PART 2: ANALYTIC TASK (OPEN-NOTE, BUT NO GEN AI)

About the data: We are using data from a national survey of outpatient health care providers (referred to here as the "outpatient visit data" (OVD)). **Health care providers across the country were randomly contacted to participate, and then they were asked to complete surveys about a random sample of visits to their office,** including the diagnoses the patients received during the visit, patient demographics and pre-existing conditions, any test or lab results from the visit, and what procedures were done during the visit. This is real data from a real national survey (I have only simplified a few variables for you, and I deleted some variables you don't need).

Our research question: The US Preventive Services Task Force (USPSTF) recommends that certain people have their blood sugar levels ("HbA1c") regularly tested to monitor for the development of diabetes. But, many providers do not actually follow this recommendation. We want to understand if there are differences in which patients do receive this screening.

Note: The data sets are both already in stata format (.dta) so you can just start with

```
use midtermLSD, replace
```

---

1. (4 points) USPSTF recommends regularly testing the blood sugar (HbA1c) of patients who are:

   - 35 years or older, AND
   - Overweight or obese (BMI >=25), AND
   - Not already diagnosed with any form of diabetes (Type 1, Type 2, or Unknown Type)

   Create a binary flag for everyone who meets all of these inclusion criteria. Then, **tell us what percent of those visits had an HbA1c test completed** during the visit or within the previous 12 months (the HbA1c testing variable is called "A1C").

2. (3 points) We need to prepare some other variables in the data set for our analyses. Please conduct the following steps to create variables that can be used in your analyses. **Provide the frequency table or summary statistics (mean, SD, min max) for each of the variables, using the inclusion criteria in Question 1:**

   - **Sex:** A 0/1 indicator for whether sex is female
   - **Past visits:** The number of past visits the patient has had with this provider in the last 12 months, top coded at 26. New patients should have 0 prior visits.
   - **Race/ethnicity:** No changes needed (already clean); just provide the frequency table

3. (4 points) We are also going to use diagnosis information. We want to group diagnoses into CCSR "body system" groups, which are broad diagnosis groupings (e.g., all diseases of the heart and circulatory system are grouped together in a body system called "CIR"). **Using the file called ICD_CCSR_key_final.dta, merge in the CCSR body system categories to your data set, by diagnosis code.** It may help to know that the diagnosis code variable in that file is called "ICD10_4digit" and that the file has one row per diagnosis code.

Briefly **explain why you chose the type of merge** you did (1:1, m:1, 1:m, m:m) and **explain which observations you will keep** in the data set.

4. (4 points) The USPSTF recommends that people from racial/ethnic groups with high diabetes prevalence be monitored for potential diabetes more carefully. We want to test whether there are differences in HbA1c screening by race/ethnicity. **Using the survey weights and the same inclusion criteria as question 1, run a regression to test whether the HbA1c testing varies by race/ethnicity groups**, controlling for age (continuous), number of prior visits (continuous), the female indicator, and CCSR Body System. In 1-2 sentences, **report your findings on whether there are differences by race/ethnicity.**

> **NOTE:** You may use a logistic regression or a linear regression for this (in the real world, either is acceptable for this type of outcome, depending on your academic field/department). Whichever you choose, though, you must appropriately interpret the output.