**Data Analysis Plan (DAP)**

1. **Background**

The Behavioral Risk Factor Surveillance System (BRFSS) is a database of ongoing health related telephone surveys that collect data on risk behaviors, chronic health conditions, health care access, and use of preventive services from adults (>= 18 years) in all 50 states in the U.S. and additional territories. The objective of the BRFSS is to collect uniform state-specific data on health related outcomes and behaviors relating to the leading causes of death and disability in the US. The researcher is interested in whether there is a significant difference in BMI between California and North Carolina. I am interested in investigating cancer outcomes (ever had cancer) in Colorado stratified by demographics such as race and education.

1. **Methods**

**2.1 Study Description**

The population of people in this study is adults 18 years or older in the U.S. Our cohort consists of N=48332 from 2022 and 2023. The researcher is only interested in California, Colorado, Connecticut, Delaware, New Mexico, North Carolina, and West Virginia in the final data set. The BRFSS survey is conducted as a randomized, representative landline/cell line telephone survey. The BRFSS assesses factors related to risk behaviors, chronic health conditions, health care access, and use of preventive services. It includes modules on diabetes, vaccination, covid, respiratory health, cancer survivorship and screening, cognitive decline, adverse childhood experiences, social determinants of health, marijuana and tobacco usage, and demographic information.

**2.2 Data Management Plan**

**Demographics Dataset**

The Demographics dataset was received as a .xlsx file. It will be imported into SAS using a PROC IMPORT statement with DBMS set to XLSX. The raw file will be stored in SAS Online under the BRFSS project and stored in the 1\_Source folder. The imported data set will be kept in the 2\_Import folder. All variables with missing values will be properly coded as such. Variables will be created as follows:

* Race/Ethnicity: A FORMAT statement will be used to code ethnicity using Boolean logic on the \_PRACE2 and \_HISPANIC variables, to create values for Non-Hispanic White, Non-Hispanic Black, Non-Hispanic American Indian or Alaskan Native, Hispanic, or Other.
* Insurance: A FORMAT statement will be used to factor and label patients’ primary insurance with the PRIMINSR variable. Final form will be Private/commercial/Medicare/Medigap, Medicaid/CHIP, Tricare, Indian Health Services, Other, and None/Uninsured/Unknown.
* Education: A FORMAT statement will be used to code education using Boolean logic on the EDUCA variable, to create values for <High School, Graduated High School, Some College, and 4+ Year College Degree.
* Age Category: a FORMAT statement will be used to code age category using Boolean logic to collapse the \_AGEG5YR variable into the desired categories of 18-44, 45-64, and 65+.
* Income: A FORMAT statement will be used to code income using Boolean logic to collapse the INCOME3 variable into the desired levels of <15k, 15-<35k, 50k-<100k, 100k-<200k, and 200K+.
* Total number of disabilities: A SUM statement will be used to create a variable representing the total of the deaf, blind, decide, diffwalk, diffdres, and diffalon variables.

**Outcomes Dataset**

The Demographics dataset was received as a .csvfile. It will be imported into SAS using a DATA and INFILE statement, specifying MISSOVER DSD and FIRSTOBS=2, and manual entering of the variable names and properties. The raw file will be stored in SAS Online under the BRFSS project and stored in the 1\_Source folder. The imported data set will be kept in the 2\_Import folder. All variables with missing values will be properly coded as such. Variables will be created as follows:

* Ever had a CHD or heart attack: A FORMAT statement will be used with Boolean logic to code ever having a CHD or heart attack using the CVDCRHD4 and CVDINFR4 variables. If the patient has a 1 for at least one of these, they will be coded as a yes. If they have 0 for one or both of them (with no 1’s) they will be coded as a No. If both variables are missing or a 7 or a 9, they will be coded as missing data.
* Asthma status: A FORMAT statement along with Boolean logic will be used to code asthma status using the ASTHMA3 and ASTHNOW, into current, former, never, or don’t know/not sure.
* BMI: Will be calculated using the appropriate equation and the WEIGHT2 and HEIGHT3 variables.
* BMI Cat: The BMI value will then be used in a FORMAT statement to place patients into categories of Underweight (BM<18.5), Normal weight (18.5<=BMI<25.0), Overweight (25.0<=BMI<30.0), and Obese(BMI>=30.0).
* Covid Positive: Patients will be placed into categories using a FORMAT statement and Boolean logic with the variable COVIDPOS, to place them into levels of Yes (any type), No, or Unknown/Refused/Missing.
* Ever had cancer: A FORMAT statement will be used along with Boolean logic using the CNCRDIFF variable to place patients into the categories of either having ever had cancer or not.
* Age cat of first cancer diagnosis: A FORMAT statement along with Boolean logic will be used on the CNCRAGE variable to place patients into categories of 0-<15, 15-<18, 18-<39, and 40+.

**Risk Factors Dataset**

The Risk Factors dataset was received as a .xlsx file. It will be imported into SAS using a PROC IMPORT statement with DBMS set to XLSX. The raw file will be stored in SAS Online under the BRFSS project and stored in the 1\_Source folder. The imported data set will be kept in the 2\_Import folder. All variables with missing values will be properly coded as such. Variables will be created as follows:

* % of 30 days with physical health not good: This score will be computed using the  PHYSHLTH variable using a DO loop.
* % of 30 days with mental health not good: same as above but using the MENTHLTH variable.
* Smoking status: A FORMAT statement using Boolean logic and the LCSFIRST variable will be used to class patients into either Current smoker, Former smoker, Never smokes, or Don’t know/refused/missing.
* Number of years smoked: a FORMAT statement using Boolean logic and the LCSFIRST variable will be used.
* Number of packs per day: The LCSNUMCG variable will be used.
* Drank at least on drink of alcohol in the past 30 days: A FORMAT statement using the LCDAY4 variable and Boolean logic to collapse groups will be used.
* How many months between first and second covid vaccine: The  time between the  COVIDFS1 and COVIDSE1 variables will be used to compute this variable using a DO statement.
* % of 30 days used marijuana: Will be calculated using the MARIJAN1 variable and a DO statement.
* Use any type of marijuana: A FORMAT statement using Boolean logic on the MARIJAN1 variable.

**Reference table Dataset**

The reference table dataset is a .SAS7BDAT and will be imported directly into SAS. The raw file will be stored in SAS Online under the BRFSS project and stored in the 1\_Source folder. The imported data set will be kept in the 2\_Import folder. All variables with missing values will be properly coded as such. This data set has already been transposed, and Date of Interview will be kept as a numeric date format, and converted into this format if not already.

*You will need to explain how you plan to merge all the datasets to create one final clean analysis dataset. To be able to merge, the variable you merge on must share the same characteristics in all data sets you want to merge.*

*(HINT: The dataset that has the correct characteristics of the merging variables is the reference table. You will want to base your mergers on those)*

**Combined data analysis dataset**

All datasets will be merged using the SEQNO variable if available (demographics, outcomes, and reference table) and otherwise id will be used (risk factors).