MAPÚA UNIVERSITY

SCHOOL OF HEALTH SCIENCES
In collaboration with
Arizona State University

Vital Signs Diagnosis Data

Baniqued, Dwayne Timothy[1], Loyola, Audrey Faith, R.[1]

[1]Students, BI120L, CON29, School of Health Sciences, Mapúa University

INTRODUCTION

Vital signs and physiological indicators, such as blood pressure, heart rate, Body Mass Index (BMI), glucose, and cholesterol levels, serve as fundamental metrics for assessing an individual's health status and identifying potential risks for various chronic diseases. The escalating global prevalence of non-communicable diseases, including hypertension, diabetes, and hypercholesterolemia, underscores the critical importance of understanding the intricate relationships between these physiological markers and lifestyle factors. Age, sex, smoking status, physical activity, stress levels, and sleep patterns are well-established determinants that collectively influence an individual's susceptibility to adverse health outcomes. Comprehensive health data analysis is crucial in public health surveillance and personalized medicine, as previous research has extensively documented the impact of these factors on cardiovascular health and metabolic function.

The analysis of large-scale health datasets provides invaluable insights into population-level trends, risk stratification, and the efficacy of health interventions. Statistical methodologies, including descriptive statistics, comparative analyses, and correlation studies, are indispensable tools for uncovering patterns, identifying significant differences between demographic groups, and quantifying the strength and direction of relationships between health variables. For instance, understanding sex-specific differences in metabolic parameters like glucose levels can inform tailored diagnostic approaches and preventive strategies. Similarly, elucidating the correlations between anthropometric measures like BMI and other vital signs can help predict disease progression and guide clinical management.

This investigation aims to contribute to the existing body of knowledge by conducting a comprehensive analysis of a vital signs diagnosis dataset. The study employs a rigorous data cleaning and preprocessing pipeline, including the conversion of categorical variables, removal of implausible data entries, and systematic outlier detection and removal, to ensure the integrity and reliability of subsequent analyses. Through descriptive statistics, the distributions and central tendencies of key physiological and lifestyle variables are characterized. Furthermore, various visualization techniques, such as histograms, bar plots, and scatter plots, are employed to visually explore the relationships between variables and identify potential sex-based disparities. Statistical tests, specifically a Welch Two Sample t-test, are utilized to formally assess differences in mean glucose levels between sexes, while a correlation matrix is computed to quantify the linear relationships among all numerical variables. The purpose of this investigation is to explore the distributions of vital signs and related health indicators, investigate sex-based differences in key physiological measures, and quantify the correlations between various demographic, lifestyle, and physiological parameters within the provided dataset.

Based on preliminary observations, it is hypothesized that female individuals will exhibit significantly higher mean glucose levels compared to male individuals within the study population. Furthermore, it is hypothesized that Body Mass Index (BMI) will be positively correlated with key physiological indicators such as systolic blood pressure, diastolic blood pressure, glucose levels, and cholesterol levels.

METHODS

The dataset utilized for this investigation was sourced from a CSV file titled "1_Vital_signs_diagnosis_data_Group_015.csv". Upon loading, initial data preprocessing steps were performed to ensure data quality and suitability for analysis. Numerical variables such as Systolic_BP, Diastolic_BP, and Hypertension were converted to numeric data types, with any non-numeric characters removed. Note that Hypertension was treated as a score ranging from 0 to 4. Categorical variables, specifically Sex and Smoking_Status, were converted into factor variables with appropriate levels and labels. Sex was recoded into "Female" (0) and "Male" (1), while Smoking_Status was categorized as "Non-Smoker" (0), "Occasional" (1), and "Chainsmoker" (2). Columns labeled as "Legend" were excluded from the analysis as they contained redundant information.

A filtering step was implemented to remove implausible BMI values, specifically those equal to or less than 10, as these were considered biologically unrealistic. To reduce the impact of extreme values on subsequent statistical analysis, an outlier removal function based on the interquartile range (IQR) method was applied to the Glucose_mg.dL, Cholesterol_mg.dL, and BMI variables. This approach removed any data points that fell more than 1.5 times the IQR below the first quartile or above the third quartile. The resulting dataset, referred to as df_clean, was used for all further analysis.

Descriptive statistics were calculated using the summary() and describe() functions to evaluate central tendencies, dispersion, and distribution characteristics. Visualizations, including histograms, bar plots, and box plots, were generated to assess distributions and group differences across sex categories. Scatter plots were constructed to investigate the linear relationships between BMI and other physiological variables such as blood pressure, glucose, and cholesterol. A Welch Two Sample t-test was conducted to determine if significant differences in mean glucose levels existed between males and females; this test was chosen for its robustness in the presence of unequal variances. Pearson's correlation coefficients were calculated to assess linear relationships among all numeric variables, and the corrplot package was employed to generate a visual correlation matrix. All analyses were conducted using the R statistical environment with the help of packages such as tidyverse, psych, GGally, and corrplot.

RESULTS AND FIGURES

The dataset, initially containing 1000 observations, underwent cleaning and preprocessing steps. This included converting Systolic_BP, Diastolic_BP, and Hypertension to numeric formats, and Sex and Smoking_Status to factor variables. Implausible BMI values (less than or equal to 10) were removed, and outliers in Glucose_mg.dL, Cholesterol_mg.dL, and BMI were removed using the interquartile range (IQR) method. After cleaning, the dataset contained 975 observations for most variables.

Descriptive statistics for the cleaned numerical variables are presented in Table 1. The mean age of participants was 55.16 years (SD = 21.43), ranging from 18 to 90 years. The mean BMI was 26.53 (SD = 4.94), with values ranging from 15.70 to 40.37. Systolic Blood Pressure had a mean of 133.47 (SD = 10.45), and Diastolic Blood Pressure had a mean of 81.89 (SD = 8.51). Glucose levels averaged 127.20 mg/dL (SD = 20.96), and Cholesterol levels averaged 188.90 mg/dL (SD = 32.92). The 'Daily_Sleeping_hours' variable exhibited a positive skewness of 0.75, suggesting a slight tail towards higher sleeping hours. There were 485 female participants and 490 male participants. Missing values were observed for Patient.ID (1), Age (3), Systolic_BP (4), Diastolic_BP (6), Hypertension (3), Smoking_Status (1), Physical_Activity_Hours_Week (1), and

Daily_Sleeping_hours (6). The Medication and Elevated.Risk columns contained all missing values.

Table 1. Descriptive Statistics for Cleaned Numerical Variables

Variable	n	Mean	SD	Median
ID	974	501.5	288.1	499.5
Age	972	55.2	21.4	56
Weight (kg)	975	62.7	9.1	63
Height (cm)	975	154.7	11.4	155
BMI	975	26.5	4.9	26.3
SBP	971	133.5	10.5	134
DBP	969	81.9	8.5	82
HTN (score)	972	3	1	3
HR (bpm)	975	99.3	18.7	99
PA (hrs/wk)	974	8	4.9	8
Stress (1–10)	975	5.2	2.2	5
Sleep (hrs)	969	5.4	1.2	5
Glucose	975	127.2	21	126
Cholesterol	975	188.9	32.9	190
Variable	**Min**	**Max**	**Skew**	**Kurt.**
ID	1	1000	0	-1.21
Age	18	90	-0.09	-1.2
Weight (kg)	43	84	0.06	-0.86
Height (cm)	130	175	-0.17	-0.88
BMI	15.7	40.4	0.32	-0.36
SBP	103	161	-0.05	-0.31
DBP	61	102	-0.06	-0.42
HTN (score)	0	4	-0.66	-0.48
HR (bpm)	50	151	0.1	-0.39
PA (hrs/wk)	0	16	0.04	-1.24
Stress (1–10)	1	10	0.12	-0.5
Sleep (hrs)	4	9	0.75	0.28
Glucose	70	184	0.2	-0.44
Cholesterol	103	274	0.06	-0.47

The distributions of BMI, Glucose_mg.dL, and Cholesterol_mg.dL by sex are illustrated in Figures 1, 2, and 3, respectively. Both female and male BMI distributions (Figure 1) appeared approximately normal, with a slight right skew for males. The distribution of Glucose_mg.dL by sex (Figure 2) also showed generally normal distributions for both groups, with females appearing to have a slightly higher concentration of values. Similarly, the distribution of Cholesterol_mg.dL by sex (Figure 3) displayed roughly normal distributions for both sexes.

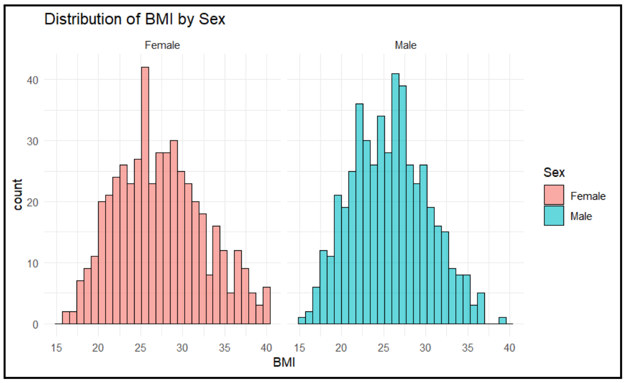

Figure 1: Distribution of BMI by Sex

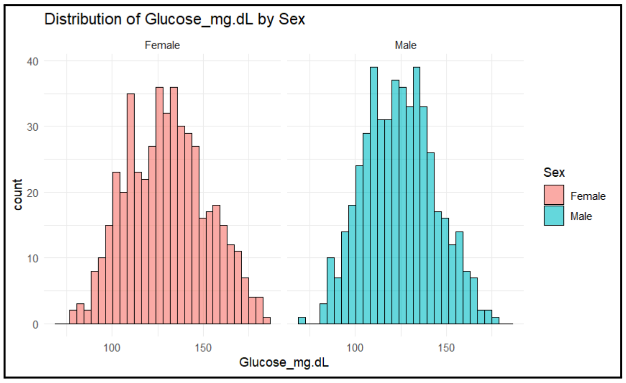

Figure 2: Distribution of Glucose by Sex

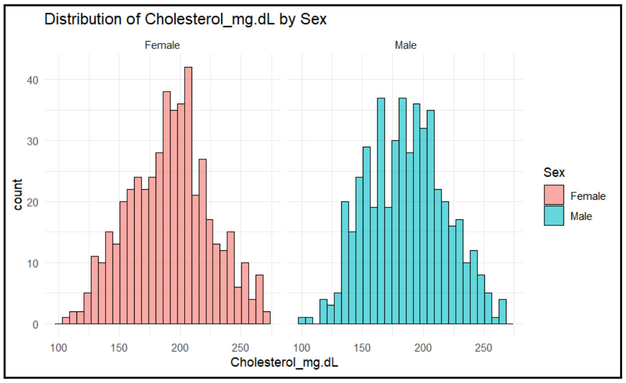

Figure 3: Distribution of Cholesterol by Sex

The frequency of smoking status categories is presented in Figure 4. "Chainsmoker" was the most frequent category with 347 individuals, followed by "Non-Smoker" with 326 individuals, and "Occasional" with 301 individuals.

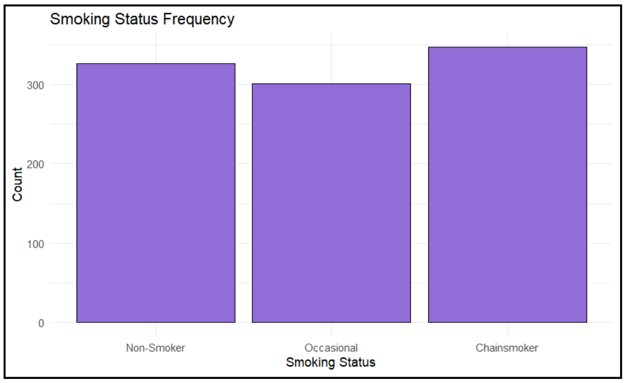

Figure 4: Smoking Status Frequency

Box plots further illustrated the distributions of key variables by sex (Figures 5, 6, and 7). For BMI (Figure 5), the median for females appeared slightly lower than that for males, though the interquartile ranges were similar. For Glucose_mg.dL (Figure 6), the median for females was visibly higher than for males, with comparable spreads. For Cholesterol_mg.dL (Figure 7), the medians for both sexes were comparable, and the distributions were similar.

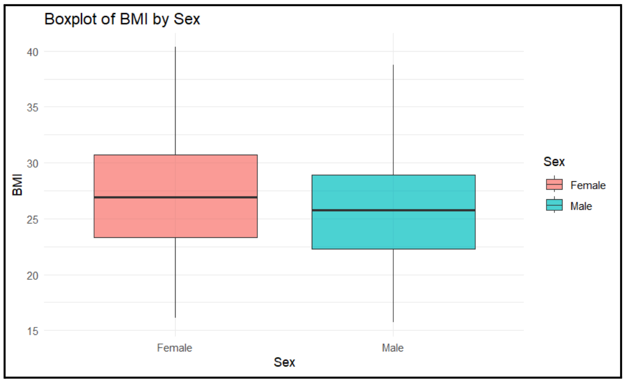

Figure 5. Boxplot of Body Mass Index (BMI) by Sex.

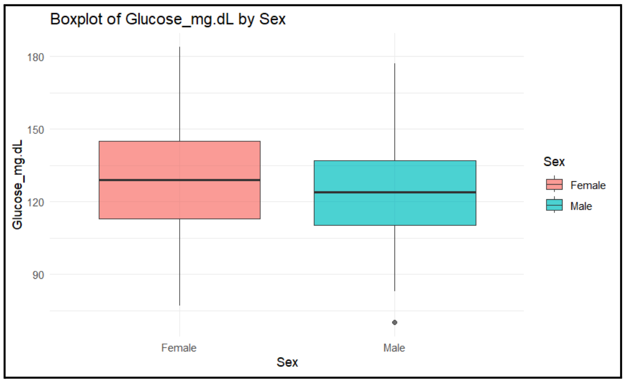

Figure 6. Boxplot of Glucose Levels by Sex.

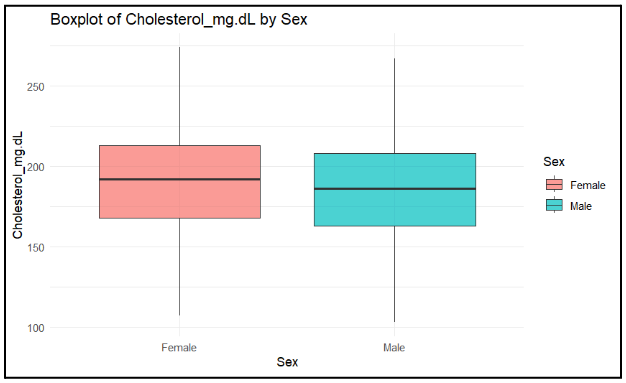

Figure 7. Boxplot of Cholesterol Levels by Sex.

Scatter plots exploring the relationship between BMI and other continuous variables (Systolic_BP, Diastolic_BP, Glucose_mg.dL, and Cholesterol_mg.dL) are presented in Figures 8-11. A consistent positive linear relationship was observed between BMI and Systolic_BP (Figure 8), Diastolic_BP (Figure 9), Glucose_mg.dL (Figure 10), and Cholesterol_mg.dL (Figure 11). This positive trend was consistent across both sexes, as indicated by the overlaid regression lines.

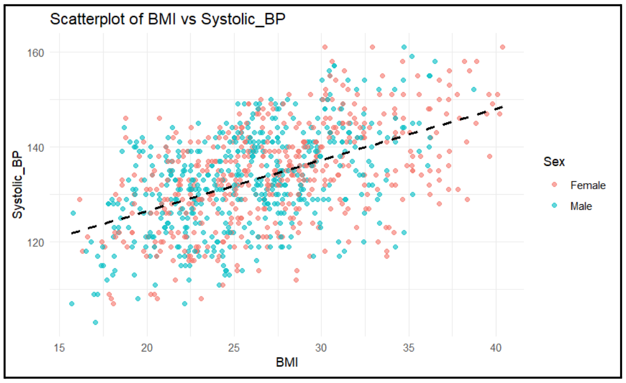

Figure 8. Scatterplot of BMI vs Systolic Blood Pressure (SBP).

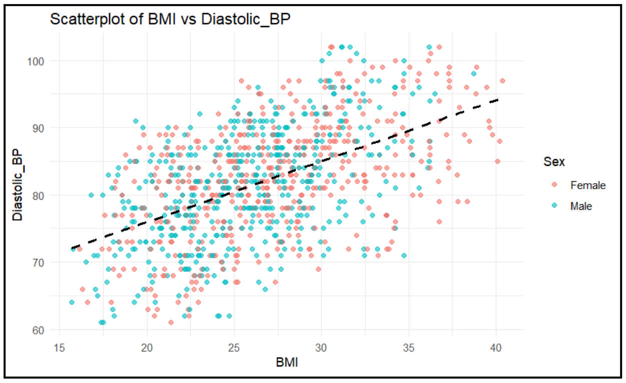

Figure 9. Scatterplot of BMI vs Diastolic Blood Pressure (DBP).

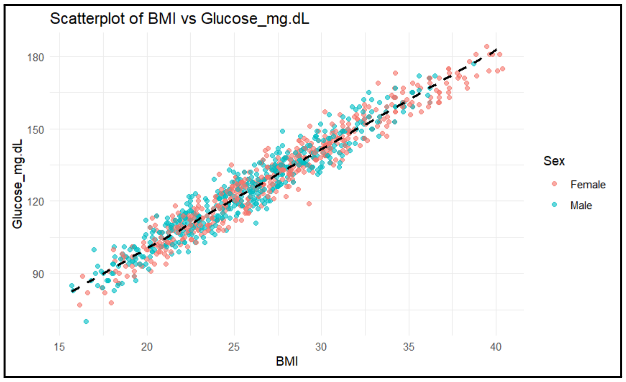

Figure 10. Scatterplot of BMI vs Glucose Levels.

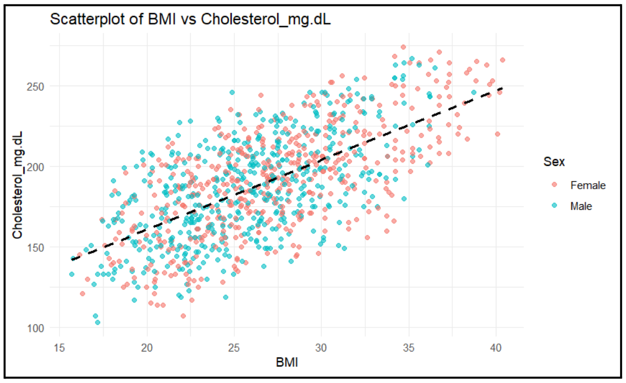

Figure 11. Scatterplot of BMI vs Cholesterol Levels.

A Welch Two Sample t-test was conducted to compare Glucose_mg.dL levels between sexes, with the results detailed in Table 2. The test indicated a statistically significant difference in mean glucose levels between females and males (t=4.4846, df=956.6, p-value = 8.192e-06). The mean glucose level for females was 130.20 mg/dL, while for males it was 124.23 mg/dL. The 95% confidence interval for the true difference in means between females and males was [3.36, 8.58].

Table 2. Welch Two Sample t-test.

```
Welch Two Sample t-test

data:  Glucose_mg.dL by Sex
t = 4.4846, df = 956.6, p-value = 8.192e-06
alternative hypothesis: true difference in means
between group Female and group Male is not equal to
0
95 percent confidence interval:
 3.356052 8.578684
sample estimates:
mean in group Female   mean in group Male
           130.2021            124.2347
```

A correlation matrix for all numeric variables was computed and is visually presented in Figure 12, with the exact Pearson correlation coefficients rounded to two decimal places. Strong positive correlations were observed between Systolic_BP and Hypertension (0.89), Systolic_BP and Diastolic_BP (0.79), Age and Cholesterol_mg.dL (0.70), and particularly between BMI and Glucose_mg.dL (0.97). Heart_rate showed a strong positive correlation with Stress_Level (0.82) and a strong negative correlation with Physical_Activity_Hours_Week (-0.57). Daily_Sleeping_hours exhibited a negative correlation with Age (-0.41) and Heart_rate (-0.41).

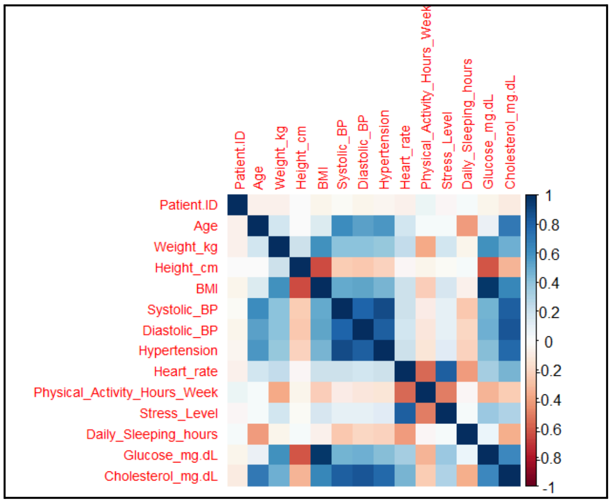

Figure 12. Correlation Matrix

INTERPRETATION AND CONCLUSION

This study achieved its primary objectives of analyzing the distribution of vital signs, assessing sex-based differences in physiological measures, and identifying significant inter-variable correlations. The hypothesis that females would have higher mean glucose levels than males was strongly supported by the statistical results, with a highly significant difference observed. This finding warrants further investigation into potential underlying biological or lifestyle factors contributing to this sex-specific difference.

The correlation analysis confirmed the second hypothesis, revealing strong positive associations between BMI and other key health indicators such as blood pressure, glucose, and cholesterol. The exceptionally strong correlation between BMI and glucose (0.97) suggests a very close relationship between these two variables within this dataset, reinforcing the well-established role of BMI as a major determinant of metabolic and cardiovascular risk. The observed relationships between age and several physiological parameters also align with known aging-related health trends, while the strong links between heart rate, stress level, and physical activity reflect expected physiological responses.

This study utilized a cross-sectional design, which limits the ability to infer causality or observe temporal patterns among the variables. Additionally, the dataset contained some missing values and the specific population from which the data was drawn might influence the generalizability of these findings.

Several recommendations arise from these results. First, further investigation into the mechanisms underlying the observed sex-specific differences in glucose levels is warranted, which could include analyses of hormonal profiles, dietary intake, or genetic markers. Second, future studies should adopt longitudinal designs to better understand causality and temporal patterns among the observed variables. Third, employing multivariate regression modeling would provide insights into the independent effects of each predictor while adjusting for confounders. Fourth, clinical practices may benefit from utilizing BMI as a screening tool to flag potential risks, given its strong associations with multiple vital signs. Lastly, addressing missing data through advanced imputation strategies could improve data completeness and analytical robustness in future work.