Name: Kelsi Dial
Course: MSBD 566
Assignment: Midterm Project Report
Title: Anemia Severity Among Reproductive-Age Women Using NHANES Data

Project Description

The goal of this project is to examine anemia severity among reproductive-age women using publicly available data from the National Health and Nutrition Examination Survey (NHANES). Anemia is a common and clinically important condition in this population and is often associated with gynecologic factors such as heavy menstrual bleeding and uterine fibroids.

Although direct fibroid diagnosis variables are not available in recent NHANES cycles, anemia represents a measurable downstream outcome that is clinically relevant and well captured in the dataset. This project aims to explore demographic patterns in anemia prevalence and assess the feasibility of predictive modeling using available features.

The focus has been on data acquisition, cleaning, exploratory data analysis, and defining appropriate modeling strategies.

Data Description

Data for this project were obtained from the **NHANES 2017–2018** cycle, a nationally representative, cross-sectional survey conducted by the Centers for Disease Control and Prevention.

Two primary datasets have been used:

1. **Demographics (DEMO_J)**: includes age, sex, and race/ethnicity

2. **Complete Blood Count (CBC_J)**: includes laboratory measurements such as hemoglobin

The datasets were merged using the participant identifier **SEQN**. The analytic sample was restricted to female participants between the ages of 18 and 50 to represent reproductive-age women. Participants with missing hemoglobin values were excluded.

NHANES data were accessed via the CDC NCHS website:
https://www.cdc.gov/nchs/nhanes/

Preliminary Data Preparation and Exploration

Initial data cleaning steps included:

1. Filtering by sex and age
2. Handling missing laboratory values
3. Creating a binary anemia indicator based on hemoglobin < 12 g/dL

Exploratory data analysis revealed that hemoglobin levels in the sample follow an approximately normal distribution with a left tail corresponding to individuals with low hemoglobin values. Preliminary results suggest that anemia affects a meaningful minority of reproductive-age women in the dataset, rather than being a rare outcome.

These exploratory findings support using anemia as a viable outcome for further modeling and analysis.

Proposed Methods and Analysis Plan

The planned analytical approach includes:

1. Logistic regression as a baseline predictive model
2. More flexible machine learning methods (e.g., neural networks) to capture potential nonlinear relationships

Predictor variables at this stage include age and race/ethnicity. These variables were selected due to their established associations with anemia risk and their relevance to gynecologic health disparities.

Model evaluation will focus on classification performance metrics such as accuracy and confusion matrices. At the midpoint of the project, model implementation is still in progress.

Limitations and Next Steps

Several limitations have been identified at this stage:

1. NHANES data are cross-sectional, limiting causal inference
2. Direct measures of fibroid diagnosis and menstrual bleeding are unavailable
3. The current predictor set is limited to demographic variables

Next steps for the project include completing predictive modeling, evaluating model performance in greater detail, and interpreting results in the context of missing clinical variables. Additional emphasis will be placed on understanding the limitations of demographic-only predictors for anemia.

Summary

At the midpoint of the project, data acquisition, cleaning, and exploratory analysis have been completed, and a clear modeling plan has been established. The preliminary findings demonstrate that anemia is a clinically meaningful outcome in this population and that further modeling is warranted.