Effect of Social Media on E-cigarette use among Youth during the COVID-19 Pandemic in 2021

by

Yufeng Zhu

BS Chemistry, University of Pittsburgh, 2021

Submitted to the Graduate Faculty of the

Department of Biostatistics

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Yufeng Zhu

It was defended on

December 12, 2022

and approved by

Thesis Advisor: Jenna C. Carlson, Assistant Processor, Biostatistics
School of Public Health, University of Pittsburgh

Ada Youk, PhD, Associate Processor, Biostatistics
School of Public Health, University of Pittsburgh

Jeanine M. Buchanich, PhD, Associate Processor, Biostatistics
School of Public Health, University of Pittsburgh

Kar-Hai Chu, PhD, Associate Professor, Behavioral and Community Health Sciences
School of Public Health, University of Pittsburgh

Copyright © by Yufeng Zhu

2022

Effect of Social Media on E-cigarette use among Youth during the COVID-19 Pandemic in 2021

Yufeng Zhu, MS

University of Pittsburgh, 2022

E-cigarette use in teens has emerged as a public health crisis in the United States. Since teens are more susceptible to e-cigarette advertising than adults, content and promotions on social media became an effective strategy for the e-cigarettes industries to maintain long-term profits by attracting teens.

The goal of this research is to explore the effect of social media on e-cigarette use among youth during the COVID-19 pandemic with demographic characteristics of age, gender, race/ethnicity. Using the data from 2021 National Youth Tobacco Survey (NYTS), weighted logistic regression models to account for the complex NYTS survey design, were performed using R.

Results showed that social media use, e-cigarette exposure through social media, and the number of social media sites with e-cigarette content were significantly associated with increased odds of ever using e-cigarettes. However, social media were not associated with the increased odds of current using e-cigarettes since addiction plays a more important role in long-term e-cigarette use.

The public health significance of this work is preventing the use of e-cigarettes and other tobacco products among youth to improve their well-beings. By studying their social media use patterns and exposure to e-cigarette-related content, we found that reducing exposure to e-cigarette promotions and related posts on social media by implementing more strict control policies, and

conducting educational campaigns could be an area of regulatory intervention to limit e-cigarette initiation among teens.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 Background

Tobacco-use in teens has recently emerged as a public health crisis in the United States. It's the leading cause of preventable disease, disability, and death. According to the CDC, 90% of smokers first try tobacco products before age 18 (CDCTobaccoFree, 2022). As tobacco products are evolving, among teens who currently used each tobacco product, 39.4% are for e-cigarettes compared with 18.9% for traditional cigarettes in 2021 (Gentzke, 2022). Furthermore, a prospective study by Pike et al. (2019) has shown that teens are more susceptible to e-cigarette advertising than adults (Pike et al., 2019). Advertisements and promotions, especially on social media, then become an effective way for the e-cigarette industries to attract teens into a potential life-long addiction for the purpose of maintaining long-term profits. Therefore, it is critical to prevent e-cigarette-related posts and content on social media sites among teens to improve teens' overall healthy development and well-being.

1.2 Previous Research

Previous research by Wulan et al (2022) has shown that lower exposure to e-cigarette advertisements on social media was significantly associated with lower smoking participation among youth (Wulan et al., 2022). However, these studies are outdated since their data are collected before the pandemic. The nationwide implementation of emergency COVID-19 guidelines in 2021 resulted in two main changes. First, research by Pandya et al. (2021) has shown that social media engagement has increased since people have limited in-person social interactions (Pandya & Lodha, 2021). Therefore, it's necessary to narrow down the area of focus to just exposure to e-cigarette-related content on social media use instead of on traditional sources such as newspapers, stores, and TV. Second, studies also have also shown that teens are more likely to quit e-cigarettes during the pandemic because they tended to spend more time with their parents (Gaiha et al., 2020). Therefore, it would be meaningful to fill gaps by re-evaluating the prevalence and correlates of youth exposure to and engagement with e-cigarette-related social media, since the National Youth Tobacco Survey (NYTS) in 2021 is the first survey that was fully conducted during the COVID-19 pandemic among a large sample.

Numerous research in the health sciences used various software to evaluate data from complicated survey designs, but they did not describe the formulation and theory for the logistics regression model. In the method section, these studies simply introduced the software and package they use instead of presenting the estimating process. Therefore, the method section in this paper explains the theory behind the logistic regression models for complex survey design in NYTS.

1.3 Public Health Impact

By studying teenagers' social media use patterns and exposure to e-cigarette related content, we can evaluate the effect of social media on e-cigarette use among them. We can reduce youth exposure to e-cigarette promotions and related posts on social media by implementing more strict e-cigarette social media policies, combined with the FDA's regulation, and increased education to resist e-cigarette use in the future to prevent the use of all tobacco products among youth.

2.0 Methods

2.1 Data

2.1.1 Data Source

The data is from the 2021 National Youth Tobacco Survey (NYTS), which was acquired from the Centers of Disease Control and Prevention (CDC). The NYTS is a cross-sectional, school-based, self-administered survey of regular public and private schools in the 50 U.S. states and the District of Columbia with students enrolled in grades 6 through 12. The survey was administered between January 18 and May 21, 2021. Out of a final sample of 25,149 students, 20,413 student questionnaires were completed, representing a sample of 508 schools, of which 279 participated (Gentzke, 2022).

However, according to the U.S. Census Bureau, in October 2021, 96% of teens between the age of 10 and 17 were enrolled in traditional schools., thus the results do not apply to the 4% of youths who have dropped out of school (Bureau, n.d.). Since the survey was administered online, the results cannot be compared to earlier NYTS survey results that were conducted in a different interview setting. This would be the main limitation of this study.

2.1.2 Outcome Variables

The outcome variables used in these analyses are ever e-cigarette use and current e-cigarette use.

2.1.2.1 Ever E-cigarettes Use

The ever e-cigarette use is a “Yes” or “No” question on whether the participants have ever used an e-cigarette, which is a two-level categorical variable. The participants who answered the question with “Yes” were then transferred to the question related to current e-cigarette use status: how many days the participants used cigarettes in the past 30 days.

2.1.2.2 Current E-cigarettes Use

For the outcome current e-cigarette use, I classified the participants into former e-cigarettes users (participants who had used e-cigarettes in the past but not within the last 30 days), and current e-cigarettes users (participants who had used e-cigarettes on at least 1 day in the last 30 days) based on the distribution of the data. Different cut-points were tested to categorize the data properly. In Figure 1, the plot at shows the distribution from 0 to 30 days, which reveals that the number of former e-cigarette users is dramatically higher than the number of former e-cigarette users. To visualize the distribution of former e-cigarette users better, I filtered out former e-cigarette users. In Figure 2, we can tell that the participants who reported had used e-cigarettes in 1 day and 30 days are relatively higher.

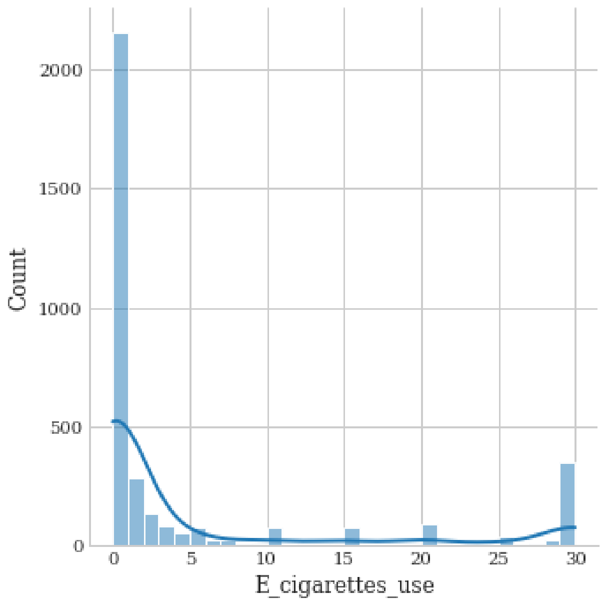

Figure 1 Distribution of E-cigarette Use in the Last 30 Days Including Former E-cigarettes Users

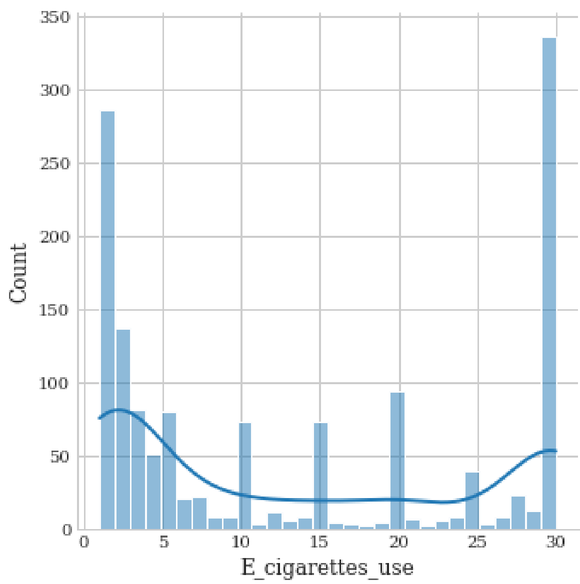

Figure 2 Distribution of E-cigarette Use in the Last 30 Days Excluding Former E-cigarettes Users

2.1.3 Explanatory Variables

The predictors considered in these analyses are how often the participants use social media, as well as how often and on how many sites the participants see posts or content related to e-cigarettes.

2.1.3.1 Social Media Use

Based on participants’ answers to the following question, the frequency of social media use was evaluated: “How often do you use social media?” Participants who reported “less than one time per week” or “about one time per week” were considered as low frequency; those who reported “a few times per week” or “less than one hour per week” were considered as medium frequency; “About 1-2 hours, daily”, “about 3-4 hours, daily” or “4 hours or more, daily” were considered as high frequency (*2021 National Youth Tobacco Survey Codebook*, n.d.).

2.1.3.2 E-cigarette Exposure through Social Media

Based on participants' answers to the following question, the frequency of exposure to e-cigarette ads was evaluated: “When you use social media, how often do you see posts or content (pictures, videos, or text) related to e-cigarettes?” Participants who reported “never” or “less than monthly” were considered as low exposure; those who reported “monthly” were considered as medium exposure; “weekly” or “daily” were considered as high exposure (*2021 National Youth Tobacco Survey Codebook*, n.d.).

2.1.3.3 Number of Social Media Sites with E-cigarette Content

Based on participants' answers to the following question, several social media sites were evaluated: “On which social media sites have you seen posts or content related to e-cigarettes?” Possible selections are Facebook, Instagram, Snapchat, TikTok, Twitter, Reddit, YouTube, and other sites. To avoid collinearity, the number of social media sites with e-cigarette content was calculated as the sum of the sites where the participants have seen posts or content related to e-cigarettes, which was modeled a continuous variable ranging from 0 to 8 (*2021 National Youth Tobacco Survey Codebook*, n.d.).

2.1.4 Covariates

The covariates are age, gender, school type, race/ethnicity, and cigarette use status. Age was modeled as a continuous variable range from nine to 19. Gender was modeled as a categorical variable of female and male. School type was modeled as a categorical variable of middle school and high school: middle school is from grade six to eight, and high school is from grade nine to 12.

The race and ethnicity responses were categorized into a single race/ethnicity group to match post-stratification weights defined by the CDC: Hispanic, non-Hispanic black, and combined (American Indian, Asian, white). For the poststratification purpose, respondents with missing race and ethnicity data, and those who reported several races were each given a distinct race and ethnicity, were classified as "Other".

I also included current cigarette use status as a covariate because there is a strong association between youth e-cigarette and traditional cigarette use (O’Brien et al., 2021). The ever cigarette use was also a “Yes” or “No” question on whether the participants have ever used cigarettes, which was a two-level categorical variable.

2.2 Weighted Survey Design

The public-use methods report for the 2021 NYTS was carefully studied to determine the type of sampling design that was applied to gather the data. The sample is a three-stage, three-level cluster sample design to produce a representative sample of students across the United States. Each student's record was given a weighting factor in order to account for nonresponse and different selection probability. Weight adjustments ensure the proportions of students in each grade matched the proportions of the national population. In the survey, three survey sampling schemes are included, which are sampling weights, primary sampling unit (PSU), and strata.

2.2.1 Sampling Weights

Sampling bias arises when certain members of a group are more likely to be chosen. Survey sampling is used to decrease the cost or effort to survey a whole community. Therefore, sample weights are used to account for the systematic variations in probability sampling. In this survey, adjustments were made to account for nonresponse, and excess weight variances, such as the conditional student weight, school sampling weight, and grade sampling weight. These adjustments align the data based on the census (CDC, 2021). As a result of the sampling design, the weight is by definition the inverse of the probability of being included in the sample, which is the inverse of the probability of selection for each responding student. The weight is calculated as the number of elements in the population divide by the number of elements in the sample.

2.2.2 Primary Sampling Unit (PSU)

The primary sampling unit (PSU) is the first unit sampled in the design that ensures the responses represent the population of interest. In the survey, the PSU is the county where the student's school was located, which was stratified by racial/ethnic makeup and urban vs. rural status. The school selection probability and corresponding weight were taken into consideration as subsampling components of the PSU weight. The grade selection that occurred within a school was the secondary sampling unit (SSU). According to the methodology report of NYTS, the weight of the PSU was the inverse of the probability of its selection. School, PSU, and stratum are denoted by the subscripts k, l, and m, respectively (*Methodology Report of the 2021 NATIONAL YOUTH TOBACCO SURVEY*, n.d.):

$$W^{P}{}_{lm} = \frac{1}{K_m}\left(\frac{MOS_{.m}}{MOS_{lm}}\right) = \frac{1}{P_{lm}^{P}} \qquad \textbf{Equation 1}$$

2.2.3 Strata

Stratification is a technique for dividing the population into several groups, frequently according to demographic factors like gender and race. Each component of the population must be a part of just one stratum exclusively. In the survey, strata are determined by the major minority

(Non-Hispanic Black or Hispanic), its location (urban or rural), and the proportion of students who belong to that minority. These strata values allow estimates based on the survey responses to be calculated and improve the precision of the response estimates.

2.3 Logistic Regression

Since each outcome variable was a binary variable with only 0 or 1, I built a logistic regression model, using the “svyglm” function in “survey” package (“survey”) can be used for the weighted logistic regression model, where p is the probability of responding “Yes” to Ever/Current e-cigarettes use, X is the predictors and covariates matrix, and β is the parameter coefficients:

$$\ln\left(\frac{p}{1-p}\right) = X\beta \qquad \textbf{Equation 2}$$

2.3.1 Weighted Logistic Regression

Suppose the population $U = \{1,2,\ldots,N\}$ is divided into $h = 1,2,\ldots,H$ strata, which are demographic factors of race/ethnicity and urban/nonurban. And each stratum is divided into $j = 1,2,\ldots,n_h$ PSUs, i.e., the counties where the student’s school was located. Each PSU is constituted by $i = 1,2,\ldots,n_{hj}$ SSUs, i.e., the grade selection that occurred within linked schools. Each SSU has n_{hji}elements. The data consist of n'_{hj} SSU was chosen from n'_{hj} PSU in the stratum. Then the total observation will be $n = \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}} n_{hji}$, and each sampling unit has a sampling weight that is the inverse of the selection $W_{hjik} = \frac{1}{P\left(Y_{hjik} = 1 \middle| X_{hjik}\right)}$. Y_{hjik} will be the binary response variable, X_{hjik} will be the covariate matrix. The survey logistic regression model will be

$$logit\{P(Y_{hjik} = 1|X_{hjik})\} \qquad \textbf{Equation 3}$$

$$= \ln\left\{\frac{P\left(Y_{hjik} = 1 \middle| X_{hjik}\right)}{1 - P(\left(Y_{hjik} = 1 \middle| X_{hjik}\right))}\right\}$$

$$= X'_{hjik}\beta$$

The parameter coefficients β are estimated by weighted maximum likelihood. The method denotes a function that approximates the likelihood function of the population, which incorporates the multiple levels of the sampling design and weights as Equation 4 shows:

$$l_p(\beta) \qquad \textbf{Equation 4}$$

$$= \sum_{h=1}^{H} \sum_{j=1}^{n'_h} \sum_{i=1}^{n'_{hj}} \sum_{k} w_{hjik} \{y_{hjik}$$

$$\times \ln\left[P\left(Y_{hjik} = 1 \middle| X_{hjik}\right)\right] + \left(1 - y_{hjik}\right)$$

$$\times \ln[1 - P\left(Y_{hjik} = 1 \middle| X_{hjik}\right)]\}$$

The estimator of β can be derived by the weighted maximum likelihood function and making it equal to 0, $(\beta) = \frac{d}{d\beta} l_p(\beta) = 0$ ((Cassy et al., 2016).

Several weighted logistic regression models were fitted for each outcome variables to verify the predictors. We can then use the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), Pseudo-R^2, and deviance to compare the models, as well as the likelihood test that measures the goodness of fit to determine the best model.

2.4 Area Under Curve (AUC)

Area under the Receiver operating characteristic (ROC) curve (AUC) was used to assess and compare the performance of binary classification models. It measures the discrimination power of the predictive classification model and determines the best cutoff value for prediction.

Sensitivity is the true positive rate (TPR), which indicates the proportion of real students who use e-cigarettes correctly detected by the model with various thresholds. While the false negative rate (FNR) is "1-specificity", which indicates the proportion of the students who use e-cigarettes got incorrectly classified. A higher TPR and a lower FNR are desirable

The R package "WeightedROC" was used to calculate AUC ("WeightedROC"). From the package "PatrickCoyle/analyzeGES", I also created a function called "predROC" to produce a data frame of ROC curve values based on a svyglm, and a function called "plotAUC" to graph the ROC curve.

3.0 Results

3.1 Summary Statistics

Tables 1 and 2 show summary statistics to describe the sample size, unweighted and weighted percent for each categorical variable. For continuous variables, mean and standard deviations are given.

For the complete dataset, a total of 20,413 observations were observed with 10 variables. According to the weighted statistics in Table 1, 19.3% of participants reported used at least one e-cigarette, and 7.5% of them reported used at least one e-cigarette in the last 30 days.

Table 1 Summary Statistics for Categorical Variables

	Variables	Details	Unweighted Frequency	Unweighted Percent	Weighted Percent
Outcome Variables	Ever E-cigarettes Use	0 No	16547	81.1%	79.9%
		1 Yes	3665	18.0%	19.3%
		Missing	201	1.0%	0.8%
	Current E-cigarette Use	0 No	2154	10.6%	11.4%
		1 Yes	1436	7.0%	7.5%
		Missing	16823	82.4%	81.1%
Explanatory Variables	Social Media Use	1 Low	831	4.1%	3.8%
		2 Medium	2244	11.0%	10.7%
		3 High	13885	68.0%	69.1%
		Missing	3453	16.9%	16.3%

	E-cigarette Exposure via Social Media	1 Low 2 Medium 3 High Missing	9055 2216 5548 3594	32.7% 15.9% 11.3% 40.1%	33.2% 16.4% 11.4% 39.0%
Covariates	Gender	1 Male 2 Female Missing	10368 9919 126	50.8% 48.6% 0.6%	52.0% 47.4% 0.6%
	School Type	1 Middle School 2 High School Missing	12155 6591 1667	59.5% 32.3% 8.2%	59.5% 32.6% 8.0%
	Ever Cigarette Use	1 Yes 2 No Missing	1558 18397 458	7.60% 90.10% 2.20%	7.9% 89.9% 2.1%
	Race/Ethnicity	1 Combined: American Indian, Asian, White	10306	50.5%	53.6%
		2 Hispanic	5056	24.8%	25.5%
		3 Non-Hispanic black	280	16.1%	12.1%
		4 Other: Missing & Non-Hispanic Multi-racial	1771	8.7%	8.8%

Table 2 Summary Statistics for Continuous Variables

	Variables	Details	Unweighted Statistics	Weighted Statistics
Explanatory Variables	Number of Social Media Sites with E-cigarette content	Range: 0 to 8	Mean 1.704 Std 1.900	Mean 1.737 Std 1.899
Covariates	Age (year)	Range: 9 to 19	Mean 14.352 Std 2.062	Mean 14.539 Std 2.071

3.2 Weighted Binary Logistic Models

The missing observations were first categorized as separate variables in Table 2 to test the validation of the weighted survey design. They were then returned from categorical value to “NA” (not available) in R, since the “svyglm” function in weighted survey package has the option “na.action” that produced complete-case analyses by default. We also checked the variance inflation factor (VIF) to ensure multicollinearity did not exist.

3.2.1 Ever E-cigarettes Use

The outcome ever e-cigarette use was first tested in the model of the single predictor. In model 1a, the predictor is social media use, which indicates how often the participants use social media. From the results in Table 3, we can tell that participants’ age, race/ethnicity, gender, and cigarette use status have P-values less than 0.05, which indicates that these variables contribute significantly to the prediction of the participants' ever e-cigarette use status. For the predictor social media use, only high frequency use was significant compared to baseline low frequency use; the odds of having ever used e-cigarettes for participants who use social media at high frequency was estimated to be 1.35 times higher than those who use social media at low frequency.

Table 3 Weighted Binary Logistic Regression for Social Media Use (Model 1a)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
Social Media Use Reference: Low	Medium	0.7 [0.48, 1.02]	0.07	3.7e-09*
	High	1.35 [1.01, 1.79]	0.04*	
Age	Age	1.43 [1.37, 1.48]	< 2.2e-16 *	< 2.2e-16*
Race/Ethnicity Reference: Combined	Hispanic	0.78 [0.64, 0.94]	0.01*	2.9e-07*
	Non-Hispanic black	0.46 [0.35, 0.61]	3.2e-07 *	
	Other	1.04 [0.81, 1.33]	0.76	
Gender Reference: Male	Female	1.15 [1.01, 1.31]	0.04*	0.04*
School Type Reference: Middle School	High School	1.13 [1, 1.27]	0.05	0.05
Cigarettes Use Reference: Yes	NO	0.07 [0.06, 0.09]	< 2.2e-16 *	< 2.2e-16*

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

In model 1b, the predictor was e-cigarette exposure via social media, which indicated how often the participants saw posts or content related to e-cigarettes. From the results in Table 4, we can tell that participants' age, race/ethnicity, and cigarette use status contributed significantly to the prediction of the participants' ever e-cigarette use status. For e-cigarette exposure via social media, both medium and high exposure were significant compared to baseline low exposure. The odds of having ever used e-cigarettes for the participants who were exposed to content related to e-cigarettes at medium frequency on social media were estimated to be 1.63 times higher than those who were exposed to content related to e-cigarettes at low frequency; while for the participants who expose at high frequency, the odds were 2.59 times higher.

Table 4 Weighted Binary Logistic Regression for E-cigarette Exposure via Social Media (Model 1b)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
E- cigarette Exposure via Social Media Reference: Low	Medium	1.63 [1.35, 1.98]	< 2.2e-16*	< 2.2e-16*
	High	2.59 [2.24, 2.99]	3.7e-06*	
Age	Age	1.42 [1.36, 1.48]	< 2.2e-16*	< 2.2e-16*
Race/Ethnicity Reference: Combined	Hispanic	0.79 [0.67, 0.95]	0.01*	1.1e-06
	Non-Hispanic black	0.48 [0.37, 0.63]	< 2.2e-16*	
	Other	1.07 [0.83, 1.37]	0.61	
Gender Reference: Male	Female	1.14 [1, 1.3]	0.06	0.06
School Type Reference: Middle School	High School	0.91 [0.8, 1.04]	0.18	0.17

Cigarettes Use Reference: Yes	NO	0.08 [0.06, 0.09]	< 2.2e-16*	< 2.2e-16

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

In model 1c, the predictor was the number of social media sites with e-cigarette content, which indicates how many sites the participants see posts or content related to e-cigarettes. From the results in Table 5, we can tell that participants' age, race/ethnicity, and cigarette use status contributed significantly to the prediction of the participants' ever e-cigarette use status. The number of social media sites with e-cigarette content is also significant, which indicates the odds of having ever used e-cigarettes for the participants who were exposed to content related to e-cigarettes on fewer number of social media sites were increased by 1.21 times for each additional one site increase in the number of social media sites with e-cigarette content.

Table 5 Weighted Binary Logistic Regression for Number of Social Media Sites with E-cigarette Content (Model 1c)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
Number of Social Media Sites with E-cigarette content	Number of Social Media Sites with E-cigarette content	1.21 [1.17, 1.24]	< 2.2e-16*	< 2.2e-16*
Age	Age	1.42 [1.36, 1.48]	< 2.2e-16*	< 2.2e-16*
Race/Ethnicity	Hispanic	0.81 [0.67, 0.97]	0.03*	2.5e-06*

Reference: Combined	Non-Hispanic black	0.5 [0.38, 0.65]	2.0e-06*	
	Other	1.06 [0.84, 1.34]	0.61	
Gender Reference: Male	Female	1.15 [1.02, 1.31]	0.03*	0.03*
School Type Reference: Middle School	High School	0.97 [0.87, 1.08]	0.63	0.63
Cigarettes Use Reference: Yes	NO	0.07 [0.06, 0.08]	< 2.2e-16*	< 2.2e-16*

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

Model 1d is the full model with all three social media predictors. From the results in Table 6, we can tell that all three social media predictors, as well as the participants' age, race/ethnicity, school type, and cigarette use status contributed significantly to the prediction of the participants' ever e-cigarette use status. The odds ratios of increased age and combined race/ethnicity, being in middle school and using cigarette are greater than one, which are associated with increased ever e-cigarette use.

Table 6 Full Weighted Binary Logistic Regression for All Three Predictors (Model 1d)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
Social Media Use Reference: Low	Medium	0.64 [0.43, 0.94]	0.03*	9.32e-4*
	High	1.01 [0.77, 1.33]	0.93	
E-cigarette Exposure via Social Media Reference: Low	Medium	1.36 [1.13, 1.65]	2.1e-3 *	< 2.2e-16*
	High	2.04 [1.75, 2.37]	9.7e-14*	
Number of Social Media Sites with E-cigarette content	Number of Social Media Sites with E-cigarette content	1.11 [1.07, 1.15]	1.8e-07*	6.4e-09*
Age	Age	1.42 [1.36, 1.48]	< 2.2e-16*	< 2.2e-16*
Race/Ethnicity Reference: Combined	Hispanic	0.8 [0.67, 0.96]	0.02*	2.8e-06*
	Non-Hispanic black	0.49 [0.38, 0.64]	2.2e-06*	
	Other	1.06 [0.83, 1.36]	0.65	
Gender Reference: Male	Female	1.1 [0.96, 1.26]	0.17	0.16
School Type Reference: Middle School	High School	0.88 [0.77, 0.99]	0.04*	0.04*
Cigarettes Use Reference: Yes	NO	0.08 [0.07, 0.09]	< 2.2e-16*	< 2.2e-16*

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

3.2.2 Current E-cigarettes Use

The outcome current e-cigarette use was also tested. The predictor in model 2a was social media use. Table 7 indicates that participants' race/ethnicity, gender, and cigarette use status contribute significantly to the prediction of the participants' current e-cigarette use status. The predictor of social media use is not significant.

Table 7 Weighted Binary Logistic Regression for Social Media Use (Model 2a)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
Social Media Use Reference: Low	Medium	0.91 [0.48, 1.74]	0.79	0.22
	High	0.75 [0.47, 1.20]	0.23	
Age	Age	1.03 [0.97, 1.09]	0.42	0.41
Race/Ethnicity Reference: Combined	Hispanic	0.74 [0.59, 0.93]	0.01*	0.05
	Non-Hispanic black	0.82 [0.55, 1.22]	0.33	
	Other	0.81 [0.56, 1.17]	0.26	
Gender Reference: Male	Female	1.38 [1.14, 1.66]	1.3e-3*	8.0e-4*
School Type Reference: Middle School	High School	0.94 [0.76, 1.17]	0.57	0.57
Cigarettes Use	NO	0.31 [0.24, 0.39]	2.7e-14*	< 2.2e-16*

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
Reference: Yes				

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

The predictor in model 2b is e-cigarette exposure via social media. Table 8 indicates that participants' race/ethnicity, gender, and cigarette use status contributed significantly to the prediction of the participants' current e-cigarette use status. The predictor of e-cigarette exposure via social media was not significant.

Table 8 Weighted Binary Logistic Regression for E-cigarette Exposure via Social Media (Model 2b)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
E-cigarette Exposure via Social Media Reference: Low	Medium	1.03 [0.81, 1.32]	0.80	0.53
	High	1.15 [0.90, 1.46]	0.27	
Age	Age	1.03 [0.97, 1.09]	0.38	0.37
Race/Ethnicity Reference: Combined	Hispanic	0.75 [0.60, 0.93]	0.01*	0.06
	Non-Hispanic black	0.85 [0.56, 1.28]	0.44	
	Other	0.81 [0.55, 1.20]	0.30*	
Gender Reference: Male	Female	1.36 [1.13, 1.64]	1.8e-3*	1.1e-3
School Type	High School	0.91 [0.74, 1.12]	0.39	0.38

Reference: Middle School				
Cigarettes Use Reference: Yes	NO	0.31 [0.25, 0.40]	4.5e-14*	< 2.2e-16

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

The predictor in model 2c was the number of social media sites with e-cigarette content. Table 9 indicates that participants' race/ethnicity, gender, and cigarette use status contributed significantly to the prediction of the participants' current e-cigarette use status. The predictor number of social media sites with e-cigarette content was not significant.

Table 9 Weighted Binary Logistic Regression for Number of Social Media Sites with E-cigarette Content (Model 2c)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
Number of Social Media Sites with E-cigarette Content	Number of Social Media Sites with E-cigarette content	0.99 [0.93, 1.04]	0.66	0.66
Age	Age	1.03 [0.97, 1.10]	0.29	0.29
Race/Ethnicity Reference: Combined	Hispanic	0.72 [0.58, 0.90]	0.01*	0.04*
	Non-Hispanic black	0.81 [0.56, 1.17]	0.27	
	Other	0.91 [0.64, 1.30]	0.62	

Gender Reference: Male	Female	1.28 [1.07, 1.53]	0.01*	0.01*
School Type Reference: Middle School	High School	0.91 [0.73, 1.13]	0.41	0.41
Cigarettes Use Reference: Yes	NO	0.32 [0.25, 0.40]	7.5e-14*	< 2.2e-16*

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

Model 2d was the full model with all three social media predictors. Table 10 indicates none of the predictors were significant, only participants' gender, and cigarette use status contribute significantly to the prediction of the participants' current e-cigarette use status.

Table 10 Full Weighted Binary Logistic Regression for All Three Predictors (Model 2d)

Variables	Details	Adj Odds Ratio [95% CI]	P-value	P-value (Global)
Social Media Use Reference: Low	Medium	0.89 [0.46, 1.72]	0.73	0.11
	High	0.71 [0.44, 1.14]	0.16	
E-cigarette Exposure via Social Media Reference: Low	Medium	1.01 [0.79, 1.30]	0.92	0.61
	High	1.13 [0.87, 1.46]	0.35	

Number of Social Media Sites with E-cigarette Content	Number of social media sites with e-cigarette content	1.03 [0.97, 1.09]	0.39	0.38
Age	Age	1.03 [0.97, 1.10]	0.31	0.30
Race/Ethnicity Reference: Combined	Hispanic	0.75 [0.6, 0.94]	0.01*	0.06
	Non-Hispanic black	0.84 [0.56, 1.26]	0.41	
	Other	0.80 [0.55, 1.18]	0.27	
Gender Reference: Male	Female	1.39 [1.15, 1.67]	8.8e-4*	5.0e-4*
School Type Reference: Middle School	High School	0.89 [0.71, 1.12]	0.32	0.32
Cigarettes Use Reference: Yes	NO	0.32 [0.25, 0.40]	5.4e-14*	< 2.2e-16*

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

3.3 Model Assessments

3.3.1 Determine the Best Model

After obtaining all eight models from the previous section, I compared models a, b, c, and d from each outcome by examining AIC, BIC, Pseudo-R^2, and residual deviance to determine the best model.

3.3.1.1 Ever E-cigarettes Use

In Table 11, model 1d had the greatest Pseudo-R^2 and smallest deviance, AIC, and BIC, which contained all three social media predictors. Model 1b was the simpler model, which also has the second best Pseudo-R^2, AIC, and deviance. Therefore, a likelihood ratio test was performed to assess the goodness of fit of the two models. The p-value was less than 0.001, which showed strong evidence that the two models significantly differ from each other. Therefore, the two extra predictors in model 1d can't be removed, and model 1d was the best model in the prediction of the outcome ever e-cigarettes use.

Table 11 Models for Outcome One Ever E-cigarette Use

Models	Predictors	AIC	BIC	Pseudo-R^2	Deviance
1a	Social Media Use	13397.19	10886.00	0.183	13360
1b	E-cigarette Exposure via Social Media	13037.56	13051.396	0.195	13000

1c	Number of Social Media Sites with E-cigarette Content	14162.24	10874.13	0.191	14120
1d	Social Media Use E-cigarette Exposure via Social Media Number of Social Media Sites with E-cigarette Content	12947.91#	13022.59#	0.200#	12900#

 # indicates the value that has the smallest error.

3.3.1.2 Current E-cigarettes Use

In Table 12, model 2a had the greatest Pseudo-R^2; Model 2b had the smallest AIC; Model 2c had the smallest BIC; model 2d has the smallest deviance. Similarly, a likelihood ratio test was performed to assess the goodness of fit of the models. The result indicates that models 2b and 2d are not significantly different from each other, while models 2a and 2c significantly differ from 2d. Therefore, the two extra predictors in model 2d can be removed. Model 2b was the best model in the prediction of the outcome current e-cigarettes use.

Table 12 Models for Outcome Two Current E-cigarettes Use

Models	Predictors	AIC	BIC	Pseudo-R^2	Deviance
2a	Social Media Use	4082.71	3575.67	0.077#	4051
2b	E-cigarette Exposure via Social Media	4034.71#	4083.10	0.075	4002

2c	Number of Social Media Sites with E-cigarette Content	4352.88	3569.35#	0.072	4322
2d	Social Media Use E-cigarette Exposure via Social Media Number of Social Media Sites with E-cigarette Content	4038.60	4100.88	0.076	3997#

 # indicates the value that has the smallest error.

3.3.2 Interaction Terms

The interactions were tested for both models 1d and 2b, and we found that some interaction terms significantly contributed to the prediction. But the results by goodness of fit test indicated the models with interaction terms are not significantly different from the original models. Therefore, the interaction terms were dropped.

3.3.3 Comparison between the Models

The coefficient and p-values were then compared to explore how different variables contributed to participants' status on ever and current e-cigarettes use. Table 13 shows that only e-cigarette exposure via social media contributed to the prediction of current e-cigarette use. Though the p-value of e-cigarette exposure via social media was greater than 0.05, the p-value of goodness of fit test showed that model 2b was not significantly different from the reduced model without predictors. Thus, we can't drop the predictor of e-cigarette exposure via social media in model 2b.

When predicting ever e-cigarettes use, all three social media predictors were significant. For the predictor of social media use, only high frequency use was significant compared to baseline low-frequency use, though model 1a indicates only medium frequency use was significant compared to baseline low frequency use, the odds ratios of high and medium frequency use in the two models were almost the same. While for the predictor of e-cigarette exposure through social media, both medium and high frequency use were significant compared to baseline low frequency use.

Furthermore, age, school type, and cigarette use status were significant when predicting ever e-cigarettes use. The odds ratio of increased age and combined race/ethnicity are greater than one, which is associated with increased odds of ever using e-cigarettes. The odds ratio of high school and no cigarette use were smaller than one, which is associated with decreased odds of ever using e-cigarettes. When predicting current e-cigarettes use, only the odds ratio of the female was greater than one.

Table 13 Comparison of Coefficients and P-values between Model 1d and 2b

		Model 1d			Model 2b		
Variables		Adj Odds Ratio	P-value	P-value (Global)	Adj Odds Ratio	P-value	P-value (Global)
Social Media Use Reference: Low	Medium	0.64	0.03*	9.32e-4*			
	High	1.01	0.93				
E-cigarette Exposure via Social Media	Medium	1.36	2.1e-3 *	< 2.2e-16*	1.02	0.80	0.53
	High	2.04	9.7e-14*		1.15	0.27	

Reference: Low							
Number of Social Media Sites with E-cigarette Content	Number of Social Media Sites with E-cigarette Content	1.11	1.8e-07*	6.4e-09*			
Age	Age	1.42	< 2.2e-16*	< 2.2e-16*	1.03	0.38	0.37
Race/Ethnicity Reference: Combined	Hispanic	0.80	0.02*	2.8e-06*	0.75	0.01*	0.06
	Non-Hispanic black	0.49	2.2e-06*		0.85	0.44	
	Other	1.06	0.65		0.81	0.30	
Gender Reference: Male	Female	1.10	0.17	0.16	1.36	1.8e-3*	1.1e-3*
School Type Reference: Middle School	High School	0.88	0.04*	0.04*	0.91	0.39	0.38
Cigarettes Use Reference: Yes	NO	0.08	< 2.2e-16*	< 2.2e-16*	0.31	4.5e-14*	< 2.2e-16

*** indicates P-value <0.05. The smallest value in R is 2.2e-16 in default.**

3.3.4 Area Under the ROC Curve (AUC)

To further evaluate the two best logistic regression models, the predictions on how well the models classify positive and negative outcomes were checked by AUC curve. Model 1d for all three social media predictors has an AUC value of 0.806 in Figure 3, which indicates good discrimination: the model ranks a random positive example over a random negative 80.6% of the time. Model 2b for e-cigarette exposure via social media has an AUC value of 0.657 in Figure 4, which indicates fair discrimination.

3.3.4.1 Ever E-cigarettes Use

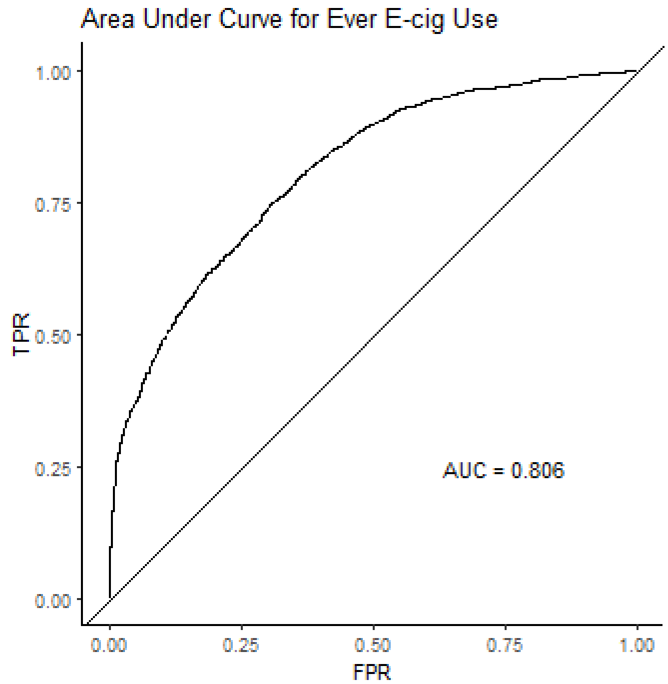

Figure 3 AUC for Model 1d for All Three Predictors

3.3.4.2 Current E-cigarettes Use

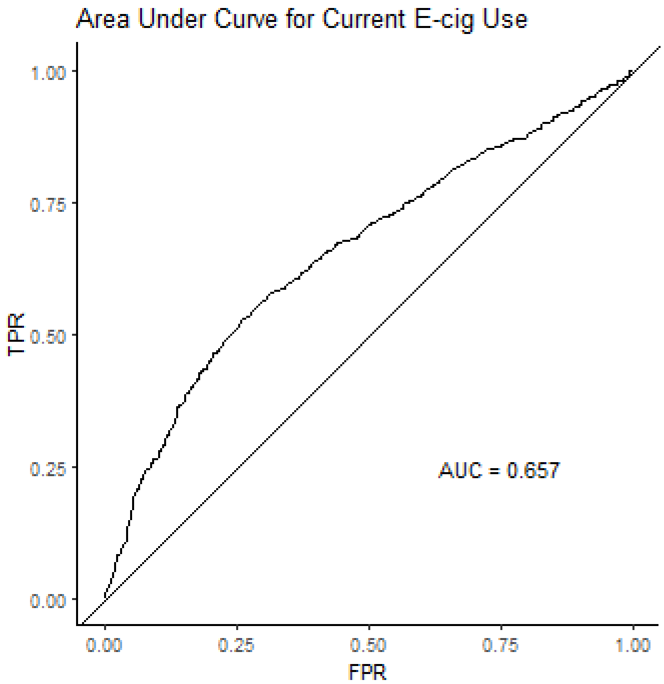

Figure 4 AUC for Model 2b for E-cigarette Exposure through Social Media

4.0 Discussion

The goal of this thesis was to explore the effect of social media on e-cigarette use among youth during the COVID-19 pandemic in 2021 with demographic characteristics of age, gender, and race/ethnicity using weighted logistic regression under a complex NYTS survey design.

The principle behind this methodology of NYTS that involves stratification, clustering, and strata are discussed in this paper, which fills the gap of lack of explanation of the weighted logistic regression model within the context of complex survey design.

Results demonstrated that social media use, e-cigarette exposure through social media, and the number of social media sites with e-cigarette content were significantly associated with increased ever using e-cigarettes.. However, the three social media predictors were not associated with the increased odds of currently using e-cigarettes. This may be because current e-cigarettes use restricts the ability to distinguish participants from regular users (current e-cig use of 30 days) since participants who reported used 1 day and 30 days fall into the same category. According to Figure 2, the frequency of regular users is the second highest. In this case, addiction may play a more important role in habitual e-cigarette use compared to the influence of social media since researchers have shown that e-cigarettes are even more addictive than traditional cigarettes (Jankowski et al., 2019).

For covariates, we identified that cigarettes use status appeared to be most significant among all the models, which indicates it is significantly associated with both increased ever and current e-cigarettes use. When predicting ever e-cigarettes use, age, race/ethnicity, and school type are significant, which indicates that participants with race/ethnicity of American Indian, Asian, and white in middle school with greater age are significantly associated with increased reports that they have ever used an e-cigarette. When predicting current e-cigarettes use, only gender was significant, which indicates that female students were more likely to report that they had used e-cigarettes on at least 1 day in the last 30 days.

This observation contradicted the finding that women have lower smoking participation than men (Chinwong et al., 2018). This is also because the categorization of current e-cigarette use restricts the ability to differentiate between participants and ordinary users, which is a limitation of the study. For future reference, different methods of categorizing current e-cigarette use can be performed to improve the result. Another limitation is that recall biases since the survey was self-reported, especially for younger participants. Lastly, The results of the 2021 NYTS survey cannot be compared to those of the previous NYTS surveys since they were performed in a different interview environment. Further research on cross-sectional data that deals with the observations on e-cigarette use in different periods can be conducted to verify the findings in this paper.

Appendix A Analysis Executed in Python

How exposure to social media and advertising affected use of e-cigarettes among youth during the pandemic?

Covariates: Age (QN1), gender (QN2), grade (QN3), race (QN5), ethnicity (QN4A-E), cigarettes use (QN38), how often see ads or promotions on internet, newspaper/magazines, stores, TV (QN128-QN131), Social Media (QN134), how often use social media (QN133), who you get from
Outcomes variable: current e-cigarettes use (how many days in the past 30 days) (QN9), Have you ever used an e-cigarette(QN6)
Method: Multinomial logistic regression
Predictors: QN128-QN138: internet, newspaper, stores, social media (what site, interaction, who)

```python
!pip install scikit-plot
!pip install mord
```

```python
import pandas as pd
import numpy as np
import scipy.stats as stats
```

```python
#These are utility tools of the DMBA book.
from dmba import regressionSummary, exhaustive_search
from dmba import backward_elimination, forward_selection, stepwise_selection
from dmba import adjusted_r2_score, AIC_score, BIC_score
from dmba import classificationSummary, gainsChart, liftChart
```

```python
# visualization and tuning the aesthetics
import matplotlib.pylab as plt
import seaborn as sns
%matplotlib inline
sns.set_style("whitegrid")
sns.set_context("notebook", font_scale=1, rc={"lines.linewidth": 2,'font.family': [u'times']})
plt.style.use('seaborn-whitegrid')
plt.rc('text', usetex = False)
plt.rc('font', family = 'serif')
plt.rc('xtick', labelsize = 10)
plt.rc('ytick', labelsize = 10)
plt.rc('font', size = 12)
plt.rc('figure', figsize = (6, 5))
```

```python
nyts = pd.read_excel('nyts2021.xlsx', na_values= ' ',
usecols=['QN1','QN2','QN3','QN4A','QN4B','QN4C','QN4D','QN4E','QN5A','QN5B','QN5C','QN5D','QN5E','QN6',

'QN20AA','QN20AB','QN20AC','QN20AD','QN20AE','QN20AF','QN20AG','QN20AH',

'QN35','QN38','QN9','QN128','QN129','QN130','QN131','QN133','QN134',

'QN135A','QN135B','QN135C','QN135D','QN135E','QN135F','QN135G','QN135H',

'psu','stratum','hsms','finwgt','v_stratum'])
```

```python
nyts
```

```python
nyts.info()
```

```python
nyts.columns
```

```python
nyts.rename(columns = {'QN1':'Age', 'QN2':'Gender', 'QN3':'Grade',
                       'QN4A':'Not_Hispanic', 'QN4B':'Yes_Mexican',
'QN4C':'Yes_Puerto','QN4D':'Yes_Cuban','QN4E':'Yes_Another',
                       'QN9':
'E_cigarettes_use','QN6':'Ever_e_cigarettes_use',

'QN20AA':'Myself','QN20AB':'Had_someone_buy','QN20AC':'Ask_someone_give','QN20AD':'Someone_offered','QN20AE':'Friend','QN20AF':'Family_member','QN20AG':'Store_people','QN20AH':'Other',
                       'QN35':'Ever_cigarettes_use',
'QN38':'Current_cigarettes_use',
                       'QN128':'Internet', 'QN129': 'Newspaper',
'QN130':'Store', 'QN131': 'TV',
'QN133':'Social_media_use','QN134':'Social_media_freq',
                       'QN135A': 'Facebook', 'QN135B':'Instgram',
'QN135C':'Snapchat','QN135D':'Tiktok', 'QN135E':'Twitter', 'QN135F':'Raddit',
'QN135G':'Youtube', 'QN135H':'Other site'},
            inplace = True)
```

```python
nyts
```

```python
nyts.describe()
```

```python
nyts.isna().sum()
```
```python
nyts['QN5A'].fillna(0, inplace=True)
nyts['QN5B'].fillna(0, inplace=True)
nyts['QN5C'].fillna(0, inplace=True)
nyts['QN5D'].fillna(0, inplace=True)
nyts['QN5E'].fillna(0, inplace=True)
conditions = [(nyts['QN5A'] == 1) & (nyts['QN5B'] != 1) & (nyts['QN5C'] != 1) & (nyts['QN5D'] != 1) & (nyts['QN5E'] != 1), (nyts['QN5B'] == 1) & (nyts['QN5A'] != 1) & (nyts['QN5C'] != 1) & (nyts['QN5D'] != 1) & (nyts['QN5E'] != 1), (nyts['QN5C'] == 1)& (nyts['QN5A'] != 1) & (nyts['QN5B'] != 1) & (nyts['QN5D'] != 1) & (nyts['QN5E'] != 1), (nyts['QN5D'] == 1)& (nyts['QN5A'] != 1) & (nyts['QN5C'] != 1) & (nyts['QN5B'] != 1) & (nyts['QN5E'] != 1), (nyts['QN5E'] == 1)& (nyts['QN5A'] != 1) & (nyts['QN5C'] != 1) & (nyts['QN5D'] != 1) & (nyts['QN5B'] != 1), nyts['QN5A'] + nyts['QN5B']+ nyts['QN5C']+ nyts['QN5D'] + nyts['QN5E'] > 1]
outputs = ['American Indian or Alaska Native', 'Asian', 'Black or African American', 'Native Hawaiian or Other Pacific Islander', 'White', 'Multi-racial']
nyts['Race'] = np.select(conditions, outputs, 'Missing')
nyts.head(20)
```

```python
nyts.rename(columns = {'QN4A':'Not_Hispanic','QN4B':'Yes_Mexican','QN4C':'Yes_Puerto','QN4D': 'Yes_Cuban', 'QN4E':'Yes_Another','QN5A':'American Indian or Alaska Native', 'QN5B':'Asian', 'QN5C':'Black or African American', 'QN5D':'Native Hawaiian or Other Pacific Islander', 'QN5E':'White'}, inplace = True)
```
Stats of Ever Cig Use

```python
nyts.Ever_cigarettes_use.describe()
```

```python
nyts.Ever_cigarettes_use.isna().sum()
```

```python
nyts.Ever_cigarettes_use.value_counts()
```

```python
nyts.Ever_cigarettes_use.value_counts()
```

```python
nyts.Ever_cigarettes_use.fillna(3, inplace=True)
```

```python
nyts.Ever_cigarettes_use.value_counts()
```

~~~
```

# Stats of Ever E-cig Use

```python
nyts.Ever_e_cigarettes_use.describe()
```

```python
nyts.Ever_e_cigarettes_use.isna().sum()
```

```python
nyts.Ever_e_cigarettes_use.value_counts()
```

```python
nyts['Ever_e_cigarettes_use']=
nyts['Ever_e_cigarettes_use'].replace([2],0)
```

```python
nyts.Ever_e_cigarettes_use.fillna(2, inplace=True)
```

```python
nyts.Ever_e_cigarettes_use.value_counts()
```

# Stats of Current E-cig Use

```python
nyts.E_cigarettes_use.describe()
```

```python
nyts.E_cigarettes_use.isna().sum()
```

```python
nyts.E_cigarettes_use.value_counts()
```

```python
sns.displot(nyts['E_cigarettes_use'], kind='hist', bins=30, kde=True,
rug=False)
```

```python
E_cigarettes_use_filter= nyts[(nyts['E_cigarettes_use']>0) &
(nyts['E_cigarettes_use']<31)]
E_cigarettes_use_filter
```

```python
~~~

```
sns.displot(E_cigarettes_use_filter['E_cigarettes_use'], kind='hist',
bins=30, kde=True, rug=False)
```

```python
E_cigarettes_use_filter.E_cigarettes_use.describe()
```

```python
E_cigarettes_days_filter1 = E_cigarettes_use_filter['E_cigarettes_use']
bins = [0.5,15.5,30.5]
bin_names = ['1', '2']
E_cigarettes_use_filter['E_cigarettes_use_category_filter1'] =
pd.cut(E_cigarettes_days_filter1,bins,labels=bin_names)
E_cigarettes_use_filter
```

```python
E_cigarettes_use_filter.E_cigarettes_use_category_filter1.value_counts(
)
```

```python
E_cigarettes_use_filter.groupby('E_cigarettes_use_category_filter1')['E
_cigarettes_use'].mean()
```

Cutpoints testing: 3 cuts

```python
E_cigarettes_days_filter2 = E_cigarettes_use_filter['E_cigarettes_use']
bins = [0.5,5.5,25.5,30.5]
bin_names = ['1', '2', '3']
E_cigarettes_use_filter['E_cigarettes_use_category_filter2'] =
pd.cut(E_cigarettes_days_filter2,bins,labels=bin_names)
E_cigarettes_use_filter
```

```python
E_cigarettes_use_filter.E_cigarettes_use_category_filter2.value_counts(
)
```

```python
E_cigarettes_use_filter.groupby('E_cigarettes_use_category_filter2')['E
_cigarettes_use'].mean()
```

```python
conditions_E_cigarettes_use = [nyts['E_cigarettes_use'] >= 1,
nyts['E_cigarettes_use'] == 0]
outputs_E_cigarettes_use = ['1', '0']
nyts['E_cigarettes_use_category'] =
np.select(conditions_E_cigarettes_use, outputs_E_cigarettes_use, '2')
nyts
```

```python
```

````
      nyts['E_cigarettes_use_category'].value_counts()
      ```

      ```python
      sns.displot(nyts['E_cigarettes_use_category'], kind='hist', bins=10,
kde=True, rug=False)
      ```

      # Stats of Age

      ```python
      nyts.Age.describe()
      ```

      ```python
      nyts.Age.isna().sum()
      ```


      ```python
      nyts.Age.value_counts()
      ```

      ```python
      nyts['Age'].fillna(12, inplace=True)
      ```

      ```python
      nyts.Age.value_counts()
      ```

      ```python
      nyts.Age.isna().sum()
      ```

      ```python
      nyts['Age_continuous'] = nyts['Age']+8
      ```

      ```python
      Age_filter= nyts[(nyts['Age_continuous']>0) &
(nyts['Age_continuous']<20)]
      Age_filter
      ```

      ```python
      Age_filter['Age_continuous'].describe()
      ```

      ```python
      AgecountByE_cigarettes_use_category =
nyts.groupby('Age')[['E_cigarettes_use_category']].count()
      AgecountByE_cigarettes_use_category
      ```


      ```python
````

pd.crosstab(nyts.Age, nyts.E_cigarettes_use_category, margins=True)
```

```

```python
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
sns.set(color_codes=True)
```

```python
sns.set(style="darkgrid")
```

```python
sns.displot(nyts['Age'], kind='hist', bins=11, rug=False)
```


```python
nyts.Age.value_counts()
```


Stats of Gender

```python
nyts.Gender.value_counts()
```


```python
nyts.Gender.isna().sum()
```

```python
conditions_Gender = [(nyts['Gender'] == 1),
                     (nyts['Gender'] == 2)]
outputs_Gender = ['1','2']
nyts['gender'] = np.select(conditions_Gender, outputs_Gender, '3')
nyts.head(20)
```

```python
pd.crosstab(nyts.Gender, nyts.E_cigarettes_use_category, margins=True)
```

Stats of Race

```python
nyts.Race.value_counts()
```

```python
nyts.Race.isna().sum()
```

Stats of Multiple Race

Note: This variable is named race_s in the public use data set. The multiple race categories are Hispanic, non-Hispanic (NH) White, non-Hispanic Black, non-Hispanic Asian, non-Hispanic American Indian or Alaskan Native (AIAN), and non-Hispanic Native Hawaiian or Pacific Islander (NHOPI).

```python
conditions_race = [(nyts['Yes_Mexican'] == 1) | (nyts['Yes_Puerto'] == 1) | (nyts['Yes_Cuban'] == 1) | (nyts['Yes_Another'] == 1),
                   (nyts['Not_Hispanic'] == 1)& (nyts['Race'] == 'White'),
                   (nyts['Not_Hispanic'] == 1)& (nyts['Race'] == 'Asian'),
                   (nyts['Not_Hispanic'] == 1)& (nyts['Race'] == 'American Indian or Alaska Native'),
                   (nyts['Not_Hispanic'] == 1)& (nyts['Race'] == 'Black or African American'),
                   (nyts['Not_Hispanic'] == 1)& (nyts['Race'] == 'Multi-racial')]
outputs_race = ['Hispanic','Non-Hispanic White', 'Non-Hispanic Asian', 'Non-Hispanic American Indian', 'Non-Hispanic Black', 'Non-Hispanic Multi-racial']
nyts['Multiple_Race'] = np.select(conditions_race, outputs_race, 'Missing')
nyts.head(20)
```

```python
nyts.Multiple_Race.value_counts()
```

For poststratification purposes, a unique race and ethnicity was assigned to respondents with missing data on race and ethnicity, those with an “Other” classification, and those reporting multiple races.

```python
conditions_race = [(nyts['Multiple_Race'] == 'Non-Hispanic Multi-racial')|(nyts['Multiple_Race'] == 'Missing'),
                   (nyts['Multiple_Race'] == 'Hispanic'),
                   (nyts['Multiple_Race'] == 'Non-Hispanic Black'),
                   (nyts['Multiple_Race'] == 'Non-Hispanic White') | (nyts['Multiple_Race'] == 'Non-Hispanic Asian') | (nyts['Multiple_Race'] == 'Non-Hispanic American Indian')]
outputs_race = ['4','2', '3', '1']
nyts['Race_Ethnicity'] = np.select(conditions_race, outputs_race, 'Missing')
nyts.head(20)
```

```python
```

````
nyts.Race_Ethnicity.value_counts()
```

```python
pd.crosstab(nyts.Race_Ethnicity, nyts.E_cigarettes_use_category,
margins=True)
```

```python
sns.displot(nyts['Race_Ethnicity'], kind='hist', bins=10, rug=False,
height=8.27, aspect=18.7/8.27)
```

```python
sns.set(rc={'figure.figsize':(20,10)})
g=sns.boxplot(x="Race_Ethnicity", y="E_cigarettes_use", data=nyts)
g=sns.stripplot(x="Race_Ethnicity", y="E_cigarettes_use", data=nyts)
```

# Stats of Education

```python
conditions_Grade = [(nyts['Internet'] >= 4)&(nyts['Internet'] < 8),
(nyts['Internet'] < 4)&(nyts['Internet'] > 0)]
outputs_Grade = ['2', '1']
nyts['Grade_category'] = np.select(conditions_Grade, outputs_Grade,
'3')
```

```python
nyts.Grade.describe()
```


```python
nyts.Grade.isna().sum()
```

```python
nyts.Grade_category.isna().sum()
```

```python
nyts.Grade_category.value_counts()
```

```python
pd.crosstab(nyts.Grade_category, nyts.E_cigarettes_use_category,
margins=True)
```

```python
g=sns.boxplot(x="Grade_category", y="E_cigarettes_use", data=nyts)
g=sns.stripplot(x="Grade_category", y="E_cigarettes_use", data=nyts)
```

```python
````

```
g=sns.boxplot(x="Grade", y="E_cigarettes_use", data=nyts)
g=sns.stripplot(x="Grade", y="E_cigarettes_use", data=nyts)
```

```python
sns.displot(nyts['Grade_category'], kind='hist', bins=10, rug=False)
```

```python
sns.displot(nyts['Grade'], kind='hist', bins=6, rug=False)
```

Stats of Social Media

```python
sns.displot(nyts['Internet'], kind='hist', bins=5, kde=False, rug=False)
```

```python
sns.displot(nyts['Newspaper'], kind='hist', bins=5, kde=False, rug=False)
```

```python
sns.displot(nyts['Store'], kind='hist', bins=5, kde=False, rug=False)
```

```python
sns.displot(nyts['TV'], kind='hist', bins=5, kde=False, rug=False)
```

```python
sns.displot(nyts['Social_media_freq'], kind='hist', bins=5, kde=False, rug=False)
```

```python
conditions2 = [nyts['Internet'] >= 3, (nyts['Internet'] < 3)&(nyts['Internet'] > 0)]
outputs2 = ['Yes', 'No']
nyts['Internet_YN'] = np.select(conditions2, outputs2, 'Missing')
```

```python
conditions3 = [nyts['Newspaper'] >= 3, (nyts['Newspaper'] < 3)&(nyts['Newspaper'] > 0)]
outputs3 = ['Yes', 'No']
nyts['Newspaper_YN'] = np.select(conditions3, outputs3, 'Missing')
```

```

```python
conditions4 = [nyts['Store'] >= 3, (nyts['Store'] < 3)&(nyts['Store'] > 0)]
outputs4 = ['Yes', 'No']
nyts['Store_YN'] = np.select(conditions4, outputs4, 'Missing')
```

```python
conditions5 = [nyts['TV'] >= 3, (nyts['TV'] < 3)&(nyts['TV'] > 0)]
outputs5 = ['Yes', 'No']
nyts['TV_YN'] = np.select(conditions5, outputs5, 'Missing')
```

```python
nyts[['Internet_YN','Newspaper_YN', 'Store_YN', 'TV_YN']].apply(pd.Series.value_counts)
```

Social Media Use (time)

```python
nyts.Social_media_use.isna().sum()
```

```python
sns.displot(nyts['Social_media_use'], kind='hist', bins=30, kde=True, rug=False)
```

```python
conditions7 = [nyts['Social_media_use'] >= 6, (nyts['Social_media_use'] < 4)&(nyts['Social_media_use'] > 1),(nyts['Social_media_use'] < 6)&(nyts['Social_media_use'] > 3)]
outputs7 = ['3', '1', '2']
nyts['Social_media_use_category'] = np.select(conditions7, outputs7, '4')
```

```python
nyts.Social_media_use_category.value_counts()
```

Social Media (exposure)

```python
sns.displot(nyts['Social_media_freq'], kind='hist', bins=30, kde=True, rug=False)
```

```python
conditions8 = [nyts['Social_media_freq'] >= 4, (nyts['Social_media_freq'] < 3)&(nyts['Social_media_freq'] > 0),nyts['Social_media_freq'] == 3]
outputs8 = ['3', '1', '2']

```
nyts['Social_media_category'] = np.select(conditions8, outputs8, '4')
```

```python
nyts.Social_media_category.value_counts()
```

Social media Site

'QN135A': 'Facebook', 'QN135B':'Instgram', 'QN135C':'Snapchat','QN135D':'Tiktok', 'QN135E':'Twitter', 'QN135F':'Raddit', 'QN135G':'Youtube', 'QN135H':'Other site'

```python
nyts['Facebook'].fillna(0, inplace=True)
nyts['Instgram'].fillna(0, inplace=True)
nyts['Snapchat'].fillna(0, inplace=True)
nyts['Tiktok'].fillna(0, inplace=True)
nyts['Twitter'].fillna(0, inplace=True)
nyts['Raddit'].fillna(0, inplace=True)
nyts['Youtube'].fillna(0, inplace=True)
nyts['Other site'].fillna(0, inplace=True)
conditions9 = [nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 0,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 1,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 2,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 3,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 4,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 5,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 6,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 7,
              nyts['Facebook'] + nyts['Instgram']+ nyts['Snapchat']+
nyts['Tiktok'] + nyts['Twitter'] + nyts['Raddit']+ nyts['Youtube'] +
nyts['Other site'] == 8,
             ]
outputs9 = ['0','1', '2', '3', '4', '5', '6', '7','8']
nyts['Social_media_sites'] = np.select(conditions9, outputs9, '9')
nyts.head(20)
```
]

```python
```

```python
nyts.Social_media_sites = pd.to_numeric(nyts.Social_media_sites)
```

```python
nyts.Social_media_sites.describe()
```

```python
nyts.Social_media_sites.value_counts()
```

```python
nyts.to_csv('nyts.csv',index=False)
```

Appendix B Analysis Executed in R

````
```{r}
library(readr)
nyts <- read_csv('nyts.csv')
nyts
```

```{r}
library(tidyverse)
library("haven")
library("survey")
library("jtools")
library("remotes")
library("svrepmisc")
library("car")
```
```{r}
str(nyts)
```
```{r}
nyts$Gender <- factor(nyts$gender)
nyts$E_cigarettes_use_category <- factor(nyts$E_cigarettes_use_category)
nyts$Ever_cigarettes_use <- factor(nyts$Ever_cigarettes_use)
nyts$Ever_e_cigarettes_use <- factor(nyts$Ever_e_cigarettes_use)
nyts$Social_media_use_category  <- factor(nyts$Social_media_use_category)
nyts$Social_media_category <- factor(nyts$Social_media_category)
nyts$Social_media_sites <- as.numeric(nyts$Social_media_sites)
nyts$Grade_category <- factor(nyts$Grade_category)
nyts$Race_Ethnicity <- factor(nyts$Race_Ethnicity)
str(nyts)
```

```{r}
#Test the data with missing values as a seperate category first
d_nyts<- svydesign(id=~psu, strata=~v_stratum, weights=~finwgt,
survey.lonely.psu = "adjust", data=nyts,
nest=TRUE)
d_nyts
```

```{r}
summary(d_nyts)
```

```{r}
# Calculate weighted stats
svyciprop(~I(gender==1), d_nyts, method="likelihood")
```
```{r}
svyciprop(~I(gender==2), d_nyts, method="likelihood")
```
````

````
```{r}
svyciprop(~I(Ever_e_cigarettes_use==0), d_nyts, method="likelihood")
svyciprop(~I(Ever_e_cigarettes_use==1), d_nyts, method="likelihood")
svyciprop(~I(Ever_e_cigarettes_use==2), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(Ever_e_cigarettes_use) %>%
  group_by(Ever_e_cigarettes_use) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```

```{r}
svyciprop(~I(E_cigarettes_use_category==0), d_nyts, method="likelihood")
svyciprop(~I(E_cigarettes_use_category==1), d_nyts, method="likelihood")
svyciprop(~I(E_cigarettes_use_category==2), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(E_cigarettes_use_category) %>%
  group_by(E_cigarettes_use_category) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```
```{r}
svyciprop(~I(Social_media_use_category==1), d_nyts, method="likelihood")
svyciprop(~I(Social_media_use_category==2), d_nyts, method="likelihood")
svyciprop(~I(Social_media_use_category==3), d_nyts, method="likelihood")
svyciprop(~I(Social_media_use_category==4), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(Social_media_use_category) %>%
  group_by(Social_media_use_category) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```
```{r}
svyciprop(~I(Social_media_category==1), d_nyts, method="likelihood")
svyciprop(~I(Social_media_category==2), d_nyts, method="likelihood")
svyciprop(~I(Social_media_category==3), d_nyts, method="likelihood")
svyciprop(~I(Social_media_category==4), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(Social_media_category) %>%
  group_by(Social_media_category) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```
```{r}
````
````

````
svyciprop(~I(Gender==1), d_nyts, method="likelihood")
svyciprop(~I(Gender==2), d_nyts, method="likelihood")
svyciprop(~I(Gender==3), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(Gender) %>%
  group_by(Gender) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```
```{r}
svyciprop(~I(Grade_category==1), d_nyts, method="likelihood")
svyciprop(~I(Grade_category==2), d_nyts, method="likelihood")
svyciprop(~I(Grade_category==3), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(Grade_category) %>%
  group_by(Grade_category) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```
```{r}
svyciprop(~I(Ever_cigarettes_use==1), d_nyts, method="likelihood")
svyciprop(~I(Ever_cigarettes_use==2), d_nyts, method="likelihood")
svyciprop(~I(Ever_cigarettes_use==3), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(Ever_cigarettes_use) %>%
  group_by(Ever_cigarettes_use) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```
```{r}
svyciprop(~I(Race_Ethnicity==1), d_nyts, method="likelihood")
svyciprop(~I(Race_Ethnicity==2), d_nyts, method="likelihood")
svyciprop(~I(Race_Ethnicity==3), d_nyts, method="likelihood")
svyciprop(~I(Race_Ethnicity==4), d_nyts, method="likelihood")
```
```{r}
nyts %>%
  drop_na(Race_Ethnicity) %>%
  group_by(Race_Ethnicity) %>%
  summarize(n=n()) %>%
  mutate(Prop=round(n/sum(n),3
                ))
```
```{r}
#Return the missing value from categorical value to NA
nyts2 <- nyts %>%
  mutate(Age=na_if(Age,12),
         Age_continuous= na_if(Age_continuous, 20),
````

````
         Gender = na_if(Gender, 3),
         Grade_category = na_if(Grade_category, 3),
         Ever_cigarettes_use = na_if(Ever_cigarettes_use, 3),
         Social_media_category = na_if(Social_media_category, 4),
         Social_media_use_category = na_if(Social_media_use_category, 4),
         Ever_e_cigarettes_use = na_if(Ever_e_cigarettes_use, 2),
             E_cigarettes_use_category = na_if(E_cigarettes_use_category, 2),
                       )
```


```{r}
#Fit the logistic regression model with the svyglm() function from the survey
package
nyts3 <- nyts2 %>%
dplyr::select(Ever_e_cigarettes_use,E_cigarettes_use_category,E_cigarettes_us
e,
Social_media_use_category,Social_media_category,Social_media_sites,Age_contin
uous,Race_Ethnicity,Gender,Grade_category,Ever_cigarettes_use,psu,finwgt,v_st
ratum)
d_nyts3<- svydesign(id=~psu, strata=~v_stratum, weights=~finwgt,
survey.lonely.psu = "adjust", data=nyts3,
nest=TRUE)
nyts4 <- nyts3 %>% dplyr::select(Ever_e_cigarettes_use,
Social_media_use_category,Social_media_category,Social_media_sites,Age_contin
uous,Race_Ethnicity,Gender,Grade_category,Ever_cigarettes_use,psu,finwgt,v_st
ratum)
d_nyts4<- svydesign(id=~psu, strata=~v_stratum, weights=~finwgt,
survey.lonely.psu = "adjust", data=nyts4,
nest=TRUE)
```
```{r}
svymean(~Age_continuous+Social_media_sites,d_nyts3, na = TRUE)
svysd(~Age_continuous+Social_media_sites,d_nyts3, na = TRUE)
svyquantile(~Age_continuous+Social_media_sites,d_nyts3, na = TRUE,
c(0,.25,.5,.75,1),ci=TRUE)
```

```{r}
svyciprop(~I(Gender==2), d_nyts4, method="likelihood")
logit_gender <- (svyglm(Ever_e_cigarettes_use~Gender, family=quasibinomial,
design=d_nyts4
, na.action = na.omit))
summary(logit_gender)
```

```{r}
#Outcome 1: Ever e-cig use
#To get around the warning regarding noninteger counts created by sample
weights, use the quasibinomial method.
logit1d <-
(svyglm(Ever_e_cigarettes_use~Social_media_use_category+Social_media_category
+Social_media_sites+Age_continuous+Race_Ethnicity+Gender+Grade_category+Ever_
cigarettes_use, family=quasibinomial, design=d_nyts4
, na.action = na.omit))
logit1d
```
````
````

````
```{r}
summ(
  logit1d,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```
```{r}
# Find adjusted odds ratio
library(basecamb)
or_model_summary(
  logit1d,
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```

```{r}
#Access the model
psrsq(logit1d, method = c("Cox-Snell"))

# The Rao-Scott approximation to the weighted loglikelihood is used to
construct AIC.
AIC(logit1d)


BIC(logit1d, maximal=logit1d)

# The Anova() function in the car package handles "svyglm"  by default type-II
Wald tests
Anova(logit1d)
```


```{r}
#Predictor 1: Social media use
logit1a <-
(svyglm(Ever_e_cigarettes_use~Social_media_use_category+Age_continuous+Race_E
thnicity+Gender+Grade_category+Ever_cigarettes_use, family=quasibinomial,
design=d_nyts4
, na.action = na.omit))
logit1a
```
````

````
```{r}
summ(
  logit1a,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default =2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```
```{r}
or_model_summary(
  logit1a,
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```

```{r}
#repeat
psrsq(logit1a, method = c("Cox-Snell"))
AIC(logit1a)
BIC(logit1a, maximal=logit1d)
Anova(logit1a)
```


```{r}
#Predictor 2: Social media exposure
logit1b <-
(svyglm(Ever_e_cigarettes_use~Social_media_category+Age_continuous+Race_Ethni
city+Gender+Grade_category+Ever_cigarettes_use, family=quasibinomial,
design=d_nyts4
, na.action = na.omit))
logit1b
```
```{r}
summ(
  logit1b,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
````
````

````
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```
```{r}
or_model_summary(
  logit1b,
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```

```{r}
psrsq(logit1b, method = c("Cox-Snell"))
AIC(logit1b)
BIC(logit1b, maximal=logit1d)
Anova(logit1b)
```

```{r}
#Predictor 3: Social media sites
logit1c <-
(svyglm(Ever_e_cigarettes_use~Social_media_sites+Age_continuous+Race_Ethnicit
y+Gender+Grade_category+Ever_cigarettes_use, family=quasibinomial,
design=d_nyts4
, na.action = na.omit))
logit1c
```
```{r}
summ(
  logit1c,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```
```{r}
or_model_summary(
  logit1c,
````

```
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```

```{r}
psrsq(logit1c, method = c("Cox-Snell"))
AIC(logit1c)
BIC(logit1c, maximal=logit1d)
Anova(logit1c)
```
```{r}
#Outcome 2: Current e-cig Use
nyts5 <- nyts3 %>% dplyr::select(E_cigarettes_use_category,
Social_media_use_category,Social_media_category,Social_media_sites,Age_contin
uous,Race_Ethnicity,Gender,Grade_category,Ever_cigarettes_use,psu,finwgt,v_st
ratum)
d_nyts5<- svydesign(id=~psu, strata=~v_stratum, weights=~finwgt,
survey.lonely.psu = "adjust", data=nyts5,
nest=TRUE)
logit2d <-
(svyglm(E_cigarettes_use_category~Social_media_use_category+Social_media_cate
gory+Social_media_sites+Age_continuous+Race_Ethnicity+Gender+Grade_category+E
ver_cigarettes_use, family=quasibinomial, design=d_nyts5
, na.action = na.omit))
logit2d
```
```{r}
summ(
  logit2d,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```

```{r}
or_model_summary(
  logit2d,
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```

````
```

```{r}
psrsq(logit2d, method = c("Cox-Snell"))
AIC(logit2d)
BIC(logit2d, maximal=logit2d)
Anova(logit2d)
```

```{r}
#Predictor 1: Social media use
logit2a <-
(svyglm(E_cigarettes_use_category~Social_media_use_category+Age_continuous+Ra
ce_Ethnicity+Gender+Grade_category+Ever_cigarettes_use, family=quasibinomial,
design=d_nyts5
, na.action = na.omit))
logit2a
```
```{r}
summ(
  logit2a,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```
```{r}
or_model_summary(
  logit2a,
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```

```{r}
psrsq(logit2a, method = c("Cox-Snell"))
AIC(logit2a)
BIC(logit2a, maximal=logit2d)
Anova(logit2a)
```

```{r}
#Predictor 2: Social media exposure
````

````
logit2b <-
(svyglm(E_cigarettes_use_category~Social_media_category+Age_continuous+Race_E
thnicity+Gender+Grade_category+Ever_cigarettes_use, family=quasibinomial,
design=d_nyts5
, na.action = na.omit))
logit2b
```
```{r}
summ(
  logit2b,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```
```{r}
or_model_summary(
  logit2b,
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```


```{r}
psrsq(logit2b, method = c("Cox-Snell"))
AIC(logit2b)
BIC(logit2b, maximal=logit2d)
Anova(logit2b)
```


```{r}
#Predictor 3: Social media sites
logit2c <-
(svyglm(E_cigarettes_use_category~Social_media_sites+Age_continuous+Race_Ethn
icity+Gender+Grade_category+Ever_cigarettes_use, family=quasibinomial,
design=d_nyts5
, na.action = na.omit))
logit2c
```
```{r}
summ(
  logit2c,
````

```
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL
)
```

```{r}
or_model_summary(
  logit2c,
  conf_int = 1.96,
  print_intercept = FALSE,
  round_est = 2,
  round_p = 4
)
```

```{r}
psrsq(logit2c, method = c("Cox-Snell"))
AIC(logit2c)
BIC(logit2c, maximal=logit2d)
Anova(logit2c)
```
```{r}
#Test if model 1 and 2 from Outcome 1 are the same
anova(logit1d,logit1b)
```

```{r}
#Test if models from Outcome 2 are the same
anova(logit2a,logit2d)
anova(logit2b,logit2d)
anova(logit2c,logit2d)
```
```{r}
# test reduced model without social media exposure
logit2b_gnf <-
(svyglm(E_cigarettes_use_category~Age_continuous+Race_Ethnicity+Gender+Grade_
category+Ever_cigarettes_use, family=quasibinomial, design=d_nyts5
, na.action = na.omit))
logit2b_gnf
anova(logit2b_gnf,logit2b)
```

```{r}
#Compare the best models from the 2 outcome variables
compareCoefs(logit2d,logit2b)
```

````
```{r}
summary(logit1d)$coefficients[,4]
summary(logit2b)$coefficients[,4]
```
```{r}
#Interactions
# All pair-wise interactions for model 1d
logit1d_int_all <- (svyglm(Ever_e_cigarettes_use~
(Social_media_use_category+Social_media_category+Social_media_sites+Age_conti
nuous+Race_Ethnicity+Gender+Grade_category+Ever_cigarettes_use)*
(Social_media_use_category+Social_media_category+Social_media_sites+Age_conti
nuous+Race_Ethnicity+Gender+Grade_category+Ever_cigarettes_use),
family=quasibinomial, design=d_nyts4
, na.action = na.omit))
summ(logit1d_int_all,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL)
anova(logit1d,logit1d_int_all)
```

```{r}
# All pair-wise interactions for model 2b
logit2b_int_all <-
(svyglm(E_cigarettes_use_category~(Social_media_category+Age_continuous+Race_
Ethnicity+Gender+Grade_category+Ever_cigarettes_use)*(Social_media_category+A
ge_continuous+Race_Ethnicity+Gender+Grade_category+Ever_cigarettes_use),
family=quasibinomial, design=d_nyts5
, na.action = na.omit))
logit2b_int_all
summ(logit1d_int,
  scale = TRUE,
  confint = getOption("summ-confint", TRUE),
  ci.width = getOption("summ-ci.width", 0.95),
  digits = getOption("jtools-digits", default = 2),
  pvals = getOption("summ-pvals", TRUE),
  n.sd = 1,
  center = FALSE,
  transform.response = FALSE,
  scale.only = FALSE,
  exp = FALSE,
  vifs = getOption("summ-vifs", TRUE),
  model.info = getOption("summ-model.info", TRUE),
  model.fit = getOption("summ-model.fit", TRUE),
  which.cols = NULL)
anova(logit2b,logit2b_int_all)
````
````

````
```



```{r}
library(WeightedROC)
predROC1 <- function (glm.obj1, newData1)
{
options(survey.lonely.psu="adjust")

  pred1 <- rep(NA, nrow(newData1)); names(pred1) <- rownames(newData1)
  model_terms1 <- attributes(glm.obj1$terms)$variables
  predictors1 <- as.character(model_terms1[3:length(model_terms1)])
  response1 <-  as.character(model_terms1[2])
  newData1 <- newData1[,c("finwgt",response1,predictors1)]
  xnn1 <- na.omit(newData1)
  pred1[-attr(xnn1, "na.action")] <- predict(glm.obj1, xnn1)
  guess1 <- 1/(1+exp(-pred1))
  dframe1 <- data.frame(response1=ifelse(newData1[, response1]==0, -1, 1),
                      guess1=guess1,
                      WEIGHT1=newData1$finwgt)
  dframe1 <- na.omit(dframe1)
  dframe1$Ever_e_cigarettes_use<-as.factor(dframe1$Ever_e_cigarettes_use)

subset(WeightedROC::WeightedROC(dframe1$guess1,
dframe1$Ever_e_cigarettes_use, weight=dframe1$WEIGHT1))
}
```

```{r}
predROC1(logit1d, nyts4)
```
```{r}
library(ggplot2)
plotAUC1 <- function (glm.obj1, newData1)
{
options(survey.lonely.psu="adjust")

  pred1 <- rep(NA, nrow(newData1)); names(pred1) <- rownames(newData1)
  model_terms1 <- attributes(glm.obj1$terms)$variables
  predictors1 <- as.character(model_terms1[3:length(model_terms1)])
  response1 <-  as.character(model_terms1[2])
  newData1 <- newData1[,c("finwgt",response1,predictors1)]
  xnn1 <- na.omit(newData1)
  pred1[-attr(xnn1, "na.action")] <- predict(glm.obj1, xnn1)
  guess1 <- 1/(1+exp(-pred1))
  dframe1 <- data.frame(response1=ifelse(newData1[, response1]==0, -1, 1),
                      guess1=guess1,
                      WEIGHT1=newData1$finwgt)
  dframe1 <- na.omit(dframe1)
  dframe1$Ever_e_cigarettes_use<-as.factor(dframe1$Ever_e_cigarettes_use)
tp.fp1 <- WeightedROC::WeightedROC(dframe1$guess1,
dframe1$Ever_e_cigarettes_use, weight=dframe1$WEIGHT1)
ggplot()+ geom_path(aes(FPR, TPR), data=tp.fp1)+
coord_equal()+theme_classic() + ggtitle("Area Under Curve for Ever E-cig
Use") + geom_abline(intercept = 0, slope = 1)+
````

```
     annotate("text", x = .75, y = .25, label = paste("AUC
=",round(WeightedAUC(tp.fp1),3)))
}
```
```{r}
png(file="C:/Users/zhulu/Desktop/BIOST 2099/AUC1.png",
width=600, height=350)
plotAUC1 (logit1d, nyts4)
dev.off()
```

```{r}
library(WeightedROC)
nyts5 <- nyts2 %>% dplyr::select(E_cigarettes_use_category,
Social_media_use_category,Social_media_category,Social_media_sites,Age_contin
uous,Race_Ethnicity,Gender,Grade_category,Ever_cigarettes_use,psu,finwgt,v_st
ratum)

```

```{r}
predROC2 <- function (glm.obj2, newData2)
{
options(survey.lonely.psu="adjust")

  pred2 <- rep(NA, nrow(newData2)); names(pred2) <- rownames(newData2)
  model_terms2 <- attributes(glm.obj2$terms)$variables
  predictors2 <- as.character(model_terms2[3:length(model_terms2)])
  response2 <-  as.character(model_terms2[2])
  newData2 <- newData2[,c("finwgt",response2,predictors2)]
  xnn2 <- na.omit(newData2)
  pred2[-attr(xnn2, "na.action")] <- predict(glm.obj2, xnn2)
  guess2 <- 1/(1+exp(-pred2))
  dframe2 <- data.frame(response2=ifelse(newData2[, response2]==0, -1, 1),
                       guess2=guess2,
                       WEIGHT2=newData2$finwgt)
  dframe2 <- na.omit(dframe2)

subset(WeightedROC::WeightedROC(dframe2$guess2,
dframe2$E_cigarettes_use_category, weight=dframe2$WEIGHT2))
}
```

```{r}
#png(file="C:/Users/zhulu/Desktop/BIOST 2099/AUC2.png",
#width=600, height=350)
predROC2 (logit2b, nyts5)
#dev.off()
```

```{r}
plotAUC2 <- function (glm.obj2, newData2)
{
options(survey.lonely.psu="adjust")

  pred2 <- rep(NA, nrow(newData2)); names(pred2) <- rownames(newData2)
  model_terms2 <- attributes(glm.obj2$terms)$variables
```

```
  predictors2 <- as.character(model_terms2[3:length(model_terms2)])
  response2 <-  as.character(model_terms2[2])
  newData2 <- newData2[,c("finwgt",response2,predictors2)]
  xnn2 <- na.omit(newData2)
  pred2[-attr(xnn2, "na.action")] <- predict(glm.obj2, xnn2)
  guess2 <- 1/(1+exp(-pred2))
  dframe2 <- data.frame(response2=ifelse(newData2[, response2]==0, -1, 1),
                       guess2=guess2,
                       WEIGHT2=newData2$finwgt)
  dframe2 <- na.omit(dframe2)
  tp.fp2 <- WeightedROC::WeightedROC(dframe2$guess2,
dframe2$E_cigarettes_use_category, weight=dframe2$WEIGHT2)

ggplot()+ geom_path(aes(FPR, TPR), data=tp.fp2)+
coord_equal()+theme_classic() + ggtitle("Area Under Curve for Current E-cig
Use") + geom_abline(intercept = 0, slope = 1)+
     annotate("text", x = .75, y = .25, label = paste("AUC
=",round(WeightedAUC(tp.fp2),3)))
}
```

```{r}
png(file="C:/Users/zhulu/Desktop/BIOST 2099/AUC2.png",
width=600, height=350)
plotAUC2(logit2b, nyts5)
dev.off()
```

Bibliography

Bureau, U. C. (n.d.). *School Enrollment in the United States: October 2020 - Detailed Tables*. Census.Gov. Retrieved November 29, 2022, from https://www.census.gov/data/tables/2020/demo/school-enrollment/2020-cps.html

Cassy, S. R., Natário, I., & Martins, M. R. (2016). Logistic Regression Modelling for Complex Survey Data with an Application for Bed Net Use in Mozambique. *Open Journal of Statistics*, *6*(5), Article 5. https://doi.org/10.4236/ojs.2016.65074

CDCTobaccoFree. (2022, November 9). *Youth and Tobacco Use*. Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/youth_data/tobacco_use/index.htm

Chinwong, D., Mookmanee, N., Chongpornchai, J., & Chinwong, S. (2018). A Comparison of Gender Differences in Smoking Behaviors, Intention to Quit, and Nicotine Dependence among Thai University Students. *Journal of Addiction*, *2018*, e8081670. https://doi.org/10.1155/2018/8081670

Gaiha, S. M., Lempert, L. K., & Halpern-Felsher, B. (2020). Underage Youth and Young Adult e-Cigarette Use and Access Before and During the Coronavirus Disease 2019 Pandemic. *JAMA Network Open*, *3*(12), e2027572. https://doi.org/10.1001/jamanetworkopen.2020.27572

Gentzke, A. S. (2022). Tobacco Product Use and Associated Factors Among Middle and High School Students—National Youth Tobacco Survey, United States, 2021. *MMWR. Surveillance Summaries*, *71*. https://doi.org/10.15585/mmwr.ss7105a1

Jankowski, M., Krzystanek, M., Zejda, J. E., Majek, P., Lubanski, J., Lawson, J. A., & Brozek, G. (2019). E-Cigarettes are More Addictive than Traditional Cigarettes—A Study in Highly Educated Young People. *International Journal of Environmental Research and Public Health*, *16*(13), 2279. https://doi.org/10.3390/ijerph16132279

Methodology report of the 2021 NATIONAL YOUTH TOBACCO SURVEY. (n.d.). 41.

O'Brien, D., Long, J., Quigley, J., Lee, C., McCarthy, A., & Kavanagh, P. (2021). Association between electronic cigarette use and tobacco cigarette smoking initiation in adolescents: A systematic review and meta-analysis. *BMC Public Health*, *21*(1), 954. https://doi.org/10.1186/s12889-021-10935-1

Pandya, A., & Lodha, P. (2021). Social Connectedness, Excessive Screen Time During COVID-19 and Mental Health: A Review of Current Evidence. *Frontiers in Human Dynamics*, *3*. https://www.frontiersin.org/articles/10.3389/fhumd.2021.684137

Pike, J. R., Tan, N., Miller, S., Cappelli, C., Xie, B., & Stacy, A. W. (2019). The Effect of E-cigarette Commercials on Youth Smoking: A Prospective Study. *American Journal of Health Behavior*, *43*(6), 1103–1118. https://doi.org/10.5993/AJHB.43.6.8

Lumley, T., *Survey.pdf*. 2022. Retrieved December 13, 2022, from https://cran.r-project.org/web/packages/survey/survey.pdf

Hocking, D. T., *WeightedROC.pdf*. 2022. Retrieved December 16, 2022, from https://cran.r-project.org/web/packages/WeightedROC/WeightedROC.pdf

Wulan, W. R., Kusuma, D., Nurjanah, N., Aprianti, A., & Ahsan, A. (2022). Is Exposure to Social Media Advertising and Promotion Associated with E-cigarette Use? Evidence from Indonesia. *Asian Pacific Journal of Cancer Prevention : APJCP*, *23*(4), 1257–1262. https://doi.org/10.31557/APJCP.2022.23.4.1257

Tatum A. Jolink, Nicholas J. Fendinger, Gabriella M. Alvarez, Mallory J. Feldman, Monica M. Gaudier-Diaz, Keely A. (2022). Muscatell, Inflammatory reactivity to the influenza vaccine is associated with changes in automatic social behavior, Brain, Behavior, and Immunity, 10.1016/j.bbi.2021.10.019, 99, (339-349)