NHANES Project

Angela Zhao

8/18/24

Introduction

The purpose of this analysis is to investigate the relationships between certain health behaviors and conditions in the U.S. population using data from the National Health and Nutrition Examination Survey (NHANES). Specifically, the analysis aims to explore associations between smoking, alcohol use, diabetes, and oral health issues, adjusting for demographic factors where appropriate. The NHANES data, which combines interviews and physical examinations, provides a comprehensive overview of health and nutritional status among U.S. adults and children.

Data

The data used for this project comes from Wave I (2015-2016) of NHANES, a survey conducted by the Centers for Disease Control and Prevention (CDC). The public release of Wave I included a total of 9971 participants between 0 and 150 years old. Participants 80 years and older were all coded as 80 years old. Extensive sociodemographic, disease, clinical, lifestyle, and overall health data were collected. We take a subset of the data field, including information on smoking and alcohol consumption, diabetes diagnosis, and oral health issues. Demographic variables such as age, gender, and education level are also included. Some data fields had high levels of missingness (40-50%), which was a key consideration throughout our analysis. We will discuss this further in the methods and results sections.

We define the data fields of interest as follows:

- Alcohol use (ALQ101): Had at least 12 alcohol drinks/1 year.
- Diabetes status (DIQ010): Doctor told you have diabetes (group borderline as yes)
- Ever smoker (SMQ020): Smoked at least 100 cigarettes in life?

- Oral health issue (OHAREC): If you are recommended care other than "Continue your regular routine care"
- Age (RIDAGEYR): Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.
- Gender (RIAGENDR): Gender of the participant.
- Education (DMDEDUC2): Education level - Adults 20+
- Told the benefits of giving up cigarettes (OHQ610): "In the past 12 months, did a dentist, hygienist or other dental professional have a direct conversation with you about the benefits of giving up cigarettes or other types of tobacco to improve your dental health?"

Methods

Data Processing

Data exploration began with examining the completeness of our variables. Using SQL, we can directly query the database. For example, we can ask the data base for the counts of missing values for diabetes and alcohol. (Code included since the instructions specifically asked us to use SQL queries.) These queries tell us that there are a total of 9575 diabetes status records, of which none are missing. On the other hand, there are a total of 5735 smoking status records, of which 527 are missing. The high missingness meant that we should explore the code books and data collection process to explain the missingness, and imputation methods would not be possible or reasonable.

Data was then prepared for analysis. We joined relevant tables from an SQLite database of the Wave I data to create a data frame comprising demographic and health variables of interest. Variables such as smoking status, alcohol use, and diabetes status were recoded into categorical factors for analysis. Participants that refused to answer, did not know, or had missing data for any data field were assigned NA for that data field.

Figure Figure 1 shows the proportion of missing data for each variable in the full data set of 9971 participants. Around 40-50% of the participants are missing Education, Smoking Status and Alcohol Use data. We would not be able to proceed with our analyses and produce reliable results from this data. Upon closer inspection, we first notice that by definition, education only includes data from participants 20 years and older. Furthermore, NHANES only collected smoking and alcohol data from participants 18 years and older as well. Importantly, in the U.S., only adults 21 years and older are allowed to purchase alcohol or cigarettes. To better understand the missingness in this data set, we look only at adults 21 years and older (Figure 2). Here, the proportion of missing data seems much more reasonable. However, because the missingness for alcohol use and oral health issues is greater than 10%, imputation may not be suitable.

The data was filtered to include participants aged 21 years and older, as smoking and alcohol use data were only available for adults and to better align with legal age requirements for alcohol and cigarette use in the U.S. A summary of the characteristics of participants in this subset were included (Table 1).

```
# SELECT
#   SUM(CASE WHEN DIQ010 IS NULL THEN 1 ELSE 0 END) AS missing_diabetes,
#   COUNT(*) AS total_records
# FROM DIQ_I;

# SELECT
#   SUM(CASE WHEN ALQ101 IS NULL THEN 1 ELSE 0 END) AS missing_alcohol,
#   COUNT(*) AS total_records
# FROM ALQ_I;
```

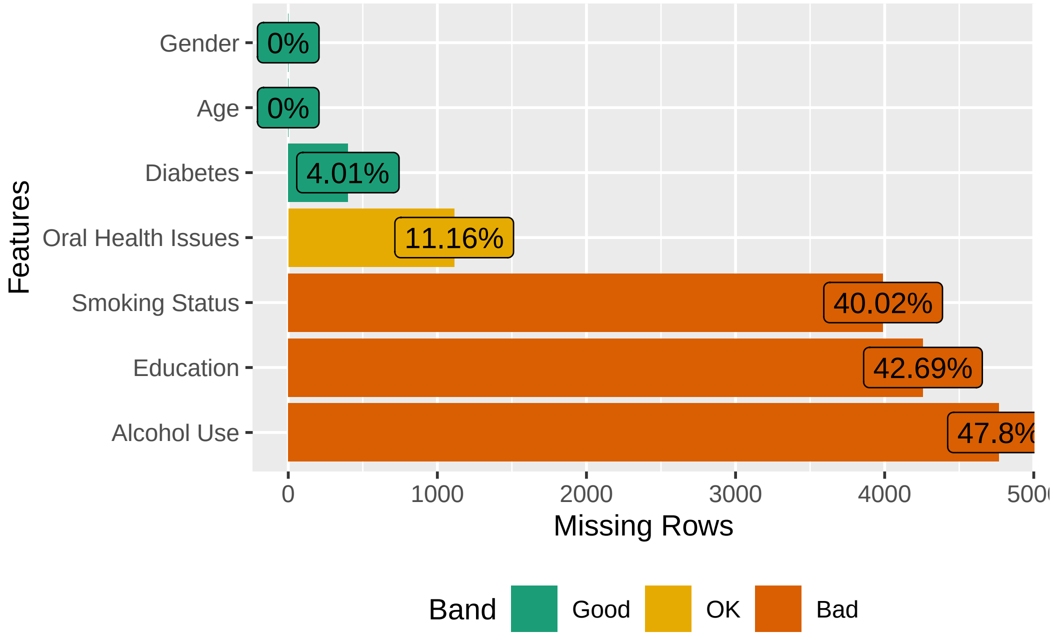

Figure 1: Missingness in Wave I

Firstly, by definition, education only includes data from participants 20 years and older. Upon closer inspection, we notice that NHANES only collected smoking and alcohol data from participants 18 years and older as well. Importantly, in the U.S., only adults 21 years and older are allowed to purchase alcohol or cigarettes. To better understand the missingness in this data set, we look only at adults 21 years and older (Figure 2). Here, the proportion of missing data seems much more reasonable. However, because the missingness for alcohol use and oral health issues is greater than 10%, imputation may not be suitable.

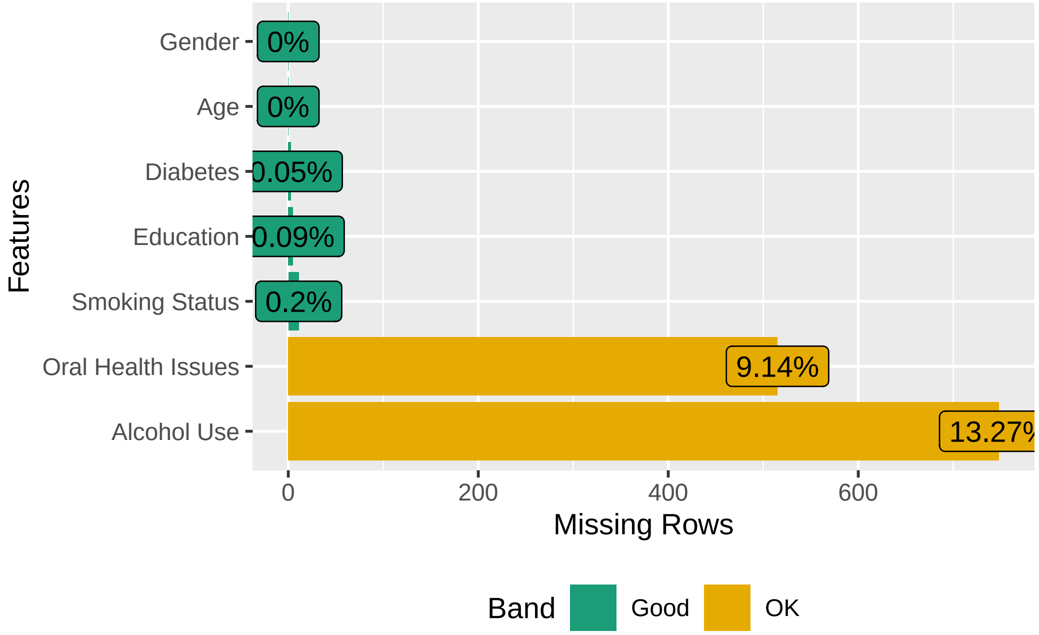

Figure 2: Missingness in Wave I, Participants 21 Years and Older Only

Table 1: Summary of Characteristics

	Overall
n	5992
Age (mean (SD))	48.12 (18.52)
Gender = Male (%)	2887 (48.2)
Education (%)	
9-11th grade (Includes 12th grade with no diploma)	676 (11.3)
College graduate or above	1422 (23.7)
Don’t Know	5 (0.1)
High school graduate/GED or equivalent	1236 (20.6)
Less than 9th grade	688 (11.5)
Some college or AA degree	1692 (28.2)
NA	273 (4.6)
Smoking Status (%)	
Don’t know	9 (0.2)
No	3559 (59.4)

	Overall
Refused	2 (0.0)
Yes	2422 (40.4)
Alcohol Use (%)	
Don't know	3 (0.1)
No	1728 (28.8)
Yes	3477 (58.0)
NA	784 (13.1)
Diabetes (%)	
Borderline	132 (2.2)
Don't know	4 (0.1)
No	5011 (83.6)
Yes	845 (14.1)
Oral Health Issues (%)	
Continue your regular routine care	3463 (57.8)
See a dentist at your earliest convenience	1540 (25.7)
See a dentist immediately	2 (0.0)
See a dentist within the next 2 weeks	456 (7.6)
NA	531 (8.9)

Statistical Analysis

Hypothesis Testing

We used hypothesis testing to explore the association between smoking status and whether participants were informed about the benefits of quitting smoking. We created a contingency table between the two covariates and used Pearson's Chi-squared test to evaluate this relationship between categorical variables. The null hypothesis was that there is no association between smoking status and being informed about quitting benefits, while the alternative hypothesis proposed that such an association exists.

Regression Model

To assess the impact of health behaviors on oral health outcomes, we used logistic regression modeling. The response variable was defined as a binary outcome indicating whether participants were recommended to see a dentist (i.e. oral health issue is present) or continue regular care (i.e. no oral health issue). The predictors were smoking status and alcohol use. We also controlled for age, gender, diabetes status, and education. We fitted a logistic regression model and specified the binomial family. The regression model can be expressed as follows:

Let $Y = 1$ if oral health issues were present and $Y = 0$ otherwise. Let X_i, $i \in 1, 2, ..., p$ for p covariates of interest. In our case, we had 6 covariates: smoking status, alcohol use, age, gender, diabetes, and education. Diabetes had 3 factor levels and education had 5 factor levels. The reamining covariates had 2 factor levels. Hence $p = 10$. β_i, $i \in 0, 1, ..., p$ refer to the estimated coefficients for the intercept and each covariate.

$$\text{logit}(P(Y = 1)) = \log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Results

Population Characteristics

Table 1 summarizes the characteristics of the study subset 21 years and older (n = 5992). The average age was 48.12 years (SD = 18.52), and 48.2% of participants were male. Education varied widely, with 23.7% having completed college and 11.5% with less than a ninth-grade education. Smoking was reported by 40.4% of participants, and alcohol use was reported in 58.0%. Oral health issues were present in just under half of the population, with 57.8% of participants being recommended to continue regular care and 33.3% instructed to see a dentist.

Hypothesis Testing

Pearson's Chi-squared test revealed a significant association between smoking status and receiving advice on quitting smoking. Table 2 is the contingency table. The p-value fromm our hypothesis testing was 5.00×10^{-4}, which means we reject the null hypothesis. This result indicates a significant association between smoking status and being advised about quitting smoking, suggesting that smokers were more likely to be counseled by healthcare professionals.

Regression Model

The logistic regression model highlighted several significant predictors of oral health issues. The AIC for the model was 5956. Figure 3 shows the odds ratios resulting from our analysis and 95% confidence intervals were shown for each odds ratio. A red line was drawn at 1 (i.e. no significant effect if the confidence band covers 1). Smoking status, alcohol use were both significant. After controlling for all demographic covariates, smokers were 48% (OR = 1.48, 95% CI: $1.29 - 1.70$, p-value = 1.30×10^{-8}) more likely to report oral health issues than non-smokers, emphasizing the negative impact of tobacco on oral health. Interestingly, participants who reported alcohol use were 21% less likely to report poor oral health compared

Table 2: Contingency Table for Smoking Status v.s. If the Participant was Told the Benefit of Quitting Smoking

	Don't know	No	Yes	NA
Don't know	0	2	1	6
No	1	1971	125	1462
Refused	0	0	0	2
Yes	1	846	263	1312
NA	0	262	9	3708

to those who did not consume alcohol (OR = 0.79, 95% CI: $0.67 - 0.92$, p-value $= 1.18 \times 10^{-3}$). This could suggest the presence of confounding factors such as higher socioeconomic status or better access to dental care among moderate drinkers. Diabetes was not found to have a significant association with oral health.

For each additional year of age above 21 years old, participants were 1% less likely to report oral health issues (OR = 0.99, 95% CI: $0.98 - 0.99$, p-value $= 8.13 \times 10^{-6}$), suggesting that older participants, while generally more prone to health issues, may also have better-managed oral health due to more consistent healthcare access and to prioritization of oral health. This is visualized in Figure 4, which shows the relationship betwen oral health issues and age. Firstly, the bars indicate the proportion of participants at each age between 21 and 80+ that have oral health issues.The jittered black dots at the top of the graph plot all the participants with oral health issues at each age and the jittered black points at the bottom of the graph plot all the participants without oral health issues at each age. The smooth blue line and grey ribbons was obtained from glm smoothing with binomial family specified and shows the probability of having oral health issues at each age along with the 95% confidence interval. The demographic covariates that we controlled for (age, gender, and education) were also significant. As our regression odds ratio suggests, the risk of having an oral health decreases slightly with age.

Men were also 31% more likely to report oral health issues compared to women (OR = 1.31, 95% CI: $1.14 - 1.50$, p-value $= 5.16 \times 10^{-5}$). Amongst the education categories, participants with some college or an AA degree were 35% less likely to report oral health issues compared to those with less education (OR = 0.65, 95% CI: $0.53 - 0.79$, p-value $= 4.59 \times 10^{-5}$), while college graduates were 68% less likely to report oral health issues (OR = 0.32, 95% CI: $0.26 - 0.39$, p-value $< 2.00 \times 10^{-16}$) compared to those with less education. This highlights the importance of health literacy and awareness in maintaining oral health.

```
Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)
```

```
data:  smoking_benefit_table
X-squared = 2740.1, df = NA, p-value = 0.0004998
```

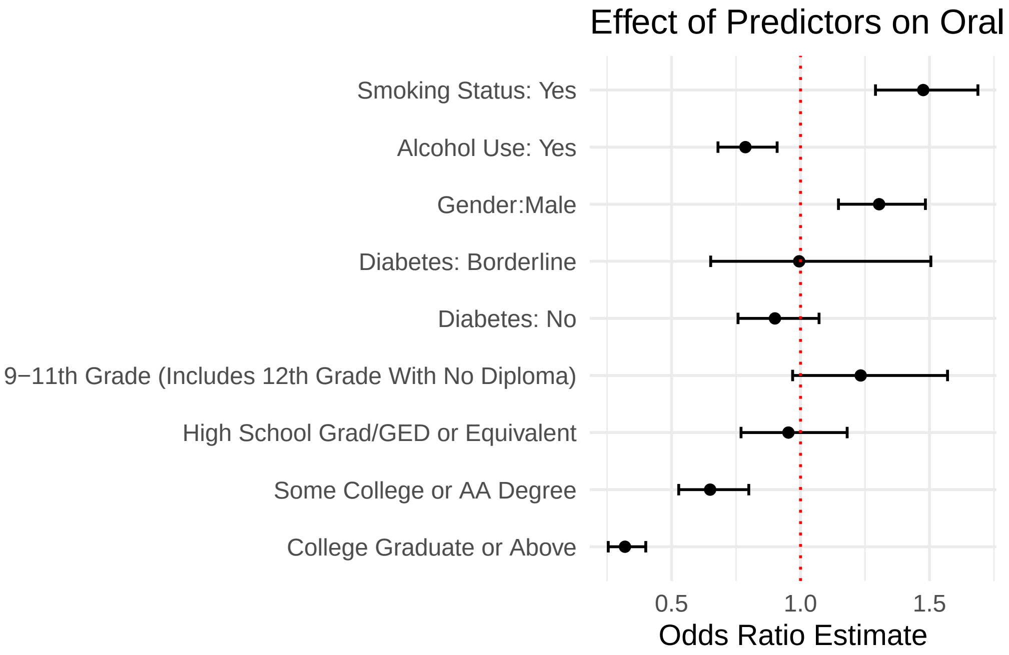

Figure 3: Odds Ratios Obtained From Logistic Regression Model (Binomial Family)

Conclusion

This analysis of NHANES data underscores significant associations between smoking, alcohol use, and oral health outcomes. Smoking is a well-known major risk factor for poor oral health and our results confirm this. We emphasize the need for targeted smoking cessation programs. Furthermore, being told the benefits of quitting smoking is significantly associated with smoking status. This suggests that actively engaging with smokers to discuss cessation can is an actionable intervention, which is a positive finding from a public health perspective. Meanwhile, alcohol use was associated with somewhat better oral health outcomes. This is a unexpected because the literature suggests that alcohol is a risk factor for poor oral health. Though, this may be influenced by other lifestyle or socioeconomic factors that we did not account for. Diabetes was also not significantly associated with oral health issues, which differs from the literature. Amongst other things, this may be partially explained by glycemic control. We do not have information on glycemic control (e.g., HbA1c levels) for every participant. Poor glycemic control is more likely to contribute to oral health issues like periodontitis. If participants with diabetes have well-managed glucose levels, they may not

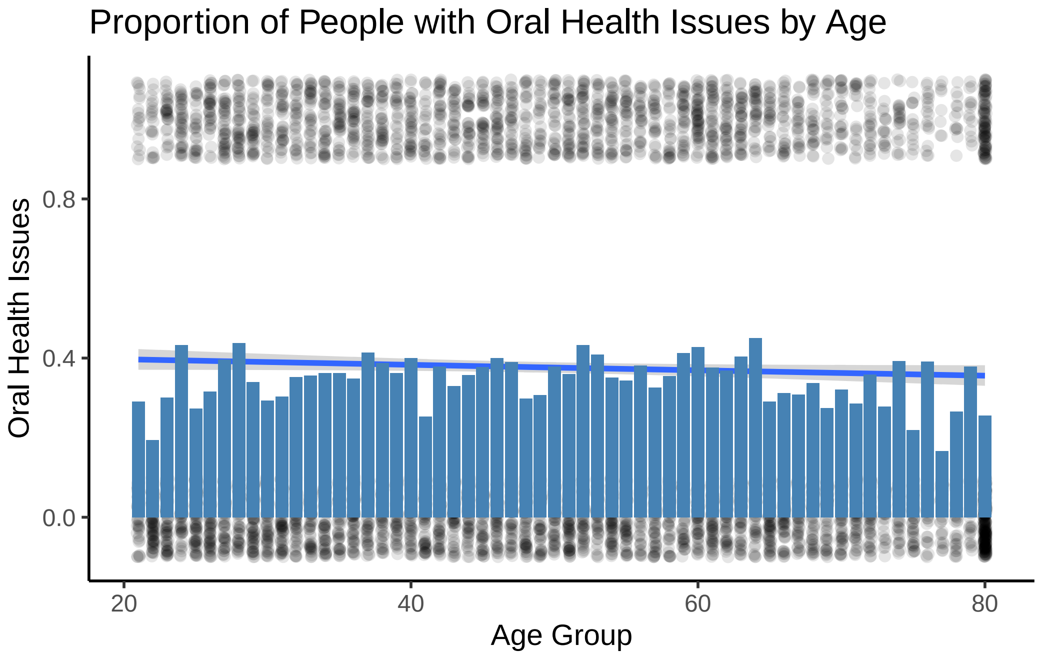

Figure 4: Associations Between Oral Health Issues and Age

exhibit the expected negative impact on oral health, leading to a weaker or non-significant association in this analysis. Age and education level also played significant roles, with older individuals and those with lower education being more at risk for dental issues.

Even after subsetting the population to 21 years and older, we still had substantial missing data ($> 10\%$). This limits the robustness of these findings. Future efforts should look for explanations for variables with higher missingness and investigate if there are ways to account for or impute them. We should also expand the study to account for socioeconomic factors. Nevertheless, the insights from this study reinforce the need for public health initiatives to improve oral health, focused on smoking cessation, health education, and increased access to dental care, particularly for vulnerable populations.