Analysis of the Behavior of Selected Sample Statistics on the Youth Risk Behavior Surveillance System Population

Sam Oliszewski

Oregon State University

ABSTRACT

This paper discusses the analysis of some sample statistics using 2003 and 2013 data from a large survey of high-school students in America, called the Youth Risk Behavior Surveillance System (YRBSS). This analysis includes the inference on the population of high-school students based on the sample of YRBSS students. First, there is a determination as to whether high-school students have increased their BMI over time. Second, there is a determination as to whether male high-school students are more likely to smoke than female high-school students. Finally, there is an estimate as to how much TV the average high-schooler watches per day. Ultimately, the information gathered from the YRBSS data can help make useful inferences about the population of high-school students. The research defined in this paper provide the context for such inferences. Ultimately, it was observed that students have increased their BMI over time and that male students are more likely to smoke than females. It is also known that students on average are watching approximately two hours of TV per day. Understanding this data allows law-makers and educators to guide student behavior in a way that could improve student health. For future research, it would be useful to model the responses from the YRBSS data to quantify the risk each student has based on their responses. This would require further questioning about the overall health and well-being of the students to be done. Building a predictive model based on risk factors would help reveal the risk that certain behaviors have in the context of real health outcomes.

Keywords: Youth Risk Behavior Surveillance System (YRBSS), two-sample t-test, two-sample proportion test

ANALYSIS OF THE BEHAVIOR OF SELECTED SAMPLE STATISTICS ON THE YOUTH RISK BEHAVIOR SURVEILLANCE SYSTEM POPULATION

This paper provides an analysis of some sample statistics using 2003 and 2013 data from a large survey of high-school students in America, called the Youth Risk Behavior Surveillance System (YRBSS), and includes the inference on the population of high-school students based on the sample of YRBSS students. There are three explanatory questions of interest addressed in this analysis. First, there is a determination as to whether high-school students are observing an increase in their BMI over time. Second, there is a determination as to whether male high-school students are more likely to smoke than female high-school students. Finally, there is an estimate as to how much TV the average high-schooler watches per day. Ultimately, the information gathered from the YRBSS data can help make useful inferences about the population of high-school students. The research defined in this paper provide the context for such inferences. The paper will begin by discussing the exploratory analysis that was performed to provide insight on the data being studied. Next, the approaches for statistical test determination are described. Finally, conclusions drawn from the analysis are reported, and caveats of the analysis are offered to aid in the overall interpretation of the study results.

Exploratory Analysis

Original Variables. The raw data for this analysis came from two separate datasets containing the YRBSS responses for the years 2003 and 2013, respectively. Each dataset was of the same structure. A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices A and B.

Reducing the Number of Variables. For this analysis, not all of the variables in the original data are relevant to the questions of interest being addressed. Therefore, the data was reduced to only contain the variables required for the analysis. This included: year, BMI, sex, q33, and q81. A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices C and D.

Visualizing the Data

Figures for Each Variable by Year. For this analysis, it is beneficial to examine plots for each variable in the data by each year the survey was performed. These plots are shown below:

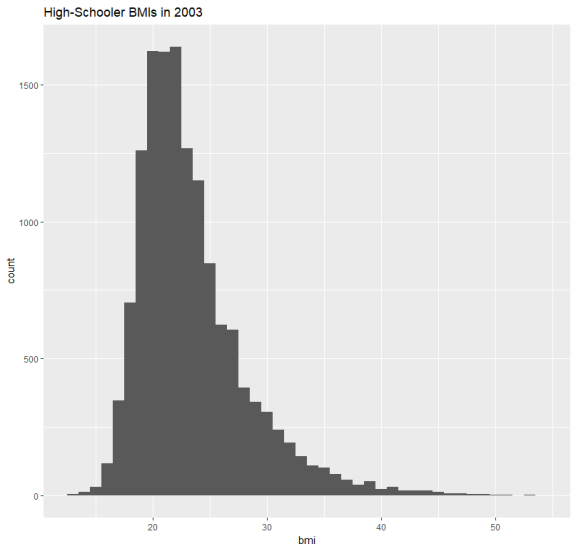

Figure 1. High School BMI in 2003

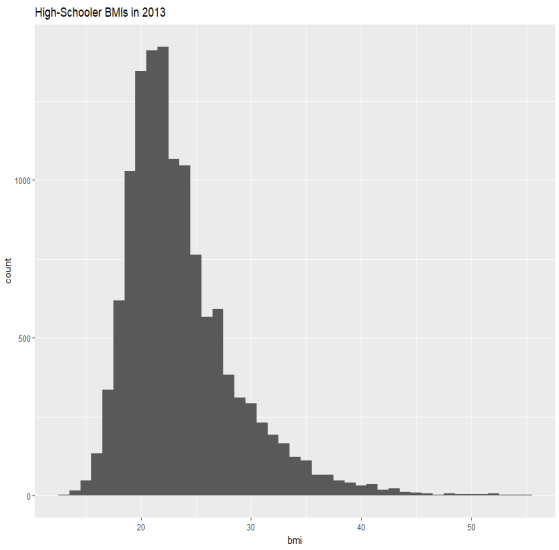

Figure 2. High School BMI in 2013

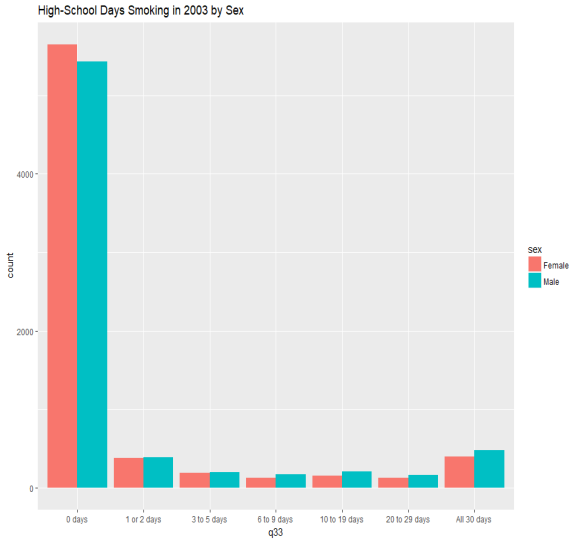

Figure 3. High School Days Smoking by Sex in 2003

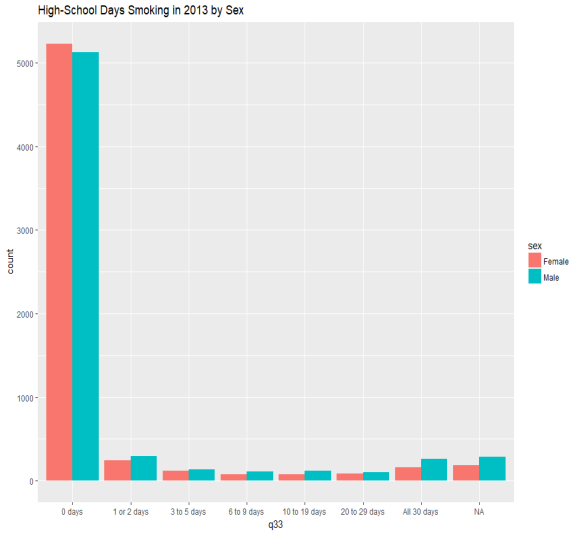

Figure 4. High School Days Smoking by Sex in 2013

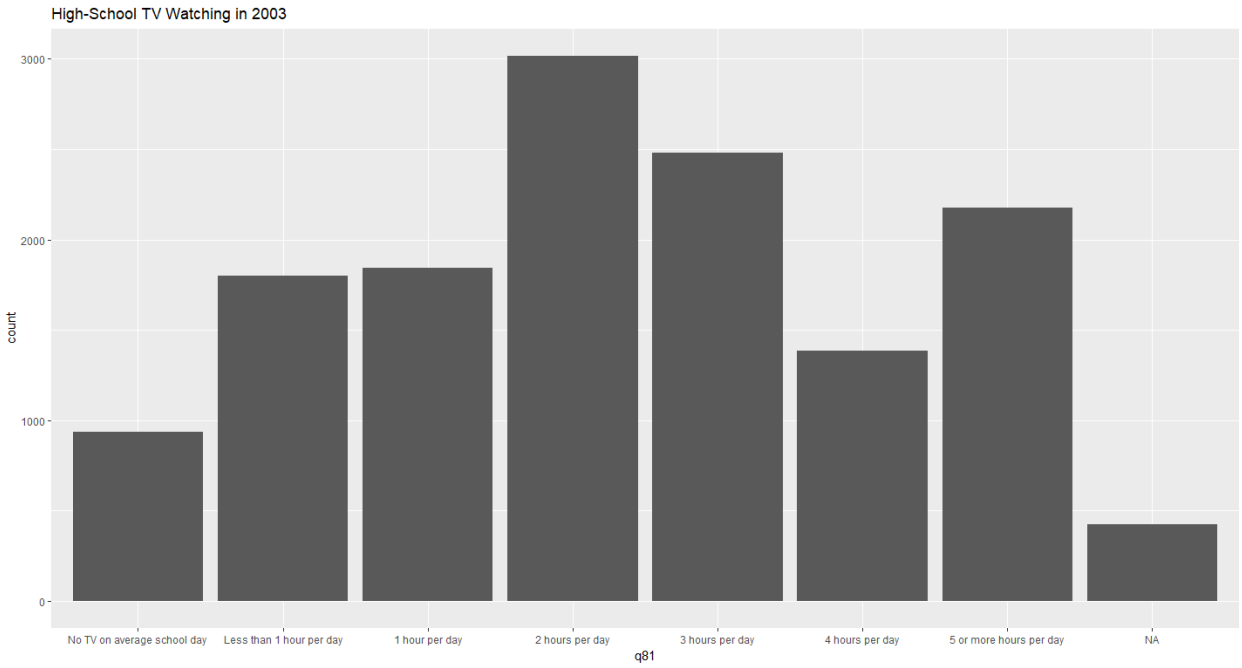

Figure 5. High School TV Watching in 2003

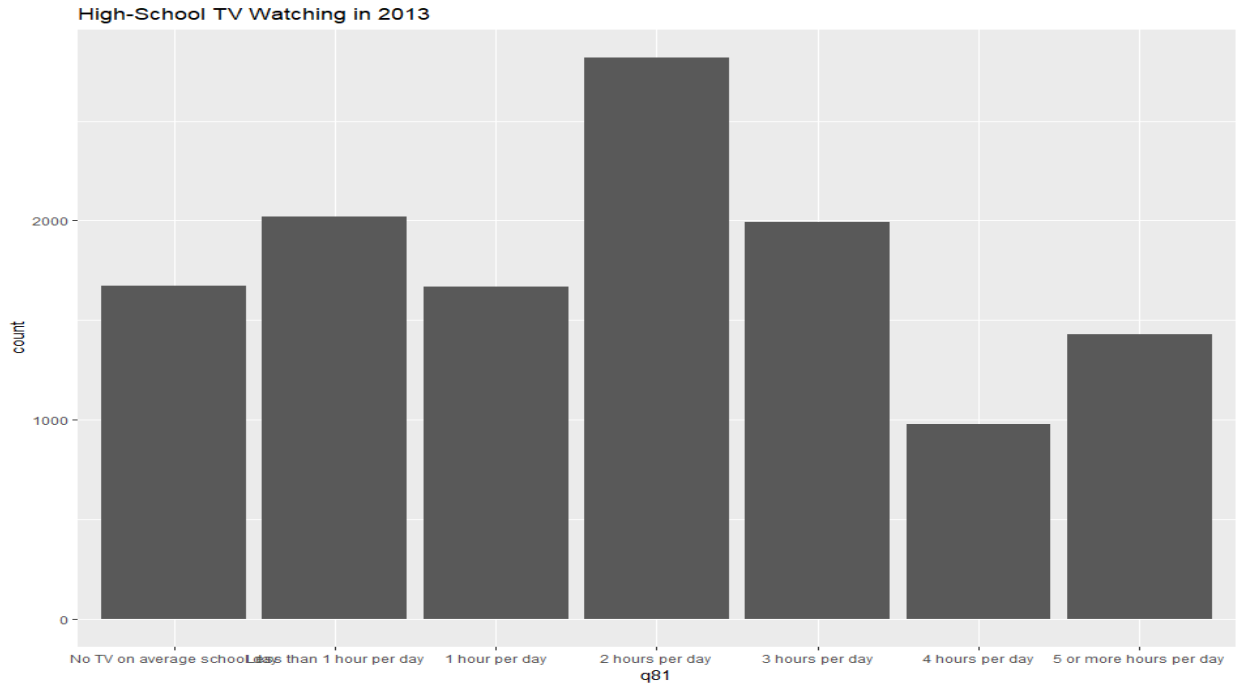

Figure 6. High School TV Watching in 2013

Based on these plots, there are small differences between 2003 and 2013 in terms of BMI; there appears to be less smoking overall between 2003 and 2013 and it is more common for male high schoolers than female; and the number of hours of TV watched per day appears to be slightly less on average, while the most common amount of time is 2 hours per day.

Handling Missing Values. Since there are a handful of missing values in the data, and the number of missing values is relatively small (<1%), median imputation will be performed on the

data. This will provide a few more observations to be available in the data. Median imputation will be performed on the variables q33 and q81 as they are the only variables with observed missing values in the reduced data.

Explanatory Problems

How Has the BMI of High-Schoolers Changed Between 2003 and 2013?

Methods. To study the BMI of high-school students between 2003 and 2013, a Welch's two-sample t-test was run with a null hypothesis that the mean BMI of students in 2003 is equal to the mean BMI of students in 2013 and an alternative hypothesis that the mean BMI of students in 2013 is greater than the mean BMI of students in 2003. The variances are assumed to be unequal between the populations since there is no clear evidence that they should be assumed equal. The data from 2003 and 2013 are also assumed to be independent because they are not tracking the same students. A two-sample t-test is an appropriate choice for this study because the parameter in question is the mean. This test can be applied because the sample size is large validating the use of the approximately normal test statistic. Further, a t-distribution is a reasonable comparison to use, since the sample is approximately normal.

Results. The two-sample t-test yielded a p-value of 8.76e-05, suggesting that there is significant evidence to reject the null hypothesis that the mean BMI of students in 2003 is equal to the mean BMI of students in 2013. It is estimated that the mean BMI of students in 2013 is 23.64 and the mean BMI of students in 2003 is 23.41. With 95% confidence, the mean BMI of students in 2013 is at least 0.13 units larger than the mean BMI of students in 2003. Additionally, the median BMIs in 2003 and 2013 are observed to be 22.29 and 22.49.

Figures. The following figure depicts the increase in high-value outliers in the 2013 BMI data that caused the increase in mean BMI from 2003 to 2013:

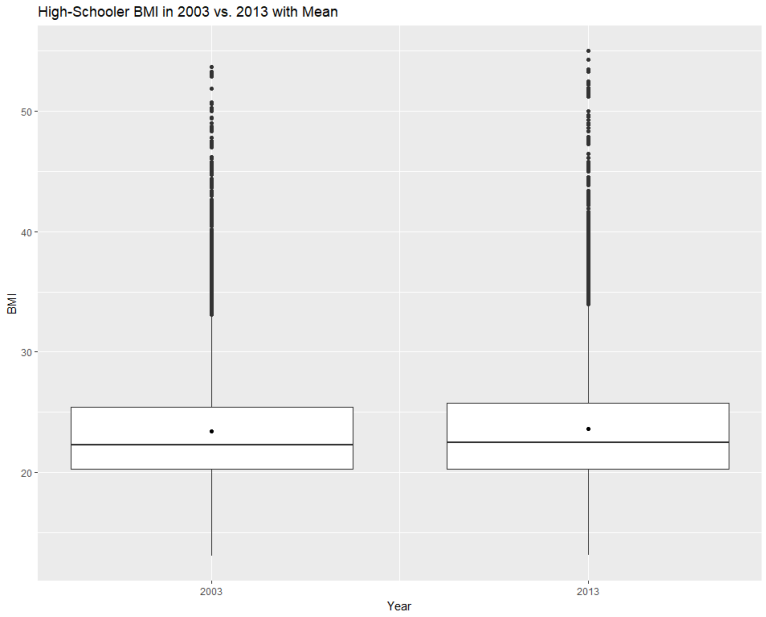

Figure 7. Boxplots of BMI Data From 2003 and 2013 with Mean Labeled.

Conclusion. The result of this test showed that high-school students in 2013 have higher BMIs than the students from 2003 (Welch's two sample t-test, t = 3.75, df = 25988, p-value = 8.76e-05). This suggests that the BMI of high-schoolers is increasing over time. Further, the median BMIs of 22.29 and 22.49 for 2003 and 2013, respectively, indicate that the BMI of high schoolers is increasing over time and that this result is not simply the result of influential outliers.

Are Male High-Schoolers More Likely to Smoke than Female High-Schoolers?

Methods. To study the likelihood that male high-schoolers are more likely to smoke than female high-schoolers, a two-sample proportion test was run with a null hypothesis that the proportion of male high-school smokers was equal to the proportion of female high-school smokers and an alternative hypothesis that the proportion of male high-school smokers is greater than the proportion of female high-school smokers. A two-sample proportion test is appropriate for this study because proportions are an appropriate tool for comparing binary data. In the case of determining whether students are smokers, the answer would be "yes" or "no" for the purpose of this question. Therefore, a proportion of "yes" responses would indicate how many students are smokers. Then, the comparison of proportions between the two sexes would indicate whether one is more likely to smoke than the other. The two-sample proportion test can be used because the following conditions are met: 1. The samples of female and male students are sufficiently large, and 2. The samples of female and male students are independent. To define a "yes" response for whether the student is a smoker, the student must have indicated for q33 that they have smoked greater than 0 days in the last 30 days, otherwise it will be assumed as a "no" response.

Results. The two-sample proportion test yielded a p-value of 1.02e-07, suggesting there is significant evidence to reject the null hypothesis that the proportion of male high-school smokers is equal to the proportion of female high-school smokers. It is estimated that the proportion of male high-school smokers is 10.99% and the proportion of female high-school smokers is 8.25%. With 95% confidence, the proportion of male high-school smokers is between 2% and 100% larger than the proportion of female high-school smokers.

Figures. The following figure depicts that male high schoolers generally smoke more than females:

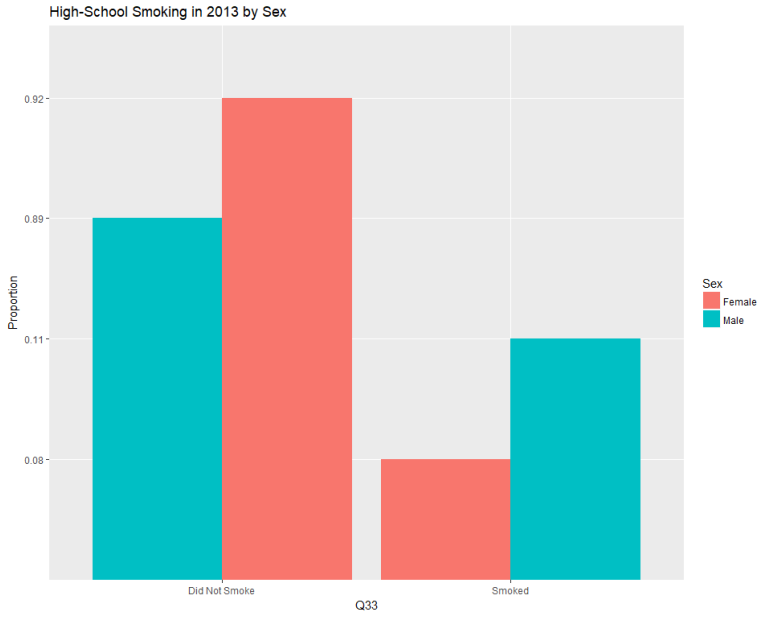

Figure 8. High School Smoking by Sex in 2013

Data. The following table displays the number of smokers versus total students and the calculated proportion of smokers by sex:

Sex	Number of Smokers	Number of Students	Proportion of Student Smokers
Male	705	6414	**10.99%**
Female	509	6166	8.25%

Table 1. High-School Student Smoker Proportions.

Summary. The result of this test showed that male high-schoolers are more likely to smoke than female high-schoolers (two sample proportion test, x-squared = 27.00, df =1, p-value = 1.02e-07).

How Much TV Do High-Schoolers Watch?

Methods. To study how much TV high-schoolers watch, the median TV time value reported by high-schoolers in 2013 was calculated. This value is assumed to reflect the population of high-schoolers in the present because it is the most recent data available. Since the responses on the YRBSS survey are categorical, the answer to this analysis will also be reported as categorical. The median is an appropriate measure of the average amount of time high-schoolers watch TV because the data recorded was in ordinal categories and therefore the median is an appropriate measure of the center of the data because other statistics, such as the mean, wouldn't apply to this type of data.

Results. The median value reported by students in 2013 was "2 hours per day".

Figures. The following figure depicts the responses by students in 2013 about how much TV they watch on average.

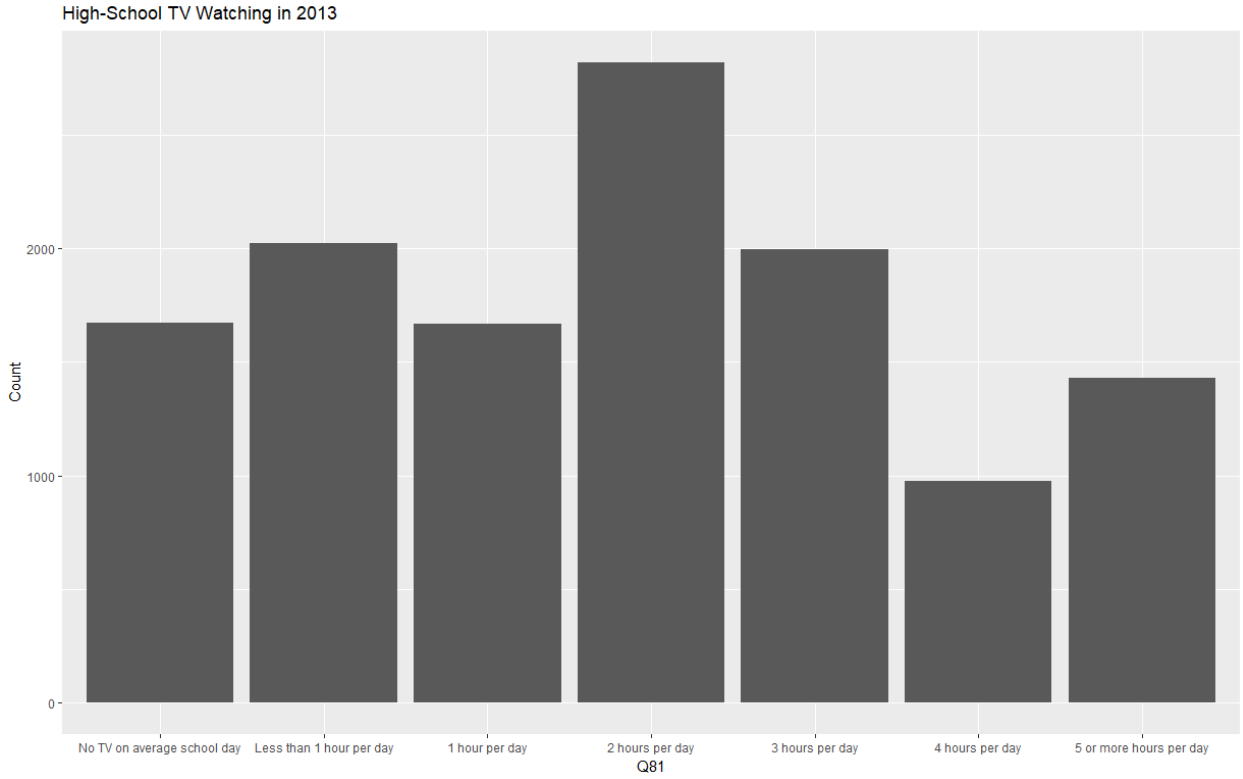

Figure 9. High School TV Watching in 2013

Data. This table displays the student responses to the amount of TV they watch per day and the frequency of each response:

TV Watching Time	Frequency
No TV on average school day	1,671
Less than 1 hour per day	2,021
1 hour per day	1,667
2 hours per day	**2,548**
3 hours per day	1,995
4 hours per day	977
5 or more hours per day	1,430

Table 2. Frequency of TV Watching Responses.

Summary. The amount of TV that high-schoolers watch is estimated to be two hours per day.

Conclusion

Based on the analysis of the YRBSS data, inferences about the population of high-school students can be drawn. It was observed that students have increased their BMI over time and that male students are more likely to smoke than females. It is also known that students on average are

watching approximately two hours of TV per day. Understanding this data allows law-makers and educators to guide student behavior in a way that could improve student health. For future research, it would be useful to model the responses from the YRBSS data to quantify the risk each student has based on their responses. This would require further questioning about the overall health and well-being of the students to be done. Building a predictive model based on risk factors would help reveal the risk that certain behaviors have in the context of real health outcomes.

APPENDIX A

Sample Data: YRBSS 2003

Year	BMI	Age	Sex	Grade	Race	Q9	Q33	Q77	Q80	Q81
2003	21.8	12 years old or younger	Male	11th	All other races	Rarely	0 days	NA	NA	5 or more hours per day
2003	21.5	12 years old or younger	Male	9th	All other races	Never	All 30 days	NA	NA	5 or more hours per day
2003	21.4	13 years old	Male	10th	Black or African American	Always	0 days	NA	NA	5 or more hours per day
2003	18.9	13 years old	Male	9th	Hispanic/Latino	Always	0 days	NA	NA	2 hours per day
2003	18.0	13 years old	Male	9th	White	Rarely	0 days	NA	NA	5 or more hours per day
2003	18.1	13 years old	Male	9th	White	Always	0 days	NA	NA	1 hour per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Age: Age of the student

Possible Values: Text

- 12 years old or younger
- 13 years old
- 14 years old
- 15 years old
- 16 years old
- 17 years old
- 18 years old or older

Sex: Sex of the student

Possible Values: Text

- Female
- Male

APPENDIX A CONTINUED

Grade: School grade of the student

Possible Values: Text

- 9th grade
- 10th grade
- 11th grade
- 12th grade
- Ungraded or other grade

Race: Race of the student

Possible Values: Text

- White
- Black or African American
- Hispanic/Latino
- All Other Races

Q9: Response to the question "How often do you wear a seat belt when riding in a car driven by someone else?"

Possible Values: Text

- Never
- Rarely
- Sometimes
- Most of the Time
- Always

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

Q77: Response to the question "During the past 7 days, how many times did you drink a can, bottle, or glass of soda or pop, such as Coke, Pepsi, or Sprite? (Do not count diet soda or diet pop.)

Possible Values: Text

- I did not drink soda or pop during the past 7 days
- 1 to 3 times during the past 7 days
- 4 to 6 times during the past 7 days
- 1 time per day
- 2 times per day
- 3 times per day
- 4 or more times per day

APPENDIX A CONTINUED

Q80: Response to the question "During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spent in any kind of physical activity that increased your heart rate and made you breathe hard some of the time.)

Possible Values: Text

- 0 days
- 1 day
- 2 days
- 3 days
- 4 days
- 5 days
- 6 days
- 7 days
-

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day

APPENDIX B

Sample Data: YRBSS 2013

Year	BMI	Age	Sex	Grade	Race	Q9	Q33	Q77	Q80	Q81
2013	22.0	12 years old or younger	Male	9th	Black or African American	Never	NA	4 or more times per day	7 days	No TV on an average school day
2013	21.5	12 years old or younger	Male	10th	Black or African American	Sometimes	0 days	Did not drink soda or pop	7 days	3 hours per day
2013	19.0	12 years old or younger	Male	12th	All Other Races	Always	NA	4 or more times per day	2 days	5 or more hours per day
2013	21.9	12 years old or younger	Male	12th	Black or African American	Always	0 days	Did not drink soda or pop	7 days	2 hours per day
2013	17.6	13 years old	Male	9th	White	Always	0 days	1 to 3 times	7 days	No TV on an average school day
2013	28.9	13 years old	Male	9th	Black or African American	Always	0 days	1 to 3 times	6 days	3 hours per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Age: Age of the student

Possible Values: Text

- 12 years old or younger
- 13 years old
- 14 years old
- 15 years old
- 16 years old
- 17 years old
- 18 years old or older

APPENDIX B CONTINUED

Sex: Sex of the student

Possible Values: Text

- Female
- Male

Grade: School grade of the student

Possible Values: Text

- 9th grade
- 10th grade
- 11th grade
- 12th grade
- Ungraded or other grade

Race: Race of the student

Possible Values: Text

- White
- Black or African American
- Hispanic/Latino
- All Other Races

Q9: Response to the question "How often do you wear a seat belt when riding in a car driven by someone else?"

Possible Values: Text

- Never
- Rarely
- Sometimes
- Most of the Time
- Always

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

APPENDIX B CONTINUED

Q77: Response to the question "During the past 7 days, how many times did you drink a can, bottle, or glass of soda or pop, such as Coke, Pepsi, or Sprite? (Do not count diet soda or diet pop.)

Possible Values: Text

- I did not drink soda or pop during the past 7 days
- 1 to 3 times during the past 7 days
- 4 to 6 times during the past 7 days
- 1 time per day
- 2 times per day
- 3 times per day
- 4 or more times per day

Q80: Response to the question "During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spent in any kind of physical activity that increased your heart rate and made you breathe hard some of the time.)

Possible Values: Text

- 0 days
- 1 day
- 2 days
- 3 days
- 4 days
- 5 days
- 6 days
- 7 days

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day

APPENDIX C

Sample Data: YRBSS 2003

Year	BMI	Sex	Q33	Q81
2003	21.8	Male	0 days	5 or more hours per day
2003	21.5	Male	All 30 days	5 or more hours per day
2003	21.4	Male	0 days	5 or more hours per day
2003	18.9	Male	0 days	2 hours per day
2003	18.0	Male	0 days	5 or more hours per day
2003	18.1	Male	0 days	1 hour per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Sex: Sex of the student

Possible Values: Text

- Female
- Male

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day

APPENDIX D

Sample Data: YRBSS 2013

Year	BMI	Sex	Q33	Q81
2013	22.0	Male	NA	No TV on an average school day
2013	21.5	Male	0 days	3 hours per day
2013	19.0	Male	NA	5 or more hours per day
2013	21.9	Male	0 days	2 hours per day
2013	17.6	Male	0 days	No TV on an average school day
2013	28.9	Male	0 days	3 hours per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Sex: Sex of the student

Possible Values: Text

- Female
- Male

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day