Tracking Data Validation Protocol

Background:

Data for the Tracking Network are received either through the secure file gateway, retrieved from a data provider public site, or sent to Tracking staff through another secure route. Once the raw data are retrieved, they undergo a series of data quality checks as they move through the data management process and phases, starting from the raw data to the calculated measures published on the Tracking Network Public Portal.

The tier one data cleaning checks are performed on the data as submitted through a secure data submission gateway or retrieved directly from the data source. Data that pass these data structure and cleaning checks are accepted into the Tracking environment. They undergo a second tier of validations for accuracy before being used to calculate measures for display on the Tracking Network Public Portal. The tier two validation checks performed on the clean data fall under four overarching themes: strange patterns, lack or excess of data, outliers or inconsistencies, and unexpected results. These checks and themes are described below.

Validation Checks by Theme and Phase:

1. **Tier One:** Data Cleaning and Completeness (Gateway Checks):
    a. Missing or invalid values for variables in the dataset needed for measure calculation (*Refer to the data dictionary and how-to guide):*
        i. Check year with corresponding metadata (if applicable)
        ii. Check for valid state IDs and/or FIPS codes
        iii. Check for valid coding against a corresponding data dictionary and/or metadata
        iv. Check for invalid county FIPS codes: compare county FIPS code values with reference table (Tracking uses 2010 Census geographies)
    b. Data structure:
        i. Check the data structure and attributes match provided schema/data dictionary
            1. Check for order of variables and table structure
            2. Check variable restraints and constraints
    c. Duplicates:
        i. Use the minimum number of variables needed to determine a unique record for each dataset to identify duplicates
2. **Tier Two:** Raw Data Validation Checks *(See provided sample code in SAS and R)*
    a. Strange patterns:
        i. Check for all even or odd values in the dataset
    b. Lack/excess of data – record level:
        i. Check all expected states are in the data
        ii. Review the number of records by state are as expected
        iii. Check the number of unique counties in the data, including 'U' if applicable. Compare number to expected number of counties per state.
        iv. Variable level frequencies for key variables, check all expected levels of the variables are present, and review the number of records by variable levels
            1. Check number of records by county: Frequency by county code

2. Verify age groups: Record frequency by age group
3. Verify sex variable: Record frequency by sex
4. Check completeness of race and ethnicity variables
5. Check records by other variables as defined in a data dictionary

v. Check total number of records by year, ensure the records by year is relatively consistent

vi. Confirm ratio of counts by variables are consistent with existing data

1. Ensure percent or ratio of counts by advanced variables is relatively consistent across years

c. Outliers/inconsistencies – values:

i. Review sum of events by year and county

ii. Review sum of events by year and age

iii. Review sum of events by year and sex

iv. Review sum of events by month and county. Look for unexpected concentrations or sparseness of counts by month inconsistent with seasonal trends (if applicable).

v. Check min/mean/max number events (outcome data) or concentration values (exposure data) by state, year, and additional stratification variables

d. Unexpected results:

i. Compare total count and % differences between new years of data compared with previous years of data (archive)

ii. Review the distribution of counts and rates across all years of data. Tracking uses boxplots to compare across years of data *(See boxplot code)*

iii. Compare crude rates between average archive data and newly submitted years of data using a Poisson rate comparison test *(See Poisson check code)*

iv. Review the distance from the archive mean by standard deviations. Flag extreme outliers for review *(See standard deviation check code)*

v. Compare/spot check cases or rates of national dataset to published data from the data source site