ASSOCIATION BETWEEN AGE, GENDER, BMI, BONE MINERAL DENSITY, ALCOHOL CONSUMPTION, SMOKING, AND INCIDENCE OF OSTEOPOROSIS

Asmatahasin Mohammed, Amulya Rangineni, Manaswini Dasari, Prathyusha Vajinepally, Sri Rishitha Dammalapati

1. INTRODUCTION:

Osteoporosis is a condition affecting the skeletal system that is characterized by a progressive reduction in bone mineral density (BMD), resulting in an increased risk of fractures. As the elderly population grows, healthcare systems will face increasing pressure, leading to elevated costs, morbidity, and mortality rates related to fractures. Hence, identifying the risk factors that lead to reduced BMD is critical to prevent, detect, and manage osteoporosis (Yao et al., 2019).

According to the National Osteoporosis Foundation (2021), osteoporosis is a significant public health issue, particularly among older adults, and is associated with a high rate of morbidity and mortality. BMD is a crucial factor in the diagnosis and management of osteoporosis, as it is used to estimate fracture risk by measuring the amount of bone minerals in each area of bone. Low BMD is linked to an increased risk of osteoporotic fractures. Therefore, early detection and monitoring of BMD are essential for preventing and managing osteoporosis (Cosman et al., 2014).

Smoking and excessive alcohol consumption are major risk factors for osteoporosis. According to Kanis et al. (2013), smoking is associated with decreased BMD and increased fracture risk. Similarly, heavy alcohol consumption is linked to decreased BMD and an increased risk of fractures (National Institute on Alcohol Abuse and Alcoholism, 2021). Excessive alcohol consumption can lead to a decline in BMD, particularly in postmenopausal women, by interfering with the bone remodeling process. Furthermore, it may cause malnutrition, impaired liver function, and hormonal imbalances, which can worsen osteoporosis (Akkawi & Zmerly, 2018).

2. RESEARCH QUESTION:

Is there any significant relationship between age, gender, BMI, Bone mineral density, alcohol consumption, smoking, and incidence of osteoporosis?

Hypothesis

Null Hypothesis:

There is no significant relationship between age, gender, BMI, bone mineral density, alcohol consumption, smoking, and incidence of Osteoporosis.

Alternative Hypothesis:

There is a significant relationship between at least one of the variables (age, gender, BMI, bone mineral density, alcohol consumption, smoking) and the incidence of Osteoporosis.

3. DATA:

3.1 Data Source:

https://www.kaggle.com/datasets/jehanbhathena/bone-mineral-density

The dataset contains 1538 observations (rows) and 40 variables (columns). The variables are classified into two categories: categorical variables and numerical variables. Nine variables

were selected for analysis, including Osteoporosis, Calcium, Magnesium, Phosphorus,

Variable	Description
Age	In years (Numerical)
Gender	1= male, 2=female (Categorical)
BMI	(Kg/m2) (Numerical)
Ca	(mmol/L) (Numerical)
P	(mmol/L) (Numerical)
Mg	(mmol/L) (Numerical)
OP	0 = No, 1 = Yes (Categorical)
Smoking	0 = No, 1 = Yes (Categorical)
Drinking	0 = No, 1 = Yes (Categorical)

Age, Gender, Smoking, Drinking and BMI.

Table 3.1 Description of the variables

3.2 Exploratory Data Analysis:

We conducted exploratory data analysis on the dataset by downloading it from its source and importing it into R Studio using the 'setwd()' function to set the working directory and the 'read.csv ()' function to import the data. We then computed and summarized key statistics for each variable in the dataset using functions such as 'head ()' and 'summary ()'. To check the structure of the data frame, we used the 'str ()' function. We identified null values in the data using the 'sum (is.na ())' function and replaced them with the mean values for each column to maintain the overall structure of the data. We used boxplots to identify outliers in the data and replaced them with median values. To assess the normality of the data, we used QQ plots and the Shapiro-Wilk normality test. Histograms were created to visualize the distribution of each variable in the dataset. We created a correlation matrix and corresponding plot to examine the relationships between variables and a pair plot was created to explore the relationships between all variables in the dataset.

3.3 Statistical Methods:

The statistical tests were performed to analyze the relationship between different variables and osteoporosis. The Shapiro Wilk test was used to check if the data is normally distributed. If the data is not normally distributed, the t-test is not appropriate and the non-parametric Wilcoxon rank-sum test or Mann Whitney U test is used instead.

The t-test was used to compare the mean values of two groups. In this case, it was used to compare the mean values of osteoporosis patients with the general population.

The Wilcoxon rank-sum test or Mann Whitney U test is a non-parametric alternative to the t-test, used when the data is not normally distributed. It was used to compare the median values of two groups.

The Kruskal-Walli's test was used to analyze the relationship between osteoporosis and gender. It is a non-parametric test used to compare more than two groups.

These tests were chosen based on the nature of the data and the research question being asked. The findings from these tests can provide insight into the relationship between different variables and osteoporosis.

Statistical Tests performed	Variables Used	Findings
Shapiro Wilk test	All variables	Not normally distributed
t-test	All variables	Null hypothesis rejected
Wilcoxon rank-sum test / Mann Whitney U test	Gender (Samples of male and female with osteoporosis) Age (Age is divided in to 4 quartiles)	p-value is 0.8942 for Gender. p-value is 0.05544 for Age
Kruskal-Walli's test	Gender (Samples of male and female with osteoporosis)	p-value is 0.8941, Chi-square value is 0.017706

Table 3.3 Description of statistical methods performed and their findings.

3.4 Limitations:

The number of rows or observations in a dataset can have a significant impact on the accuracy and generalizability of statistical analyses. In this case, having a limited number of rows in the dataset can result in reduced statistics, making it difficult to detect significant effects or establish correlations between variables. Additionally, normal distribution is an important assumption for many statistical methods. If the data does not follow a normal distribution, it can result in biased or inaccurate results. In this case, the lack of normal distribution for most of the data in the dataset means that many statistical methods may not be appropriate or may result in less accurate results. As a result of these limitations, the analysis may be less accurate or may require additional data or alternative methods to overcome these limitations. It is important to carefully consider the limitations of the dataset when interpreting the results and drawing conclusions from the analysis.

3.5 Appropriateness:

The use of non-parametric tests (Shapiro Wilk, Wilcoxon rank-sum test/Mann Whitney U test, Kruskal-Walli's test) is appropriate since the normality assumption is violated for some variables, and the data are not normally distributed. The use of a logistic regression model is appropriate for determining the effect of predictor variables on the binary outcome of osteoporosis.

4. Findings:

To address missing values in our dataset, we replaced them with the mean values of the respective columns. We also identified and replaced outliers with the median values to reduce their impact. We observed that none of the columns in our dataset followed a normal distribution, which was confirmed by a non-linear relationship between the dataset's quantiles and those of the standard normal distribution. Additionally, our analysis showed a weak correlation between Age and OP, as indicated by both Spearman and Pearson correlation coefficients. The histogram showed that Gender, OP, Smoking, and Drinking variables had highly skewed data, while the correlation plot revealed correlations between drinking and smoking, BMI and Mg, and Ca and P. We performed two non-parametric tests, the Kruskal-Walli's test, and the Mann-Whitney U test, to examine the significance of osteoporosis in different groups based on gender and age, respectively. Both tests

indicated no significant difference between the groups.

We then conducted three logistic regression models using binomial family to assess the relationship between the onset of osteoporosis and various predictors. In the first model, BMI and Gender were considered, but neither predictor showed a significant effect on OP. The second model included Smoking, Drinking, and BMI as predictors, and only Smoking and Drinking showed a statistically significant relationship with OP, while BMI did not. In the third model, Ca, Mg, and P were tested as predictors, and only Ca showed a weak negative relationship with OP, while Mg and P did not have any significant relationship. We trained a Random Forest model on 70% of the dataset and tested it on the remaining 30%, achieving an overall accuracy of approximately 60.17%. Using a grid search approach, we tuned the hyperparameters of the model and achieved an accuracy of 62.77%. However, the specificity of the model decreased while its sensitivity increased after tuning the hyperparameters. Based on our analysis, we concluded that Smoking, Drinking, and Ca might be useful predictors for the onset of OP, while BMI, Gender, Mg, and P may not be significant predictors.

REFERENCES:

Akkawi, I., & Zmerly, H. (2018). Osteoporosis: Current concepts. *Joints, 6*(2), 122-127. https://doi.org/10.1055/s-0038-1660790

Cosman, F., de Beur, S. J., LeBoff, M. S., Lewiecki, E. M., Tanner, B., Randall, S., Lindsay, R., & National Osteoporosis, F. (2014, Oct). Clinician's guide to prevention and treatment of osteoporosis. *Osteoporosis International, 25*(10), 2359-2381. https://doi.org/10.1007/s00198-014-2794-2

Kanis, J. A., Johnell, O., Oden, A., Johansson, H., & McCloskey, E. (2008, Apr). FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporosis International, 19*(4), 385-397. https://doi.org/10.1007/s00198-007-0543-5

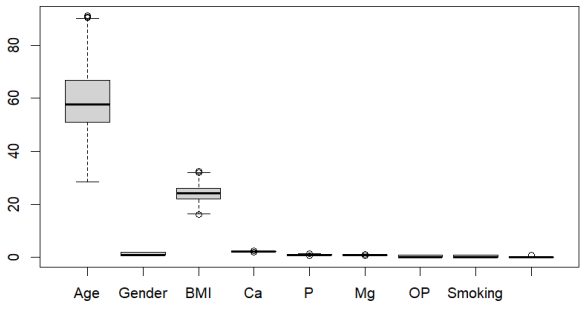
National Institute on Alcohol Abuse and Alcoholism. (2021). *Alcohol's Effects on the Body.* Retrieved May 3, 2023, from Alcohol's Effects on the Body | National Institute on Alcohol Abuse and Alcoholism (NIAAA) (nih.gov)

Bone Health & Osteoporosis Foundation. *What is osteoporosis and what causes it?* (2023, February 24). Retrieved May 3, 2023, from https://www.bonehealthandosteoporosis.org/patients/what-is-osteoporosis/#:~:text=Osteoporosis%20is%20a%20bone%20disease,bone%20looks%20like%20a%20honeycomb

Yao, X., Chen, L., Xu, H., & Zhu, Z. (2019). The Association between Serum Uric Acid and Bone Mineral Density in Older Adults. *International Journal of Endocrinology, 2020*. https://doi.org/10.1155/2020/3082318

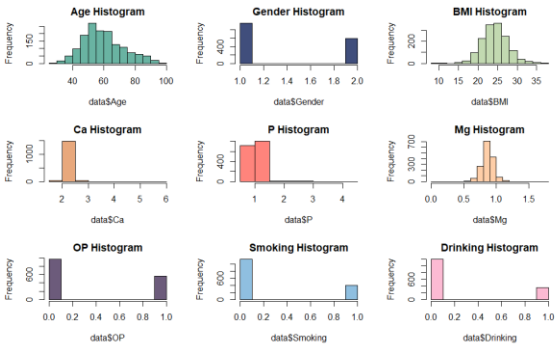
APPENDIX:

```
# Checking outliers after cleaning
```{r}
boxplot(data, outline = TRUE)
```
```



Histogram

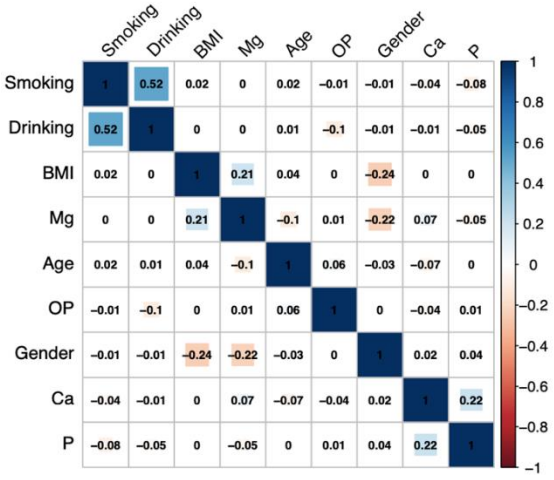
```
par(mfrow=c(3,3))
hist(data$Age, col = "#69b3a2", main="Age Histogram")
hist(data$Gender,col = "#404e7c", main="Gender Histogram")
hist(data$BMI,col = "#c2d9b1", main="BMI Histogram")
hist(data$Ca,col = "#e8a87c", main="Ca Histogram")
hist(data$P,col = "#ff847c", main="P Histogram")
hist(data$Mg,col = "#fecea8", main="Mg Histogram")
hist(data$OP,col = "#6c5b7b", main="OP Histogram")
hist(data$Smoking,col = "#8fbfe0", main="Smoking Histogram")
hist(data$Drinking,col = "#fcbad3", main="Drinking Histogram")
```



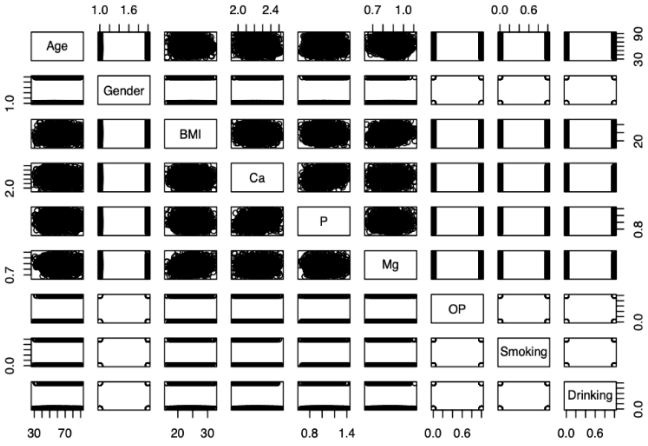

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(data), method = "square", order = "hclust",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.7)
```




```
pairs(data)
```




```
# Variable importance measures
varImpPlot(rf_model)
```