**For final project proposal**

**Team Hungary – Sri Lanka**

**Members:** Choi, Jane, Hee, Da, Mestre, Luis Miguel, Watson, Katharine, & Xu, Xiao

**Dataset:** NHANES I Epidemiological Follow-up Study

The National Health and Nutrition Examination Survey (NHANES) I Epidemiological Follow-up Study (NHEFS) dataset includes non-institutionalized civilian United States aged 50-74 years old. The cohort was started in 1972 and 96% of the cohort was traced to the latest, 1996, follow-up. This dataset includes *demographics variables* including the biological sex of participant, age, education, marital status, annual income for the household, poverty ratio, and many other demographic variables. The NHANES also contains *physical variables* such as body weight, length, height, BMI, pulse rate, systolic and diastolic blood pressure; *health variables* including total cholesterol, diabetes status, urine flow, pregnancy, sleep of hours at night and *lifestyle variables* as alcohol consumption, smoking consumption, physical activity, marijuana use, use of hard drugs, sexual activity, and number of sexual partners.

2) **Goal:** To measure the association between diabetes and obesity. BMI will be the measured obesity variable. We will assume diabetes mellitus (from now referred to as “diabetes”) is a time-to event outcome and a dichotomous variable. The Longitudinal NHANES does not classify diabetes as type I or II, so we will not make this distinction.

3) **Approach:**

* Explore which variables in the NHANES dataset might influence this association.
* We aim to demonstrate how EDA can be applied in an epidemiologic study/analysis.
* Our outcome will be diabetes and our exposure will be BMI. BMI will represent the different categories of body mass from underweight to obesity.
* Compare logistic regression analysis (marginal and conditional Odds Ratios (OR)) with survival analysis regression (Hazard Ratios (HRs) and conditional HR) and how those interpretations might change. We will show and interpret any observed limitations based on our outcomes (e.g., when the disease is not rare, how the Odds Ratio can be interpreted).
* **How will the exposure, Obesity will be measured?**
  + Those with a BMI lower than 18.5 kg/m2 are considered **underweight**.
  + Those with a BMI greater or equal than 18.5 kg/m2 but lower than 25 kg/m2 are **normal weight**
  + Those with a BMI greater or equal 25 kg/m2 but lower than 30 kg/m2 are **overweight**
  + Those with a BMI greater or equal than 30 kg/m2 are **obese**
* **How will the outcome, diabetes will be measured?**
  + With the NHANES dataset, diabetes will be measured as a dichotomous variable. Participants will be categorized with diabetes as follows:
    - Affirmative response to the survey question “Do you have diabetes?”

AND/OR

* + - Fasting blood glucose (BG) > 125 mg/dL
  + We will also categorize diabetic participants as having controlled or uncontrolled diabetes.
    - Controlled diabetes
      * Fasting BG 70-125-mg/dL
      * Being categorized (as above) as having diabetes.
* Uncontrolled diabetes
  + Fasting BG greater than 125 mg/dL