INFO-B518 APPLIED STATISTICAL METHODS FOR BIOMEDICAL INFORMATICS

The Effects of Lifestyle and Hormonal Factors on PCOS

Instructor: Gopikrishnan Chandrasekharan

Team members:
Architha Singamsetty
Jamsetta Quiqui
Treyden Stansfield
Veera Venkata Satyavathi Surapureddy.

Abstract

Polycystic Ovary Syndrome (PCOS) is a common hormonal disorder in women, often leading to infertility and other health complications. Despite its prevalence, the causes remain unclear, making early detection critical. This study investigated the relationship between lifestyle factors (BMI, weight gain, fast food consumption, exercise) and hormonal markers (progesterone levels, blood pressure) to identify predictors of PCOS. Using a dataset of 2,000 women from Kaggle, we applied logistic regression and chi-squared tests to analyze the variables.

The results revealed that weight gain, frequent fast food consumption, and regular exercise significantly increased the risk of developing PCOS, with coefficients of 1.63, 1.33, and 0.675, respectively. Higher progesterone levels were associated with reduced risk, with a coefficient of –0.819. Chi-squared tests further confirmed the strong association between PCOS and categorical lifestyle factors.

These findings suggest that weight-related lifestyle factors are more influential predictors of PCOS than hormonal markers, highlighting the importance of monitoring weight gain and dietary habits for early intervention. This research could improve screening protocols and inform preventive care strategies for women at risk of PCOS.

Keywords: PCOS, Polycystic Ovary Syndrome, Lifestyle factors, Hormonal factors, Weight gain, Progesterone, Fast food consumption, Exercise habits, Logistic regression, Chi-squared test.

Introduction:

Polycystic Ovary Syndrome (PCOS) is a prevalent hormonal disorder in women, characterized by delayed menstrual cycles, painful ovarian cysts, and fertility challenges. It results from elevated levels of male hormones, known as androgens, in reproductive-age women (Mayo Clinic, 2022). PCOS impacts approximately 13% of the global female population and is a leading cause of infertility, often accompanied by significant mental and physical health complications (World Health Organization, 2023). While early diagnosis and intervention are more effective than delayed treatment, the exact etiology of PCOS remains uncertain (Rashid et al., 2022). Identifying risk factors is therefore critical to improving early detection and management strategies.

This study seeks to explore the lifestyle and hormonal factors associated with an increased risk of developing PCOS. Using a dataset from Kaggle comprising 2,000 female patients with documented PCOS or negative status, as well as detailed lifestyle, physiological, and hormonal data (Divya, 2022), the research aims to identify key predictors of PCOS. The findings are intended to inform strategies for faster diagnosis and intervention, though the study is limited by its focus on a subset of lifestyle and hormonal variables.

Data Description:

Our dataset is a set of 2000 women's medical records from PCOS patients and non-PCOS patients. It contains a variety of lifestyle and physiological datapoints including categorical and continuous data points (Divya, 2022). This dataset was taken from Kaggle (https://www.kaggle.com/datasets/cm037divya/pcos-dataset).

The variables analyzed in this study include PCOS status (categorical, dependent variable), BMI (continuous), the ratio of Follicle-Stimulating Hormone (FSH) to Luteinizing Hormone (LH) levels (continuous), systolic and diastolic blood pressure (continuous), resting blood sugar (RBS) (continuous), progesterone (PRG) levels (continuous), weight gain (categorical), fast food consumption (categorical), and regular exercise (categorical).

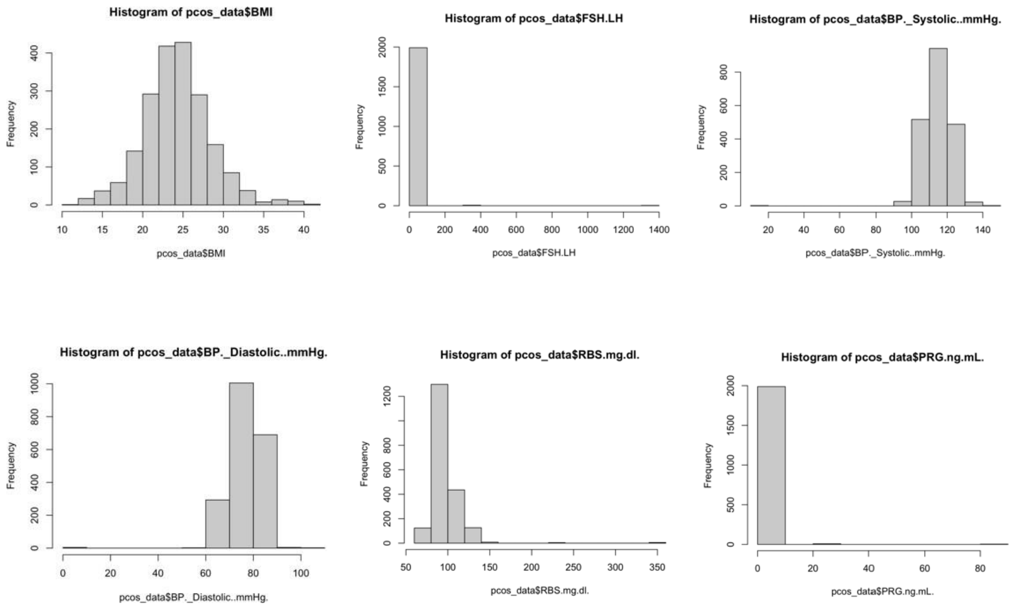
The lifestyle-related variables like BMI**,** weight gain (Yes/No)**,** fast food consumption (Yes/No)**,** regular exercise (Yes/No), and blood pressure were included to examine the impact of lifestyle choices on PCOS risk. The hormonal variables RBS**,** FSH/LH ratio, and PRG levels were chosen for their physiological relevance. Elevated FSH/LH ratios are associated with PCOS, making it an essential variable to study (Saadia, 2020). Progesterone, a critical female hormone, was selected due to its association with androgen levels. Additionally, resting blood sugar levels provide an important connection between dietary habits and hormonal regulation, offering insights into the interplay of lifestyle and hormonal factors.

Preprocessing was performed to ensure data quality and suitability for analysis. Categorical variables such as weight gain, fast food consumption, regular exercise, and PCOS status were converted into factors using R. Continuous variables had no missing values and thus did not require imputation. The hist function in R was utilized to visualize the distribution of variables.

Exploratory analysis revealed that BMI followed a normal distribution, while the FSH/LH ratio**,** RBS, and PRG levels exhibited right-skewed distributions with outliers. Both systolic and diastolic blood pressure were left-skewed. Outliers in the continuous variables were identified and removed, reducing the dataset size from 2,000 to 1,951 records. Despite this reduction, the dataset maintained a sufficiently large sample size for robust analysis. After preprocessing, the data distributions were normalized, ensuring that the variables were appropriate for the subsequent statistical methods.

Figure 1

Visualizations of the Continuous Variables

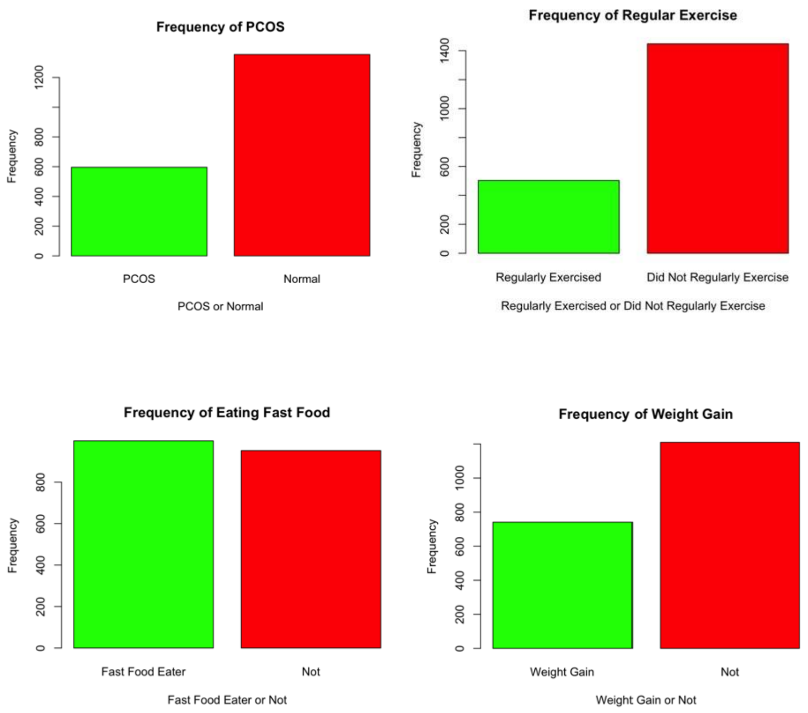


Note: From left to right the variables are: BMI, FSH/LH, BP Systolic, BP Diastolic, RBS levels, and PRG levels.

Next, categorical variables were visualized using bar graphs, as shown in Figure 2. The data reveal that the number of PCOS patients is approximately half that of non-PCOS patients. The majority of patients reported not engaging in regular exercise. Additionally, there was an even distribution between individuals who frequently consumed fast food and those who did not. Lastly, approximately twice as many patients reported no significant weight gain compared to those who experienced weight gain.

Figure 2

Distributions of Categorical Variables



Note: From left to right the variables are: PCOS status, Regular Exercise (Y/N), Frequent Fast-Food Consumption (Y/N), and Weight Gained (Y/N).

Statistical Methods:

Logistic regression was employed in our study to identify variables significantly contributing to the probability of developing Polycystic Ovary Syndrome (PCOS). PCOS status, a binary variable (Yes/No), was modelled as the dependent variable, while lifestyle and hormonal factors served as predictors. Logistic regression was chosen because it is specifically designed to handle binary outcomes and provides odds ratios that quantify the strength and direction of associations between predictors and the outcome. This makes it well-suited for assessing risk factors and their relative contributions to PCOS.

In addition to logistic regression, chi-squared tests were used to examine the relationship between individual categorical variables and PCOS status. These tests evaluated whether significant associations existed, with the null hypothesis stating no relationship between the variables and PCOS status, and the alternative hypothesis indicating a significant relationship. A Bonferroni correction was applied to account for multiple comparisons, adjusting the significance threshold to reduce the likelihood of Type I errors.

The assumptions of logistic regression were thoroughly assessed. The absence of multicollinearity among predictors, tested using the R package corrplot, confirmed no significant collinearity, supporting the model's validity, we performed the variance inflation factor (VIF) analysis. The chi-squared tests assumed that both variables were categorical and that expected frequencies met the required thresholds, ensuring reliable results.

The combination of logistic regression and chi-squared tests provided robust insights into the relationships between lifestyle, hormonal factors, and PCOS risk, supporting the study’s objectives.

Results:

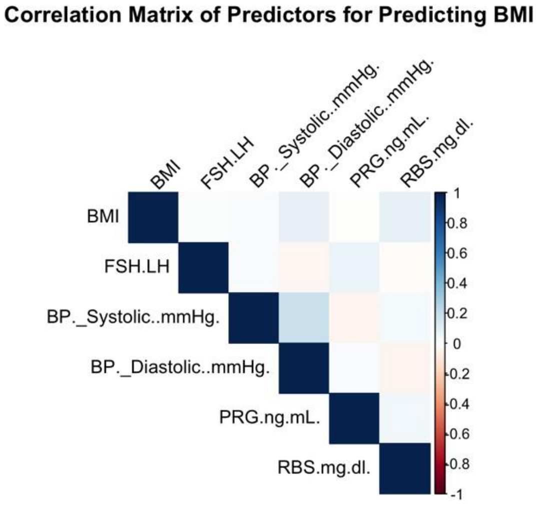
The logistic regression analysis yielded a model with an AIC of 1899.5. Significant predictors included Weight.gain.Y.N. Yes**,** Fast.food..Y.N. Yes**,** Reg.Exercise.Y.N. Yes, and PRG.ng.mL. The coefficient for Weight.gain.Y.N. Yes was 1.63, indicating that individuals reporting weight gain had a 163% higher likelihood of developing PCOS. Fast.food..Y.N. Yes had a coefficient of 1.33, suggesting that frequent fast food consumption increased the risk of PCOS by 133%. Reg.Exercise.Y.N. Yes exhibited a coefficient of 0.675, corresponding to a 68% higher risk, while PRG.ng.mL demonstrated a protective effect with a coefficient of –0.819, indicating an 82% reduction in risk per unit increase in progesterone levels.

To evaluate categorical predictors, three chi-squared tests were performed, with a Bonferroni correction applied to account for multiple comparisons. Using an alpha level of 0.05, the corrected p-critical value was calculated as 0.0167. The first chi-squared test, comparing weight gain (Yes/No) and PCOS status, yielded an X^2 value of 370.17 and a p-value of < 2.2e-16, rejecting the null hypothesis of no association. The second test, analyzing fast food consumption (Yes/No) and PCOS status, returned an X^2 value of 293.69 and a p-value of < 2.2e-16, similarly rejecting the null hypothesis. The third test, examining regular exercise (Yes/No) and PCOS status, produced an X^2 value of 8.9703 and a p-value of 0.002, also rejecting the null hypothesis. As all p-values were below the corrected threshold, significant associations were confirmed for all three variables.

Logistic regression relies on two primary assumptions: the dependent variable must be binary, and continuous predictors must exhibit no multicollinearity. These assumptions were verified using the corrplot package in R (Wei & Simko, 2024). The correlation plot results, presented in Figure 3, indicated no significant multicollinearity among the variables, confirming that the model's assumptions were met.

Figure 3

Multicolineraity Test



Note: The legend on the right is the amount of linearity between the two variables.

Our dataset successfully passed the multicollinearity test, ensuring that all assumptions for logistic regression were satisfied. No significant linear relationships were observed between any continuous variables, validating the model's suitability. Additionally, the chi-squared test, a non-parametric statistical method, does not require any specific data distribution assumptions. Its primary requirement, that both variables being analyzed are categorical was fully met in this study.

Discussion:

The logistic regression analysis revealed that the variable Weight.gain.Y.N. Yes had the largest coefficient of 1.63, indicating that weight gain increases the likelihood of developing PCOS by 163%. The second largest coefficient was for Fast.food.Y.N. Yes, with a value of 1.33, meaning that consuming fast food regularly increases the likelihood of developing PCOS by 133%. The coefficient for PRG.ng.mL was –0.819, suggesting that an increase in progesterone levels reduces the probability of developing PCOS by 82%. The smallest coefficient was for Reg.Exercise.Y.N. Yes, at 0.675, indicating that regular exercise is associated with a 68% higher likelihood of developing PCOS.

These findings suggest that lifestyle factors, particularly weight gain, play a more significant role in predicting PCOS risk compared to hormonal factors, as evidenced by the larger coefficients and significance of lifestyle-related variables. This conclusion is further supported by the results of three chi-squared tests, which demonstrated significant associations between PCOS status and each lifestyle factor. This finding is consistent with existing literature, which has shown that higher progesterone levels are associated with a reduced risk of PCOS, as women with higher progesterone tend to exhibit lower levels of androgens.

The positive association between regular exercise and PCOS risk warrants further investigation. This may seem counterintuitive, given that regular exercise is generally known to reduce the risk of PCOS, while the other variables suggest that unhealthy behaviors such as weight gain and poor diet increase risk. It is possible that women who gain weight may begin exercising as a response, making exercise a correlate of weight gain rather than an independent risk factor for PCOS. This hypothesis aligns with previous research linking weight gain to an increased risk of PCOS (Barber et al., 2019).

Conclusion:

PCOS is a life-upsetting and painful disease which causes problems for many women across the globe. With no known cure and no definite cause, the best method of treatment is early lifestyle and pharmaceutical intervention. Our research has illuminated 3 key lifestyle risk factors for PCOS: weight gain, frequent fast food consumption, and regular exercise with higher progesterone levels having a protective effect.

We hope that care providers will use this information to screen for PCOS as weight gain starts occurring, for a faster and more effective intervention. Still more research will need to be done to see if other factors also impact PCOS risk, and to figure out if there are any conditional relationships between these variables. In conclusion, it seems that lifestyle factors related to weight gain are a better predictor for PCOS than hormonal factors.

References

Divya, C. M. (2022). PCOS Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/cm037divya/pcos-dataset

Barber, T. M., Hanson, P., Weickert, M. O., & Franks, S. (2019). Obesity and Polycystic Ovary Syndrome: Implications for Pathogenesis and Novel Management Strategies. Clinical medicine insights. Reproductive health, 13, 1179558119874042. https://doi.org/10.1177/1179558119874042

Mayo Clinic. (2022, September 8). Polycystic ovary syndrome (PCOS) - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/pcos/symptoms-causes/syc-20353439

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rashid, R., Mir, S. A., Kareem, O., Ali, T., Ara, R., Malik, A., Amin, F., & Bader, G. N. (2022). Polycystic ovarian syndrome-current pharmacotherapy and clinical implications. Taiwanese journal of obstetrics & gynecology, 61(1), 40–50. https://doi.org/10.1016/j.tjog.2021.11.009

Saadia Z. (2020). Follicle Stimulating Hormone (LH: FSH) Ratio in Polycystic Ovary Syndrome (PCOS) - Obese vs. Non- Obese Women. Medical archives (Sarajevo, Bosnia and Herzegovina), 74(4), 289–293. https://doi.org/10.5455/medarh.2020.74.289-293

Wei, T., & Simko, V. (2024). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.95). https://github.com/taiyun/corrplot

World Health Organization. (2023, February 15). Polycystic ovary syndrome.

https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome