Prepared for:

**Centers for Disease Control and Prevention**

**Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare (Health FFRDC) – A Federally Funded Research and Development Center**

Childhood Obesity Data Initiative

**Task Order No. 200-2016-F-89363**

Prevalence Query Implementation Guide

For the Denver, Colorado Pilot (2019–2020)

Draft

Version DRAFT

July 23, 2020

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as official government position, policy, or decision unless so designated by other documentation. This guide may serve as a reference or framework for others implementing similar childhood obesity data solutions. However, the CODI Implementation Guide does not represent official views or guidance of CDC.

© 2020, The MITRE Corporation. All Rights Reserved.

Record of Changes

| Version | Date | Author / Owner | Description of Change |
| --- | --- | --- | --- |
| DRAFT | July 9, 2020 | Erin Tanenbaum / Health FFRDC | Initial Version |
| DRAFT | July 23, 2020 | Erin Tanenbaum / Health FFRDC | Edited |

Table of Contents

[1. Introduction 1](#_Toc46395229)

[1.1 Background 1](#_Toc46395230)

[1.2 Purpose 2](#_Toc46395231)

[1.3 Scope 2](#_Toc46395232)

[1.4 Audience 3](#_Toc46395233)

[1.5 Limitations 3](#_Toc46395234)

[1.6 Document Organization 4](#_Toc46395235)

[2. Data Pre-processing 5](#_Toc46395236)

[2.1 Growthcleanr 5](#_Toc46395237)

[2.2 Pre-processing CODI-PQ Algorithm 5](#_Toc46395238)

[2.2.1 American Community Survey Data 5](#_Toc46395239)

[2.2.2 EHR Data 6](#_Toc46395240)

[3. Prevalence Estimation Overview 7](#_Toc46395241)

[3.1 Estimate Outputs 7](#_Toc46395242)

[3.2 Importance of Statistical Weights 7](#_Toc46395243)

[3.3 Frequency with Which Processes Can and Should Be Run 7](#_Toc46395244)

[4. Algorithm Use 8](#_Toc46395245)

[4.1 User Inputs 8](#_Toc46395246)

[4.2 Understanding the Output 9](#_Toc46395247)

[4.3 Error Codes 11](#_Toc46395248)

[Appendix A. Methodology Notes 13](#_Toc46395249)

[A.1 Acronyms 13](#_Toc46395250)

[A.2 Age Adjustment 13](#_Toc46395251)

[A.3 Body Mass Index 14](#_Toc46395252)

[A.4 Prevalence 15](#_Toc46395253)

[A.5 Race Exclusion 15](#_Toc46395254)

[A.6 Race Imputation 15](#_Toc46395255)

[A.7 Statistical Weights 15](#_Toc46395256)

[A.8 Social Determinants of Health 17](#_Toc46395257)

[A.9 Standard Error 17](#_Toc46395258)

[A.10 Suppression 17](#_Toc46395259)

[A.11 Variance Estimation 20](#_Toc46395260)

[A.12 Weight Category 21](#_Toc46395261)

[A.13 ZCTA-3 21](#_Toc46395262)

[Appendix B. FIPS Codes and State Names 23](#_Toc46395263)

[References 25](#_Toc46395264)

[NOTICE 26](#_Toc46395265)

List of Figures

[Figure 1. NCHS Suppression Standards 20](#_Toc46314476)

List of Tables

[Table 1. CODI-PQ Output Data Dictionary 9](#_Toc46395266)

[Table 2. Example Synthetic Data 10](#_Toc46395267)

[Table 3. Projected Year 2000 U.S. Population Proportion Distribution by Age for Age Adjusting 14](#_Toc46395268)

[Table 4. Final List of Prioritized Social Determinants of Health 17](#_Toc46395269)

[Table 5. NCHS Data Presentation Standards for Proportions 18](#_Toc46395270)

# 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 Childhood Obesity Data Initiative (CODI). CODI will expand the ability to capture, standardize, integrate, and query existing patient-level electronic health record (EHR) and non-clinical community data. To accomplish this, open source tools and resources were updated to support childhood obesity research using distributed health data network (DHDN) infrastructure. DHDNs are traditionally networks of clinical organizations that map their EHR data concepts to a common data model so that researchers can query similar information across organizations.

As part of this effort, the Health FFRDC developed open source queries that can be used to estimate prevalence among children and teens by weight category using convenience samples (i.e., non-probability samples) typical of data generated solely from clinical settings and from EHRs. This document describes how to use the CODI prevalence queries using the selected technology as well as the nuances within the queries that generate estimates for three different geographic levels (state, state and ZIP code tabulation area [three digit] [ZCTA-3], and state and county). The queries will be referenced as the CODI prevalence queries (CODI-PQs).

## Background

Individuals are likely to have data at multiple organizations. It is critical to be able to use linked information gathered from different places to construct a complete picture of an individual for research purposes.

CODI-PQ is a process whereby researchers create childhood obesity prevalence estimates from non-probability samples derived from EHR data. Some researchers criticize the use of EHR data for characterizing population health and assert that EHR data sets are created from convenience samples (Flood et al., 2015). See the Limitations section for more information. Several strategies are used to overcome possible biases within the CODI prevalence estimation, including implementation of unique patient identifier, adoption of standardized rules for inclusion/exclusion criteria, as well as statistical procedures for data harmonization, weighting, and analysis (Bower, Patel, Rudy, & Felix, 2017).

The algorithms described in this document, were designed to

1. impute race for children and teens missing race information, and
2. create estimate childhood obesity prevalence by weight categories, including:

* **Underweight**: less than 5th percentile
* **Healthy Weight**: 5th percentile to less than the 85th percentile
* **Overweight**: 85th to less than the 95th percentile
* **Obesity**[[1]](#footnote-2): 95th percentile to less than 120 percent of the body mass index (BMI) value for the 95th percentile
* **Severe Obesity**: 120 percent or greater of the BMI value for the 95th percentile.

Weight categories are assigned to each child and teen based on their BMI percentile[[2]](#footnote-3) value.

Throughout this document, we refer to two different data types: (1) a large sample of data drawn from EHRs across the United States, which we refer to as IQVIA, based on the data set with which we tested the algorithm; and (2) a smaller set of data derived from a distributed health data network operating CODI infrastructure, which we refer to as CODI. Our algorithms can generate prevalence estimates derived from a large sample of data from across the United States (e.g., IQVIA data) at either the state or ZCTA-3 level. The CODI algorithms generate prevalence estimates at the county level.

## Purpose

CODI-PQ was developed in partnership between the Health FFRDC and CDC, with feedback from leading health researchers, to support the creation of weighted prevalence estimates.

The purpose of this document is to provide the guidance necessary for researchers to utilize CODI-PQ. It provides a description of:

* CODI-PQ data inputs and pre-processing
* CODI-PQ processing
* CODI-PQ analytic output

## Scope

This document provides implementation guidance for CODI-PQ. The algorithms leverage de-identified EHR height and weight information for children and teens ages 2 to 19. The CODI-PQ assumes users include all available EHR data for a geography and/or subpopulation.

The CODI-PQ solution is designed to operate with a single height and weight observation per year for each child or teen. This project used the open source R package called growthcleanr[[3]](#footnote-4) to identify height and weight values that were likely biologically implausible, and then randomly selected a paired height and weight for each child. Users may wish to use the CODI-modified growthcleanr tool[[4]](#footnote-5) to similarly clean their data prior to running the CODI-PQ.

The CODI-PQ and pre-processing algorithms described in this document were created and tested with IQVIA’s Ambulatory Electronic Medical Record (AEMR) data[[5]](#footnote-6) and synthetic data generated for CODI using Synthea.[[6]](#footnote-7) 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 source programs.

## Audience

The primary audience for this document is the researcher using CODI-PQ algorithms or output. This document is written with public health scientists and health services researchers as a primary focus, and is applicable to other efforts seeking a CODI-PQ solution. Please refer to GitHub for further technical details regarding the implementation of the algorithms.[[7]](#footnote-8)

## Limitations

The findings from CODI-PQ are subject to at least eight limitations. First, the findings might differ from those based on the National Health and Nutrition Examination Survey (NHANES), a probability-based survey that might be more representative of the general population. Obesity among national 2016 IQVIA estimates differed from the population-based sample in NHANES in the 2015 to 2016 wave (e.g., 17.2% obesity among IQVIA 2016 versus 18.5% in the NHANES for the 2015–16 season). Second, additional comparisons with IQVIA and CODI over multiple years are needed to determine whether the EHR-based estimates provide valid assessments of trends. Inclusion or exclusion of individual practices as well as growth or decline in participating practices may vary from year to year. Third, the CODI sample was not randomly selected from children and teens in the United States. Data derived from EHRs such as IQVIA and CODI uses a non-probability sample of medical encounters, where children and teens without a medical encounter in a year would be excluded. Children and teens captured in an EHR system may not be representative of the general population, due to issues such as informed presence bias. Informed presence is the belief that children and teens do not randomly go to the doctor’s office and thus are not randomly present in EHR data (Goldstein, Bhavsar, Phelan, & Pencina, 2016). Instead, critics suggest that children and teens who are ill or have a chronic condition are more likely to be present in EHR data. If critics are correct, children and teens in the healthcare system may be systematically different than non-patients. Critics also suggest that preventive care interactions and their EHR records may offer an exception to the rule and reduce informed presence. Yet selection bias concerns still remain because of factors that may influence the use of primary care services, such as education, health insurance coverage, and transportation (Christopher et al., 2016; Romo et al., 2016). Fourth, the records in IQVIA and CODI are shared after creation of data-sharing agreements with individual clinics. Not all medical clinics may share data, introducing additional potential for bias of children and teens who receive medical care from non-participating clinics. For this reason, post-stratification weighting is strongly recommended. However, weighting the results based on this non-probability sample might not be fully representative of childhood obesity in the United States. Fifth, each child’s estimate was based on clinical records from a single visit and was not verified. Human error, software glitches, and user errors could be present, and thus the actual height and weight of the child or teen may differ from the EHR record. Sixth, height and weight measurement protocols may differ between medical providers, even with clear recommendations aimed to increase consistency between medical professionals (Best & Shepherd, 2020). Seventh, the 2015 to 2016 NHANES results might not be directly comparable with the 2016 IQVIA estimates because they represent different populations. NHANES estimates obesity for the non-institutionalized civilian residential population of the United States using an address-based sample. NHANES sample frame does not include persons residing in nursing homes, members of the armed forces, institutionalized persons, or U.S. nationals living abroad. Childhood obesity estimates from NHANES exclude pregnant females from all published tables, whereas CODI estimates obesity for all children and teens who visit a participating clinic in the United States and does not exclude pregnant females or institutionalized children and teens, children and teens of active duty personnel, or children and teens who do not reside in a home (e.g., homeless and those visiting from abroad). Post-stratification weighting does adjust the count of children and teens within CODI to represent all U.S. residents. Eighth, IQVIA provides the birth year of each child and teen, whereas NHANES relies on proxy reported date of birth and CODI relies on the date of birth in medical records. Since date of birth is not available in IQVIA, all children and teens are assigned the same arbitrary birth date, and thus their age is an approximation, but not known.

## Document Organization

This document is organized as follows:

* Section 2 – Data Pre-processing
* Section 3 – Reference Data Sets
* Section 4 – Assumptions
* Section 5 – Algorithm Use
* Appendix A – Methodology Notes
* Appendix B – FIPS Codes and State Names
* References

# Data Pre-processing

This section describes pre-processing steps conducted by an EHR algorithm used to clean pediatric height and weight data from EHRs, the data source used to provide population-level estimates at the state, state and ZCTA3, and state and county levels, and how this data is merged to the EHR data.

## Growthcleanr

The growthcleanr function is an R package for cleaning data from EHR systems, focused on cleaning height and weight measurements. Growthcleanr implements the Daymont et al. (2017) algorithm, as specified in Supplemental File 3 within the Supplementary Material published with that paper. For more information, please visit <https://github.com/mitre/growthcleanr>. Data inputs must be pre-processed using either growthcleanr or something that will produce EHR data with similar assumptions (one record per person per year, identical variable names and format) prior to CODI-PQ.

## Pre-processing CODI-PQ Algorithm

The pre-processing CODI-PQ algorithm 1) imports EHR and American Community Survey (ACS) data, 2) compiles the data, and 3) imputes race for children and teens missing race information. CODI-PQ calculates statistical weights based on age, sex, race, and geographic location (education) for all children and teens, yet clinical data includes children and teens whose race information is unknown. This algorithm is required to pre-process all data before calculating prevalence estimates.

### American Community Survey Data

CODI-PQ uses the American Community Survey 2018 five-year estimate public use file (ACS five-year PUF) for race imputation and statistical weighting to population totals. EHR data uses population estimates at the state and ZCTA-3[[8]](#footnote-9) levels. In contrast, CODI-PQ estimates from CODI-derived samples for prevalence values at the county level utilize population estimates at the state, county, and ZCTA-3 levels.

The ACS is a series of monthly surveys that are compiled to create annual estimates at a geographic level (state and ZCTA-3 or state and county) that are released as five-year estimates. The sample size for the ACS is about 3.54 million addresses per year, and the survey contains various questions regarding demographic and socioeconomic status information about the household. The variables used in these programs pertain to a combination of age, sex, race, and education within a geographic area. The PUF population by age estimates do not isolate children aged newborn to 1 year of age from those 2 to 5 years of age. Thus, the CODI-PQ discounts the PUF estimate of children aged 0 to 5 years by 32.6%, as approximately 67.4% of the age 0 to 5 population in 2018 were age 2 to 5 according to the ACS one-year PUF.

Our sampled data from IQVIA included the child or teen’s state and ZCTA-3; therefore, the pre-processing algorithm merges ACS data based on state and ZCTA-3. Data is released by the Census Bureau at the state and ZCTA levels. Census data was aggregated to the state and ZCTA-3 levels and made available at <add link here>.

CODI data includes the child or teen’s state and county; therefore, ACS data is merged to CODI-derived data based on state and county within the algorithm. Census data for CODI is made available at <add link here>.

### EHR Data

Process all EHR data with the growthcleanr[[9]](#footnote-10) and pre-processing algorithm before creating the prevalence estimates. The output of the pre-processing algorithm will include state and location (state + ZCTA-3 or state + county), age, sex, race/ethnicity, chronic conditions, height, weight, weight category, and count of visits to a provider for each child or teen.

# Prevalence Estimation Overview

This section provides an overview of the algorithm output formats, the importance of statistical weights, and the frequency with which the pre-processing and CODI-PQ algorithms can and should be run.

## Estimate Outputs

The final prevalence estimates are calculated at each weight category (1 through 4 and 4a) for crude estimates, weighted estimates, and age-adjusted estimates (optional). All prevalence estimates are reviewed to ensure alignment with the National Center for Health Statistics (NCHS) Suppression Standards (see Appendix A. Suppression). A final report is generated and exported as a comma separated value (csv) file. If all numeric prevalence estimates are blank (.), then one or more error reasons will be provided (see Error Codes).

## Importance of Statistical Weights

Data derived from EHRs represents a non-probability sample.

For CODI-PQ estimates, the use of statistical weights is recommended 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 estimates and overstated significance levels.**

Learning about the features of the weighting will help ensure that the results of analyses represent unbiased estimates with accurate statistical significance levels. See Appendix A Statistical Weights for more information.

A statistical weight is assigned to each child or teen per year. If prevalence reports are created across years and a child or teen has an estimate in more than one year, then a statistical weight is created for the most recent year only, and all other years are removed from the estimate. Statistical weights can be thought of as denoting the number of children and teens in the population represented by that sample person in EHR, accounting for differences between the distribution of the sample and total populations.

## Frequency with Which Processes Can and Should Be Run

The CODI-PQ may be processed multiple times on the same data file with different inclusion criteria (e.g., age, sex). The pre-processing algorithms should be processed only once per set of EHR records prior to creating one or more CODI-PQ prevalence tables. If EHR records for the same year are available at a later time and appended to a previously created EHR file, the growthcleanr and pre-processing algorithms will need to be completed again to ensure the correct health encounter is retained. The race of children and teens with unknown race may be changed each time the pre-processing algorithm is processed.

The CODI-PQ was published in the summer of 2020. Statistical programs require review and updates over time. We recommend statistical review of the programs each year. In addition, we strongly recommend updating the race imputation model after the release of 2020 Census data.

# Algorithm Use

This section describes the user inputs and provides guidance on interpreting the CODI-PQ output. Once geographic and demographic criteria are selected, optional algorithms are chosen, and file paths are defined, the CODI-PQ subsets the data file based on the inclusion criteria, generates statistical weights, calculates prevalence estimates, and suppresses estimates as needed before generating the report.

## User Inputs

The CODI-PQ algorithm requires user inputs to identify the year(s), geographic area(s), age group(s), sex, and race value(s) to be included in the analyses. The user may also select two optional inclusions to the algorithm: race imputation and age adjusting. More details on these options are listed below. The user must also define the computer folder where data and programs are stored.

* **Year**: The user may select one or more years for inclusion in the estimates. Selection of multiple years pools the records across years to create one estimate of prevalence, not one estimate per year.
* **Geographic area**: For CODI, geographic area is defined based on the state and county. For IQVIA, geographic area is defined by either the state or the state and ZCTA-3.
* **Age groups**: In both CODI and IQVIA, age groups include ages 2 to 4, 5 to 9, 10 to 14, 15 and 17, and 18 to 19 years of age.
* **Sex**: Sex is defined as male or female. At the time of publication, there is not a way to calculate weight category for non-binary children or teens. For this reason, children with non-binary or missing sex information are excluded from analysis.
* **Race**: In both CODI and IQVIA, race is defined as White, Black/African American, Asian, or Other.
* **Race imputation:** Setting race imputation to yes allows the algorithms to include all available EHR data for children and teens even if the medical record did not include a known race. The availability of race information for children and teens varies by geography and data provider. When creating estimates by race, it is recommended to exclude imputed race.
  + *Cautions:*
    - Race is used to weight all children and teens, which means the weighted estimates are subject to increased potential for bias as the percentage of imputed race increases. Excluding imputed race decreases the sample size, increases the variance, and may reduce bias. Conversely, including imputed race increases the sample size, reduces the variance, and may increase bias.
    - If race imputation is selected, the prevalence output table provides the percentage of the children and teens with imputed race. Applying statistical weights for all estimates is highly recommended; however, users should employ caution. Although there is not a fixed cutoff point and researcher review is strongly recommended, reporting estimates with 40% or more children and teens with imputed race is not recommended.
* **Age adjusting:** Setting age adjusting to yes allows the algorithm to include two additional variables in the output: age-adjusted prevalence and age-adjusted standard error. Researchers age-adjust to eliminate differences in observed estimates that result from differences in the age distribution of the population among geographies. See Appendix A. Age Adjusting for more information.

## Understanding the Output

Once complete, CODI-PQ generates a prevalence table as a csv file. Table 1 provides an overview of the variables included, and Table 2 provides example output based on synthetic data. Note in the sample provided in Table 2, descriptive information about the CODI-PQ user inputs, error codes, sources of technical documentation, caveats, and a recommendation 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 1. CODI-PQ Output Data Dictionary

|  |  |
| --- | --- |
| **Column** | **Description** |
| Order | Row order |
| Weight Category | The weight category based on BMI percentile. |
| Sample | The observed (or unadjusted, or crude) count of children and teens in the study population. |
| Population | The weighted (or adjusted) count of the study population. |
| Crude Prevalence | The observed (or unadjusted, or crude) prevalence in the study population. |
| Crude Prevalence Standard Error | The observed (or unadjusted, or crude) standard error in the study population. |
| Weighted Prevalence | Prevalence estimate based on weighted counts.  The use of sample weights is highly recommended. A sample weight is assigned to each sample person. It is a measure of the number of children and teens in the population represented by that sample person. See Appendix A. Sample Weights for more information. |
| Weighted Prevalence Standard Error | Standard error estimates based on weighted counts. See Appendix A. Variance Estimation for more information. |
| Age-adjusted Prevalence | Prevalence estimate based on weighted, age-adjusted counts. See Appendix A. Age Adjustment for more information. |
| Age-adjusted Prevalence Standard Error | Standard error estimates based on weighted, age-adjusted counts. |

Table 2. Example Synthetic Data

| Order | Weight Category | Sample | Population | Crude Prevalence | Crude Prevalence Standard Error | Weighted Prevalence | Weighted Prevalence Standard Error |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | (1) Underweight (<5th percentile) | 61 | 5,801 | 4.16 | 0.25 | 4.80 | 0.82 |
| 1 | (2) Healthy Weight (5th to <85th percentile) | 824 | 82,020 | 56.17 | 1.3 | 67.88 | 3.58 |
| 1 | (3) Overweight (85th to <95th percentile) | 251 | 14,752 | 17.11 | 0.98 | 12.21 | 2.42 |
| 1 | (4) Obesity (>95th percentile) | 331 | 18,255 | 22.56 | 1.09 | 15.11 | 2.89 |
| 1 | (4b) Severe Obesity (>120% of the 95th percentile) | 100 | 6,085 | 7.52 | 1.01 | 5.25 | 2.22 |
| 2 | Totals: | 1,467 | 120,828 |  |  |  |  |
| 3 | Dataset: IQVIA, 2016-2018 |  |  |  |  |  |  |
| 4 | Query Parameters: AGE RACE SEX GEOGRAPHY YEAR |  |  |  |  |  |  |
| 5 | AGE: (02 - 04, 05 - 09, 10 - 14, 15 - 17, 18 - 19) |  |  |  |  |  |  |
| 6 | SEX: (Male, Female) |  |  |  |  |  |  |
| 7 | RACE: (White, Black, Asian, Other) |  |  |  |  |  |  |
| 8 | RACE Suppressed: (Other) See https://github.com/NORC-UChicago/CODI-PQ for more information. |  |  |  |  |  |  |
| 9 | RACE Imputed: People with unknown race were excluded. |  |  |  |  |  |  |
| 10 | Geography: (XX) State\_Name |  |  |  |  |  |  |
| 11 | Year: 2018 |  |  |  |  |  |  |
| 12 | Weighting cells were collapsed for: (Geography) |  |  |  |  |  |  |
| 13 | AGE adjusted?: (No) |  |  |  |  |  |  |
| 13 | Error Codes: (None) |  |  |  |  |  |  |
| 14 | Technical Documentation: See https://github.com/NORC-UChicago/CODI-PQ for more information and full details on data sources and methodologies. |  |  |  |  |  |  |
| 15 | Query Date: Friday, 1 July 2020 4:09:46 PM |  |  |  |  |  |  |
| 15 | Suggested Citation: AEMR-US version 5 OMOP 5 [Aug 2019 Release] accessed through the E360TM Software-as-a-Service (SaaS) Platform. Accessed through prevalence query on Friday, 1 January 1960 4:09:46 PM |  |  |  |  |  |  |
| 16 | Caveats |  |  |  |  |  |  |
| 17 | Children with either missing or invalid age, sex, height, weight, or geography are not included in counts and prevalence estimates. |  |  |  |  |  |  |
| 18 | The method used to calculate the standard errors are documented in the technical documentation. |  |  |  |  |  |  |
| 19 | The population estimates are based on age-race-sex-location specific counts from the 2014-2018 American Community Survey Five-year Estimates released by the Census Bureau on December 19, 2019. |  |  |  |  |  |  |

## Error Codes

There are several reasons the algorithm may fail to generate estimates and produce an error code, including:

1. One or more demographic or geographic category has no groups selected. One or more group must be selected in each category. Ensure that each demographic and geographic category has one or more groups selected (e.g., age group, select an age range for inclusion).
2. Years are out of scope for either IQVIA or CODI.

* Acceptable years include 2016, 2017, and 2018 for IQVIA.
* Acceptable years include 2016, 2017, 2018, and 2019 for CODI.

1. Geographic level (GEO\_GROUP) has been left blank or has been set to an unacceptable value. To remedy the issue, update the GEO\_GROUP variable to either STATE, ZCTA-3, or county.
2. State, county, or ZCTA-3 is incorrectly specified. Review the lists and ensure each value is:

* Surrounded by quotations,
* Comma delimited, and/or
* The correct length (e.g., “001,” “002,” “003”).

1. Current selection criteria return an insufficient number of children and teens and do not meet minimum threshold to estimate sample weights. Ensure that selections are correct (e.g., correct list of state codes, ZCTA-3 values, or county codes) or include additional geographic or demographic categories (e.g., add additional communities or include additional or all races, age groups, sex, etc.).
2. Iterative proportional fitting weighting routine has failed to converge. Please revise selection criteria and rerun algorithm.
3. One or more rows has suppressed results. Totals and percentages are not available for all results due to suppression constraints.
4. A SAS error has occurred within the algorithm. Review the SAS log or contact a system administrator for further assistance.
5. Methodology Notes
   1. Acronyms

|  |  |
| --- | --- |
| **Term** | Description |
| **ACS** | American Community Survey |
| **BMI** | Body Mass Index |
| **CDC** | Centers for Disease Control and Prevention |
| **CODI** | Childhood Obesity Data Initiative |
| **CODI-PQ** | Childhood Obesity 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 |
| **IQVIA** | IQVIA’s Ambulatory EMR (AEMR) |
| **MSE** | Mean Square Error |
| **NCHS** | National Center for Health Statistics |
| **NHANES** | National Health and Nutrition Examination Survey |
| **PUF** | Public Use File |
| **SDOH** | Social Determinants of Health |

* 1. Age Adjustment

Data is age-adjusted in order to eliminate differences in observed estimates 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 (Klein & Schoenborn, 2001). The specific age groups used for age adjustment are 2 to 4 years, 5 to 14 years, and 15 to 19 years. Age adjusted values may be similar to the weighted values since age is used within the weighting algorithm.

Age adjustment, using the direct method, is the application of age-specific rates in a population of interest to a standardized age distribution in order to eliminate differences in observed rates 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 rate.

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

Table 3. Projected Year 2000 U.S. Population Proportion Distribution by Age for Age Adjusting

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

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

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

* 1. Body Mass Index

BMI is a person’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 diagnostic of an individual’s body fatness or health.

For children and teens, BMI is age- and sex-specific and is often referred to as BMI-for-age. In children and teens, a high amount of body fat can lead to weight-related diseases and other health issues. Being underweight can also put individuals at risk for health issues.

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

* 1. Prevalence

A prevalence rate is either:

* Crude: the proportion of the sample that has a health condition at a point in time.
* Weighted: the estimated proportion of the population that has a health condition at a point in time. See Appendix A. Statistical Weights for more information.
* Age-adjusted: the estimated proportion of the population (adjusted by three age categories) that has a health condition at a point in time. See Appendix A. Age Adjustment for more information.
  1. Race Exclusion

One or more races may be excluded from prevalence estimates. A race may be excluded by the algorithm if the sample includes 20 or fewer children or teens of a particular race.

* 1. Race Imputation

Race is a required input for CODI-PQ. The pre-processing algorithm imputes race for each child or teen missing race information. The algorithm operates sequentially in three phases, imputing race for children and teens who:

* Have a chronic condition,
* Are identified as Hispanic and do not have a chronic condition, or
* Neither have a chronic condition nor are identified as Hispanic.

The algorithm requires a combination of clinical and ACS data.

Once complete, the output from each phase is combined with each child or teen with an EHR-provided race, an imputed race, or categorized as “unknown.”

A child’s race may be missing after race imputation for one of four reasons:

* The child’s location (state and ZCTA-3) is either invalid or did not have a population count in the 2018 ACS.
* The child’s age is outside of the scope of the algorithm or is unknown. Only children ages 2 to 19 are in scope.
* The sex of the child or teen is unknown.
* The height of the child or teen is unknown.
  1. Statistical Weights

CODI and IQVIA data is derived from EHRs. As described in the Limitations section, applying statistical weights is strongly recommended to overcome potential biases introduced by the EHR sampling methodology. Ratio adjustments are applied to all sampled children and teens. Ratio adjustment is a statistical weighting technique aimed to improve the accuracy of survey estimates by both reducing bias and increasing precision (Little, 1993). One method for accomplishing 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 estimates.

Implementing raking procedures requires the specification of appropriate weighting classes or cells. Data used to form classes for adjustments must be available for both sample and the population. CODI-PQ raking includes social determinant of health categories – age, sex, race, and education levels in the surrounding area (based on percentage of adults in the community with a bachelor’s degree or higher). See Appendix A. Social Determinants of Health. Once formed, the weighting classes are assessed, and cells with small sample counts are collapsed with their nearest neighbor to reduce variability in the estimates. The collapsing follows these guide points:

* **Age** = age category less than or greater than current
* **Sex** = do not collapse
* **Race** = do not collapse, instead exclude small cell categories from prevalence estimates
* **Education**= community with a similar education level

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 estimates would be adjusted by sex groups. This procedure continues 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 estimates caused by unusually high or low weights in some categories.

The fundamental objective of our CODI-PQ is to provide 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 square error (MSE) of the primary statistics of interest as low as possible. The MSE of a survey estimate 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 estimates when compared with crude variances.

The increases in variance result from the added variability in the sampling weights due to the adjustments. Thus, the researcher 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.

* 1. Social Determinants of Health

The statistical weights cell formulation is based on the Childhood Obesity Data Initiative – Social Determinants of Health (SDOH) Prioritization report[[10]](#footnote-11) as well as other artifacts. The SDOH report was developed in partnership between the Health FFRDC and CDC, with feedback from leading health researchers. Table 4 presents the finalized list of obesity-related social determinants of health used in the CODI-PQ.

Table 4. Final List of Prioritized Social Determinants of Health

|  |  |  |
| --- | --- | --- |
| **Concept** | **Measure** | **Source** |
| Age | Age of child, continuous | EHR |
| Race | Individual race, categorical | EHR |
| Sex | Individual sex, categorical | EHR |
| Education | Educational attainment of population age 25 to 64 years of age | ACS |

* 1. Standard Error

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

* 1. Suppression

Prevalence estimates may be suppressed. The CODI-PQ method for data suppression is based on the NCHS data presentation standards for reporting proportions in NCHS reports and data products (Parker et al., 2017). This method was developed by the Data Suppression Workgroup at NCHS.

The multistep NCHS Data Presentation Standards for Proportions are based on a minimum denominator sample size and on the absolute and relative widths of a confidence interval calculated using the Clopper-Pearson method. The NCHS Data Presentation Standards for Proportions are applied to all CODI-PQ output. Using these standards, some estimates will be identified as suppressed. Based on the criteria described in the standards, others may be identified as unreliable, and some estimates may be identified for statistical review. The CODI-PQ algorithm suppresses only estimates. It is up to the analyst to identify estimates as unreliable and/or to perform statistical review prior to publication and use. The Table 5 and Figure 1 briefly describe the Standards.

If one or more rows is suppressed, the user may choose to increase their research criteria by including additional years of data, increasing the geography, or including more age, race, or sex categories.

Table 5. NCHS Data Presentation Standards for Proportions

| **Statistic** | **Standard** |
| --- | --- |
| Sample Size | Estimated proportions should be based on a minimum denominator sample size and effective denominator sample size (when applicable) of 30. Estimates with either a denominator sample size or an effective denominator sample size (when applicable) less than 30 should be suppressed. If the number of events is 0 (or its complement[[11]](#footnote-12)), then the denominator sample size should be used to obtain confidence intervals. If all other criteria are met for presentation, an estimate based on 0 events (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 events is greater than 0 and the degrees of freedom criterion (below) is met. If the number of events is 0 (or its complement) or the degrees of freedom criterion is not met, then the estimate 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 estimate 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. |

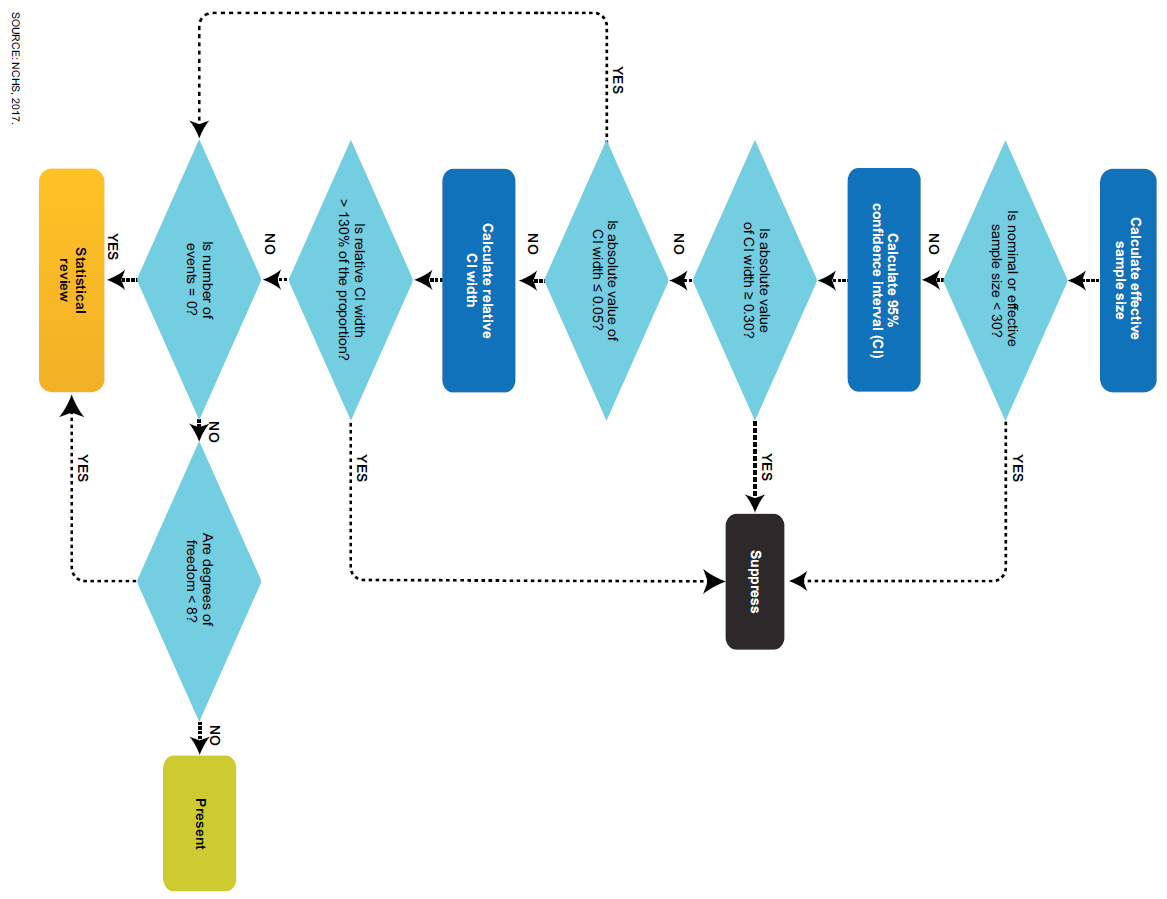


Figure 1. NCHS Suppression Standards

* 1. Variance Estimation

Estimated obesity prevalence is derived using the sample weights and data on obesity status. Obesity prevalence is a ratio, and the ratio estimator, , corresponds to a population parameter, , such as the true but unknown obesity rate. To define the population parameter, let

= the number of children and teens in stratum (), where stratum refers to state-ZCTA-3 for IQVIA data and to (ZIP code, census tract?) for CODI data

= the value of for child/teen of stratum (often the possible values of are 0 and 1, as when indicates whether a child/teen is obese or in a specified weight category)

= 0 or 1, indicating whether child/teen 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 children and teens in stratum

= the sampling weight for child/teen in stratum

= the value of for child/teen in stratum

= the value of the domain indicator for child/teen in stratum

The distinction between and and between and is merely that for and the subscript refers to sampled children and teens within stratum *,* whereas for and they refer to children and teens in the population in stratum . Then, the combined ratio estimator for is   
To estimate the variance of, a Taylor-series approximation is used (Wolter, 2007). Within stratum, linearization yields the new variable.

Then, letting

the Taylor-series approximation to the variance of is

See Appendix A. Suppression for more information.

* 1. Weight Category

Prevalence estimates are calculated from a child’s body mass index. Each child and teen with a height, weight, sex, and age may be classified as one (or more for severe obesity) of the following categories:

1. Underweight: less than 5th percentile
2. Healthy Weight: 5th percentile to less than the 85th percentile
3. Overweight: 85th to less than the 95th percentile
4. Obesity[[12]](#footnote-13): 95th percentile to less than 120 percent of the BMI value for the 95th percentile
   1. Severe Obesity: 120 percent or greater of the BMI value for the 95th percentile.

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

* 1. 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.[[13]](#footnote-14)

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 IQVIA and ACS data.

1. FIPS Codes and State Names

| **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 |

References

1. 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
2. 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.
3. 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
4. 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.
5. 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
6. Best, C., & Shepherd, E. (2020). Accurate measurement of weight and height 2: Calculating height and BMI. *Nursing Times* [online]; 116: 5, 42-44.
7. 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>
8. 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).
9. Little, R. (1993). Post-stratification: A modeler’s perspective. *Journal of the American Statistical Association*, 88(423), 1001-1012. doi:10.2307/2290792
10. Klein, R. J., & Schoenborn, C. A. (2001). Age adjustment using the 2000 projected U.S. population. *Healthy People 2000 statistical notes*, (20), 1–9.
11. 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).
12. 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.

© 2020 The MITRE Corporation.

1. Note: prevalence estimates of obesity will include two categories: those that are category 4 and 4a. [↑](#footnote-ref-2)
2. Based on randomly selected paired height and weight data that are processed to extract biologically implausible values. [↑](#footnote-ref-3)
3. 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, *JAMIA* 24(6): 1080–1087. [↑](#footnote-ref-4)
4. The MITRE Corporation. mitre/growthcleanr. Available: https://github.com/mitre/growthcleanr. [↑](#footnote-ref-5)
5. IQVIA’s Ambulatory EMR (AEMR) data contains the clinical data of approximately 74 million patients since 2006 and 22 million per year from all 50 states recorded by over 100,000 providers who are affiliated with over 800 ambulatory large practices and physician networks. About 40% of the contributing physicians are primary care practitioners, and the remaining 60% are specialists. These records capture key clinical variables such as laboratory values, weight, and blood pressure to prescriptions, diagnoses, hospital metrics, and therapeutic outcomes. AEMR is also used to connect patient vitals, health behaviors, and risk factors to diagnosis and treatment and develop insights based on provider treatment decisions and prescribed medications (in contrast to dispensed medications).

   Source: AEMR-US version 5 OMOP 5 (Aug 2019 release) accessed through the E360TM Software-as-a-Service (SaaS) Platform. [↑](#footnote-ref-6)
6. <https://synthetichealth.github.io/synthea/> [↑](#footnote-ref-7)
7. <https://github.com/NORC-UChicago/CODI-PQ> [↑](#footnote-ref-8)
8. <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html> [↑](#footnote-ref-9)
9. <https://github.com/mitre/growthcleanr> [↑](#footnote-ref-10)
10. <https://github.com/NORC-UChicago/CODI-PQ> [↑](#footnote-ref-11)
11. The complement of a proportion p is (1 – p). The complement of the number of events in the numerator for p is the number of events in the numerator for (1 – p). [↑](#footnote-ref-12)
12. Note: prevalence estimates of obesity will include two categories: those that are category 4 and 4a. [↑](#footnote-ref-13)
13. <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html> [↑](#footnote-ref-14)