Data 698: Project Proposal
Leticia Salazar
February 28, 2023

Using Machine Learning Algorithm to Predict the likelihood of PCOS based on Demographic, Clinical and Lifestyle Factors

Introduction:

Polycystic Ovary Syndrome (PCOS) is a 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 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. DataSet for PCOS:
   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. Polycystic Ovary Syndrome (PCOS):
   PCOS dataset contains all physical and clinical parameters of patients from 10 different hospitals across Kerala, India.
3. Menstrual Cycle Data:
   Randomized Comparison of Two Internet –Supported Natural Family Planning Methods.

Methodology:

This project will investigate mainly publicly available datasets except for one dataset from NICHD Dash that will be used to create predictive models on markers in routine test results to make a diagnosis. Some variables that are included in these datasets are:

- Age
- Weight
- BMI
- Race/ethnicity
- Family history of PCOS
- Menstrual cycle irregularity
- Hormone levels (e.g., testosterone, LH, FSH)
- Insulin resistance
- Physical activity level
- Diet

I will be using R (statistical performing language) to perform exploratory data analysis to process and analyze the data to check for structural errors and be able to create graphs and perform tests with minimal errors. Once the data is ready to use, I will be splitting the data into a training and test set to be able to use a machine learning algorithm such as logistic regression to create a predictive model. The predictive model will be based on markers (variables mentioned above) used to identify individuals who are at high risk for PCOS and target interventions to manage the condition.

To answer my research question:

1. The datasets publicly available and the NICHD Dash looking to obtain all have PCOS and non-PCOS patients including demographic information, medical history, and laboratory tests. Preprocess the data by removing missing values, outliers, and redundant variables. Perform feature selection to identify the most informative variables for prediction.

2. Split the dataset into training, validation, and testing sets. The training set is used to train the machine learning algorithm, the validation set is used to tune hyperparameters and prevent overfitting, and the testing set is used to evaluate the performance of the final model.
3. Select an appropriate machine learning algorithm for the task at hand, such as logistic regression, decision trees, random forests, support vector machines, or neural networks. Train the algorithm on the training set using various techniques, such as cross-validation and regularization, to optimize its performance.
4. Evaluate the performance of the trained model on the validation set using various metrics, such as accuracy, precision, recall, F1 score, and area under the curve. Use feature importance analysis to identify the most influential variables for prediction.
5. Tune the hyperparameters of the machine learning algorithm using grid search, random search, or Bayesian optimization to improve its performance on the validation set.
6. Select the final model based on its performance on the validation set. Evaluate its performance on the testing set to assess its generalization ability.
7. Interpret the results of the machine learning algorithm using various techniques, such as decision trees, feature importance analysis, and partial dependence plots. Visualize the results using graphs, charts, and heatmaps to facilitate understanding and communication.
8. Deploy the trained model on new data and disseminate the findings through scientific publications, presentations, and online platforms.

Note: this methodology plan is not exhaustive and may vary depending on the specific research question, dataset, and machine learning algorithm used.

Note 2: data has already been collected, there is no need for me to gather participants, perform exams (such as bloodwork), use medical equipment to collect the data, perform surveys, have a location to perform a study, etc. I will be the sole person studying the data set and conducting the analysis.

Objectives:

My objective for this project is to predict a PCOS diagnosis using machine learning algorithms like logistic regression on markers presented on bloodwork. Machine learning algorithms have shown promise in advancing our understanding of the disease and improving its diagnosis and treatment.

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?

Note: this methodology plan is not exhaustive and may vary depending on the specific research question, dataset, and machine learning algorithm used.

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. Yet there are justifications exploring PCOS in depth:

1. Identify diagnostic biomarkers that can distinguish PCOS patients from healthy individuals or those with other disorders. These biomarkers can aid in earlier diagnosis and better management of the disease.
2. Predict the likelihood of disease progression and the risk of developing complications, such as diabetes and cardiovascular disease, in PCOS patients. This information can guide treatment decisions and improve patient outcomes.
3. Develop personalized treatment plans for PCOS patients based on their individual characteristics and medical history. This approach can lead to more effective and targeted interventions.
4. Integrate data from various sources, such as electronic health records, imaging studies, and genetic analyses, to provide a more comprehensive understanding of PCOS. This can help identify new pathways involved in the disease and potential targets for therapy.
5. Aid in the design and analysis of clinical trials, leading to more efficient and informative studies. This can accelerate the development of new treatments for PCOS.

Early diagnosis and management of PCOS can lead to better health outcomes, improved quality of life, and reduced long-term health risks. Therefore, predicting PCOS diagnosis can have several societal benefits, including:

1. Predicting PCOS diagnosis can help healthcare providers identify women at risk of developing PCOS and intervene early with appropriate treatment, such as lifestyle modifications and medication, to prevent or minimize the long-term health consequences of the disorder.
2. Early diagnosis and treatment of PCOS can help manage symptoms such as irregular periods, infertility, acne, and excess hair growth, leading to improved physical and mental health outcomes for affected women.
3. By predicting PCOS diagnosis and intervening early, healthcare providers can prevent or reduce the need for more expensive treatments or surgeries later in life, resulting in cost savings for individuals, healthcare systems, and society.

4. Predicting PCOS diagnosis can increase awareness of the disorder among healthcare providers, patients, and the public, leading to more education, research, and advocacy efforts aimed at improving PCOS diagnosis, treatment, and management.
5. Early intervention and management of PCOS can improve the quality of life for affected women, leading to increased productivity, better mental health, and greater overall well-being.

Overall, I'll be able to explore the insights into PCOS pathophysiology, diagnosis, and treatment. Their use in PCOS research can lead to more personalized and effective care for patients with this complex disorder.

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/