR file must:

* Have a unique identifier for all patients and the identifier is consistent between years.
* Include a **maximum** of one record per patient per year.
* Include a valid height and weight value obtained on the same day which was used to calculate the BMI for all patients (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[[8]](#footnote-9).
* Have reconciled demographic characteristics across years, including:
  + Sex
  + Race

Additional variables on the file must be included in the data file but may be left blankas they are optional and not required. These variables include:

* A sickle cell disease[[9]](#footnote-10) indicator
* A pregnancy indicator
* A diabetes spectrum indicator (prediabetes, diabetes, no evidence of diabetes).

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

### STEP 3: Link Population (Pre-processing)

Open the “Quickstart-Pre\_Processing\_CODI\_APQ\_GEO3” SAS program stored in “\1\_SAS\_Programs\Pre-Processing” and change the selection in Section 1, 2, and 3 per the steps outlined in the tables below.

Note that the pre-prevalence program should be submitted once and only once per file. As such, include the start and end years for the full file (three years maximum). 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 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 . Change Specifications, Processing Steps.

| **Order** | **Description** | **Details** |
| --- | --- | --- |
| **1** | Open the Pre-processing Quickstart program | The Quickstart program is stored in the folder:  "..\1\_SAS\_Programs\Pre-Processing” |
| **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: |

Table . Change Specifications, SAS Macro Variables

|  |  |  |
| --- | --- | --- |
| **SAS Macro Variable** | **Details** | **Example** |
| ROOT\_PRE | The core folder name where CODI-APQ-GEO3 are saved (see part 2.1.1). | %let ROOT\_PRE = P:\CODI-APQ-master\CODI-APQ-GEO3; |
| PROGS\_PRE | The folder name of the pre-processing SAS programs (e.g. core folder name plus “\1\_SAS\_Programs\Pre-Processing”). | %let PROGS\_PRE = &ROOTPRE.\1\_SAS\_Programs\Pre-Processing;    /\* for programs stored in P:\CODI-APQ-master\CODI-APQ-GEO3\1\_SAS\_Programs\Pre-Processing \*/ |
| PRE\_DEST | The folder name for results inside the “\2\_Output” | %let PRE\_DEST = CODI\_APQ\_GEO3;    /\* output would be stored in P:\CODI-APQ-master\CODI-APQ-GEO3\2\_Output\Pre\_Processed\_CODI\_APQ\_GEO3\*/ |
| ACS\_FILENAME | The American Community Survey file name from part 2.2 or 2.3. The file is in csv format. Do not include the extension in the file name. | %let ACS\_FILENAME = ACS\_State\_COUNTY;  Or  %let ACS\_FILENAME = ACS\_State\_ZCTA3; |
| EHR\_FileNAME | The adult level EHR file from part 2.2 or 2.3. Do not include the extension. | %let EHR\_FileNAME = EHR\_filename; |
| LOG\_NAME\_PRE | The name of the SAS log file. Quickstart\_Pre\_Processing\_CODI\_APQ. Users have the option to rename the log file name before it is created. | %let LOG\_NAME\_PRE = LogName;    /\*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\*/ |

Table . Change Specifications, Processing Steps, Continued

| **Order** | **Description** | **Details** |
| --- | --- | --- |
| **3** | Edit the SAS program within “SECTION 2: Beginning and End Year of longitudinal EHR data” | Edit the start and end year to include as appropriate, include all years within your file as ***pre-processing start and end years should include all years on the file***. In contrast, the prevalence program will request the specific year for each prevalence analysis (Part 4 below). |
| **4** | Edit the SAS program within “SECTION 3: Optional Results File Name Suffix” | The name of the results file. For example:  %LET EHR\_PRE\_Out = NEW;  /\*following the same example as above, the results from pre-processing will be stored as a SAS data file in P:\CODI-APQ-master\CODI-APQ-GEO3\2\_Output\ Pre\_Processed\_CODI\_APQ\_GEO3 \*/ |
| **5** | Edit the SAS program within “Section 4: County or ZCTA3 data (REQUIRED)” | Edit with a Y or N to indicate if the file is at the county or ZCTA-3 level, respectively. |
| **6** | Save the Quickstart program. | SAS encourages saving all files before submitting the program. |

Table . Pre-processing CODI-APQ Program Execution Processing 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, warning, and uninitialized. Assuming no errors, continue to Part 4. In the event of errors, reassess the location of the files and the file formats. |

### STEP 4: Generate Results

Open the “Quickstart-CODI\_APQ\_GEO3” SAS program stored in “\1\_SAS\_Programs\Adult Prevalence Query” and change the selections in Section 1, 2, and 3 within the program per the steps outlined in the table below.

The final results (CODI-APQ results) will be generated in .csv or Excel format and saved in “\2\_Output.” Appendix E provides an example of the results. Note that results are for the group of patients selected by the user. To calculate results for multiple geographic or demographic characteristics (e.g., first for 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 . Change Specifications, Processing Steps

|  |  |  |
| --- | --- | --- |
| **Order** | **Description** | **Details** |
| **1** | Open the Quickstart program | The Quickstart program is stored in the folder:  "..\1\_SAS\_Programs\CODI PQ” |
| **2** | Edit the SAS program within  “SECTION 1: Input Folder and file names”;  “SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE” | Follow the SAS programs and update the macro variable specifications. |

Table . Change Specifications, Macro Variables

| **SAS Macro Variable** | **Details** | **Example** |
| --- | --- | --- |
| **SECTION 1: Input Folder and file names** | | |
| ROOT\_PQ | The core folder name (see section 2.1.1). | %let ROOT\_PQ = P:\ CODI-APQ-master\CODI-APQ-GEO3; |
| PROGS\_PQ | The folder name of CODI-APQ SAS programs. | %let PROGS\_PQ= &ROOT\_PQ.\1\_SAS\_Programs\Adult Prevalence Query;    /\* for programs saved in P:\ CODI-APQ-master\CODI-APQ-GEO3\1\_SAS\_Programs\Adult Prevalence Query \*/ |
| PRE\_DEST | The folder name for pre-processed data which is the input for CODI-APQ programs. This should equal the same value as PRE\_DEST in section 3.1=. | %let PRE\_DEST = CODI\_APQ\_GEO3;    /\* output would be stored in P:\CODI-APQ-master\CODI-APQ-GEO3\2\_Output\Pre\_Processed\_CODI\_APQ\_GEO3\*/ |
| EHR\_PRE\_OUT | The adult level EHR file from section 3.1 and 3.2. This should equal the same value as what was specified in section 3.1. | %let EHR\_PRE\_OUT = NEW;    /\*following the same example as above, the output from pre-processing will be stored as a SAS data file in  P:\CODI-APQ-master\CODI-APQ-GEO3\2\_Output\ Pre\_Processed\_CODI\_APQ\_ZC \*/ |
| LOG\_NAME | The name of the resulting SAS log. Users have the option to rename the log file name before it is created. | %let LOG\_NAME = LogName;  /\*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\*/ |
| FileOUT\_Name | The prefix for the resulting .csv or Excel file. | %LET FileOUT\_Name = File\_name;  /\*the .csv or Excel file will be stored in: P:\CODI-APQ-master\CODI-APQ-GEO3\2\_Output. Note, the program automatically includes the date and time in all results file names\*/ |
| **SAS Macro Variable** | Details | Example |
| **SECTION 2: Subset data based on specifications INCLUDING YEAR, GEOGRAPHY, STATE** | | |
| BEG\_YEAR | Subsets the prevalence to medical encounters in this year for diabetes or BMI prevalence. The prevalence will include adult EHRs from this year and after. | /\*\*\*/ %LET BEG\_YEAR = 2016; /\*@Note: Beginning year of analysis (ACCEPTED VALUES: 4-Digit numeric, 2015-2019) \*\*\*/ |
| END\_YEAR | Subsets the prevalence to medical encounters through this year for diabetes or BMI prevalence. The prevalence will include adult EHRs from this year and before. | /\*\*\*/ %LET END\_YEAR = 2016; /\*@Note: End year of analysis (ACCEPTED VALUES: 4-Digit numeric, 2015-2019) \*\*\*/ |
| GEO\_GROUP | At what level will the prevalence be created? | /\*\*\*/ %LET GEO\_GROUP = STATE; /\*@Note: Level of geography (ACCEPTED VALUES: STATE/ZCTA3/COUNTY) \*\*\*/ |
| GEO\_LIST | If GEO\_GROUP= STATE; then by default the program will include all geographical locations listed. If GEO\_GROUP= COUNTY; then by default the program will subset the prevalence based on the state+county codes listed. If GEO\_GROUP= ZCTA3; then by default the program will subset the prevalence based on the state+ZCTA-3 codes listed. | **If** GEO\_GROUP= STATE**;** /\*\*\*/ %LET GEO\_LIST = %STR(‘08’);  **If** GEO\_GROUP= COUNTY**;** /\*\*\*/ %LET GEO\_LIST = %STR(‘08059’, ’08125’);  **If** GEO\_GROUP= ZCTA3**;** /\*\*\*/ %LET GEO\_LIST = %STR(‘01350’, ’08805’);    /\*@Note: IF STATE="Y" then generate with State FIPS code(s), otherwise generate with FIPS State+FIPS County code(s) (ACCEPTED VALUES: 2-digit state FIPS or 5-digit state FIPS+county FIPS (Must be surrounded by single quotation and comma delimited)) \*\*\*/ |
| ALL\_STATES | Includes all states (including D.C.) 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) | /\*@Note: Include all geographical locations in file? (ACCEPTED VALUES: Y/N) \*\*\*/ |
| ACSCOUNTY | At what level is the ACS file? County (Y) or ZCTA-3 (N) | /\*\*\*/ %LET ACSCOUNTY = N; /\*@Note: Is the ACS data at the county or ZCTA3 level? (ACCEPTABLE VALUES: Y for County level data, N for ZCTA3 level data \*\*\*/ |
| 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; /\*@Note: Include all age ranges? (ACCEPTED VALUES: Y/N) \*\*\*/ |
| 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; /\*@Note: Include all sex values? (ACCEPTED VALUES: Y/N) \*\*\*/ |
| 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 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; /\*@Note: Include all race categories? (ACCEPTED VALUES: Y/N) \*\*\*/ |
| **SECTION 3: Additional Flags** | | |
| ACS\_COUNTY | Specifies the geographic level of the ACS data | /\*\*\*/ %LET ACSCOUNTY = N; /\*@Note: Is the ACS data at the county or ZCTA3 level? (ACCEPTABLE VALUES: Y for County level data, N for ZCTA3 level data) \*\*\*/ |
| INCLUDE\_PREGNANCY | Excludes all patients from the prevalence where the pregnancy flag is equal to 1. | /\*\*\*/ %LET INCLUDE\_PREGNANCY = Y; /\*@Note: Include pregnant patients (ACCEPTED VALUES: Y/N) \*\*\*/ |
| SAMPLE\_CHECK | Executes an optional review of the sample size by strata. All strata with insufficient sample size (n = 20) will be displayed in the SAS output window. | /\*\*\*/ %LET SAMPLE\_CHECK = Y; /\*@Note: Run check for sufficient sample size of query (ACCEPTED VALUES: Y/N) \*\*\*/ |
| CO\_OCCURRING | Set to yes if diabetes prevalence is requested. | /\*\*\*/ %LET CO\_OCCURRING = Y; /\*@Note: Include co-occurring conditions (ACCEPTED VALUES: Y/N) \*\*\*/ |
| CO\_OCCURING\_COND\_VAR | Set this equal to the variable on the file with the diabetes variable. | /\*\*\*/ %LET CO\_OCCURING\_COND\_VAR =DIABETES\_SPECTRUM; /\*@Note: Include co-occurring condition variable \*\*\*/ |

Table . Change Specifications, Processing Steps, 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” | **EHR\_PRE\_Out** by default, the EHRs file after pre-processing will be named CODI\_APQ. Users have the option to rename this file before it is created. |

Table . Change Specifications, SAS Macro Variables

| **SAS Macro Variable Category** | **Details** | **Example** |
| --- | --- | --- |
| **SECTION 4: Only complete section 3 for any "N" values listed in section 2** | | |
| If ALL\_STATES = N | GEO\_GROUP informs the program what level of geography is to be used in the GEO\_LIST macro variable. GEO\_LIST subsets the prevalence based on the geographic location of the adult. If GEO\_GROUP=STATE; then the program defaults to using state FIPS codes. If GEO\_GROUP=ZCTA3; then the program defaults to using state State FIPS+ZCTA-3 codes. If GEO\_GROUP=County; then the program defaults to using state FIPS+ County codes. Of note, values should be surrounded by single quotes and comma delimited. | /\*IF ALL\_STATES= 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 ALL\_STATES = N THEN SELECT ONE OR MORE AGE CATEGORIES BELOW: \*\*\*/  /\*\*\*/ %LET GEO\_GROUP = STATE; /\*@Note: Level of geography (ACCEPTED VALUES: STATE/ZCTA3, or STATE/COUNTY) \*\*\*/  /\*\*\*/ %LET GEO\_LIST = %STR('08'); /\*@Note: IF GEO\_GROUP="STATE" then generate with State FIPS code(s), If GEO\_GROUP="GEO3" then generate with FIPS State+FIPS County code(s) or FIPS State+ZCTA3 values \*\*\*/ |
| If ALL\_AGES = N and CO\_OCCURRING = N; | If ALL\_AGES is set to no and CO\_OCCURRING is set to no, the age macros (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, then the SAS program does not review 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; |
| If ALL\_RACES = N; | If ALL\_RACES is set to no, the race macros (White, Black, Asian, Other) subset the prevalence based on the race or imputed race of the adult and the responses to each individual age macro. Note that if ALL\_RACES is set to yes, then the SAS program does not review the race-specific macros. | %LET RACE\_WHITE = N;  %LET RACE\_BLACK = Y;  %LET RACE\_ASIAN = Y;  %LET RACE\_OTHER = Y; |
| If ALL\_SEXES = N; | If ALL\_SEXES is set to no, the sex macros (male, female) subset the prevalence based on the sex of the adult and the responses to each individual sex macro. Note that if ALL\_SEXES is set to yes, then the SAS program does not review the sex-specific macros. | %LET SEX\_MALE = N;  %LET SEX\_FEMALE = Y; |
| **SAS Macro Variable Category** | Details | Example |
| **SECTION 4: Methodological option selections** | | |
| Race Imputation | If IMP\_RACES is set to yes, then the program includes adults with imputed race values. Otherwise, if IMP\_RACES is set to no, then the patients with imputed races are excluded. | %LET IMP\_RACES = Y; |
| Age Adjustment | If AGE\_ADJ is set to yes, then the program generates age adjusted prevalence and standard errors. Otherwise, if AGE\_ADJ is set to no, age adjusted prevalence is not generate. | %LET AGE\_ADJ = Y; |

Table . 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 . 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, warning, and uninitialized. Assuming no errors, continue to step 3. In the event of errors, reassess the location of the files and the file formats. |
| **3** | Review the results. | Review the results for possible data suppression or errors. Consider a statistical review based on the NCHS data presentation standards. In the event of errors reassess the choices described in part 4.1 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 more years of data. |

### Review BMI prevalence Results

CODI-APQ generates prevalence outputs as a csv 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 3-19.

Table . CODI-APQ BMI prevalence Results Data Dictionary

| Column | Description |
| --- | --- |
| Order | Row order |
| Weight Category | The weight 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 implementation guide, Appendix A. Sample Weights for more information. |
| Weighted Prevalence Standard Error | Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information. |
| Age-adjusted Prevalence | Prevalence based on weighted, age-adjusted counts (optional). See implementation guide, Appendix A. Age Adjustment for more information. |
| Age-Adjusted Prevalence Standard Error | Standard error based on weighted, age-adjusted counts. See implementation guide, See implementation guide, Appendix A. Age Adjustment for more information. |

* Review BMI and Co-Occurring Conditions Prevalence Results

CODI-APQ generates 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 . 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 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 . 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. |
| 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 implementation guide, Appendix A. Sample Weights for more information. |
| BMI prevalence | Weighted Prevalence Standard Error | Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information. |
| BMI prevalence | Age-adjusted Prevalence | Prevalence based on weighted, age-adjusted counts (optional). See implementation guide, Appendix A. 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 implementation guide, Appendix A. 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 patient. It is a measure of the number of adults in the population represented by that sample patient. See implementation guide, Appendix A. Sample Weights for more information. |
| Diabetes prevalence, BMI by diabetes prevalence | Weighted Prevalence Standard Error | Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information. |
| Diabetes prevalence, BMI by diabetes prevalence | Age-adjusted Prevalence | Prevalence based on weighted, age-adjusted counts (optional). See implementation guide, Appendix A. 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 implementation guide, See implementation guide, Appendix A. Age Adjustment for more information. |

## 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

###### Analysis Details

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.[[10]](#footnote-11) 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:

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

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.

Body Mass Index

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

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

For more information, see

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

Data Sources (Inputs)

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

* EHRs (data in csv format, provided by user) provided by the user, and
* American Community Survey (ACS) data file (provided by the Health FFRDC[[11]](#footnote-12)).

CODI-APQ are intended for use with all available EHRs for a geography or subpopulation. The programs were created and tested with the Ambulatory Electronic Medical Record (AEMR)[[12]](#footnote-13) data and synthetic data generated for CODI using Synthea.[[13]](#footnote-14) The guide provided in this document is implemented through open-access programs.

The programs were tested using EHRs and AEMRs. Both provide a non-probability sample of longitudinally linked patients’ medical records from within the United States. CODI-APQ subsets 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)[[14]](#footnote-15) 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 EHRs pre-processed using ‘growthcleanr.’ The ‘growthcleanr’ package is a publicly available program for identifying biological implausible height and weight measurements in longitudinal files at <https://github.com/mitre/growthcleanr-web>. The program evaluates data against published growth trajectory charts for youth, teens and adults and flags measurements for plausibility ([Daymont et al., 2017](#_ENREF_9)).

To statistically weight EHRs 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.

Diabetes Spectrum

Prevalence is calculated from an optional indicator for diabetes spectrum. Diabetes indicators are based on phenotype and each patient can have must one diabetes 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 includes 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.

Prevalence

A prevalence is either:

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

Race

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

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

Race Exclusion

Adults from one or more races may be excluded from prevalence results. 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.

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 . 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% |

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, or
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 four reasons:

1. The patient’s geography is either invalid or did not have a population count in the 2019 ACS.
2. The patient’s age is outside of the scope of the program or is unknown. Only persons age 20 to 64 are in scope.
3. The sex of the patient is unknown.

CODI-APQ assigns 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 reassigns 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 US. To reduce these disparities, high-quality data on race are needed. However, these data are often missing in some portion of EHRs. CODI-APQ imputes 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 EHRs 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 . 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% |

Statistical Weights

CODI and National AEMR data are derived from EHRs. As described in Appendix A, applying statistical weights is often used to reduce potential biases introduced by the EHRs 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.[[15]](#footnote-16) One way to accomplish this goal is known as iterative proportional fitting or raking. Raking adjusts the data so that groups that are underrepresented in the sample can be accurately represented in the final data set. Raking accurately matches sample distributions to known demographic characteristics of populations. The use of raking reduces nonresponse bias and has been shown to reduce error within sample results.

Implementing raking programs require the specification of appropriate weighting classes or cells. Data used to form classes for adjustments must be available for both sample and the population. CODI-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

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.

The five-year estimates of ACS do not include an age group of 2 to 4 years. Thus, CODI-APQ calculates the population of this age group by multiplying the count of persons aged 5 and under within each geography by the percentage of the national population that is aged 2 to 4, given that they are under the age of 5, based on Annual Estimates of the Resident Population by Single Year of Age and Sex for the United States, July 1, 2018, U.S. Census Bureau. Population counts by race for those age less than 5 are also adjusted from Annual Estimates of the Resident Population using the same adjustment.

Similarly, for diabetes prevalence the ACS population aged 12 to 14 years is calculated by multiplying the count of persons aged 10 to 14 within each state and GEO3 by the percentage of the national population that is aged 12 to 14, given that they are aged 10 to 14, based on Annual Estimates of the Resident Population by Single Year of Age and Sex for the United States, July 1, 2018, U.S. Census Bureau. Population counts by race and sex for those aged 10 to 14 are also adjusted from Annual Estimates of the Resident Population using the same adjustment.

ACS race is categorized to match the EHRs 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 United States fall above 20%, and 48% fall below.

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.[[16]](#footnote-17)

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

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.

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,[[17]](#footnote-18) developed by the Data Suppression Workgroup at NCHS.

The multistep NCHS Data Presentation Standards for Proportions are based on a minimum denominator sample size and on the absolute and relative widths of a confidence interval calculated using the Clopper-Pearson method. The National Center for Health Statistics (NCHS) Data Presentation Standards for Proportions are applied to all CODI-APQ results. The Presentation Standards also provide guidance for identifying results for statistical review, CODI-APQ does not identify records for statistical review and leave this step for the user. The data presentation standards are described in Table 19 and Figure 6.

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

Table . 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[[18]](#footnote-19)), then the denominator sample size should be used to obtain confidence intervals. If all other criteria are met for presentation, a result based on 0 encounters (or its complement) should be flagged for statistical review by the clearance official. The review could result in either the presentation or the suppression of the proportion. |
| **Confidence interval** | If the sample size criterion is met, calculate a 95% two-sided confidence interval using the Clopper-Pearson method, or the Korn-Graubard method for complex surveys, and obtain its width. |
| **Small absolute confidence interval width** | If the absolute confidence interval width is greater than 0.00 and less than or equal to 0.05, then the proportion can be presented if the number of encounters is greater than 0 and the degrees of freedom criterion (below) is met. If the number of encounters is 0 (or its complement) or the degrees of freedom criterion is not met, then the result should be flagged for statistical review by the clearance official. The review could result in either the presentation or the suppression of the proportion. |
| **Large absolute confidence interval width** | If the absolute confidence interval width is greater than or equal to 0.30, then the proportion should be suppressed. |
| **Relative confidence interval width** | If the absolute confidence interval width is between 0.05 and 0.30 and the relative confidence interval width is more than 130%, then the proportion should be suppressed. |
| **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. |

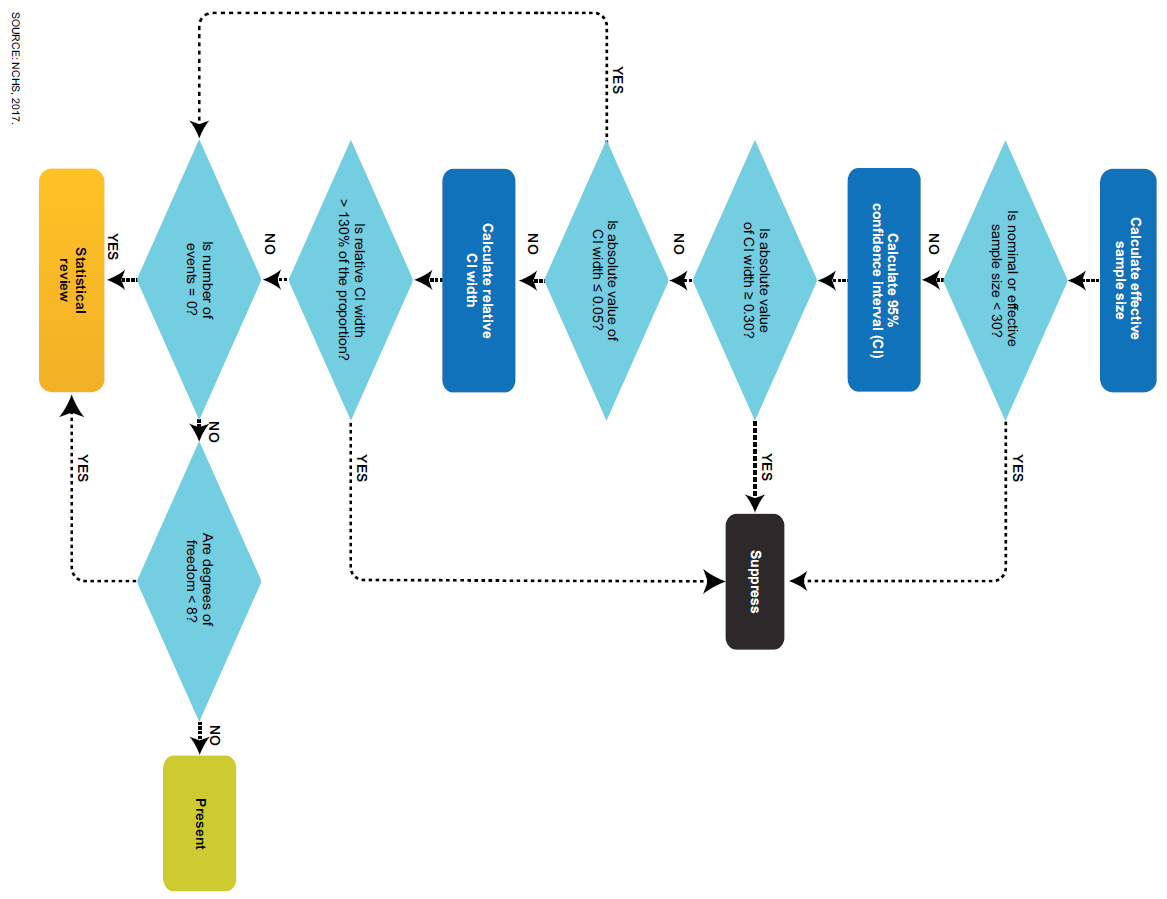


Figure . NCHS Suppression Standards

Variance

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, , corresponds to a population parameter, , such as the true but unknown BMI or diabetes prevalence. To define the population parameter, let

= the number of adults in stratum (), where stratum refers to state-GEO3

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

= 0 or 1, indicating whether adult of stratum belongs to a particular domain (such as a specified race)

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

In the sample, let

= the number of sample adults in stratum

= the sampling weight for adults in stratum

= the value of for adult in stratum

= the value of the domain indicator for adult in stratum

The distinction between and and between and is merely that for and the subscript refers to sampled adults within stratum *,* whereas for and they refer to adults in the population in stratum . Then, the ratio estimator for is   
To calculate the variance of, a Taylor-series approximation is used.[[19]](#footnote-20) Within stratum, linearization yields the new variable.

Then, letting

the Taylor-series approximation to the variance of is

BMI

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

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

For more information, visit

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

ZCTA-3

A ZCTA is a statistical geographic entity that approximates the delivery area for a U.S. Postal Service five-digit (ZCTA) ZIP code. ZCTAs are aggregations of census blocks that have the same predominant ZIP code associated with the residential mailing addresses in the U.S. Census Bureau’s Master Address File. ZCTAs do not precisely depict ZIP code delivery areas, and do not include all ZIP codes used for mail delivery. The U.S. Census Bureau has established ZCTAs as a new geographic entity similar to, but replacing, data tabulations for ZIP codes undertaken in conjunction with the 1990 and earlier censuses. For more information, refer to census.gov.[[20]](#footnote-21)

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.

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 – EHRs 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 EHRs, 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 EHRs, 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 requires the use of a phenotype to estimate the prevalence of prediabetes and diabetes within a population. The population and phenotype require careful consideration of the pro’s and con’s. For example, whether the population be subset to wellness visit only can greatly impact the denominator and thus prevalence. Similarly, one standard phenotype is not available for diabetes spectrum. Thus, the user may consider different versions of phenotypes used in public health.
* Measurement error - Height and weight measurement protocols may differ between medical providers, even with clear protocols aimed to increase consistency between medical professionals,[[21]](#footnote-22) leading to potential measurement error. Additionally, height and weight values in EHRs 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 suppresses results based on published small sample guidelines using the National Center for Health Statistics Data Presentation Standards for Proportions[[22]](#footnote-23).

###### ACS File Layouts

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, and in the order listed below. Variable geoid (option 1) is for ZCTA-3 files only and geoid (option 2) is for the County file only.

Table . ACS Input File Layout, CSV File

| Variable Name | Label | Description | Format | Example |
| --- | --- | --- | --- | --- |
| Geoid | Geoid (option 1) | 3 digits ZIP Code Tabulation Areas (ZCTAs) followed by two-letter State Abbreviations | Character | 221 |
| Geoid | Geoid (option 2) | 3-digit County FIPS Code | Character | 059 |
| State\_code | State FIPS code | 2-digit State Code | Character | 08 |
| b01001f\_001 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Total: Total: | Population count | Number | 8199 |
| b01001f\_003 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: Under 5 years | Population count | Number | 256 |
| b01001f\_004 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 5 to 9 years | Population count | Number | 246 |
| b01001f\_005 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 10 to 14 years | Population count | Number | 495 |
| b01001f\_006 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 15 to 17 years | Population count | Number | 297 |
| b01001f\_007 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Male: 18 and 19 years | Population count | Number | 145 |
| b01001f\_018 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: Under 5 years | Population count | Number | 271 |
| b01001f\_019 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 5 to 9 years | Population count | Number | 188 |
| b01001f\_020 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 10 to 14 years | Population count | Number | 267 |
| b01001f\_021 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 15 to 17 years | Population count | Number | 139 |
| b01001f\_022 | SEX BY AGE (SOME OTHER RACE ALONE); Universe: People who are Some Other Race alone; Female: 18 and 19 years | Population count | Number | 134 |
| b01001e\_001 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Total: Total: | Population count | Number | 278 |
| b01001e\_003 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: Under 5 years | Population count | Number | 0 |
| b01001e\_004 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 5 to 9 years | Population count | Number | 28 |
| b01001e\_005 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 10 to 14 years | Population count | Number | 28 |
| b01001e\_006 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 15 to 17 years | Population count | Number | 0 |
| b01001e\_007 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Male: 18 and 19 years | Population count | Number | 0 |
| b01001e\_018 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: Under 5 years | Population count | Number | 0 |
| b01001e\_019 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 5 to 9 years | Population count | Number | 0 |
| b01001e\_020 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 10 to 14 years | Population count | Number | 0 |
| b01001e\_021 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 15 to 17 years | Population count | Number | 0 |
| b01001e\_022 | SEX BY AGE (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE); Universe: People who are Native Hawaiian and Other Pacific Islander alone; Female: 18 and 19 years | Population count | Number | 10 |
| b01001d\_001 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Total: Total: | Population count | Number | 14824 |
| b01001d\_003 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: Under 5 years | Population count | Number | 277 |
| b01001d\_004 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 5 to 9 years | Population count | Number | 222 |
| b01001d\_005 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 10 to 14 years | Population count | Number | 276 |
| b01001d\_006 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 15 to 17 years | Population count | Number | 263 |
| b01001d\_007 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Male: 18 and 19 years | Population count | Number | 774 |
| b01001d\_018 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: Under 5 years | Population count | Number | 101 |
| b01001d\_019 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 5 to 9 years | Population count | Number | 237 |
| b01001d\_020 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 10 to 14 years | Population count | Number | 355 |
| b01001d\_021 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 15 to 17 years | Population count | Number | 242 |
| b01001d\_022 | SEX BY AGE (ASIAN ALONE); Universe: People who are Asian alone; Female: 18 and 19 years | Population count | Number | 1404 |
| b01001c\_001 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Total: Total: | Population count | Number | 755 |
| b01001c\_003 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: Under 5 years | Population count | Number | 36 |
| b01001c\_004 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 5 to 9 years | Population count | Number | 44 |
| b01001c\_005 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 10 to 14 years | Population count | Number | 3 |
| b01001c\_006 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 15 to 17 years | Population count | Number | 22 |
| b01001c\_007 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Male: 18 and 19 years | Population count | Number | 37 |
| b01001c\_018 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: Under 5 years | Population count | Number | 0 |
| b01001c\_019 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 5 to 9 years | Population count | Number | 14 |
| b01001c\_020 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 10 to 14 years | Population count | Number | 11 |
| b01001c\_021 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 15 to 17 years | Population count | Number | 9 |
| b01001c\_022 | SEX BY AGE (AMERICAN INDIAN AND ALASKA NATIVE ALONE); Universe: People who are American Indian and Alaska Native alone; Female: 18 and 19 years | Population count | Number | 26 |
| b01001b\_001 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Total: Total: | Population count | Number | 13407 |
| b01001b\_003 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: Under 5 years | Population count | Number | 283 |
| b01001b\_004 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 5 to 9 years | Population count | Number | 222 |
| b01001b\_005 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 10 to 14 years | Population count | Number | 425 |
| b01001b\_006 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 15 to 17 years | Population count | Number | 439 |
| b01001b\_007 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Male: 18 and 19 years | Population count | Number | 429 |
| b01001b\_018 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: Under 5 years | Population count | Number | 485 |
| b01001b\_019 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 5 to 9 years | Population count | Number | 254 |
| b01001b\_020 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 10 to 14 years | Population count | Number | 359 |
| b01001b\_021 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 15 to 17 years | Population count | Number | 189 |
| b01001b\_022 | SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE); Universe: Black or African American alone; Female: 18 and 19 years | Population count | Number | 561 |
| b01001a\_001 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Total: Total: | Population count | Number | 426175 |
| b01001a\_003 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: Under 5 years | Population count | Number | 9461 |
| b01001a\_004 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 5 to 9 years | Population count | Number | 9446 |
| b01001a\_005 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 10 to 14 years | Population count | Number | 11409 |
| b01001a\_006 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 15 to 17 years | Population count | Number | 7231 |
| b01001a\_007 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Male: 18 and 19 years | Population count | Number | 8352 |
| b01001a\_018 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: Under 5 years | Population count | Number | 8754 |
| b01001a\_019 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 5 to 9 years | Population count | Number | 10226 |
| b01001a\_020 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 10 to 14 years | Population count | Number | 10640 |
| b01001a\_021 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 15 to 17 years | Population count | Number | 7731 |
| b01001a\_022 | SEX BY AGE (WHITE ALONE); Universe: People who are White alone; Female: 18 and 19 years | Population count | Number | 9368 |
| b01001g\_001 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Total: Total: | Population count | Number | 12014 |
| b01001g\_003 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: Under 5 years | Population count | Number | 732 |
| b01001g\_004 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 5 to 9 years | Population count | Number | 659 |
| b01001g\_005 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 10 to 14 years | Population count | Number | 823 |
| b01001g\_006 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 15 to 17 years | Population count | Number | 491 |
| b01001g\_007 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Male: 18 and 19 years | Population count | Number | 501 |
| b01001g\_018 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: Under 5 years | Population count | Number | 683 |
| b01001g\_019 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 5 to 9 years | Population count | Number | 650 |
| b01001g\_020 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 10 to 14 years | Population count | Number | 652 |
| b01001g\_021 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 15 to 17 years | Population count | Number | 410 |
| b01001g\_022 | SEX BY AGE (TWO OR MORE RACES); Universe: People who are Two or More Races; Female: 18 and 19 years | Population count | Number | 651 |
| b03002\_012 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Hispanic or Latino: | Population count | Number | 56886 |
| b03002\_013 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: White alone | Population count | Number | 43689 |
| b03002\_014 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Black or African American alone | Population count | Number | 1753 |
| b03002\_015 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: American Indian and Alaska Native alone | Population count | Number | 196 |
| b03002\_016 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Asian alone | Population count | Number | 253 |
| b03002\_017 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone | Population count | Number | 150 |
| b03002\_018 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Hispanic or Latino: Some other race alone | Population count | Number | 7913 |
| b03002\_019 | HISPANIC OR LATINO ORIGIN BY RACE; Universe: Total population; Two or more races: Two or more races: | Population count | Number | 2932 |
| b15001\_011 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: 25 to 34 years: | Population count | Number | 28989 |
| b15001\_017 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Bachelor's degree | Population count | Number | 7476 |
| b15001\_018 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Graduate or professional degree | Population count | Number | 2604 |
| b15001\_019 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: 35 to 44 years: | Population count | Number | 24797 |
| b15001\_025 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Bachelor's degree | Population count | Number | 4891 |
| b15001\_026 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Graduate or professional degree | Population count | Number | 3258 |
| b15001\_027 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: 45 to 64 years: | Population count | Number | 62253 |
| b15001\_033 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Bachelor's degree | Population count | Number | 11482 |
| b15001\_034 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Graduate or professional degree | Population count | Number | 8394 |
| b15001\_058 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Bachelor's degree | Population count | Number | 8124 |
| b15001\_059 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 25 to 34 years: Graduate or professional degree | Population count | Number | 5237 |
| b15001\_060 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: 35 to 44 years: | Population count | Number | 26186 |
| b15001\_066 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Bachelor's degree | Population count | Number | 6646 |
| b15001\_067 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 35 to 44 years: Graduate or professional degree | Population count | Number | 5657 |
| b15001\_068 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: 45 to 64 years: | Population count | Number | 67764 |
| b15001\_074 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Bachelor's degree | Population count | Number | 14062 |
| b15001\_075 | SEX BY AGE BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 18 YEARS AND OVER; Universe: Population 18 years and over; 45 to 64 years: Graduate or professional degree | Population count | Number | 11986 |

ACS for Use with GEO3 Data

Table . ACS Pre-processing Results File Layout – GEO3

| Variable Name | Format | Length |
| --- | --- | --- |
| Geography | Character | 5 |
| GEO3 | Character | 3 |
| State\_FIPS | Character | 2 |
| TOTAL\_ACS\_POPULATION | Number | 8 |
| AGE\_L5\_MALE\_WHITE | Number | 8 |
| AGE\_5\_9\_MALE\_WHITE | Number | 8 |
| AGE\_10\_14\_MALE\_WHITE | Number | 8 |
| AGE\_15\_17\_MALE\_WHITE | Number | 8 |
| AGE\_18\_19\_MALE\_WHITE | Number | 8 |
| AGE\_L5\_FEMALE\_WHITE | Number | 8 |
| AGE\_5\_9\_FEMALE\_WHITE | Number | 8 |
| AGE\_10\_14\_FEMALE\_WHITE | Number | 8 |
| AGE\_15\_17\_FEMALE\_WHITE | Number | 8 |
| AGE\_18\_19\_FEMALE\_WHITE | Number | 8 |
| AGE\_L5\_MALE\_BLACK | Number | 8 |
| AGE\_5\_9\_MALE\_BLACK | Number | 8 |
| AGE\_10\_14\_MALE\_BLACK | Number | 8 |
| AGE\_15\_17\_MALE\_BLACK | Number | 8 |
| AGE\_18\_19\_MALE\_BLACK | Number | 8 |
| AGE\_L5\_FEMALE\_BLACK | Number | 8 |
| AGE\_5\_9\_FEMALE\_BLACK | Number | 8 |
| AGE\_10\_14\_FEMALE\_BLACK | Number | 8 |
| AGE\_15\_17\_FEMALE\_BLACK | Number | 8 |
| AGE\_18\_19\_FEMALE\_BLACK | Number | 8 |
| AGE\_L5\_MALE\_ASIAN | Number | 8 |
| AGE\_5\_9\_MALE\_ASIAN | Number | 8 |
| AGE\_10\_14\_MALE\_ASIAN | Number | 8 |
| AGE\_15\_17\_MALE\_ASIAN | Number | 8 |
| AGE\_18\_19\_MALE\_ASIAN | Number | 8 |
| AGE\_L5\_FEMALE\_ASIAN | Number | 8 |
| AGE\_5\_9\_FEMALE\_ASIAN | Number | 8 |
| AGE\_10\_14\_FEMALE\_ASIAN | Number | 8 |
| AGE\_15\_17\_FEMALE\_ASIAN | Number | 8 |
| AGE\_18\_19\_FEMALE\_ASIAN | Number | 8 |
| AGE\_L5\_MALE\_OTHER | Number | 8 |
| AGE\_5\_9\_MALE\_OTHER | Number | 8 |
| AGE\_10\_14\_MALE\_OTHER | Number | 8 |
| AGE\_15\_17\_MALE\_OTHER | Number | 8 |
| AGE\_18\_19\_MALE\_OTHER | Number | 8 |
| AGE\_L5\_FEMALE\_OTHER | Number | 8 |
| AGE\_5\_9\_FEMALE\_OTHER | Number | 8 |
| AGE\_10\_14\_FEMALE\_OTHER | Number | 8 |
| AGE\_15\_17\_FEMALE\_OTHER | Number | 8 |
| AGE\_18\_19\_FEMALE\_OTHER | Number | 8 |
| AGE\_25\_64\_BACH\_GRAD | Number | 8 |
| AGE\_25\_64\_BACH\_GRAD\_GTR20PERC | Number | 8 |
| TOTAL\_LATIN | Number | 8 |
| LATIN\_WHITE | Number | 8 |
| LATIN\_BLACK | Number | 8 |
| LATIN\_ASIAN | Number | 8 |
| LATIN\_OTHER | Number | 8 |

###### EHR File Layouts

EHR Input File Layout

EHR GEO3 Data

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

Table . EHR Input File Layout for GEO3-Level Programs, CSV File[[23]](#footnote-24)

| 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 | 11 |
| RACE\_ETH | Patient’s race if known or ethnicity when race is not known | Character | “AFRICAN AMERICAN”  “ASIAN  “CAUCASIAN”  “HISPANIC”  “OTHER”  “UNKNOWN” | CAUCASIAN |
| STATE\_ABR | Patient’s residential state, two-letter state abbreviations | Character | any postal abbreviation of state | MI |
| GEO3 | Either: Patient’s residential a) county code or b) ZIP-3 | Number | Any numeric value | 059 |
| WEIGHT\_CATEGORY | Patient’s BMI | 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. | Number | <blank> = No Evidence of Diabetes  1 = Prediabetes,  2 = Diabetes, or  3 = No Evidence of Diabetes | 1 |
| SCDCNT | Sickle-Cell indicator | Number | Program treats patients with a count of 1 or higher as having sickle cell disease. If sickle cell disease information is not available, set this value to blank or zero. | 2 |
| PREGNANCY\_FLAG | Pregnancy flag | Number | 0/1  If pregnancy information is not available, set this value to zero or blank. | 1 |
| ZIP | 5-digit ZIP code | Character | 5-digit ZIP code. Required for county level records. Optional for ZCTA-3 level records. | 20814 |

EHR Results File Layout for GEO3

Table . GEO3

|  |  |  |
| --- | --- | --- |
| Variable Name | Format | Example |
| SUBJID | Character | 12345626 |
| Ageyr | Number | 16 |
| AGE\_CATEGORIES | Character | 15 – 17 |
| WTCAT | Character | (2) Healthy Weight (5th to <85th percentile) |
| STATE\_ALPHA | Character | MI |
| STATE\_FIPS | Character | 48 |
| ZIP | Character | 20184 |
| GEO3 | Character | 100 |
| Geography | Character | 48100 |
| DIABETES SPECTRUM | Number | 1 |
| SCDCNT | Number | 0 |
| PREGNANCY\_FLAG | Number | 1 |
| Race | Character | Black |
| Sex | Character | Female |
| Year | Number | 2018 |
| SCD | Number | 0 |
| Imputed\_Race | Character | Black |
| Race\_Imputed | Number | 0 |

###### CODI-APQ-GEO3 Example SAS Programs

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 EHRs between 2015 and 2018 and creates a SAS file named CODI\_Ready stored in the folder P:\Example. The SAS log is stored in P:\Example\2\_Output\SAS LOG\This\_is\_the\_Name\_for\_the\_Log <plus date and time information>.log.

/\*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 PROGS\_PRE = &Root\_PRE.\1\_SAS\_Programs\;/\*@Note: where SAS programs are stored (ACCEPTABLE VALUES: computer directory name)\*\*\*/

/\*\*\*/ %LET PRE\_DEST = CODI\_APQ\_GEO3; /\*@Note: Suffix name for EHR Output folder (ACCEPTABLE VALUES: folder name (no puctuation)\*\*\*/

/\*\*\*/ %LET ACS\_FILENAME = ACS\_ZCTA3\_FILE; /\*@Note: ACS file name (ACCEPTABLE VALUES: file name, do not include ".csv")\*\*\*/

/\*\*\*/ %LET EHR\_FILENAME = EHR\_ZCTA3\_FILE; /\*@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\_ZCTA3; /\*@Note: EHR output file name (ACCEPTABLE VALUES: SAS file name (no punctuation)\*\*\*/

/\*SECTION 4: County or ZCTA3 data (REQUIRED)\*\*\*/

/\*\*\*/ %LET COUNTY = N; /\*@Note: County/ZCTA3 indicator (ACCEPTABLE VALUES: Y for County level data, N for ZCTA3 level data\*\*\*/

/\*\*\*Note: ROOT\_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 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. \*/

Generate Results Example with GEO3 Data

Appendix D2. includes a program excerpt to generate prevalence results using the Quickstart program and data inputs: ACS\_COUNTY and CODI\_Ready. This example uses COUNTY data.

The file processes EHRs 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 [ROOT\_PQ] \2\_Output\Example\_Only. The SAS log is stored in [ROOT\_PQ]\ 2\_Output\SAS LOGS\Prevalence\_2017.log.

Subpopulation: EHR records from 2017 including patients 2 to 14 years of age who are either white or Asian, living in Jefferson County (059) Colorado (FIPS code = 08) or Yuma County (125) Colorado (FIPS code = 08) see: <https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697> to determine the correct value (08059 is combined state and County). Note that Sex\_Male and Sex\_Female have not been edited since the ALL\_SEXES is turned on to yes.

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

/\*Note: subsection of the full program. Be sure to only edit this section, but submit the full program. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* -- USER SELECTION CRITERA SECTIONS 1 through 4 -- \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* -- 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 PROGS\_PQ = &Root\_PQ.\1\_SAS\_Programs\; /\*@Note: Location of SAS programs (ACCEPTABLE VALUES: computer directory name)\*\*\*/

/\*\*\*/ %LET PRE\_DEST = CODI\_APQ\_GEO3; /\*@Note: Suffix name of pre-processing output folder (ACCEPTABLE VALUES: folder name (no punctuations)) \*\*\*/

/\*\*\*/ %LET EHR\_PRE\_OUT = CODI\_APQ\_ZCTA3; /\*@Note: Suffix name of pre-processing output file (ACCEPTABLE VALUES: file name (no punctuations)) \*\*\*/

/\*\*\*/ %LET LOG\_NAME = CODI\_ZCTA3\_APQ; /\*@Note: Name for SAS log storage location\*\*\*/

/\*\*\*/ %LET FileOUT\_Name = CODI\_ZCTA3\_APQ; /\*@Note: Output file name\*\*\*/

/\*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 numeric, 2015-2019)\*\*\*/

/\*\*\*/ %LET END\_YEAR = 2017; /\*@Note: End year of analysis (ACCEPTED VALUES: 4-Digit numeric, 2015-2019)\*\*\*/

/\*\*\*/ %LET ALL\_STATES = Y; /\*@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 = Y; /\*@Note: Include all race categories? (ACCEPTED VALUES: Y/N)\*\*\*/

/\*SECTION 3: Additional flags\*\*\*/

/\*\*\*/ %LET ACSCOUNTY = N; /\*@Note: Is the ACS data at the county or ZCTA3 level? (ACCEPTABLE VALUES: Y for County level data, N for ZCTA3 level data)\*\*\*/

/\*\*\*/ %LET INCLUDE\_PREGNANCY = Y; /\*@Note: Include pregnant individuals (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\*\*\*/

/\*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 = STATE; /\*@Note: Level of geography (ACCEPTED VALUES: STATE/ZCTA3)\*\*\*/

/\*\*\*/ %LET GEO\_LIST = %STR('08'); /\*@Note: IF GEO\_GROUP="STATE" then populate with State FIPS code(s), If GEO\_GROUP="ZCTA3" then populate with FIPS State+FIPS County code(s)\*\*\*/

/\*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 = Y; /\*@Note: Black/African American (ACCEPTED VALUES: Y/N)\*\*\*/

/\*\*\*/ %LET RACE\_ASIAN = Y; /\*@Note: Asian (ACCEPTED VALUES: Y/N)\*\*\*/

/\*\*\*/ %LET RACE\_OTHER = Y; /\*@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 4: Methodological option selections\*\*\*/

/\*\*\*/ %LET IMP\_RACES = Y; /\*@Note: Include imputed race values? (ACCEPTED VALUES: Y/N)\*\*\*/

/\*\*\*/ %LET AGE\_ADJ = Y; /\*@Note: Produce age-adjusted estimates? (ACCEPTED VALUES: Y/N) \*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*Note: Root directory includes subfolders:

"..\0\_Raw\_Data"

"..\1\_SAS\_Programs"

"..\2\_Output" and

"..\2\_Output\SAS LOGS"\*\*\*/

/\*\*\*NOTE: SAS programs must be stored in the PROGS directory including:

Macro1-CODI\_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 the specified FileOUT\_Name\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*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\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Note: subsection of the full program. Be sure to only edit this section but submit the full program. \*/

###### CODI-APQ Results

Example BMI prevalence

Once complete, CODI-APQ generates prevalence results as either a csv file or csv files and Excel files. Table 32 provides an overview of the variables included when diabetes prevalence is set to no, and Table 33 provides example results based on synthetic data. Note in the sample provided in Table 34 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 19. The number of rows output will vary based on the criteria selected.

Table . 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 implementation guide, Appendix A. Sample Weights for more information. |
| Weighted Prevalence Standard Error | Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information. |
| Age-adjusted Prevalence | Prevalence based on weighted, age-adjusted counts. See implementation guide, Appendix A. Age Adjustment for more information. |

Table . Results Example from Synthetic Data[[24]](#footnote-25)

| **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** | 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, 55 - 64) |  |  |  |  |  |  |  |  |
| **6** | SEX: (Male, Female) |  |  |  |  |  |  |  |  |
| **7** | Pregnancy: included in analysis |  |  |  |  |  |  |  |  |
| **8** | RACE: (White, Black, Asian, Other) |  |  |  |  |  |  |  |  |
| **9** | RACE Suppressed: (None) |  |  |  |  |  |  |  |  |
| **10** | RACE Imputed: People with unknown race were excluded. |  |  |  |  |  |  |  |  |
| **11** | Geography: (99999) |  |  |  |  |  |  |  |  |
| **12** | Year: 2017 |  |  |  |  |  |  |  |  |
| **13** | Weighting cells were collapsed for: (None) |  |  |  |  |  |  |  |  |
| **14** | AGE adjusted: (Yes) |  |  |  |  |  |  |  |  |
| **15** | Error Codes: (None) |  |  |  |  |  |  |  |  |
| **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, November 3, 2021 4:43:51 PM |  |  |  |  |  |  |  |  |
| **18** | Suggested Citation: Tanenbaum, E., Campbell, S., Chelluri, D., Zalsha, S., Boim, J., Paddock, S., Copeland, K. (2021). Clinical and Community Data Initiative Prevalence Query (CODI-PQ) SAS programs (version 2015-2019). |  |  |  |  |  |  |  |  |
| **18** | 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, November 3, 2021 4:43:51 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. |  |  |  |  |  |  |  |  |
| **22** | The population is calculated from age-race-sex-geography specific counts from the user provided American Community Survey Five-year Estimates. |  |  |  |  |  |  |  |  |
| **23** | The age-adjusted prevalence calculations are documented in the Implementation Guide. |  |  |  |  |  |  |  |  |

Table . Example Results with Errors (insufficient sample size), error messages are shown in row order

| **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** | Version: CODI-APQ GEO3 2015-2019 |  |  |  |  |  |  |  |  |
| **4** | Query Parameters: AGE RACE SEX GEOGRAPHY YEAR |  |  |  |  |  |  |  |  |
| **5** | AGE: (25 - 29, 30 - 34, 35 - 44) |  |  |  |  |  |  |  |  |
| **6** | SEX: (Female) |  |  |  |  |  |  |  |  |
| **7** | Pregnancy: included in analysis |  |  |  |  |  |  |  |  |
| **8** | RACE: (White, Black, Asian, Other) |  |  |  |  |  |  |  |  |
| **9** | RACE Suppressed: (Error) |  |  |  |  |  |  |  |  |
| **10** | Imputed race: People with unknown race were excluded. |  |  |  |  |  |  |  |  |
| **11** | Geography: (99999) |  |  |  |  |  |  |  |  |
| **12** | Year: 2017 |  |  |  |  |  |  |  |  |
| **13** | Weighting cells were consolidated for: (Error) |  |  |  |  |  |  |  |  |
| **14** | AGE adjusted: (Yes) |  |  |  |  |  |  |  |  |
| **15** | Error Codes: (Current selections return an insufficient number of patients and do 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, November 3, 2021 3:30:52 PM |  |  |  |  |  |  |  |  |
| **18** | Suggested Citation: Tanenbaum, E., Campbell, S., Chelluri, D., Zalsha, S., Boim, J., Paddock, S., Copeland, K. (2021). Clinical and Community Data Initiative Prevalence Query (CODI-PQ) SAS programs (version 2015-2019). |  |  |  |  |  |  |  |  |
| **18** | 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, November 3, 2021 3:30:52 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. |  |  |  |  |  |  |  |  |
| **22** | The population is calculated from age-race-sex-geography specific counts from the user provided American Community Survey Five-year Estimates. |  |  |  |  |  |  |  |  |
| **23** | The age-adjusted prevalence calculations are documented in the Implementation Guide. |  |  |  |  |  |  |  |  |

Table . Example Results with Errors (out of scope years), error messages are shown in order = 15

| **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** | 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, 55 - 64) |  |  |  |  |  |  |  |  |
| **6** | SEX: (Female) |  |  |  |  |  |  |  |  |
| **7** | Pregnancy: included in analysis |  |  |  |  |  |  |  |  |
| **8** | RACE: (White, Black, Asian, Other) |  |  |  |  |  |  |  |  |
| **9** | RACE Suppressed: (Error) |  |  |  |  |  |  |  |  |
| **10** | Imputed race: People with unknown race were excluded. |  |  |  |  |  |  |  |  |
| **11** | Geography: (99999) |  |  |  |  |  |  |  |  |
| **12** | Year: 2020 |  |  |  |  |  |  |  |  |
| **13** | Weighting cells were consolidated for: (Error) |  |  |  |  |  |  |  |  |
| **14** | AGE adjusted: (Yes) |  |  |  |  |  |  |  |  |
| **15** | Error Codes: (Years are out of scope for IQVIA. Acceptable years include 2015, 2016, 2017, 2018, and 2019 for IQVIA.) |  |  |  |  |  |  |  |  |
| **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, November 3, 2021 3:52:32 PM |  |  |  |  |  |  |  |  |
| **18** | Suggested Citation: Tanenbaum, E., Campbell, S., Chelluri, D., Zalsha, S., Boim, J., Paddock, S., Copeland, K. (2021). Clinical and Community Data Initiative Prevalence Query (CODI-PQ) SAS programs (version 2015-2019). |  |  |  |  |  |  |  |  |
| **18** | 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, November 3, 2021 3:52:32 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. |  |  |  |  |  |  |  |  |
| **22** | The population is calculated from age-race-sex-geography specific counts from the user provided American Community Survey Five-year Estimates. |  |  |  |  |  |  |  |  |
| **23** | The age-adjusted prevalence calculations are documented in the Implementation Guide. |  |  |  |  |  |  |  |  |

Diabetes Prevalence

CODI-APQ generates optional diabetes prevalence results as an Excel file. Table 36 provides an overview of the Excel Worksheets and Table 37 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 . 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 . 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. |
| 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 implementation guide, Appendix A. Sample Weights for more information. |
| BMI prevalence | Weighted Prevalence Standard Error | Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information. |
| BMI prevalence | Age-adjusted Prevalence | Prevalence based on weighted, age-adjusted counts (optional). See implementation guide, Appendix A. 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 implementation guide, Appendix A. 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 patient. It is a measure of the number of adults in the population represented by that sample patient. See implementation guide, Appendix A. Sample Weights for more information. |
| Diabetes prevalence, BMI by diabetes prevalence | Weighted Prevalence Standard Error | Standard error based on weighted counts. See implementation guide, Appendix A. Variance for more information. |
| Diabetes prevalence, BMI by diabetes prevalence | Age-adjusted Prevalence | Prevalence based on weighted, age-adjusted counts (optional). See implementation guide, Appendix A. 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 implementation guide, See implementation guide, Appendix A. Age Adjustment for more information. |

###### 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 . State FIPS Codes

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

###### Glossary

**ACS** – American Community Survey. CODI-APQ relies 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 query. (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 EHRs 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 EHRs.

**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 Network’s.

**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 EHRs 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.

###### 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-APQ | Clinical and Community Data Initiative Prevalence Query |
| CSV | Comma Separated Value |
| DHDN | Distributed Health Data Network |
| EHR | Electronic Health Record |
| EMR | Electronic Medical Record |
| FFRDC | Federally Funded Research and Development Center |
| HHS | U.S. Department of Health and Human Services |
| IG | Implementation Guide |
| IPW | Inverse-Probability Weighting |
| MSE | Mean Square Error |
| NCHS | National Center for Health Statistics |
| NHANES | National Health and Nutrition Examination Survey |
| PUF | Public Use File |
| SAS | A Statistical Software Suite |
| SDOH | Social Determinants of Health |
| SFTP | Secured File Transfer Protocol |
| ZCTA | ZIP Code Tabulation Area |

###### 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>

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

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.

**NOTICE**

This document was produced for the U. S. Government under Contract Number HHSM-5000-2012-00008I, 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.

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

© 2021 The MITRE Corporation.

1. DHDNs are traditionally networks of medical organizations that map their EHRs concepts to a common data model so that users can query similar information across organizations. [↑](#footnote-ref-2)
2. [About Adult BMI](https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html). [↑](#footnote-ref-3)
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. [↑](#footnote-ref-4)
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 United States 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). [↑](#footnote-ref-5)
5. <https://synthetichealth.github.io/synthea/> [↑](#footnote-ref-6)
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. [↑](#footnote-ref-7)
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. [↑](#footnote-ref-8)
8. 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. [↑](#footnote-ref-9)
9. 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. [↑](#footnote-ref-10)
10. Klein & Schoenborn, 2001. [↑](#footnote-ref-11)
11. 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. [↑](#footnote-ref-12)
12. CDC provided Ambulator Electronic Medical Record data under a Data Use Agreement with the Health FFRDC. [↑](#footnote-ref-13)
13. The Synthea package is based on Walonoski, et al., 2017 and is available at <https://synthetichealth.github.io/synthea/>. [↑](#footnote-ref-14)
14. 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. [↑](#footnote-ref-15)
15. Little, 1993. [↑](#footnote-ref-16)
16. Wolter, 2007. [↑](#footnote-ref-17)
17. Parker et al., 2017. [↑](#footnote-ref-18)
18. 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). [↑](#footnote-ref-19)
19. Wolter, 2007. [↑](#footnote-ref-20)
20. <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html> [↑](#footnote-ref-21)
21. Best & Shepherd, 2020. [↑](#footnote-ref-22)
22. Parker JD, Talih M, Malec DJ, et al, 2017. [↑](#footnote-ref-23)
23. One record per patient per year, thus a patient may be included multiple times in the EHR. [↑](#footnote-ref-24)
24. Note: borders and shading are for demonstration purposes only. CSV exports columns separated with a comma. The results can be imported into Excel. [↑](#footnote-ref-25)