**DSC 510 Final Project**

**Data Analysis with NHANES Dataset**

**Group Members:**

William Ranick

Nibedita Bal

Hima Spandana Barla

**Abstract:**

This paper presents 3 various approaches to predictive analysis of the National Health and Nutrition Examination Survey (NHANES) Dataset. An overlaying goal behind the data manipulation is to predict the Systolic Blood pressure, cancer diagnosis, diabetes diagnosis within the set of 10,000 observations. This data set is projected from a survey unique in that it combines interviews and physical examinations.NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation. Some reasons behind our drive to study the different disease is due to the underlying fact that findings from this survey will be used to determine the prevalence of major diseases and risk factors for diseases and that information will be used to assess nutritional status and its association with health promotion and disease prevention. Also, NHANES findings are also the basis for national standards for such measurements as height, weight, and blood pressure. Data from this survey will be used in epidemiological studies and health sciences research, which helps develop sound public health policy, direct and design health programs and services, and expand the health knowledge for the Nation. The results derived from our analysis of NHANES is to help explain and predict the factors that could be responsible for and how we can further support our results from other previous studies on health and disease control with similar attributes. For this analysis, R was used to perform data pre-processing, data exploration, linear regression, model, evaluation, and model visualization.

**Introduction:**

Data on the prevalence of chronic disorders in the population has been obtained for the NHANES programme. Estimates are generated through the survey for previously undiagnosed disorders, as well as those identified to and reported by respondents. NHANES discusses risk factors, certain components of the lifestyle, constitution, heredity, or climate of a person that could increase the likelihood of contracting a certain illness or condition. Smoking, alcohol consumption, sexual practices, drug use, physical fitness and activity, weight, and dietary intake are studied. The diseases, medical conditions, and health indicators that are studied in the project are blood pressure, diabetes and cancer respectively. According to *World Health Organisation*, hypertension (systolic blood pressure greater than or equal to 130 mmHg or diastolic blood pressure greater than or equal to 80 mmHg or current high blood pressure medication) is a public health issue in the United States as it directly raises the risk of cardiovascular disease and people with high blood pressure have been shown to be at an elevated risk of developing type 2 diabetes by around 50 percent. According to *MAYO Clinic*, the second most prevalent chronic illness and cause of death in the United States is cancer. Our aim was to examine the correlations of sedentary activity and nutrient consumption with overall and cancer-specific mortality. The Healthy Eating Index (HEI) is associated with lower risk of chronic diseases and mortality in different populations Higher diet quality was associated with lower all-cause, cardiovascular disease, and cancer mortality risk. Diabetes is one of the most prevalent chronic diseases in the United States and worldwide, along with the various complications presented in this compendium, resulting in increased impairment. The basis for more studies on risk factor and intervention investigations as well as health care preparation is to understand the extent of diabetes incidence in the U.S. population and in subgroups in which there is an extremely high burden.

**Literature Reviews:**

In *Association between body size and blood pressure in children from different ethnic origins,* five ethnic groups were examined. Linear regression was used to examine non-linear associations between body size and blood pressure, adjusted for age, sex, height and birth weight.In general, BMI, WHtR and FMI had a relatively weak association with BP in Black- Caribbean children; most associations were non-significant and tended towards the null.This cross-sectional study explored ethnic-specific associations between three measures of body size in children aged 5-6 years( Hoog, Eijsden, Stronks , Gemke and Vrijkotte ,2012). The data revealed ethnic differences in the association between body size and BP, first, body size had a relatively large effect on BP in Turkish children, indicating higher BP levels with increasing BMI and FMI, second is, Black- Caribbean and Moroccan children showed high BP at low BMI, whereas at low FMI this effect was seen only in Moroccan children. Moroccan children also showed higher SBP at high BMI and FMI, third is ,in general, the strongest associations with BP were found for BMI in all ethnic groups.Among the various measures of body size, BMI appeared to be the most reliable determinant of BP in any ethnic group( Hoog, Eijsden, Stronks , Gemke and Vrijkotte ,2012).

In *Relationship between perceived stress and dietary intakes in type 2 diabetes patients,* the major aspect of lifestyle, dietary patterns are concerned with health issues in all conditions including stress (Dehghan, Pourmoradian, Mahdavi, Sarmadi, Mehralizadeh, 2016). Stress was the independent variable.The aim of this study was to explore the relationship between dietary intake and perceived stress in type 2 diabetics.This study was to examine separately negative/positive perceived stress effect on dietary intake. Carbohydrate rich foods may raise glucose and trigger insulin secretion, which facilitates the uptake of most large neutral amino acids, except tryptophan, into peripheral tissues. It is widely accepted that eating behavior in humans changes according to changes in their emotional feelings such as stress, anger, joy, depression and other moods. A positive perception of stress may lead to healthy food choices and negative perceived stress to the unhealthy dietary patterns. A diet with adequate vitamin B complex, vitamin E and antioxidant minerals may be a positive factor in coping with stress and help positive interpretation of stressful events(Dehghan, Pourmoradian, Mahdavi, Sarmadi, Mehralizadeh, 2016).

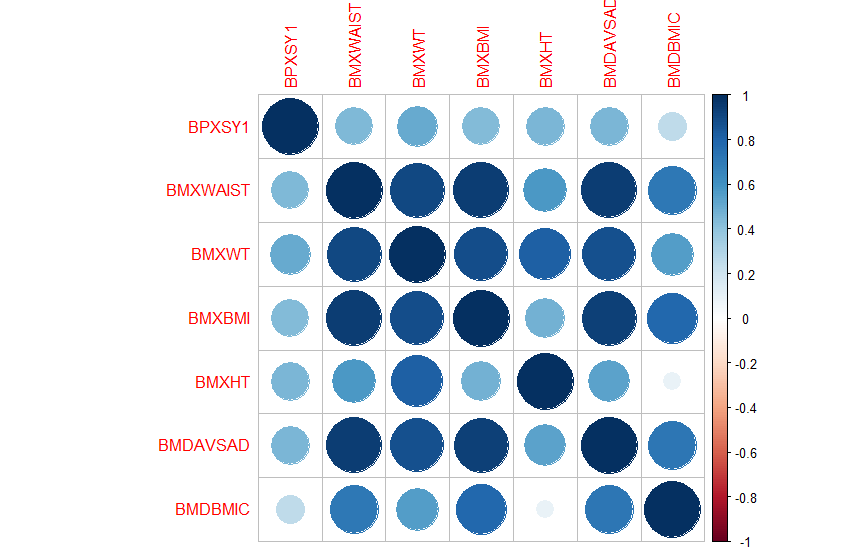
In *The association between circulating levels of vitamin D and inflammatory markers in the first 2 years after colorectal cancer diagnosis*, study was conducted to observe the correlation between serum vitamin D levels and the progression of colorectal cancer. Data was collected from colorectal cancer patients over the course of 2 years of treatment. By performing regression on data of participants’ serum vitamin D levels and the existence/levels of cancer-associated cytokines in the blood, it was found that a higher level serum vitamin D was correlated with decreases in the levels of these cytokines, specifically IL-6(Balvers, Bours, Wilt, Witkamp, Baar, Geijsen, Halteren, Keulen, Kok, Kouwenhoven, Ouweland,Zutphen, Weijenberg, Duijnhoven, 2020). Statistically significant correlations, however, were not found between the serum levels of vitamin D and other cytokines (besides IL-6) which are associated with colorectal cancer progression. It is also noted in the study that its limitations include low levels of these cytokines, generally, in all of the patients. Therefore, it may have been difficult to observe significant changes in cytokine levels in response to vitamin D, and conversely, may have allowed the effect of vitamin D to have been possibly exaggerated in regards to levels of cytokine IL-6(Balvers, Bours, Wilt, Witkamp, Baar, Geijsen, Halteren, Keulen, Kok, Kouwenhoven, Ouweland,Zutphen, Weijenberg, Duijnhoven, 2020).

**Methods and Results:**

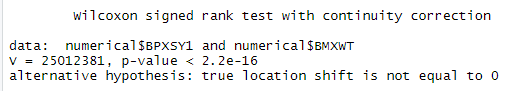
**Does Systolic Blood pressure explain BMI?**

For knowing the relationship between Systolic blood pressure and BMI,below are the variables used are BPXSY1: Systolic: Blood pressure (first reading) mm Hg - Dependent Variable, BMDAVSAD: Average Sagittal Abdominal Diameter (cm), BMXHT:Standing Height, BMXBMI: Body Mass Index (kg/m\*\*2), BMXWAIST:Waist Circumference (cm), BMXWT: Weight (kg). For Systolic blood pressure and BMI, exploratory data analysis is done with data cleaning, data preprocessing. Some statistical methods are used such as normalization, selecting significant variables by stepwise regression, multicollinearity checked by VIF and visualize the relationship among significant variables.

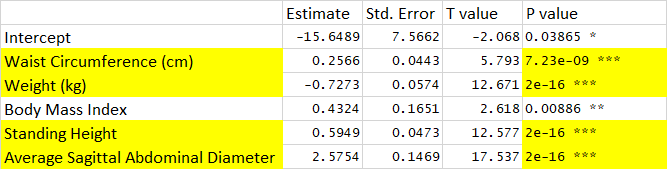
Collinearity among numerical variables is checked and it is shown that most of the variables are positive and linearly correlated;however, the vif values are greater than 5,so the variables are highly correlated.

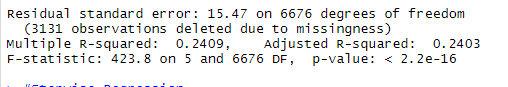


Normality of Systolic blood pressure and other independent variables are checked by Shapiro-Wilk test. The P value is less than .05 which means the data is not normal which means it cant reject the null hypothesis. As the data is not normal and the sample size is big , wilcox test was chosen for this variables.



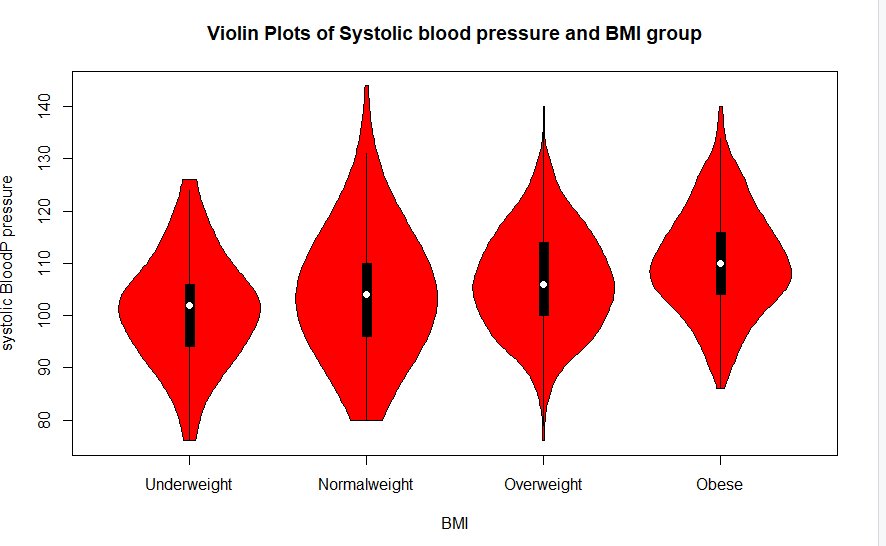
After removing highly correlated variables, we build a linear regression model. Below are the findings we have gotten. It is shown that Waist Circumference,Weight (kg), Height,Sagittal Abdominal Diameter, and Body Mass Index are highly significant which means these variables impact on systolic blood pressure. From this result, it is assumed that if a person has larger belly, higher BMI which means the person is overweight or obese , he is more prone to systolic blood pressure.The P value is significant which means the null hypothesis is rejected and alternative hypothesis is accepted which tells BMDAVSAD,BMXHT, BMXWAIST, BMXWT and BMXBMI have surely impact on blood pressure.





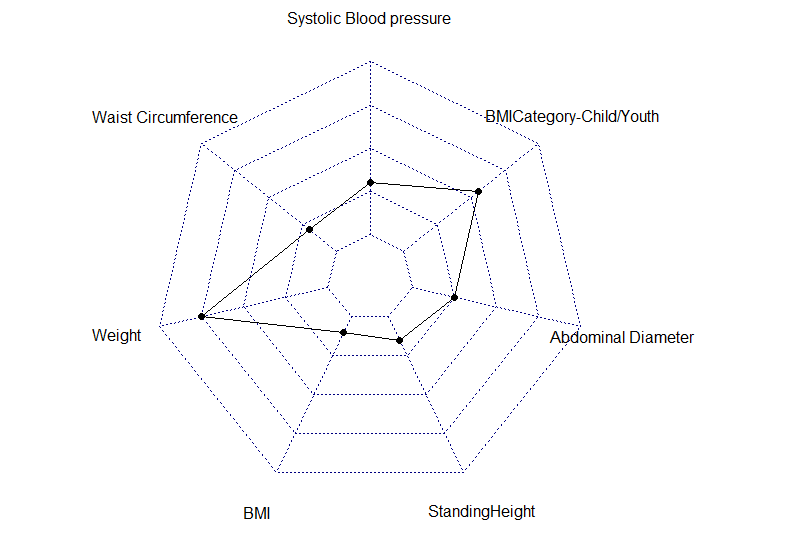
Below is a visualization among blood pressure and age group which tells that blood pressure increases as the BMI increases. It is clearly visible that obese plot has higher median and skewd towards high that means obese people suffer systolic blood pressure.

**Fig:1**



Below is the Radar plot of the data which shows that weight has more positive impact on systolic blood pressure. In other words, it can be explained as the people having more weight(obese, over weight), they are more prone to suffer blood pressure. The second variable is BMI category in child and youth. It is usually observed that children who are obese, in their eatrly youth stage, they suffer blood pressure early rather than youth who having good BMI.

**Fig2:**

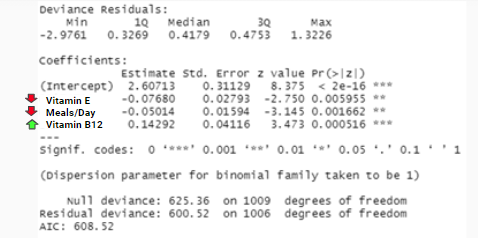


**The intake of which vitamins and organic compounds reduce/increase the risk of cancer diagnosis?**

First, relevant variables were pulled from two different datasets (the Questions dataset and Diet dataset). These variables included a binary variable, if the patient had been diagnosed with any type of cancer. Additionally, the age of diagnosis of several different types of gastrointestinal cancers were taken from the dataset and transformed into a single binary variable (for whether or not the patient had been diagnosed with any of these cancers). 64 quantitative variables were taken from the diet dataset, which represented the intake of 64 compounds on a single measured day. The existence multicollinearity was checked between these quantitative values with the intent of removing the offending variables from the final model if applicable.

Backwards elimination was performed from a full model of all variables to generate a logistic regression model of only the most correlated variables. None of the variables present in this model exhibited multicollinearity with one another.The creation of a similar model, but for only gastrointestinal cancers, was attempted. However, there were too few positive instances of any of the gastrointestinal cancers (less than 30 total) to perform logistic regression. The first logistic regression was then trained and tested against the data set to assess its ability to predict cancer diagnosis.

Shown below, is the final logistic model that was created to predict the odds (log odds) of being diagnosed with any type of cancer based on the intake of different vitamins and organic compounds. This model included vitamin E (shown as DR1TATOA) and the total number of meals eaten daily (shown as DR1TNUMF) as demonstrating a slightly *increased* risk of being diagnosed with cancer, while vitamin B12 (shown as DR1TVB12) seemed to reduce the log odds of being diagnosed with cancer. All of these variables demonstrated very low coefficient estimates (the largest being that ingesting 1mcg of vitamin B12 decreased the log odds of being diagnosed with cancer by approximately 0.143). Also, all of the variables demonstrated disappointingly low significance - with only vitamin B12 exhibiting acceptable significance. As expected, the trained model performed very poorly against test data, accurately assessing the patient’s risk of being diagnosed with cancer in only 0.03% of cases.

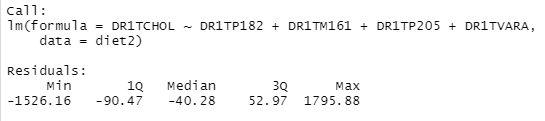


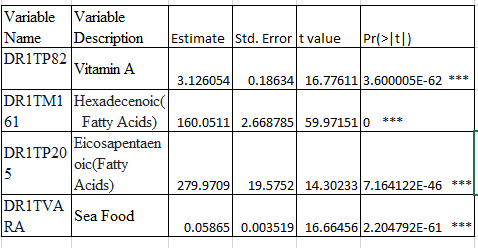
**How well can diabetes diagnosis be predicted from the diet and vitamins of the patients?**

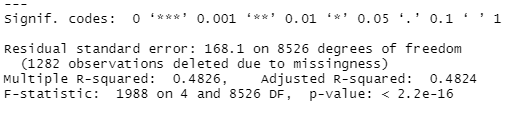
Diabetes and high cholesterol often occur together. The American Heart Association (AHA) states that diabetes often lowers HDL (good) cholesterol levels and raises triglycerides and LDL (bad) cholesterol levels. Both of these increase the risk for heart disease and stroke.

Multiple regression model is used for analysis, where the DR1TCHOL(diabetes /cholesterol) is the independent variable and other dependent variable are DR1TP182 PFA 18:2 (Octadecadienoic) (gm), DR1TM161 MFA 16:1 (Hexadecenoic) (gm), DR1TP205 PFA 20:5 (Eicosapentaenoic) (gm), DR1TVARA- Vitamin A as retinol activity equivalents (mcg) etc. I have tried both training and testing methods to find the patients who were suffering from diabetic and used R to show the regression values and tableau to average the number of times sea food was eaten in 30 days by diabetics patients. Before performing the multiple regression model I have checked for collinearity and all the variables are correlated to each other.

For a given predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero. It can be seen that changes in DR1TP82 and DR1TM161, DR1TP205,DR1TVARA are significantly associated with changes in diabetes. The multiple r square and the adjusted r square are almost same that is 48% and the p value is <2.2e-16



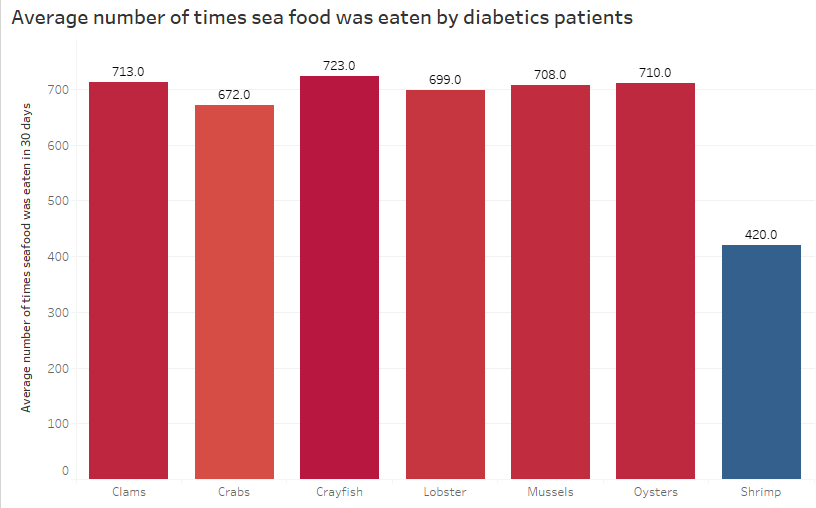
****



Performed multiple regression to figure out how the vitamins used by the patients in their diet has helped them to reduce the cholesterol levels in their body.Hence we can reject the null hypothesis and accept the alternative.

Data visualization using Tableau

**Fig3:**

****

The above visualization gives basic information of the average number of patients in the NHANES survey showed interest in having seafood at least once in their meal in 30 days. Lobster (DRD350D), clams(DRD350E), crabs(DRD350B), mussels(DRD350E), oysters(DRD350F), Crayfish (DRD350C) and shrimp(DRD350H) all pack a ton of flavor and not much fat, which makes them great choices for managing diabetes.

Protein helps the body to grow new tissue, therefore helping to build muscle and repair damage to the body. Protein can also be broken down by the body into glucose and used for energy. A person with diabetes can help stabilize their blood glucose levels by eating protein as part of a healthy meal. The protein helps slow down carbohydrate absorption, so glucose spikes are minimized.

**Limitations of the research :**

For the research question, Systolic blood pressure and BMI, the dataset has overall 2 years of data but the dataset is incomplete which means the data covers only one or two conditions which we are trying to model-as when a model is built for knowing the relationship between BP and BMI, we need a long term data but here the small duration data might not give us accurate model. In addition to this, the data attributes collected are very ambiguous which takes a lot of time to understand and analyze.

For diet/cancer analysis, the online dictionary of all variables include many variables that were not actually present in the data set, such as measures of the amount of supplements taken monthly. The only data available to use for this analysis were measurements of how much of these compounds were consumed for only 1 day of study participation. This single day of dietary intake is very unlikely to be representative of the patient’s overall diet. Additionally, the data lacked sufficient numbers of participants with gastrointestinal cancers. It is possible that these types of cancers are the most likely to be affected by a patient’s diet.

Several limitations are there with the study of cancer and diet. The dietary data was logged as estimated vitamin intake for only a single day of patient measurement. While exploring this question, the summary of all variables on kaggle was consulted. Unfortunately, some of the variables listed on the kaggle summary (which included measurements of a second day as well as monthly dietary estimates) did not actually appear in the dataset that was downloaded. Having these variables would have certainly increased the chance that the dietary data used was more representative of the patients’ overall diet.

Additionally, there were very few instances of patients included in the dataset who had been diagnosed with gastrointestinal cancer. Several of the different cancers that would fall into this category had *zero* instances of patients reporting that they had been diagnosed with that cancer. Despite combining all gastrointestinal cancers into one binary variable, there were still not enough positive cases to perform regression.

**Future work that could be conducted on the dataset :**

For future research, the use of other methods such as generalized canonical correlation, principal component analysis, factor analysis could be used to analyze random variables, other than regression. Furthermore, some other mathematical methods that are used in performance evaluation and classification accuracy can be applied in the analysis to know the accuracy of the method. Consequently, further studies can explore the relationship between ‘blood pressure’ and ‘bmi’ , diabetics and diet, cancer and vitamin intake based on the different machine learning techniques.

**Conclusion:**

To conclude, systolic blood pressure depends on five variables, such as Waist area,weight, BMI, abdominal diameter do have impact on blood pressure. If a person has a larger belly, higher BMI which means the person is overweight or obese , he is more prone to systolic blood pressure. Also, children who are obese, in their eatrly youth stage, they suffer blood pressure early rather than youth who having good BMI. For cancer diagnosis, in the future, more thorough data should be collected and selected for analyses in order to determine if the intake of several vitamins and organic compounds may negatively or positively impact cancer diagnosis, particularly for gastrointestinal cancers. In Diabetics research, the long-term trend shows a persistent rise in occurrence, though a slowing in more recent years, needs continued surveillance. The fact that one-quarter to one-third of diabetes is undiagnosed and that another third of the total population has prediabetes emphasizes the importance of sustained monitoring and improvements in health care delivery.

**References:**

1. <https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/diagnosis-treatment/drc-20373417>
2. <https://www.who.int/news-room/q-a-detail/noncommunicable-diseases-hypertension>
3. de Hoog, M. L. A., van Eijsden, M., Stronks, K., Gemke, R. J. B. J., & Vrijkotte, T. G. M. (2012). Association between body size and blood pressure in children from different ethnic origins. *Cardiovascular Diabetology*, *11*(1), 136–145. <https://doi-org.ezproxy.depaul.edu/10.1186/1475-2840-11-136>
4. Wesselink, E., Balvers, M., Bours, M. J. L., de Wilt, J. H. W., Witkamp, R. F., van Baar, H., Geijsen, A. J. M. R., van Halteren, H., Keulen, E. T. P., Kok, D. E., Kouwenhoven, E. A., van den Ouweland, J., van Zutphen, M., Weijenberg, M. P., Kampman, E., & van Duijnhoven, F. J. B. (2020). The association between circulating levels of vitamin D and inflammatory markers in the first 2 years after colorectal cancer diagnosis. *Therapeutic Advances in Gastroenterology*, *13*, 1–15. <https://doi-org.ezproxy.depaul.edu/10.1177/1756284820923922>
5. Dehghan, P., Pourmoradian, S., Mahdavi, A. M., Sarmadi, B., & Mehralizadeh, S. (2016). Relationship between Perceived Stress and Dietary Intakes in Type 2 Diabetic Patients. *Current Topics in Nutraceutical Research*, *14*(3), 199–205.