Modeling Past Diagnosis of Sleep Disorder

David Noonan

More than 50 million Americans suffer from a sleep disorder (Institute of Medicine), with societal impacts that range from decreased prodctivity in the workplace to fatal car accidents on the road. Especially in an ever more technologically dependent economy, where cognitive tasks are more common place, the loss of executive function (the capacity to plan and organize complex tasks) caused by sleep disorders (Naegele) will have an ever increasing drag on the entire economy. An understanding of the causes of sleep disorders will be an important stone in the path the future will lead us down. We start here with a simple analysis of factors that are associated with having been diagnosed with a sleep disorder, ones who's nature can be further illuminated with future research.

Dataset

The dataset we use comes from the National Health and Nutrition Examination Survey (NHANES) program from the Center for Disease Control (citation). The NHANES program obtains a nationally representative sample from people in United States, which includes observations from questionnaires, laboratory exams, medical exams, and demographics. The program collects a set of these observations for two-year time-periods. Here we use the latest completed time-period for this study: 2013-2014. Our outcome variable comes from the questionnaire data, where subjects were asked the question: "Were you ever told by a doctor that you have a sleep disorder?" The answer is binary: "yes" or "no."

Variable

Before building a model, we need to select a subset of variables from the NHANES dataset. We begin by setting a limit for the number of parameters we allow in our model. Guidelines suggest there should be more than 10 observations of each outcome for every parameter allowed in the model (Agresti p. 138). The least common outcome in our data is the answer "Yes," which has 566 observations. Our limit is then 56 variables to in our model. This presents a challenge, because there are more than 1000 variables to select from in NHANES, which potentially involves too many parameters to justify.

We limit our variable pool to two categories: questionnaire, and demographic observations. We remove laboratory and medical examination data from consideration because they mostly indicate status in the present, whereas our income variable indicates observations in the past (diagnosis of sleep disorder in the past).

We further restrict the variable pool by excluding "follow-up" questions. These questions are only asked if a subject answered a previous question in a certain way. For example, if a person answers "yes" to a question about whether they get regular physically exercise or not, they are then asked whether this is from bicycling or walking. We remove these questions because they are asked only to a fraction of the sample, and this would restrict our study to representing only those who answered a particular way on the "original" question.

Finally, we "hand-pick" variables based on our own arbitrary criteria based on whether we suspect a variable has an effect on sleep disorders. Our selection is broad, and contains demographic controls, health-condition controls, and variables pertaining to living conditions. In total, we start with 32 variables. We do not consider interaction effects here, because for two variable interactions alone, there are $\binom{32}{2} = 496$ possible interactions, which is beyond our limit of 56.

Data Cleaning

Before model building, we need to format our dataset. Three main changes need to be made. The first comes from the way that questions in NHANES are answered. When a subject refused to answer a question, the observation was recorded in the data as "7," "77," or "777." Similarly, when the answer was "I don't know," the observation was recorded as "9," "99," or "999." We recode these answers as missing data.

A second issue is that many questions are only answered by subjects who are 20 years of age or older. To address this, we exclude all observations from subjects who are less than 20 years of age. This removes many observations recorded as missing, but it changes the interpretation of our model because it reflects a subset of the population, persons who are aged 20 years or older.

The final issue is an assessment of the proportion of missing values for each variable. Guidelines in the NHANES literature suggest that no greater than 10% of the data should be missing to avoid complicated remedial measures. In our dataset, two variables relating to pesticide and herbicide usage at home had slightly more than 10% missing. Here we decide to keep those variables, because they may provide useful information to the model.

Model building

Our model needs to explain a binary outcome: whether someone had been told by a medical professional that they had a sleep disorder. The answer to this is "yes" or "no." Here we decide to use logistic regression to model this outcome. We assemble the model using the Logistic procedure in SAS. We begin by creating a model using all 32 variables. The result looks encouraging: many of the variables are significant, but some are not.

We improve the model by employing the backward selection process. This process removes variables sequentially using the highest p-value as the criteria. The result is 18 variables removed, leaving a simpler model. Our new model has 14 explanatory variables corresponding to 24 parameters including the intercept. For illustration, here is the model as produced by SAS:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1259	0.4748	0.0703	0.7909
RIAGENDR	Female	1	-0.5903	0.1103	28.6565	<.0001
RIDAGEYR		1	-0.00938	0.00430	4.7540	0.0292
RIDRETH3	Mexican American	1	-0.7789	0.1958	15.8197	<.0001
RIDRETH3	Non-Hispanic Asian	1	-1.0190	0.2803	13.2172	0.0003
RIDRETH3	Non-Hispanic Black	1	-0.4364	0.1429	9.3307	0.0023
RIDRETH3	Other Hispanic	1	-0.2987	0.1968	2.3049	0.1290
RIDRETH3	Other Race - Including Multi-Racial	1	0.0303	0.2808	0.0117	0.9140
BPQ020	Yes High Blood Pressure	1	0.2924	0.1244	5.5257	0.0187
BPQ080	Yes High Cholesterol	1	0.3286	0.1201	7.4929	0.0062
HSD010	Excellent General Health	1	-0.8119	0.3177	6.5333	0.0106
HSD010	Fair General Health	1	-0.3125	0.2193	2.0309	0.1541
HSD010	Good General Health	1	-0.6049	0.2175	7.7357	0.0054
HSD010	Very Good General Health	1	-0.9919	0.2424	16.7418	<.0001
DIQ010	Borderline Diabetes	1	0.1614	0.2641	0.3733	0.5412

RIDAGEYR is a continuous variable representing the individual's age.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
DIQ010	Yes Diabetes	1	0.4430	0.1413	9.8279	0.0017
HIQ011	Yes Health Insurance	1	0.3733	0.1617	5.3276	0.0210
HOQ065	Owned or Being Bought	1	-0.7179	0.2814	6.5086	0.0107
HOQ065	Rented	1	-0.6301	0.2863	4.8415	0.0278
MCQ010	Yes Asthma	1	0.5395	0.1276	17.8792	<.0001
MCQ080	Yes Overweight	1	0.8898	0.1163	58.5731	<.0001
MCQ160A	Yes Arthritis	1	0.5950	0.1256	22.4393	<.0001
OCD150	Employed	1	-0.3296	0.1230	7.1849	0.0074
SLD010H		1	-0.1876	0.0358	27.5077	<.0001

SLD010H is a continuous variable representing the current number of hours of sleep the individual gets.

Model Checking

We first check whether our number of parameters is appropriate for the data. The rule of thumb we used in the variable selection section says we must have at least 10 observations of every outcome type for each explanatory variable we have. In our starting model with all predictors, we have 32 explanatory variables total. Our response variable has two outcomes (excluding missing data): diagnosis of sleep disorder (566 total) and no diagnosis of sleep disorder (5191). Our limit for the number of appropriate explanatory variables to include is then:

$$\frac{566}{10} = 56.6 \approx 56$$

Our starting number of variables can then be considered appropriate. We should note that our starting variables do not include any interactions. Adding interactions would exceed the appropriate number of predictors, so we did not include interactions.

After the model selection process, we have 345 observations for diagnosis of sleep disorder, and 3591 of the other type. In this case, we have a maximum limit of parameters:

$$\frac{345}{10} = 34.5 \approx 34$$

We find our model after selection to have an appropriate number of parameters (24).

Our second validation process is to check for lack of fit. This is to test whether the parameters in the model are equal to zero or not. We apply this to the final model after selection. Using the likelihood ratio test, which tests the null hypothesis that the parameters are all zero:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{24} = 0 \qquad H_a: \text{At least one} \beta_i \text{is not zero}$$

Our test statistic is 452.6252, which is chi-squared distributed with degrees of freedom equal to 23. The p-value is <.0001, so we can conclude that at least one of the parameters is not equal to zero.

Our final validation test is whether the model fits. For this we use the HosmerLemeshow test for goodness of fit, which compares fitted values to observed ones. In this test, the null hypothesis is that the model fits, and the alternative is that it doesn't. Our test statistic for the Hosmer-Lemeshow test is 5.5152, which is chi-squared distributed with 8 degrees of freedom, which corresponds to a p-value of 0.7013. We fail to reject the null hypothesis that the model fits, and conclude that the model fits.

Interpretation

Some of the results are intuitive, but others are not. For each factor, we first take a look at the general distribution of those who have been and have not been told they have a sleep disorder. Sometimes the pattern observed corresponds to the disparity between odds for each value of a factor and other times they do not. It is important to keep in mind that it is the odds ratios that fully take into accout the effect of all other variables in the model. As a summary, factors that decrease or increase the odds of having been told they have a sleep disorder are displayed in a table below.

Demographic Factors

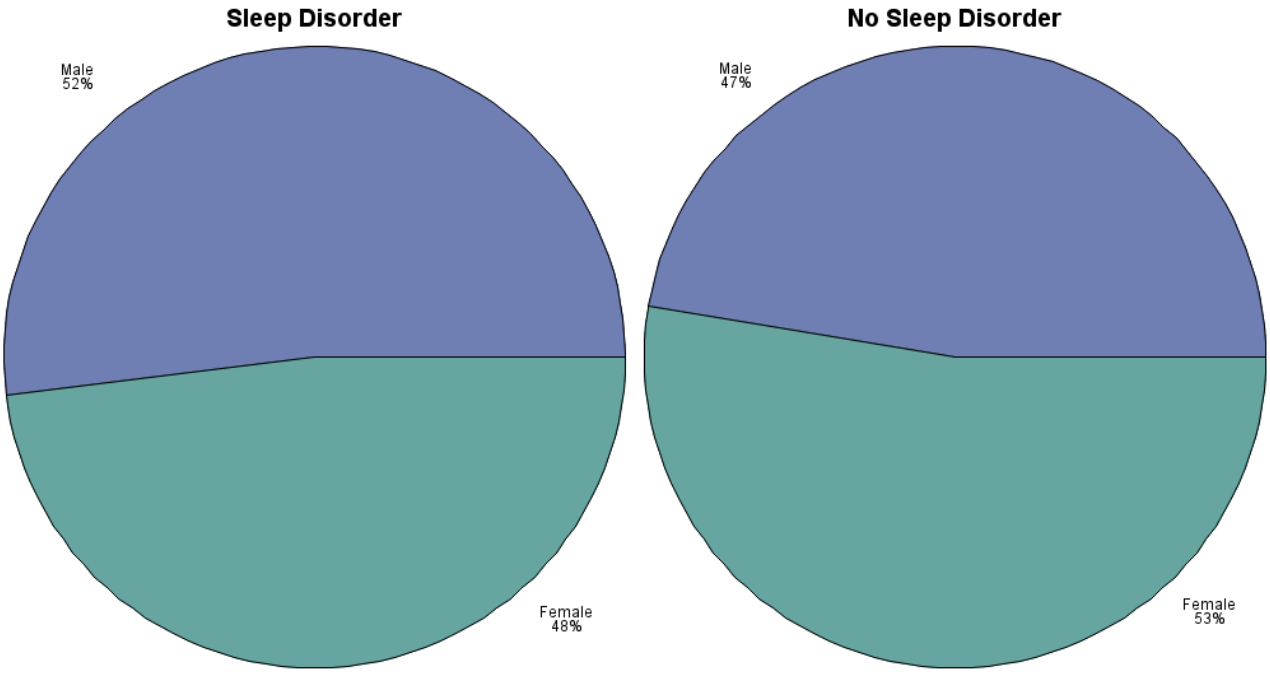
Variable	Comparison	Odds of	95% Confidence Interval
Gender	Female vs Male	Lower by 44.6%	Between 31.2% and 55.4%
Age	One Year Increase	Lower by 0.9%	Between 0.1% and 1.8%
Race	Mexican vs White	Lower by 54.1%	Between 32.6% and 68.7%
	Asian vs White	Lower by 63.9%	Between 37.5% and 79.2%
	Black vs White	Lower by 35.4%	Between 14.5% and 51.1%
	Hispanic vs White	Not significantly different from White.	
	Other vs White	Not significantly different from White.	

Medical Factors

Variable	Comparison	Odds of	95% Confidence Interval
General Health	1 - Excellent vs Poor	Lower by 55.6%	Between 17.2% and 76.2%
	2 - Very Good vs Poor	Lower by 62.9%	Between 40.4% and 76.9%
	3 - Good vs Poor	Lower by 45.4%	Between 16.4% and 64.3%
	4 - Fair vs Poor	Not significantly different from Poor	
Diabetes	Yes vs No	Higher by 55.7%	Between 18.1% and 105.5%
	Borderline vs No	Not significantly different from No Diabetes	
High Blood Pressure	Yes vs No	Higher by 34.0%	Between 5.0% and 70.9%
High Cholesterol	Yes vs No	Higher by 38.9%	Between 9.8% and 75.8%
Asthma	Yes vs No	Higher by 71.5%	Between 33.6% and 120.2%
Overweight	Yes vs No	Higher by 143.5%	Between 93.9% and 205.8%
Arthritis	Yes vs No	Higher by 81.3%	Between 41.7% and 131.9%

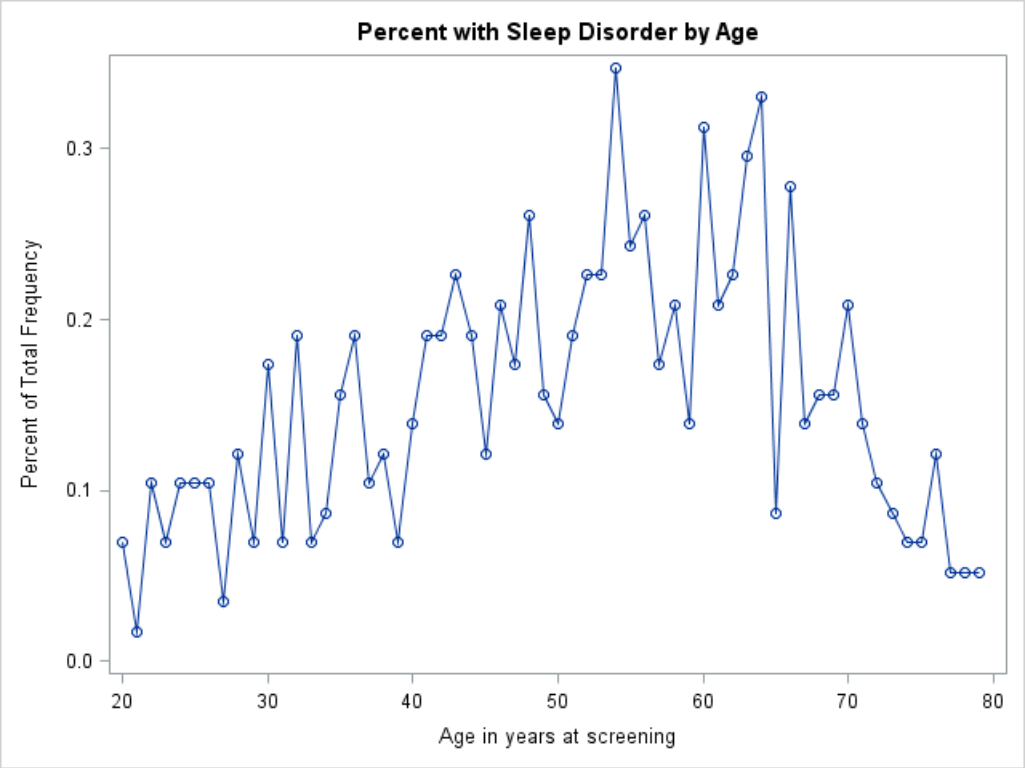
Other General Factors

Variable	Comparison	Odds of	95% Confidence Interval
Employment	Employed vs Unemployed	Lower by 28.1%	Between 8.5% and 43.5%
Hours of Sleep	One Hour Increase	Lower by 17.1%	Between 11.1% and 22.7%
Living Arrangement	Owned/Buying vs Other	Lower by 51.2%	Between 15.3% and 71.9%
	Renting vs Other	Lower by 46.7%	Between 6.7% and 69.6%
Health Insurance	Yes vs No	Higher by 45.3%	Between 5.8% and 99.4%



Overall, Males are more likely to have a sleep disorder than Females. This holds even when accounting for all other variables in the model. The model shows that, holding all other variables constant, Females have between 31.2% and 55.4% lower odds of having been told by a doctor that they have a sleep disorder compared to Men.

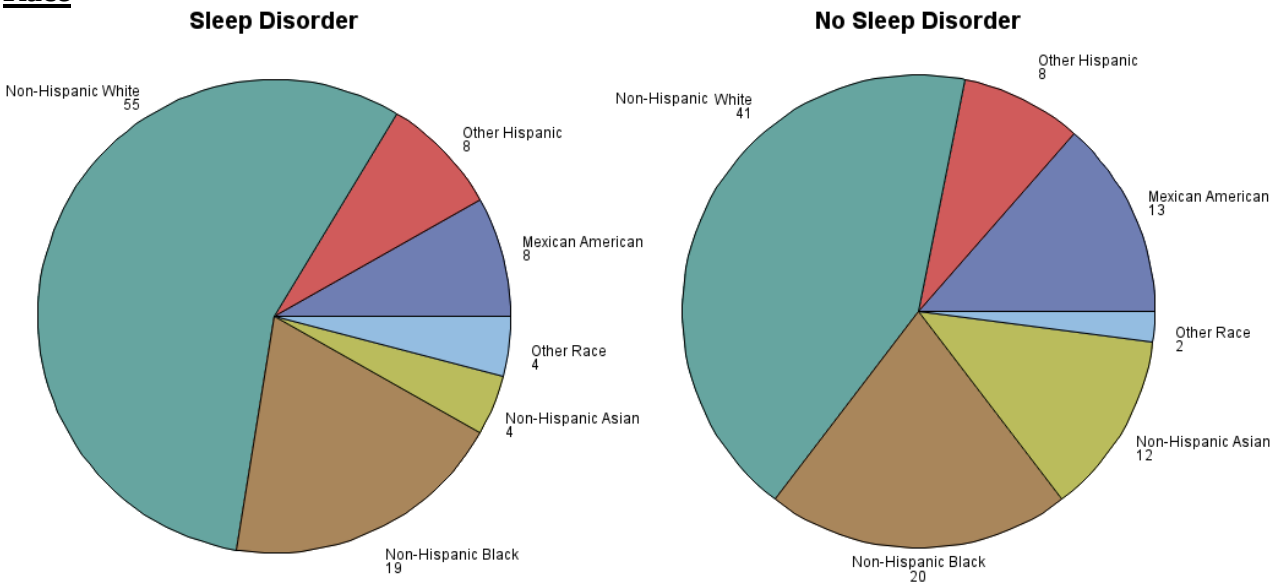
Age



Overall, the chances of having a sleep disorder increase as people get older and seem to peak somewhere around 60 years old, then decrease after that. However, accounting for all other variables in the model gives a very different result. The model shows that, holding all other variables constant, for each increase in age by 1 year, the odds of having been diagnosed with a sleep disorder *decrease* by between 0.1% and 1.8%.

Now, considering that the question asks if a person has *ever* been told they have a sleep disorder by a doctor, we have to keep in mind that a diagnosis 20 years ago is counted the same in the data as a diagnosis last year. There are two possible causes that come to mind to explain the data; age itself may be related to any *current* diagnoses of a sleep disorder, and the overall diagnosis rate may itself be *changing over time*. This makes any interpretation of the decrease in odds, accounting for other variables, very difficult. Age should be thought of as simply a control variable in our model.

Race



Overall, Non-hispanic Whites and Other Races (including Multi-Racial) seem to have an increased chance of having a sleep disorder, while Mexican Americans, Asians, and Blacks seem to have a decreased chance of having a sleep disorder. This mostly holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant,

(1) Mexican American's have between 32.6% and 68% lower odds of having been told by a doctor that they have a sleep disorder compared to Whites.

(2) Asians have between 37.5% and 79.2% lower odds of having been told by a doctor that they have a sleep disorder compared to Whites.

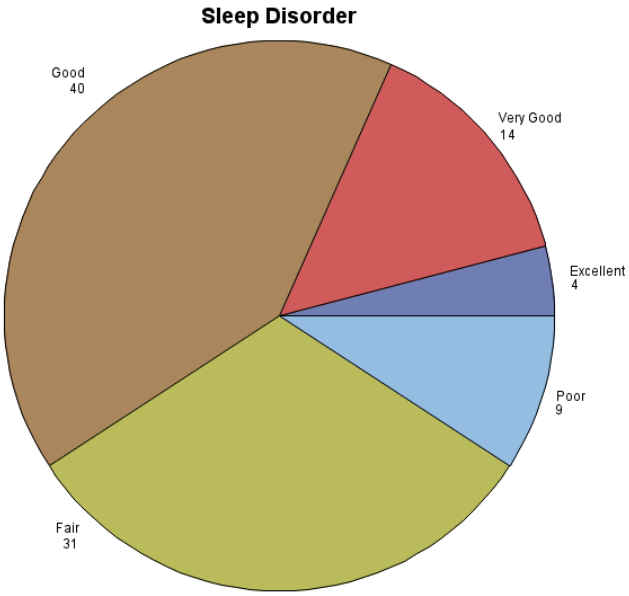
(3) Blacks have between 14.5% and 51.1% lower odds of having been told by a doctor that they have a sleep disorder compare to Whites.

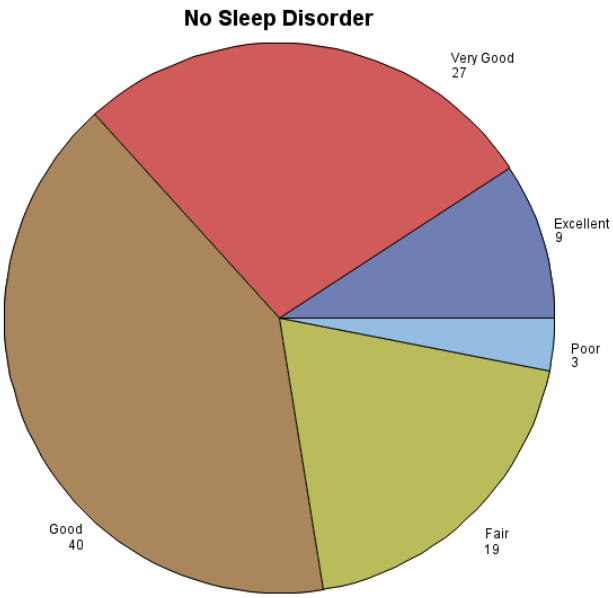
(4) Non-Mexican Hispanics and Other Races (including Multi-Racial) do not have significantly different odds of having been told by a doctor that they have a sleep disorder compared to Whites.

Why these differences in odds exist isn't clear. This can't be ascertained simply from this retrospective study using a questionaire. Further studies might be able to illuminate why these differences exist, but for this model, Race can be thought of a simply a control variable.

Now we will consider some health factors that were found to be associated with having been told by a doctor that a person has a sleeping disorder. This includes a self-reported level of general health, along with the association with diabetes, high blood pressure, high cholesterol, asthma, being overweight, and arthritis.

General Health





Overall, more people *without* a sleep disorder self-rate themselves as having either Excellent or Very Good general health. The percentages are the same for those that self-rate themselves as having Good general health. Also, more people *with* a sleep disorder self-rate themselves as having either Fair or Poor general health. This pattern generally holds even when accounting for all other variables in the model. The model shows that, holding all other variables constant,

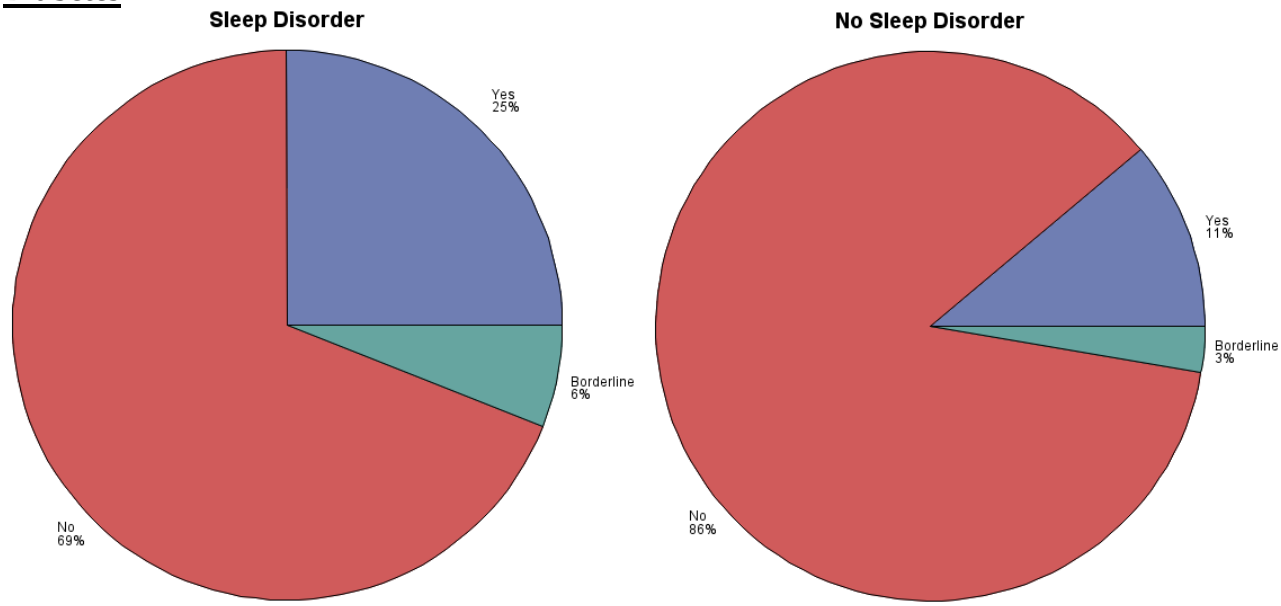
(1) People that self-rate themselves as having Excellent general health have between 17.2% and 76.2% lower odds of having been told by a doctor that they have a sleep disorder compared to people that self-rate themselves as having Poor general health.

(2) People that self-rate themselves as having Very Good general health have between 40.4% and 76.9% lower odds of having been told by a doctor that they have a sleep disorder compared to people that self-rate themselves as having Poor general health.

(3) People that self-rate themselves as having Good general health have between 16.4% and 64.3% lower odds of having been told by a doctor that they have a sleep disorder compared to people that self-rate themselves as having Poor general health.

(4) People that self-rate themselves as having Fair general health do not have significantly different odds of having been told by a doctor that they have a sleep disorder compared to people that self-rate themselves as having Poor general health.

Diabetes

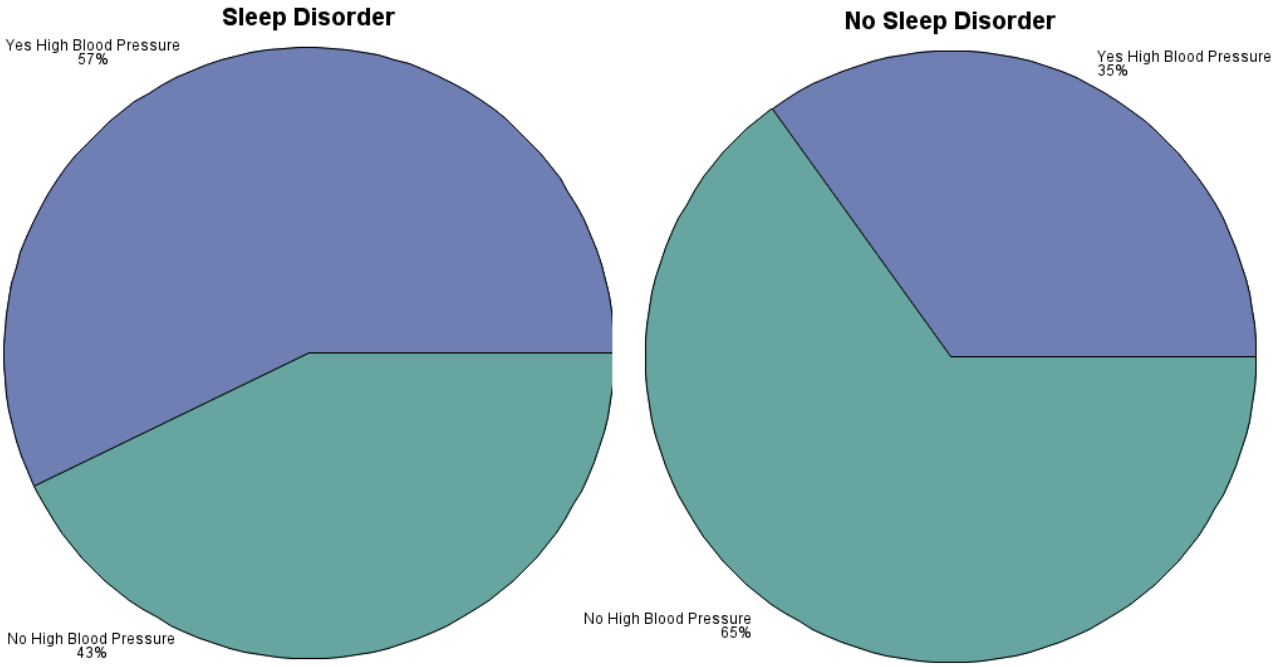


Overall, more people with a sleep disorder have also been told by a doctor that they have Diabetes compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant,

(1) Those who have been told by a doctor that they have diabetes have between 18.1% and 105.5% higher odds of also having been told they have a sleep disorder compared to those who have not been told they have diabetes.

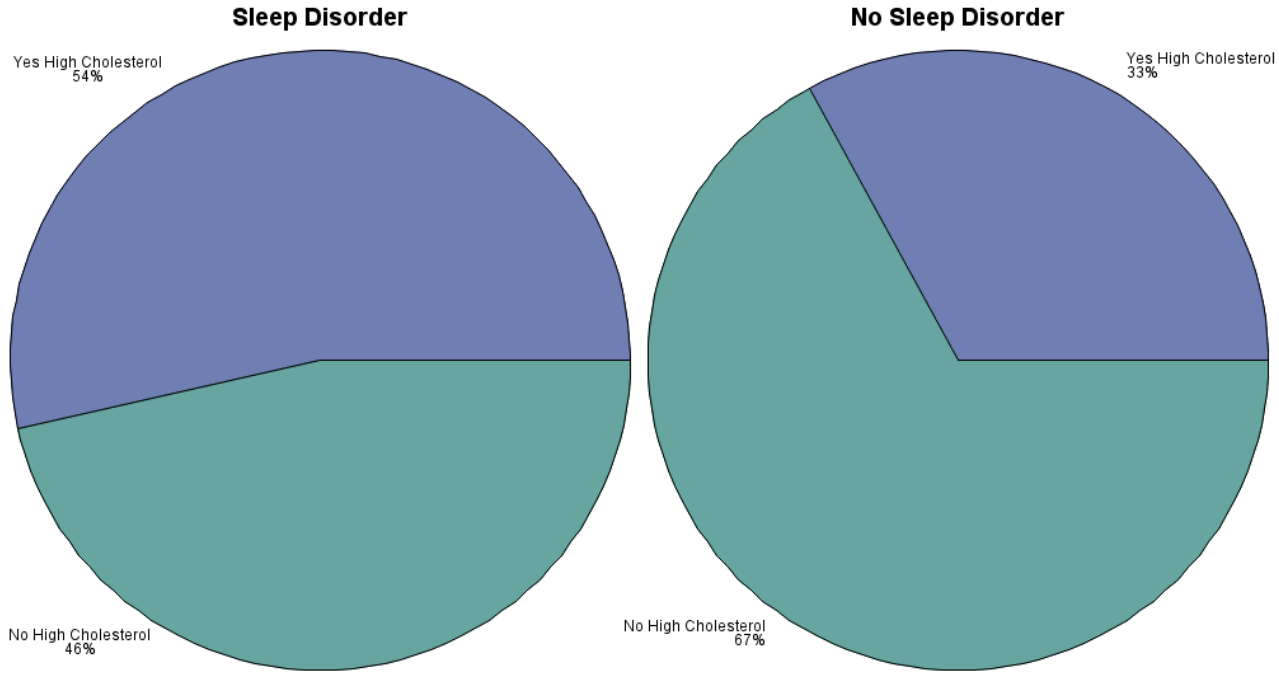
(2) Those who have been told by a doctor that they are borderline diabetic do not have significantly different odds of having been told by a doctor that they have a sleep disorder compared those that have not been told they have diabetes.

High Blood Pressure



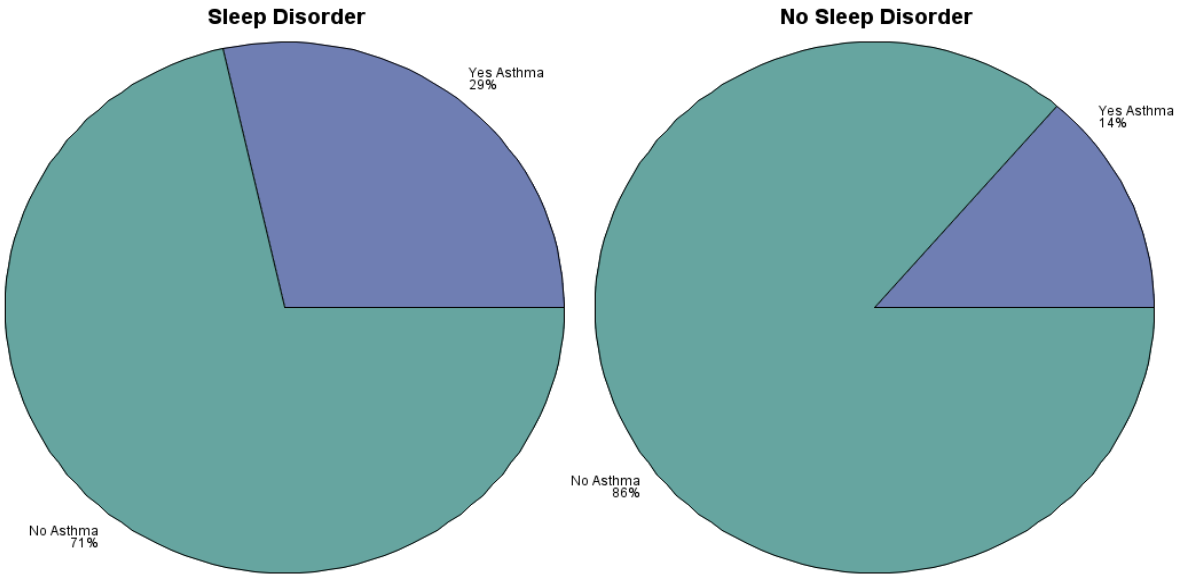
Overall, more people with a sleep disorder have also been told by a doctor that they have high blood pressure compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant, people who have been told by a doctor that they have high blood pressure have between 5% and 70.9% higher odds of also having been told that they have a sleep disorder compared to those without high blood pressure.

High Cholesterol



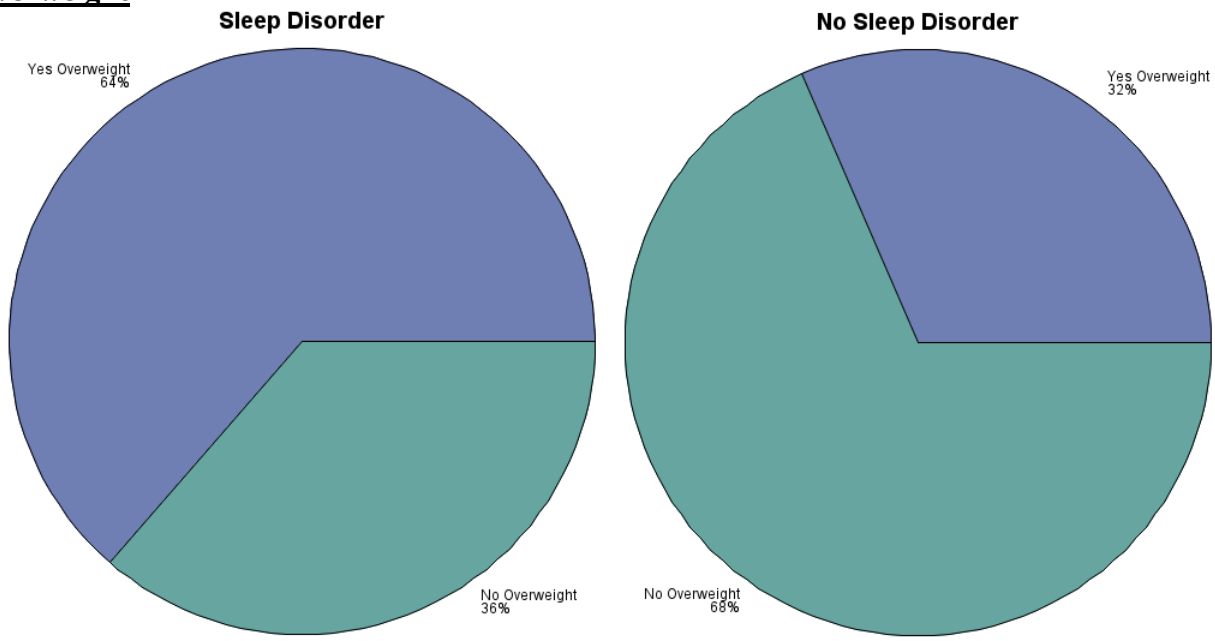
Overall, more people with a sleep disorder have also been told by a doctor that they have high cholesterol compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant, people who have been told by a doctor that they have high cholesterol have between 9.8% and 75.8% higher odds of also having been told that they have a sleep disorder compared to those without high cholesterol.

Asthma



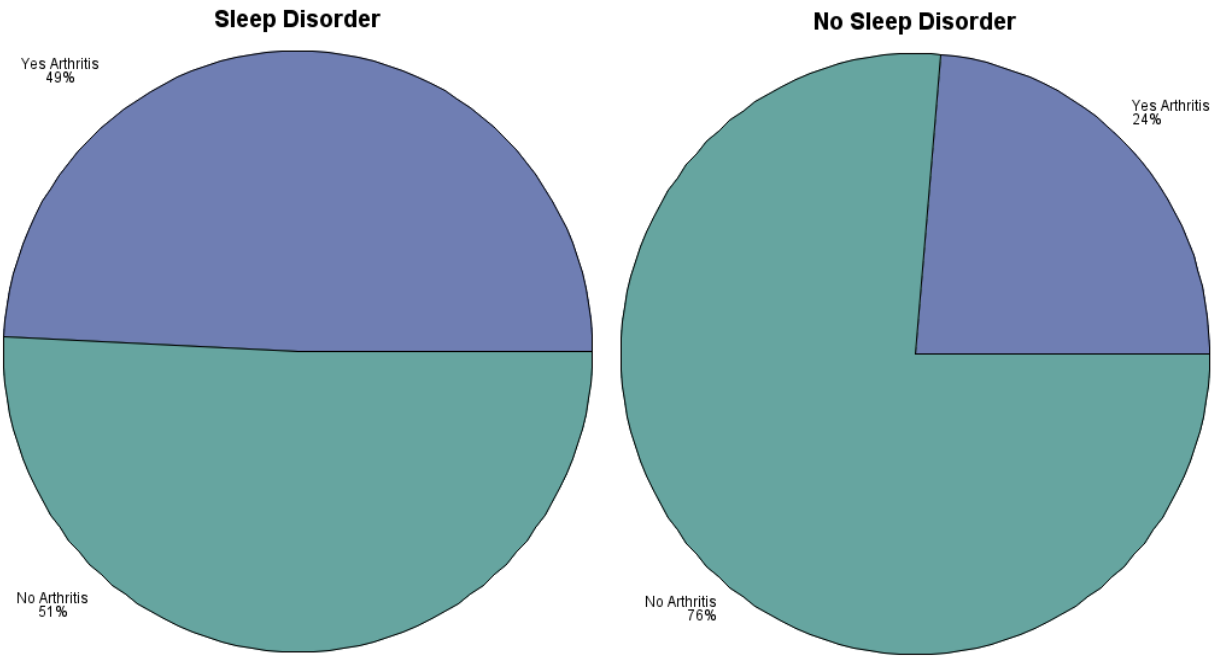
Overall, more people with a sleep disorder have also been told by a doctor that they have asthma compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant, people who have been told by a doctor that they have asthma have between 33.6% and 120.2% higher odds of also having been told that they have a sleep disorder compared to those without asthma

Overweight



Overall, more people with a sleep disorder have also been told by a doctor that they are overweight compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant, people who have been told by a doctor that they are overweight have between 93.9% and 205.8% higher odds of also having been told that they have a sleep disorder compared to those that are not overweight.

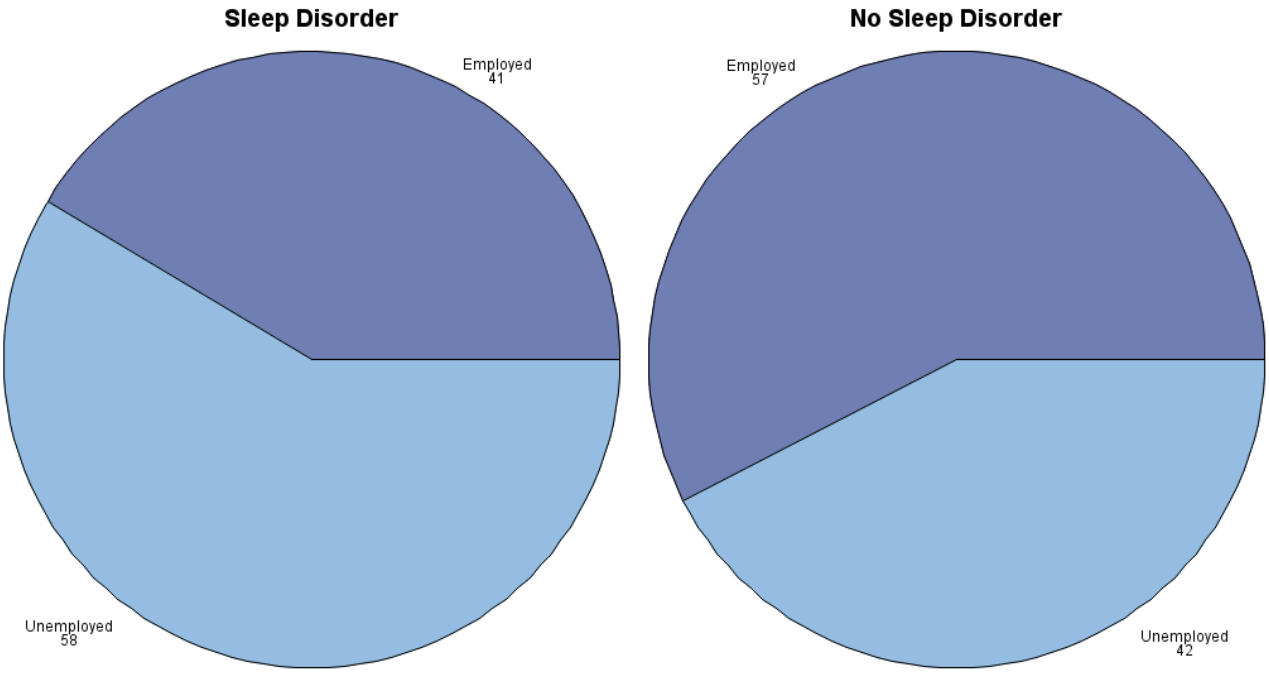
Arthritis



Overall, more people with a sleep disorder have also been told by a doctor that they have arthritis compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant, people who have been told by a doctor that they have arthritis have between 41.7% and 131.9% higher odds of also having been told that they have a sleep disorder compared to those without arthritis.

We now turn our attention to some other general factors that were found to be associated with having ever been told by a doctor that they have a sleep disorder.

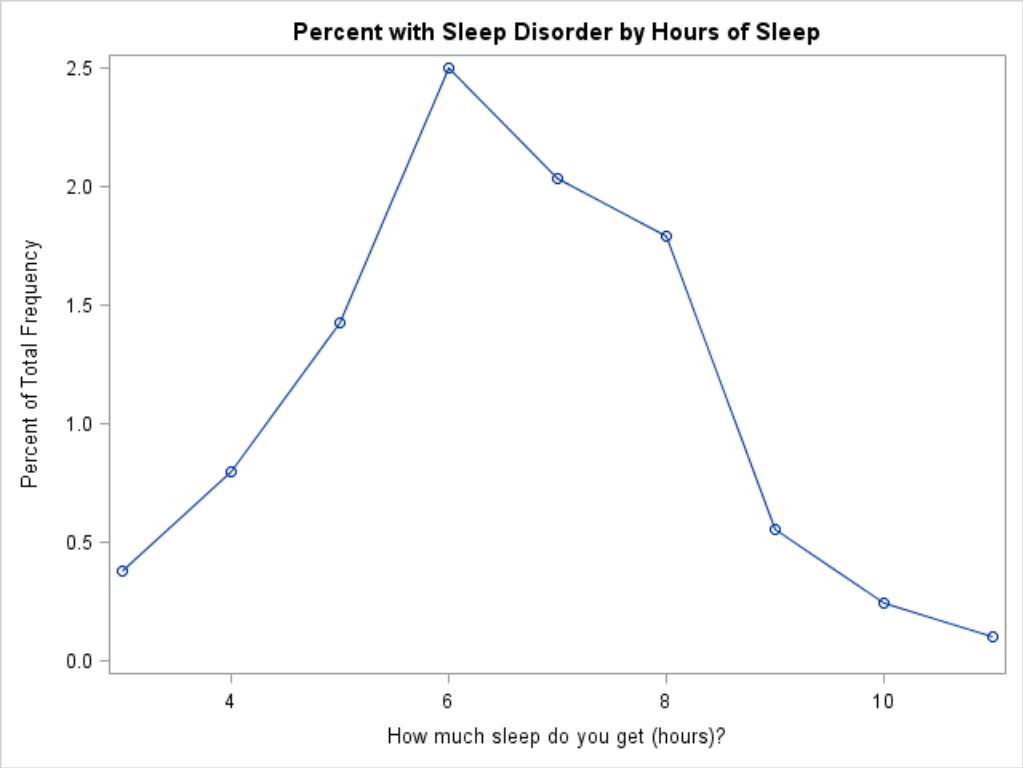
Employment



Overall, less people with a sleep disorder are employed compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant, people who were employed as of the time of the survey have between 41.7% and 131.9% lower odds of also having been told that they have a sleep disorder compared to those that are unemployed.

It is important to understand that *current* unemployment could not possibly be a cause of having been diagnosed with a sleep disorder in the *past*. There are many possible explanations for this association. Working in fields that lead to unemployment more often that other fields could be a cause of the sleep disorder. It also could be that having a sleep disorder causes people to end up being unemployed for some reason. Any causal relationships are not clear simply from this model. Further studies would be needed explain the association between these two variables.

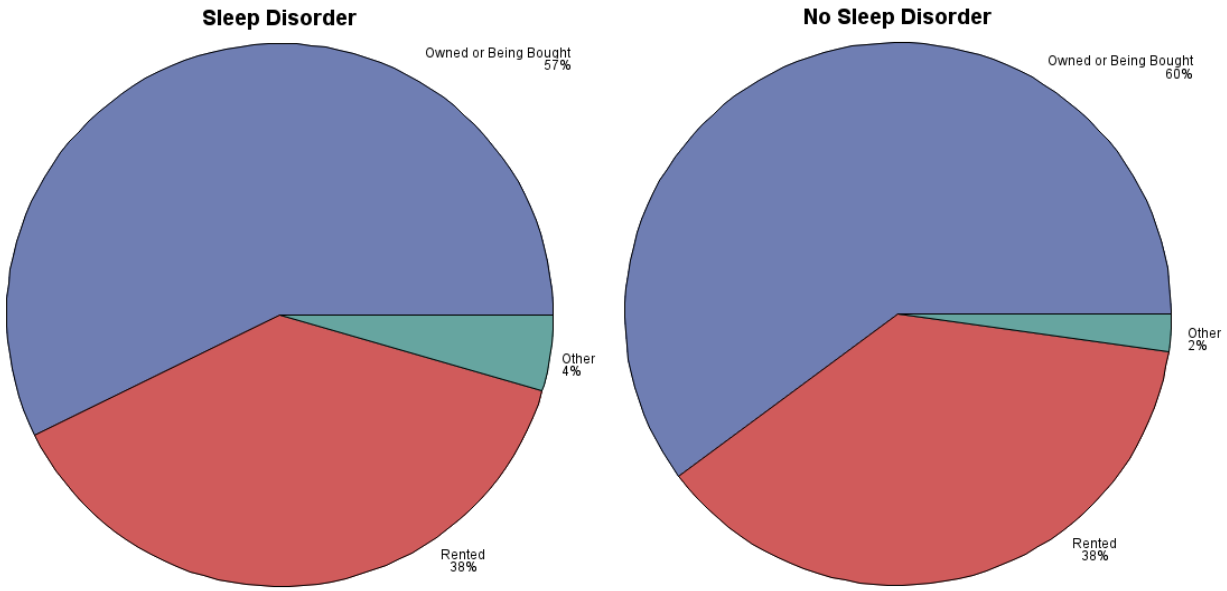
Hours of Sleep



Overall, there is no clear, one directional trend between the hours of sleep the person is getting *as of when the questionaire was done* and whether or not theh person has ever been told by a doctor that they have a sleep disorder. It seems peculiar that of those who are getting less sleep, less of them have been diagnosed with a sleep disorder. This may not be a pattern that would be replicated if the survey was redone because the number of people reporting very little sleep each night was very small.

However, when accounting for all other variables in the model, a trend was detected. The model shows that, holding all other variables constant, for each increase in sleep by 1 hour, the odds of having been diagnosed with a sleep disorder *decrease* by between 11.1% and 22.7%. This makes intuitive sense, whatever the causal relationship might be.

Living Arrangement



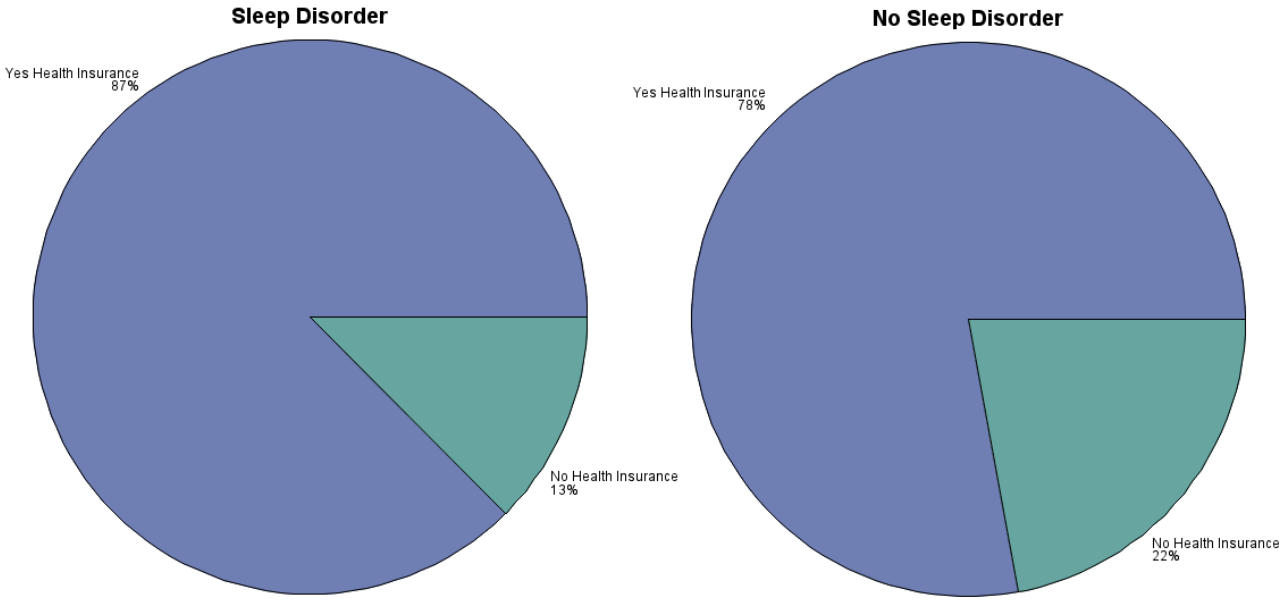
Overall, more people with a sleep disorder have "Other" living arrangments compared to those who were not diagnosed with a sleep disorder. Also, slightly less people with a sleep disorder own or are buying the place that they live compared to those without a sleep disorder. This not very strong pattern still holds, even when accounting for all other variables in the model. The model shows that,

(1) People who own or are buying the place that they live have between 15.3% and 71.9% lower odds of having been told by a doctor that they have a sleep disorder compared to those who have "other" living arrangements.

(2) People who are renting the place that they live have between 6.7% and 69.6% lower odds of having been told by a doctor that they have a sleep disorder compared to those who have "other" living arrangements.

Both of these seem intuitive in that an unstable living arrangement could lead to a sleep disorder. However, the causal relationship cannot be ascertained from this survey alone. Additionally, since the sleep disorder could have been diagnosed years before, it could be that the sleep disorder is causing an inability to hold a steady job, on average. This may cause more individuals with sleep disorders to lose their job and need to find "other" living arrangements. The causal relationship is not very clear what it could be. Further studies would be needed to investigate this relationship further.

Health Insurance



Overall, *more* people with a sleep disorder have health insurance compared to those without a sleep disorder. This cursory look suggests a possible association between the two. This holds, even when accounting for all other variables in the model. The model shows that, holding all other variables constant, people who had health insurance as of the time of the survey have between 5.8% and 99.4% higher odds of also having been told that they have a sleep disorder compared to those without health insurance.

It is important to understand that *current* insurance could not possibly be a cause of having been diagnosed with a sleep disorder in the *past*. There are many possible explanations for this association. Having health insurance would make it easier for an individual to go to a doctor and be diagnosed. Those without insurance may have a disorder, but not yet be diagnosed. But since the health insurance is *current* insurance, this could not be the case for every individual. Any causal relationship would need to be investigated with further non-retrospective studies.

Conclusion

Our model found interesting associations, but further refinement can enhance our conclusions. For one, we did not include interactions between the variables in our model. This is not because we thought the interactions were not important, but because there are too many combinations to include all of them. What is possible to include however, is a smaller subset of plausible interactions. If significant, these interactive effects could provide better estimates for the effects of each disease, as well as indicating further research into those interactions.

Another improvement that could be made to our model is to resolve chronological issues with our dataset and outcome variable. Our outcome variable refers to an event in the past (sleep disorder diagnosis). We do not know when the sleep disorder occurred. We do not know the order of events with regard to other variables in our dataset. Our interpretation could improve if we found a dataset that accounted for the chronology of events. For example, what if patients diagnosed with a certain disease in the past were more likely to experience a sleep disorder, but not the reverse? This would direct us to investigate a causal relationship. These are questions we cannot ask given the data available, but they would be important to understanding the nature of these relationships.

Finally, we conclude with suggestions for further research directions light of our findings. Given that our model is an exploratory study, we lack information about causal associations. However, we can guide further research by using the associations we found. We found relationships between certain diseases such as diabetes and sleep disorders, which can point to interesting questions for further research. For example, do these diseases cause sleep disorders, or does another factor cause both? Could sleep disorders be indicators for certain diseases? This information could be important to the treatment and prevention of those disorders. What is more, we uncovered an relationship between living conditions and sleep disorders. This suggests a direction into the role of housing, income, and socioeconomic status in sleep disorders. We are hopeful that these directions will uncover useful insights.

References

Institute of Medicine. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem. Washington, DC: The National Academies Press; 2006.

Naegele B, Thouvard V, Pepin JL, Levy P, Bonnet C, Perret JE, Pellat J, Feuerstein C. Deficits of cognitive executive functions in patients with sleep apnea syndrome. Sleep. 1995;18(1):43–52.

Alan Agresti (2007). An Introduction to Categorical Data Analysis. Hoboken: John Wiley and Sons.