Draft title for CSC 595 Report

Authors:

CSC 595

...

1 Introduction and Data Description

1.1 Data Introduction and Overall Summary

The NHANES is a large scale longitudinal database created by the Centers for Disease Control and Prevention (CDC) to collected data from a sub-sampling of the United States population. The data includes interview data, physical examinations to include vital signs (systolic and diastolic blood pressure, height, weight, BMI, ...) as dental information, laboratory measurements (i.e. glucose or HBA1C values for diabetes), and demographics information. The purpose of this data is set national standards (i.e. BMI percentile measurements standardized by age and sex), track health data for diseases like cardiovascular disease or diabetes to shape public policy, or provide data for international research organizations and academic institutions for many purposes. Please see the following example of areas which have used this data: Iranpour at al (2019) looked at the inverse relationship between amount of caffeine consumed and symptoms of depression; Wang at al (2018) looked at the relationship between the cadmium in the blood and both blood pressure and obesity; Howell et al (2017) looked at the mortality rate of Mexican American adults in relation to their BMI/BMI percentile as well as demographic information.

The focus of the interactive tool as well as this report is to provide summaries of obesity data collected in NHANES. Obesity is a serious problem as documented by the CDC and World Health Organization (WHO) which has led to an increase in many diseases ranging from cardiovascular disease to diabetes worldwide. Hales, Carrol, Fryar, and Ogden (2017) documented the incidence of obesity in the united states based on different demographic data based on a publication within that organization from 20165- 2016. This paper presented this information through bar charts to overall counts by demographic groups as well as line charts to show the incidence rates within the time frame of 2015-2016. They also showed line plots overtime from 1999 to put these results in a general context. Thompson et al (2007) explored the impact of childhood obesity upon later health effects and identified risk factors for long term health. In this study the long terms impacts for how the development of cardiovascular disease were examined to see impact upon children as well as effects within

adulthood. The Centers for Disease Control and Prevention looked at incidence of diabetes with regards to many risk factors including obesity (2011). Ford ES (2005) summarized the risk factors for both cardiovascular disease and diabetes and indicated the link between both diseases and obesity.

This interactive tool will provide a summary of the prevalence of obesity using the NHANES data. This will due this for different demographics variables to show the prevalence based on variables like age. Variables were also generated showing the continuous variables like age, height, and weight were binned for use in tree or treemap plots. It was felt that from the clinical perspective these were good as they mimicked clinical flow diagrams. Thus this would be good visualization approaches based on how clinicians look at diagnosis strategies based on flow charts. The purpose of the data was the emphasize summary of obesity in a way meant to reflect clinical judgement. Once they clicked on an element of the space filling plot they could see the sub-statistics for the additional selectable variables through bar charts to gain a better understanding of the relationship between the two.

1.2 Data Description

The data included in our combined dataset includes the following:

Summary of Variables				
Variable Name	Units	Type	Labels	Description
Sequence Number		Categorical		Primary key for all csv files for NHANES
Age	Years	Integer		Age at time of the vital sign measurement in years
Age Grouping		Categorical	21-30, 31-40,...	Derived based on ten year intervals
Height	cm	Float with 1-decimal		

Height Grouping		Categorical	151-175, 176-200, ...	Derived
Weight	kg	Float with 1-decimal		
Weight Grouping		Categorical	71-80, 81-90, ...	Derived
BMI	$\frac{kg}{m^2}$	Float with 1-decimal		Derived based on height and weight
Systolic BP	$mmHg$	Float with 1-decimal		
Diastolic BP	$mmHg$	Float with 1-decimal		
Total Cholesterol	$\frac{mg}{dL}$	Integer		
Obesity		Boolean	TRUE, FALSE	TRUE = Obese, FALSE = Not Obese

The data contains the overall key as well as important vital sign variables such as height, weight, BMI, and blood pressure. Some variables have been derived to create categorical groupings. This makes sense as there may be much change year to year for most people under 40 (in the elderly population there would be more impact upon health even from year to year or even month to month). Yet there are significant medical issue which appear from decade to decade as the health problems for a 20 year old and different than a 40 year old. They also allow for these variables to be used in our visualizations such as treemaps. The same is true for the height and weight groupings. The categorical bins allow for the binning of subjects in the space filling approaches to expand to more detailed data on obesity within those groupings. The treemaps would show the counts based on space filling for variables like height, weight, or age by bins whereas the bar charts would show the relationship between to the factors based on percentages to obesity and cholesterol.

A few things to note on the relationship between obesity and the variables. There is a clear relationship between BMI, age, gender, and obesity. Using the CDC website [9] obesity can be defined as a function of BMI with BMIs 30 or higher classified as obese and BMIs between 25.0 and 29.9 as overweight. For teens and

children BMI is not calculated in the same manner as the CDC indicates this is a function of age and gender. Obesity is based on the BMI percentile for children and a teen or child would be considered obese if their BMI percentile was above the 95^{th} percentile for their age and gender. This is a standard value based on a teen or child's age or gender. It is also clear that height and weight impact obesity as BMI is weight divided by height squared. There should be a correlation between blood pressure and obesity but not as strong as the other variables. There should be a positive correlation between hypertension (high blood pressure) and obesity. Similarly there should be a positive correlation between total cholesterol (especially triglycerides) and obesity. High and extreme levels of cholesterol (dyslipidemia, hyperlidemia, or hypertriglyceridemia) should have an increased risk for obesity. Nonetheless both cholesterol and blood pressure should have a weaker relationship as there other factors impacting each such as stress and genetics. Choi (2014) has an interesting paper looking at the relationship between obesity and different complication like diabetes and cholesterol issues based on a biological perspective. This would explain biologically the correlations between the various issues.

1.3 Data Cleaning

Even though the NHANES data is contained is various databases as .csv files (lab data, vital signs data, concomitant medication data, demographics data, ...) not much was required for data cleaning. All databases had a common linkage based on the sequence number (primary key) which connected all the databases with demographics easily. All that was needed was simple merging to get whatever demographics variables as needed. For some continuous variables it was necessary to create a new categorical variables based on binning. For example to make the different bins for cholesterol a new ordinal variable was created for each bin so they could be incorporated into the visualization easier. No other data cleaning was done as the focus was on the visualization as opposed to outlier diagnostics or checking for data integrity.

2 Description of Interactive Visualization Approach

References

[1] Centers of Disease Control and Prevention. National diabetes factor sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. US Department of Health and Human Services, Centers for Disease Control and Prevention, 1(1), 2568-69.

[2] Choi, Jung U (2014). Obesity and its metabolic complications: the role of adipokines, inflammation, insulin resistance, dyslipidemia, and nonalcoholic fatty liver disease. International Journal of Molecular Sciences, 15(4), 6184-6223.

[3] Ford ES (2005). Risks for all-cause mortality, cardiovascular disease, and diabetes associated with metabolic syndrome: a summary of the evidence, Diabetes Cares, 28(7), 1769-1778.

[4] Hales, CM, Carrol, MD, Fryar, CD, and Ogden, CL (2017). Prevalence of obesity among adults and youth: United States, 2015-2016.

[5] Howell, CR, Fontaine, K, Ejima, K, Ness, KK, et al (2017). Maximum lifetime body mass index and mortality in Mexican American adults: the National Health and Nutritional Examination Survey III (1988-1994) and NHANES 1999-2010. Preventing Chronic Disease, 1545-1151(14).

[6] Iranpour, S. and Sabour S (2019). Inverse Association between caffeine intake and depression symptoms in US adults: data from National Health and Nutritional Examination Survey (NHANES) 2005-2006. Psychiatry Research, 1872-7123(271), 732-739.

[7] Thompson, DR, Obarzanek, E, Franko, DL, Barton, BA, Morrison, J, Biro, FM, Daniels, SR, Striegel-Moore, RH (2007). Childhood overweight and cardiovascular disease risk factors: the National Heart, Lung, and Blood Institute Growth and Health Study. The Journal of Pediatrics, 150(1), 18-25.

[8] Wang Q, Wei S (2018). Cadmium affects blood pressure and negatively interacts with obesity: Findings in the NHANES 1999-2014. The Science of the Total Environment, 1879-1026(643), 270-276.

[9] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Appendix Code

If we want to have code or snippets of code in an appendix