Data 698: Project Proposal
Leticia Salazar
February 28, 2023

Using predictive analytics to diagnose Polycystic Ovary Syndrome (PCOS)

Introduction:

Polycystic Ovary Syndrome (PCOS) is a reproductive hormonal imbalance disorder affecting women of reproductive age. The exact number of women affected by PCOS worldwide is difficult to determine as many cases go undiagnosed but according to the World Health Organization (WHO), there's 3.4% of women who are affected [1]. It may not seem as much, but women make up 49.7% of the today's population [2], almost 13% of those women are of reproductive age [3], leaving about 17.5 million women who reported to suffer from PCOS. Aside from the toll PCOS takes on your physical and emotional health, it causes problems in the ovaries, making it difficult for women to have a healthy menstrual cycle leading to the development of cysts and infertility. It's important to note that the prevalence of PCOS varies by region and ethnic groups, with some studies suggesting higher rates of PCOS in certain populations [4]. This project will investigate publicly available datasets that will allow us to create predictive models on markers in routine test results to make a diagnosis.

Literature Review:

The following articles will be resourceful for this project:

1. **Racial and ethnic differences in the metabolic response of polycystic ovary syndrome:**
This article speaks on the racial and ethnic disparities in the metabolic dysfunction suffered by PCOS and whether markers of metabolic function differ in nondiabetic Asian American (AS), African American (AA), Hispanic White (HW), compared to non-Hispanic White (NHW) women with PCOS.

2. **A review: Brief insight into Polycystic Ovarian syndrome:**
This review highlights a brief overview of risk and pathophysiological treatment with drugs acting on ovulation, infertility plus clinical symptoms of PCOS.

3. **DHEA, DHEAS and PCOS:**
This article examines the effect of excess adrenal precursor androgen (APA) production has on women with PCOS. The extra-adrenal factors, including obesity, insulin and glucose levels, and ovarian secretions, play a limited role in the increased APA production observed in PCOS.

Datasets:

The following datasets will be used in this project:

1. <u>DataSet for PCOS</u>:
   These datasets provide the results of an untargeted metabolomic survey was conducted on the metabolites in the FF of 35 patients with PCOS and 37 age-matched individuals as control
2. <u>Polycystic Ovary Syndrome (PCOS)</u>:
   PCOS dataset contains all physical and clinical parameters of patients from 10 different hospitals across Kerala, India.
3. <u>Menstrual Cycle Data:</u>
   Randomized Comparison of Two Internet –Supported Natural Family Planning Methods.

Methodology:

This project hopes to predict a PCOS diagnosis using logistic regression models using R. First, load the datasets and perform an exploratory data analysis on each to identify any statistical information that can be insightful to the research. Starting with the *DataSet for PCOS* to visualize the differences between the controlled group and the patients with PCOS. Then compare these results with the *Polycystic Ovary Syndrome (PCOS)* dataset I obtained from my previous project. The *Menstrual Cycle Data* will be used to view differences in fertility option woman from different race / ethnic background have. There is still one more dataset I want to analyze and have yet to find; ideally this dataset has information on women with and without PCOS from different racial/ethnic backgrounds worldwide or from the United States. After having performed EDA on these datasets, I'd then create logistic regression models to evaluate the likeliness of predicting a PCOS diagnosis based on particular markers in test results from patients.

Objectives:

I anticipate answering the following questions with my data:

1. Are there commonalities women with and without PCOS have that can be easily dismissed as normal?
2. Are there differences for women of different race/ethnic background when it comes to having PCOS? What about women without PCOS?
3. What is the likelihood of a woman developing PCOS based on her age, ethnicity, and BMI history?
4. Can we predict the risk of insulin resistance, diabetes, and cardiovascular disease in women with PCOS based on their medical history, hormone levels, and lifestyle factors?
5. Can we predict the likelihood of successful pregnancy outcomes in women with PCOS based on their age, weight, hormone levels, and treatment history?
6. Can we predict the long-term health outcomes and quality of life of women with PCOS based on their age, lifestyle factors, hormone levels, and treatment history?

Assumptions:

While there's limited information available in the medical field and even less data sets available to analyze, I have some concerns on being successful in predicting a PCOS diagnosis.

References:

1. Bulsara, J., Patel, P., Soni, A., & Acharya, S. (2021, February 10). A review: Brief insight into polycystic ovarian syndrome. Endocrine and Metabolic Science. Retrieved February 23, 2023, from https://www.sciencedirect.com

2. *World female population, 1960-2022*. Knoema. (2022). Retrieved February 24, 2023, from https://knoema.com.

3. MarchofDimes. (2022). *Population of women 15-44 years by age: United States, 2020*. March of Dimes | PeriStats. Retrieved February 24, 2023, from https://www.marchofdimes.org

4. Engmann, L., Jin, S., Sun, F., Legro, R. S., Polotsky, A. J., Hansen, K. R., Coutifaris, C., Diamond, M. P., Eisenberg, E., Zhang, H., Santoro, N., & Reproductive Medicine Network (2017). Racial and ethnic differences in the polycystic ovary syndrome metabolic phenotype. American journal of obstetrics and gynecology, 216(5), 493.e1–493.e13. https://doi.org

5. Fehring, Richard J., "Menstrual Cycle Data" (2012). *Randomized Comparison of Two Internet-Supported Methods of Natural Family Planning*. 7. https://epublications.marquette.edu

6. Khan, M. J., Ullah, A., & Basit, S. (2019). Genetic Basis of Polycystic Ovary Syndrome (PCOS): Current Perspectives. The application of clinical genetics, 12, 249–260. https://doi.org/