

# Project 2 Data analysis plan

Shuai Zhu

2024-11-06

## 1 Introduction

The National Health and Nutrition Examination Survey (NHANES) is a program of studies aimed at assessing the health and nutritional status of adults and children in the United States. This project leverages the NHANES 2011-2014 dataset, which provides data on physical activity, mortality, and various demographic factors. The primary objective is to identify risk factors associated with all-cause mortality and to evaluate the relative predictive and discriminative value of several activity and circadian features derived from wearable accelerometers.

## 2 Method

### 2.1 Data cleaning

After filtering for participants aged 40 to 80 with at least three valid days of accelerometer data, the dataset was further narrowed down to complete cases for subsequent analysis. This resulted in a reduction from 19,141 subjects to 5,564 subjects.

## 2.2 Data analysis

The analysis will utilize time-to-event methods to account for potential non-linear associations. Both univariable and multivariable Cox proportional hazards models will be employed to assess the impact of various physical activity metrics on mortality risk. In the multivariate Cox model, each physical activity measure will be included alongside other covariates, including BMI, age, and medical history of congestive heart failure and coronary heart disease. Both linear and non-linear relationships between physical activity metrics and mortality will be explored. Non-linear associations will be assessed using penalized splines (psplines) to capture complex patterns. Separate models will be fitted for males and females to account for potential gender differences in these associations.

Model performance will be evaluated using 10-fold cross-validation, with the average concordance index (C-index) used to rank the predictive accuracy of the models. All analysis will be conducted in R 4.3.3