**Project: A random forest approach to evaluating risk factor trends on obesity among U.S. youth with cardiovascular disease**

**Biostatistical Mentors: Samantha Bothwell, Emily Cooper**

**Clinical Collaborator: Jill Kaar**

**Motivation:**

Primary risk factors for cardiovascular disease (CVD) in youth are sociodemographics, low physical activity, high sedentary behavior, high screen usage, and insufficient sleep. These risk factors are also prevalent among youth classified as obese and those with diagnosed mental health conditions. The goal of this project is to assess the relationship among known risk factors for CVD and obesity by using machine learning methodology to identify patterns that are predictive of obesity among U.S. youth with CVD.

This project will leverage a large sample of parent-reported data on youth health and behavior obtained as part of the National Survey of Children’s Health from 2018-2019. This project will utilize random forest, a machine learning method, to predict obesity classification in children aged 13-18 with CVD, based on sociodemographics, screen usage, physical activity, sleep, and diagnosed mental health conditions. In addition to obesity class prediction, random forest provides interpretable estimates of variable importance and partial dependence associations for each predictor, which will aid in identifying leading risk factors for obesity among this population.

Covered topics

* Data Wrangling
* Machine Learning
* Random Forest
* Data Visualization