**Research purpose & Data construction**

Our main research purpose towards this age prediction task is to indicate that beyond blood biomarkers, data variables that come from chronic disease, self-reported functional limitation and cognition measures may help increase the accuracy when applying machine learning algorithms to predict biological age.

Our datasets consist data from 2011(sample size = 10027) and 2015(sample size = 9668). Both of these datasets contain anonymous IDs and householdIDs to help identify each data provider, and 61 features with blood biomarkers (19 features: white blood cell in thousands, MCV, blood urea nitrogen … etc.), chronic disease (14 features: ever had high blood pressure, ever had diabetes, ever had cancer … etc.), self-reported functional limitation (20 features: Diff-Dressing, Diff-Bathing, Diff-Eating … etc.) and cognition (8 features: CESD Score, immediate word recall, able to draw assign picture … etc.).

**Methodology**

The general steps to perform machine learning algorithms are:

1. Data cleaning

First of all, we filter our age column values by applying a rule of no less than 40 years old and no greater than 85 years old. Secondly, we drop duplication samples if there are any. Thirdly, we calculate the missing ratio of each column and decide whether to leave or drop the corresponding columns. Finally, we check all the data types of each columns to make sure that no mistakes to take numerical features as categorical features or the opposite.

1. Feature engineering