Milestone #5

Rachael Baartmans, Lara Petalio, Christine Truong

2022-11-28

Problem Statement

For this project, we aimed to investigate how tobacco use primarily impacts mental illness among smokers in California in 2011, as well as explore how race and location of cigarette purchase can impact disease status. In addition, we were interested in utilizing the 2011 CSC data to help us understand tobacco consumption in terms of "pack-years," which is the product of the number of packs of cigarettes smoked per day and the years a person has smoked, and comparing the average number of pack-years to asthma, heart disease, diabetes, and mental illness. Based on our observations from our analyses, we hope to provide some insight and suggestions for CDPH on how to redirect resources that can strengthen smoking cessation strategies from this project.

Methods

Data Source

Sponsored by the State of California's Department of Public Health (CDPH), the data source for this project is the 2011 California Smokers Cohort (CSC), which is a part of the California Tobacco Surveys (CTS) that collected information on the prevalence of tobacco use in California and behaviors among smokers in order to inform tobacco prevention efforts. To help CDPH better assess the effectiveness of smoking cessation strategies, the 2011 CSC data specifically investigates characters associated with quitting behavior among only smokers identified through telephone contacts purchased from data brokers, as well as through the California Health Interview Survey Longitudinal Smokers Survey (CLSS) between July 8, 2011 and December 8, 2011. This data from the surveys conducted were split into two separate data sets, with the first containing information regarding each participant's smoking status, behaviors associated with smoking, and demographics; the second data set includes information regarding each participant's race and disease outcomes. The data sets were stored as data frames called `smoker_data` and `race_data`, respectively.

Data Wrangling

Variables Kept During Import

Our group decided to keep only a select few variables from each of the two data sets during the importing process. The variables kept from each data set are as follows:

From the `smoker_data` data set:

`psraid`

`smokstat`

`WHEREBUY`

`BUYCALIF`

`HOWMANY`

`SMOK6NUM`

`SMOK6UNI`

From the `race_data` data set:

`ID`

`NERVOUS`

`WORRYING`

`PROBINTR`

`PROBDOWN`

`ASTHMA`

`HEARTDIS`

`DIABETES`

`OTHMENILL`

`race01`

`race02`

`race03`

`race04`

`race05`

`race06`

`race07`

`race08`

`race09`

`race10`

`race11`

`race12`

`race13`

`race14`

`race15`

Cleaning the Data

During the import of our data, we also included the argument of na in our read_csv() functions for both data sets in order to indicate that any missing values that show up as blank or "(DO NOT READ)"/"n/a"/any variations of N/A are identified as NA values in R.

Following the importing stage, we changed the casing for all variables to lowercase in both data sets, as well as re-coded some character values to be error free, such as "In military commissariess, or" to "In military commissaries" and "Somewhere else (SPECIFY)?" to "Somewhere else" for the variable `wherebuy` from the data frame of `smoker_data`. Our group also re-coded the value of "100 or more cigarettes" to "100" for the variable `howmany` in the `race_data` data frame, which allowed us to convert the data type of the variable from character to numeric; the purpose of this conversion was to facilitate our calculation for pack-years later on in our analysis.

In order to join the two data frames together, our group had to first re-code the values of the variable `id` in the `race_data` data frame to show only numbers and no character strings. The purpose of this process is so that the values of `id` would match those of the variable `psraid` in the `smoker_data` data frame since they both represent participant IDs. We then converted the data type of `psraid` from numeric to character; with the same data type for the identical key variables of `id` = `psraid`, we were then able to join the two data sets together into a single data frame.

Creating New Variables

Our group created a new variable called `race` to combine variables `race01` through `race15` into a single column shown in our final joined data frame; this was done by using conditional statements when a participant answered "Yes" to any particular race variable, which also allowed us to rename `race 01` through `race15` as the race categories they represent, such as White, Black, Japanese, and so on for better comprehension of the race of each individual in our data set at first glance.

Another variable we created was `pack_years`, which was created by multiplying the number of cigarette packs per day by the length of time a participant has been smoking on a daily basis. Before multiplying the length of time smoked by the number of cigarette packs per day, our group made sure to convert the length of time smoked to years based on the unit of time reported, such as dividing the length of time by 365 for time reported in days and by 12 for time reported in months; these different calculations by unit of time reported were essentially conditional statements that led us to create `pack_years`.

Additional Data Wrangling for Visualizations

For all of our visualizations, we dropped all NA values based on variables presented in tables and graphs because we found these NA values to be unhelpful toward our analysis of the data.

Results

Table 1

Table 1: <center><b><h4 style="color: black;">Average Number of Pack-years by Disease Outcome Among Smokers</h4></b></center>

Disease	Average Number of Pack-years
Asthma	25
Diabetes	25
Heart Disease	28
Mental Illness	17

Data Source: 2011 California Smokers Cohort, CA Dept. of Health

This table demonstrates the average number of pack-years per disease type for smokers who reported having asthma, diabetes, heart disease, and/or mental illness in the 2011 California Smokers Cohort study.

Among smokers who have reported having asthma, heart disease, diabetes, and/or mental illness, those with heart disease have the highest number of average pack-years (28), while those with mental illness have the lowest number of average pack-years (17).

Table 2

Table 2: <center><b><h4 style="color: black;">Mental Illness Status by Race</h4></b></center>

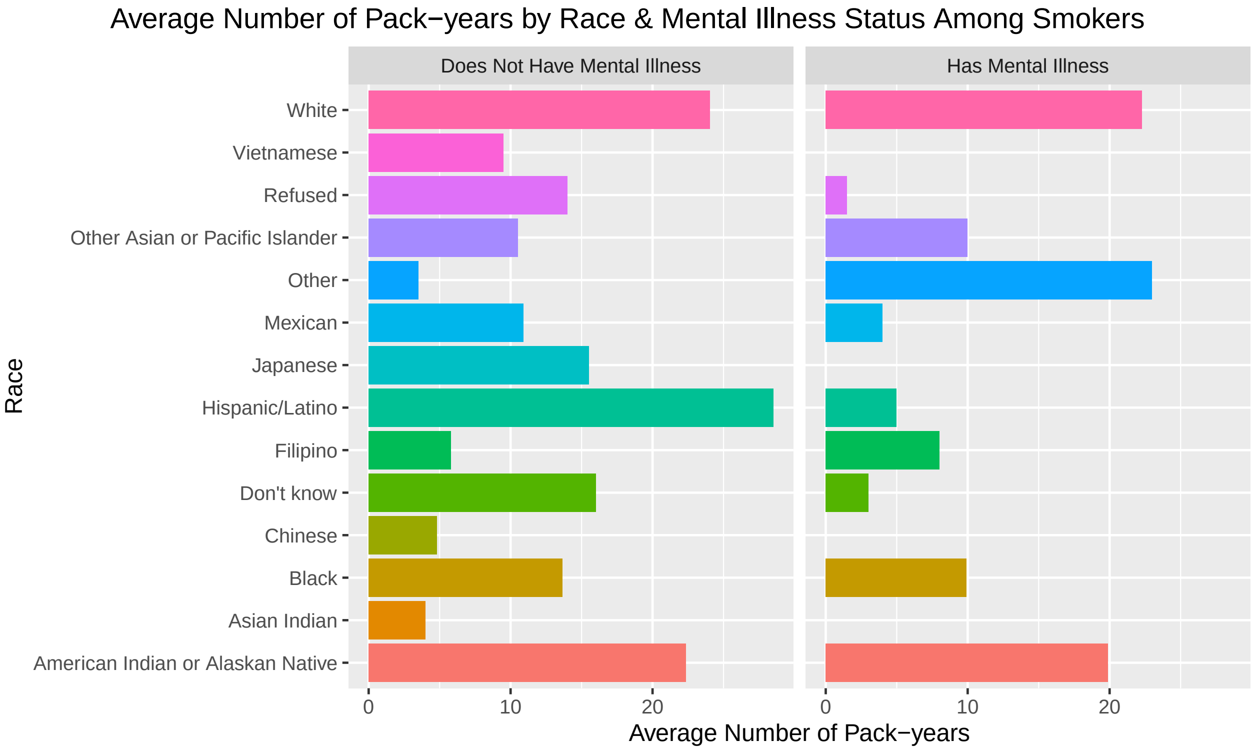
Race	Diagnosed Mental Illness	No Diagnosed Mental Illness
White	137	660
Black	13	64
American Indian or Alaskan Native	10	29
Refused	3	4
Filipino	2	6
Mexican	2	17
Don't know	1	1
Other	1	2
Other Asian or Pacific Islander	1	5
Hispanic/Latino	1	16
Asian Indian	NA	1
Vietnamese	NA	2
Japanese	NA	6
Chinese	NA	7

Data Source: 2011 California Smokers Cohort, CA Dept. of Health

This table displays the number of participants who reported having diagnosed mental illness or no diagnosed mental illness separated by racial background of the participants.

Participants who identified themselves as "White" appear to have the greatest number of both diagnosed mental illness and no diagnosed mental illness cases.

Figure 1



Average Number of Pack−years by Race & Mental Illness Status Among Smokers

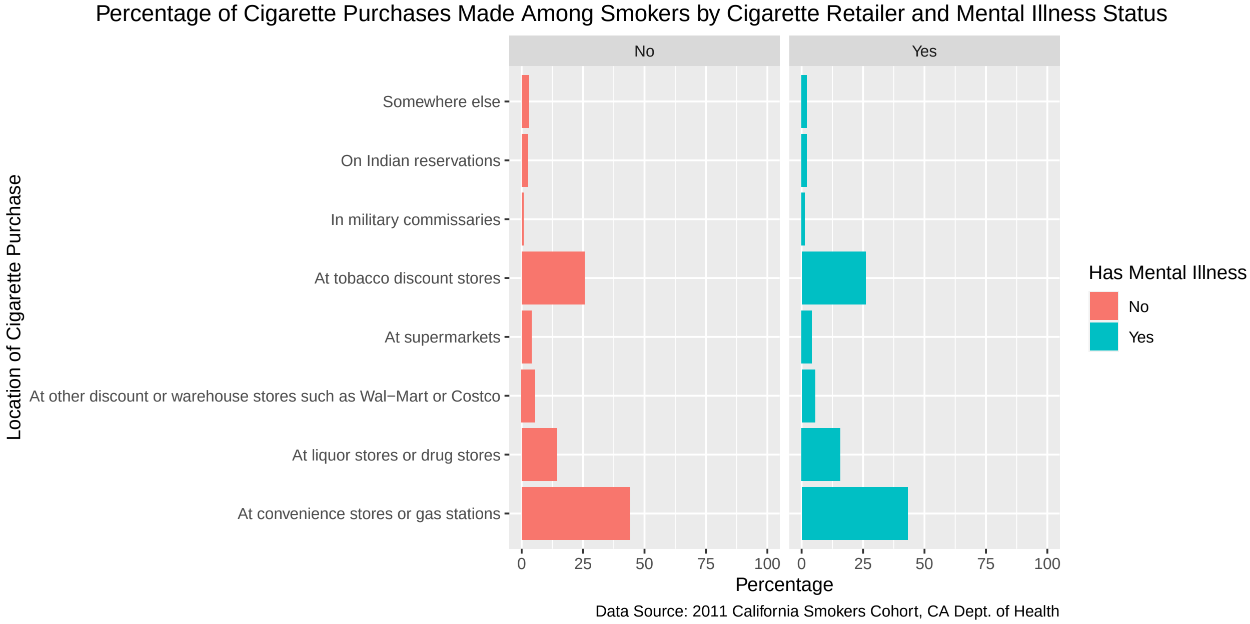
Data Source: 2011 California Smokers Cohort, CA Dept. of Health

This graph exhibits the number of average pack-years for each race category and by reported mental illness status of smokers in the 2011 California Smokers Cohort study.

Among smokers who have reported having no mental illness, those who identified as “Hispanic/Latino” by race appear to have the greatest number of average pack-years, followed by “White and”American Indian or Alaskan Native”, out of all race categories in the 2011 California Smokers Cohort.

Among smokers who have reported having mental illness, those who identified as “Other” by race appear to have the greatest number of average pack-years compared to other races in the 2011 California Smokers Cohort, with “White” and “American Indian or Alaskan Native” following closely behind.

Figure 2



This bar graph explores the relationship between cigarette purchase location frequencies by smokers and reported mental illness status among smokers in the 2011 California Smokers Cohort study.

Among smokers with reported mental illness, the greatest percentage of cigarette purchases were made at convenience stores or gas stations, followed by the location of tobacco discount stores; these are the two locations that also have the highest percentages for buying cigarettes at among smokers with no reported mental illness.

Discussion

Based on our results from the analysis of our data, the impact of race on diagnosed mental illness status is inconclusive. This conclusion is supported by the conflicting information given by Table 2 and Figure 1; as shown in Table 2, smokers who identified as “White” had the highest number of reported mental illness diagnoses, but in Figure 1, the race category of “Other” surpassed “White” in terms of tobacco consumption (average pack-years) and had the greatest number of average pack-years for those who reported diagnosed mental illnesses. However, it is notable that the “White” race category’s average number of pack-years is still very close to that of “Other”; the two race categories’ bar lengths are similar in length, as seen in Figure 1. This interesting observation warrants further research into the relationship between race and reported mental illness outcomes, with the recommendation that potential confounders and sources of potential bias are accounted for, which were not in this analysis. For example, given the data we have here, we do not know how accurate participants’ reports of diagnosed mental illnesses are, which may skew the length of the bars per race by reported mental illness status category in Figure 1.

In addition our group has concluded that there is a possible association between cigarette purchasing locations and mental illness outcomes. As exhibited by Figure 2, convenience stores/gas had the highest percentage of cigarette purchases made by smokers with reported mental illness diagnoses, followed by tobacco discount stores. In order to strengthen this conclusion, further research and thorough statistical analyses must also be performed in order to minimize confounding and account for factors that may distort the relationship between cigarette purchasing locations and mental illness outcomes.

Lastly, we conclude that out of the four diseases of asthma, diabetes, heart disease, and mental illness, tobacco consumption in terms of pack-years was the lowest among smokers who reported having mental illness and the highest among smokers who reported having heart disease; this conclusion is supported by the information exhibited in Table 1.

Overall, our group recommends that CDPH further investigate the cigarette purchasing locations of convenience stores, gas stations, and tobacco discount stores and their relationship to mental illness diagnoses. CDPH could look into potential tobacco product(s) specifically sold at these locations that smokers cannot get anywhere else, as well as the potency of the specific product(s). In addition, we recommend CDPH to look into the unique characteristics of the demographic that frequent these places. The findings of these investigations could then lead to more concrete conclusions of our group's analyses in order to further develop tobacco cessation programs and interventions. If financial budgets permit, CDPH could additionally conduct further studies on the impact of race on mental illness outcomes, possibly with different methods to measure the outcomes of mental illness instead of relying on participant reporting that would minimize bias in order to arrive at more concrete conclusions.