Development of Prediction Model for Depression Risk among Noninstitutionalized US Citizens: a Retrospective Study

Bowen Zhang[1], Jiachen Ai[1], Lang Guo[1]
[1]Department of Biostatistics, University of California, Los Angeles

I. INTRODUCTION

The impact of employment and the intricate balance between work, recreation, and exercise on mental health has garnered increasing attention in contemporary research[1]. While employment offers financial security and a sense of purpose, it can also prolong the sedentary period and engender stressors that potentially contribute to the onset or exacerbation of mental health disorders, particularly depression. Besides, according to a study, around 70% of the population in Australia actively partakes in sports, thereby offering a promising prospect for the promotion of mental well-being[2]. Understanding the nuanced dynamics of this relationship is crucial for early screening for depression, devising tailored interventions and support systems for individuals grappling with mental health challenges[3].

Recent scientific inquiry has emphasized the relationship between employment and mental health, highlighting the influence of recreational activities and exercise on this intricate nexus. The National Health and Nutrition Examination Survey (NHANES) in the United States provides a unique platform for delving into these complexities, integrating self-reported mental health assessments.

We conducted this retrospective study on the basis of pre-pandemic data from NHANES. We aim to develop a logistic model to predict if an adult is at risk for depression only using some characteristics of demographics, employment, and work-recreation-exercise equilibrium information.

II. METHODS & MATERIALS

A. *Data Set*

The National Health and Nutrition Examination Survey (NHANES) is a cross-sectional survey conducted by the National Center for Health Statistics (NCHS) and designed to monitor the health and nutrition of the civilian, noninstitutionalized resident U.S. population. NHANES uses a complex, multistage, stratified probability sampling design. Participants included those who had reached the age of maturity in their state (usually age 18) and children aged 7–17 years with a parent or guardian's permission[4].

This retrospective study used data from NHANES 2017 - March 2020 pre-pandemic data (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020), focusing on demographic data and questionnaire data. For demographic data (P_DEMO), five variables were kept, including sequence number, gender, age, race, and the ratio of family income to poverty. For questionnaire data, three data sets are used: depression screener (P_DPQ), occupation (P_OCQ), and physical activity (P_PAQ). Depression screener data contains the result of PHQ-9. PHQ-9 is a nine-item depression screening instrument that consists of the actual 9 criteria on which the diagnosis of DSM-IV depressive disorders is based[5]. A higher score indicates a higher risk for depression. Occupation data contains the employment status and working hours. Physical activity data has information about activities during work and recreation. Those four data sets were merged by the sequence number.

There are many missing data in the questionnaire data sets. Those with any missing values in PHQ-9 questions were excluded. Missing value in occupation and physical activity data sets is mostly due to the interview system. For example, if a participant answers “No” to if they have a job, the interview system will automatically skip the question “hours worked last week”. This kind of missing data is filled by logic according to the previous questions. Details are covered in the appendix.

For the rest of the missing values, it is assumed that the data were missing at random. Multiple imputation is applied. The fully conditional method (FCS) was used to impute the missing variables 10 times, and we used the mean (continuous variables) or the mode (categorial variables) to impute the missing value. During the imputation, the discriminant function was used for categorical variables, and linear regression was used for continuous variables.

B. *Study Population*

15,560 US citizens were participating in NHANES 2017 – March 2020, among which 5,867 were under 18 years old. The result of PHQ-9 for people below 18 is restricted, and thus was excluded from this study. Participants requiring a proxy were not eligible because of the sensitive nature of the questions. Participants who did not answer all the questions in PHQ-9 were excluded too. Finally, 8,276 adults were eligible for the study.

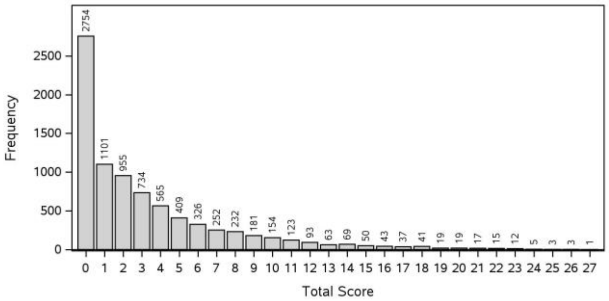


Figure 1 Histogram of PHQ-9 Score

Figure 1 gives the histogram of the total PHQ-9 score in the study population. The level of depression severity is traditionally divided into 5 groups according to the PHQ-9 score. A score below 5 indicates "minimal severity", where only 2.5% have depressive disorder in Kurt and Robert's original study[6], while 22.5% in the next level, "mild severity", have depressive disorder. As a result, we chose the score of 5 as the threshold, dividing the study population into two groups: people with scores 0-4 are classified as low-risk for depression (Non-depression), and people with scores above 5 are classified as high-risk for depression (Depression), which is the outcome of the prediction model. Therefore, there are 6,109 participants in the non-depression group and 2,167 in the depression group. Figure 2 presents the flow chart of the overall population.

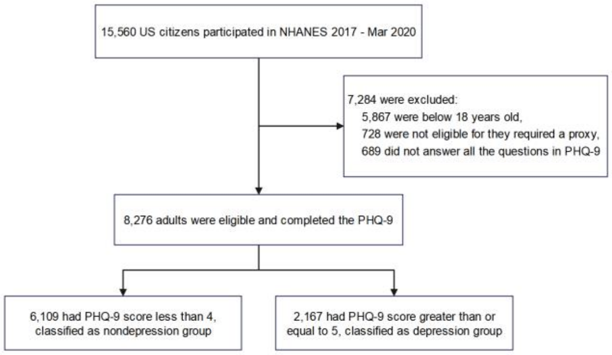

15,560 US citizens participated in NHANES 2017 - Mar 2020
7,284 were excluded:
5,867 were below 18 years old,
728 were not eligible for they required a proxy,
689 did not answer all the questions in PHQ-9
8,276 adults were eligible and completed the PHQ-9
6,109 had PHQ-9 score less than 4, classified as nondepression group
2,167 had PHQ-9 score greater than or equal to 5, classified as depression group

Figure 2 Flow Chart of Overall Population

Table 1 presents the descriptive statistics of each predictor in non-depression and depression groups. Hypothesis testing was applied to compare the characteristics between the two groups. For continuous variables, it is natural to assume that variables of those two groups have unequal variances. Therefore, Welch-Satterthwaite t-test was applied. For categorical variables, Chi-Square test was applied.

Table 1 Characteristics of Two Groups

Characteristic	Depression (N=2,167)	Non-depression (N=6,109)	P-value
Demographics			
Age — yr	48.86 ± 18.21	49.36 ± 18.36	0.26
Gender: male — no. (%)	878 (40.52)	3,193 (52.27)	<0.01**
Race — no. (%)			
Mexican American	247 (11.40)	732 (11.98)	<0.01**
Other Hispanic	261 (12.04)	589 (9.64)	
White	804 (37.10)	2,131 (37.10)	
Black	574 (26.49)	1,603 (26.24)	
Other Race	281 (12.97)	1,054 (17.25)	
Employment			
If worked last week: Yes — no. (%)	939 (43.33)	1,228 (56.67)	<0.01**
Hours worked last week — hr	16.63 ± 21.29	23.63 ± 22.26	<0.01**
Ratio of income to poverty	2.14 ± 1.46	2.74 ± 1.56	<0.01**
Physical activities			
Vigorous work activity: Yes — no. (%)	597 (27.55)	1,532 (25.08)	0.02*
Moderate work in a week — days	2.17 ± 2.46	1.99 ± 2.41	<0.01**
Moderate work in a day — mins	95.05 ± 141.34	85.58 ± 135.84	0.01**
Walk or ride in a week — days	1.05 ± 2.13	1.05 ± 2.43	0.86
Vigorous recreational activity: Yes — no. (%)	396 (18.27)	1,724 (28.22)	<0.01**
Moderate recreation in a week — days	1.35 ± 2.01	1.80 ± 2.12	<0.01**
Moderate recreation in a day — mins	27.41 ± 52.28	34.47 ± 53.90	<0.01**
Sedentary activity in a day — mins	351.44 ± 219.06	329.38 ± 195.34	<0.01**

Plus-minus values are means ± SD; * denotes significant; ** denotes extremely significant

C. Statistical Methods

The data analysis, including data management and statistical analysis, for this paper was generated using SAS software, Version 9.04 of SAS® OnDemand for Academics (SAS Institute).

We performed statistical analysis using data that had been multiply imputed. Logistic regression was conducted to assess the impact of predictors on the risk of depression. The regression model rendered the odds ratio (OR) estimate and 95% confidence interval (CI). In the case of over-fitting, we used step-wise regression to select the variables. Akaike information criterion (AIC) was chosen as the selection criteria. ROC curve and AUC were used to evaluate the performance of this prediction model.

III. RESULTS

After the step-wise logistic regression, 11 predictors were selected in the model; the coefficients and confidence interval are given in the Table 2.

Table 2 Multivariate Step-wise Logistic Model

Characteristic	Odds ratio	95% CI	P-value
Demographics			
Age	0.99	[0.99, 1.00]	<0.01**
Gender (Female)	1.55	[1.40, 1.72]	<0.01**
Race			
Mexican American	0.91	[0.76, 1.09]	0.52
Other Hispanic	1.20	[1.00, 1.43]	<0.01**
Black	0.89	[0.78, 1.01]	0.15
Other Race	0.81	[0.69, 0.95]	<0.01**
White	1.00	—	—
Employment			
If worked last week (Not work)	1.85	[1.65, 2.08]	<0.01**
Ratio of income to poverty	0.82	[0.79, 0.85]	<0.01**
Physical activities			
Vigorous work activity (No)	0.84	[0.73, 0.97]	0.02*
Moderate work in a week	1.03	[1.00, 1.06]	0.05*
Moderate work in a day	1.00	[1.00, 1.00]	0.02*
Vigorous recreational activity (No)	1.31	[1.13, 1.53]	<0.01**
Moderate recreation in a week	0.94	[0.92, 0.97]	<0.01**
Sedentary activity in a day	1.00	[1.00, 1.00]	<0.01**

* denotes significant; ** denotes extremely significant

According to Table 2, females are more likely to have depression than males. For employment status, unemployed individuals and those in low-income groups are more susceptible to experiencing depression.

There is also an association between physical activity and risk of depression. Jobs involving vigorous physical labor or prolonged activity can increase the risk of developing depression. In contrast, spending more time on physical activities during recreation indicates a lower risk for depression.

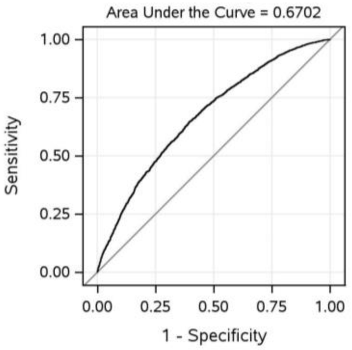


Figure 3 ROC Curve for Logistic Model

Figure 3 shows the ROC curve for the prediction model. The AUC is 0.6702, indicating that this logistic regression model exhibits good discrimination, and it can distinguish between high depression risk and low depression risk to some extent.

IV. CONCLUSIONS

In this retrospective study, we divided the study population into depression and non-depression groups according to the PHQ-9 score. We hope to develop a prediction model that can differentiate high-risk for depression only using demographic data and information related to employment and physical activities. By conducting stepwise logistic regression, we found that females have a higher risk of suffering from depression. Unemployment and low income indicate a higher risk of depression. Generally speaking, longer-time sedentary activities increase the risk of depression. However, physical activity during work and leisure time has opposite effects, where physical activities during work actually increase the risk.

The model we developed has an AUC of 0.6702, thus having good discrimination and great potential for practical applications. This model can assess people's depression risk using only employment and activity-related information, which is of significant importance for early screening of depression.

V. DICUSSION

According to the characteristics of two groups and the regression result, there are many interesting findings. Firstly, in the high-depression-risk group, the proportion of females is significantly higher than that of males. This phenomenon requires urgent attention. Sedentary activity, physical activity during work, and physical activity during recreation have a complex relationship with depression. Groups engaged in prolonged desk work and those involved in strenuous physical labor for extended periods need to pay closer attention to their mental health.

This study has several limitations. First, as a prediction model study, it only involves the development of the model. Internal validation (such as Bootstrapping) and external validation (such as using post-pandemic data to test the model) are needed. Second, we only use step-wise logistic regression to develop the model. Other models like LASSO, machine learning methods are all worth taking into consideration. Third, the performance evaluation measurements for the model are relatively limited. We can also evaluate the model with a calibration curve and decision curve analysis.

REFERENCES

1. Evans, J., Repper, J., “Employment, social inclusion and mental health”, Journal of psychiatric and mental health nursing, Vol. 7, No. 1, pp. 15-24, 2000.
2. Liddle, S. K., Deane, F. P., Vella, S. A., “Addressing mental health through sport: a review of sporting organizations' websites”, Early intervention in psychiatry, Vol. 11, No. 2, pp. 93-103, 2017.

3. Veldman, K., Reijneveld, S. A., Ortiz, J. A., Verhulst, F. C., Bültmann, U., “Mental health trajectories from childhood to young adulthood affect the educational and employment status of young adults: results from the TRAILS study”, J Epidemiol Community Health, Vol. 69, No. 6, pp. 588-593, 2015.
4. Stierman, Bryan et al, "National Health and Nutrition Examination Survey 2017–March 2020 Prepandemic Data Files Development of Files and Prevalence Estimates for Selected Health Outcomes", National Center for Health Statistics, No. 158. 2021.
5. Kroenke, K., Spitzer, R. L., “The PHQ-9: a new depression diagnostic and severity measure”, Psychiatric annals, Vol. 32 No. 9, pp. 509-515, 2002.
6. Kroenke, K., Spitzer, R. L., & Williams, J. B., "The PHQ-9: validity of a brief depression severity measure", Journal of general internal medicine, Vol. 16, No. 9, pp. 606-613, 2001.

APPENDIX

I. Managing Missing Value due to Interview System

The result of PHQ-9 was sourced from the "P_DPQ.XPT" file, encompassing 8965 individual data points about mental health. To ensure data integrity, missing values, and responses denoted as 7 or 9, representing instances where participants refused to respond to specific questionnaire items, were removed from the dataset.

For the employment data “P_OCQ.XPT” file, if a person does not have a job (answers “No” to the question “Did you work last week”), then the question “Hours worked last week” will be skipped. If “Hours worked last week” is missing and the answer to “Did you work last week” is “No”, the missing value is filled by 0.

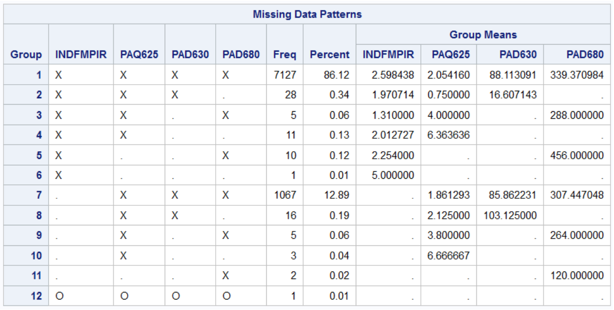
For the physical activity data “P_PAQ.XPT” file, we want to use 3 variables to describe the physical activities during work: If there is vigorous activity, days of moderate or vigorous activity in a week, minutes of moderate or vigorous activity in a day. Therefore, we do the following transformation: If the answer to “Do you have vigorous activities during work” is “Yes” and “Do you have moderate activities during work” is “No”, we fill in the missing value “Days of moderate work in a week” and “Minutes of moderate work in a day” with the days and minutes of vigorous work. Physical activities during recreation are the same.

In addition, if a person claims that he or she does not walk or ride to the workplace, the missing value of “days of walk or ride to work in a week” is filled by 0.

II. Missing Pattern

Before the imputation of missing value, we try to visualize the missing pattern. We choose the variables which contains more than 10 missing values, and the result is shown in Table A.1, where INDFMPIR denotes the ratio of income to poverty, PAQ625 denotes days of moderate work in a week, PAD630 denotes the minutes of moderate work in a day, and PAD680 denotes the minutes of sedentary work in a day.

Table A.1 Missing Pattern



Missing Data Patterns										
							Group Means			
Group	INDFMPIR	PAQ625	PAD630	PAD680	Freq	Percent	INDFMPIR	PAQ625	PAD630	PAD680
1	X	X	X	X	7127	86.12	2.598438	2.054160	88.113091	339.370984
2	X	X	X	.	28	0.34	1.970714	0.750000	16.607143	.
3	X	X	.	X	5	0.06	1.310000	4.000000	.	288.000000
4	X	X	.	.	11	0.13	2.012727	6.363636	.	.
5	X	.	.	X	10	0.12	2.254000	.	.	456.000000
6	X	.	.	.	1	0.01	5.000000	.	.	.
7	.	X	X	X	1067	12.89	.	1.861293	85.862231	307.447048
8	.	X	X	.	16	0.19	.	2.125000	103.125000	.
9	.	X	.	X	5	0.06	.	3.800000	.	264.000000
10	.	X	.	.	3	0.04	.	6.666667	.	.
11	.	.	.	X	2	0.02	.	.	.	120.000000
12	O	O	O	O	1	0.01	.	.	.	.

Notice that the ratio of income to poverty has more than 1,000 missing values. So, we compare the characteristics of observations whose INDFMPIR is missing to observations whose INDFMPIR is not missing, in order to analyze possible influencing factors for the missing. We use chi-squared test for category variables and the t-test for continuous variables. The result is given in the Table A.2.

Table A. 2 Comparison of Missing and Nonmissing Groups

Ratio of income to poverty	Missing (N=1,094)	Non-missing (N=7,182)	P-value
Demographics			
Age — yr	48.84 ± 18.62	49.29 ± 18.28	0.45
Gender: male — no. (%)	564 (51.55)	3,507 (48.83)	0.09
Race — no. (%)			
Mexican American	164 (14.99)	815 (11.35)	<0.01**
Other Hispanic	151 (13.80)	699 (9.73)	
White	266 (24.31)	2,699 (37.16)	
Black	339 (30.99)	1,838 (25.59)	
Other Race	174 (15.90)	1,161 (16.17)	
Employment			
If worked last week: Yes — no. (%)	583 (53.29)	3,995 (55.64)	0.15
Hours worked last week — hr	21.15 ± 22.38	21.89 ± 22.20	0.31
Physical activities			
Vigorous work activity: Yes — no. (%)	293 (26.86)	1,836 (25.57)	0.37
Moderate work in a week — days	1.89 ± 2.41	2.06 ± 2.42	0.03*
Moderate work in a day — mins	86.12 ± 143.60	87.83 ± 136.40	0.71
Walk or ride in a week — days	1.19 ± 2.23	1.03 ± 2.12	0.03*
Vigorous recreational activity: Yes — no. (%)	302 (27.61)	1,818 (25.31)	0.11
Moderate recreation in a week — days	1.63 ± 2.04	1.69 ± 2.11	0.33
Moderate recreation in a day — mins	37.04 ± 65.12	31.91 ± 51.56	0.013*
Sedentary activity in a day — mins	306.90 ± 199.50	339.50 ± 202.70	<0.01**

Plus-minus values are means ± SD; * significant; ** denotes extremely significant

The missing “ratio of income to poverty” seems strongly relative to “race” and “sedentary activity in a day (mins)” and has relations with “moderate work in a week (days)”, “walk or ride in a week (days)”, and “moderate recreation in a day (mins)”. We can infer that these factors, especially race or the time a person does sedentary activity a day, might result in an unwillingness to answer the income question.

III. Correlation Analysis

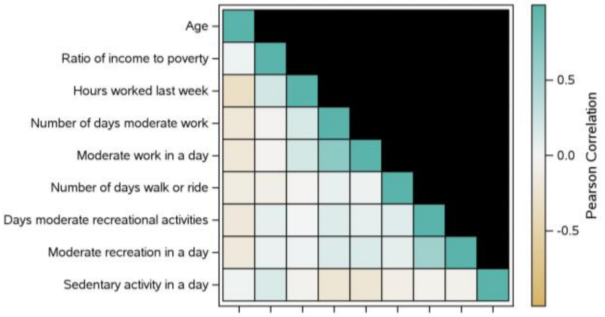


Figure A.1 Correlation Matrix of Continuous Variables

We also compute the Pearson correlation between continuous variables of the study population. Figure A.1 presents the correlation matrix. After performing calculations, the absolute values of the correlations between any two variables are all less than 0.6, indicating that there are no strongly correlated variables. The number of days of physical activity per week and the duration of activity per day are moderately positively correlated, which is quite natural. The age and the working hours have a moderate negative correlation.

BIOSTAT 203A Midterm SAS

```
* BIOSTAT 203A GROUP 6 MIDTERM PROJECT;
* This sas file is written by Bowen Zhang;
* Import all the data;

LIBNAME xptfile1 XPORT "/home/u63618391/Biostat203A/midterm/XPT/P_DEMO.XPT";
LIBNAME xptfile2 XPORT "/home/u63618391/Biostat203A/midterm/XPT/P_DPQ.XPT";
LIBNAME xptfile3 XPORT "/home/u63618391/Biostat203A/midterm/XPT/P_OCQ.XPT";
LIBNAME xptfile4 XPORT "/home/u63618391/Biostat203A/midterm/XPT/P_PAQ.XPT";
LIBNAME xptread "/home/u63618391/Biostat203A/midterm/XPTread";

PROC COPY in=xptfile1 out=xptread memtype=data;
RUN;

PROC COPY in=xptfile2 out=xptread memtype=data;
RUN;

PROC COPY in=xptfile3 out=xptread memtype=data;
RUN;

PROC COPY in=xptfile4 out=xptread memtype=data;
RUN;

* Data cleaning of P_DPQ
/*
This data includes A nine-item depression screening instrument (also called PHQ-9),
which determines the frequency of depression symptoms over the past 2 weeks.
For each symptom question, points ranging from 0~3, are associated with the response
categories "not at all," "several days," "more than half the days," and "nearly every day.”
Except that, 7 denotes "Refused to answer", 9 donotes "Don't know".
*/

* Missing data;
* If an observation contains any missing value (including refused and don't know), we drop them.;

DATA xptread.p_dpq;
    SET xptread.p_dpq;

    IF DPQ010 >=0 AND DPQ010 <=3 AND DPQ020 >=0 AND DPQ020 <=3 AND DPQ030 >=0 AND
        DPQ030 <=3 AND DPQ040 >=0 AND DPQ040 <=3 AND DPQ050 >=0 AND DPQ050 <=3 AND
        DPQ060 >=0 AND DPQ060 <=3 AND DPQ070 >=0 AND DPQ070 <=3 AND DPQ080 >=0 AND
        DPQ080 <=3 AND DPQ090 >=0 AND DPQ090 <=3;
RUN;

* We drop the column DPQ100, which is not included in further analysis;
* And we calculate the total score by simply add DPQ010 - DPQ090 up.;

DATA xptread.p_dpq;
```

```
    SET xptread.p_dpq;
    DROP DPQ100;
    score=DPQ010+DPQ020+DPQ030+DPQ040+DPQ050+DPQ060+DPQ070+DPQ080+DPQ090;
RUN;

* Data cleaning of P_DEMO;

/*
This data set includes some demographic variables. We are only interested in several variables
among them, which are:
RIAGENDR - gender
RIDAGEYR - age (More than 80 is denoted by 80)
RIDRETH1 - race
INDFMPIR - Ratio of family income to poverty (More than 5 is denoted by 5)
*/
* Select the columns we are interested in;

DATA xptread.p_demo;
    SET xptread.p_demo;
    KEEP SEQN RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR;
RUN;

* Data cleaning of P_OCQ;

/*
This data set contains survey participant interview data on employment.
The primary focus of the OCQ is the participant’s work within the previous week.
The participant’s Current Job is defined as the main paid job worked within the last week,
including work on a family farm.

It has the following variables that we are interested in:
OCD150 - Type of work done last week (If answer is not work, will skip questions below)
OCQ180 - Hours worked last week in total all jobs
OCQ670 - Overall work schedule past 3 months
*/
* Select the columns we are interested in;

DATA xptread.p_ocq;
    SET xptread.p_ocq;
    KEEP SEQN OCD150 OCQ180 OCQ670;
RUN;

* Quality check for P_OCQ

* For Type of work OCD150, 7 denotes refused, 9 denoted don't know, we turn them into missing;
* Since type 2(have job but not work) and 3(looking for job) have small proportion, turn them into 4(not work);

DATA xptread.p_ocq;
    SET xptread.p_ocq;

    IF OCD150=7 OR OCD150=9 THEN
        OCD150=.;

    IF OCD150=2 OR OCD150=3 THEN
        OCD150=4;
RUN;

* For Hours worked last week OCQ180, there are many missing value, for unemployed people do not answer this question;
* Turn missing value to 0 if the person answer 4 in OCD150;
* Turn the "refused" and "don't know" to missing;

DATA xptread.p_ocq;
    SET xptread.p_ocq;

    IF OCQ180=. AND OCD150=4 THEN
        OCQ180=0;

    IF OCQ180=77777 OR OCQ180=99999 THEN
        OCQ180=.;
RUN;

* For work scheduals OCQ670, 7 denotes refused, 9 denoted don't know, we turn them into missing;
* Turn missing value to 0 if the person answer 4 in OCD150;

DATA xptread.p_ocq;
    SET xptread.p_ocq;

    IF OCQ670=. AND OCD150=4 THEN
        OCQ670=0;

    IF OCQ670=7 OR OCQ670=9 THEN
        OCQ670=.;
RUN;

* Data cleaning of P_PAQ
/*
This data set contains questions based on the Global Physical Activity Questionnaire (GPAQ),
it provides respondent-level interview data on physical activities.

We are interested in the following variables:

* Part 1 Work activity
PAQ605 - Vigorous work activity
PAQ620 - Moderate work activity
PAQ625 - Number of days moderate work
PAD630 - Minutes moderate-intensity work

* Part 2 Bike or walk
PAQ635 - Walk or bicycle
PAQ640 - Number of days walk or bicycle

* Part 3 Recreational activities
PAQ650 - Vigorous recreational activities
```

```
PAQ665 - Moderate recreational activities
PAQ670 - Days moderate recreational activities
PAD675 - Minutes moderate recreational activities

* Part 4
PAD680 - Minutes sedentary activity
*/

* Data management of part 1 and part 3;
* If PAQ605 is "Yes" and PAQ620 is "No", turn PAQ620 to "Yes";
* Turn the "don't know" in PAQ605 and PAQ620 to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAQ605=1 AND PAQ620=2 THEN
        DO;
            PAQ620=1;
            PAQ625=PAQ610;
            PAD630=PAD615;
        END;

    IF PAQ605=7 OR PAQ605=9 THEN
        PAQ605=.;

    IF PAQ620=7 OR PAQ620=9 THEN
```

```
        PAQ620=.;
RUN;

* If PAQ625 is missing and PAQ620=2, turn PAQ625 to 0;
* If PAQ625 is refused or don't know, turn to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAQ625=. AND PAQ620=2 THEN
        PAQ625=0;

    IF PAQ625=77 OR PAQ625=99 THEN
        PAQ625=.;
RUN;

* If PAD630 is missing and PAQ620 = 2, turn PAD630 to 0.


        * If PAD630 is refused or don't know, turn to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAQ620=2 AND PAD630=.
    THEN
```

```
        PAD630=0;

    IF PAD630=7777 OR PAD630=9999 THEN
        PAD630=.;
RUN;

* If PAQ650 is "Yes" and PAQ665 is "No", turn PAQ620 to "Yes";
* Turn the "don't know" in PAQ650 and PAQ665 to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAQ650=1 AND PAQ665=2 THEN
        DO;
            PAQ665=1;
            PAQ670=PAQ655;
            PAD675=PAD660;
        END;

    IF PAQ650=7 OR PAQ650=9 THEN
        PAQ650=.;

    IF PAQ665=7 OR PAQ665=9 THEN
        PAQ665=.;
RUN;

```

```
* If PAQ670 is missing and PAQ665=2, turn PAQ625 to 0;
* If PAQ670 is refused or don't know, turn to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAQ670=. AND PAQ665=2 THEN
        PAQ670=0;

    IF PAQ670=77 OR PAQ670=99 THEN
        PAQ670=.;
RUN;

* If PAD675 is missing and PAQ650 = 2, turn PAD675 to 0;
* If PAD675 is refused or don't know, turn to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAQ650=2 AND PAD675=.
    THEN
        PAD675=0;

    IF PAD675=7777 OR PAD675=9999 THEN
        PAD675=.;
RUN;
```

```

* Data management of part 2;
* If PAQ640 is missing and PAQ635 is 'No', then PAQ640 = 0;
* If PAQ640 is refused or don't know, turn to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAQ640=. AND PAQ635=2 THEN
        PAQ640=0;

    IF PAQ640=77 OR PAQ640=99 THEN
        PAQ640=.;
RUN;

* Data management of part 4;
* If PAD680 is refused or don't know, turn to missing;

DATA xptread.p_paq;
    SET xptread.p_paq;

    IF PAD680=7777 OR PAD680=9999 THEN
        PAD680=.;
RUN;

DATA xptread.p_paq;
```

```
    SET xptread.p_paq;
    KEEP SEQN PAQ605 PAQ625 PAD630 PAQ640 PAQ650 PAQ670 PAD675 PAD680;
RUN;

* Merge data;

DATA merged_data;
    MERGE xptread.p_demo xptread.p_dpq xptread.p_ocq xptread.p_paq;
    BY SEQN;
RUN;

* Output merged_data;

PROC EXPORT DATA=merged_data
        OUTFILE="/home/u63618391/Biostat203A/midterm/merged_data.xlsx" DBMS=XLSX
        REPLACE;
    SHEET="MergedData";
RUN;

/* Now we have the data set we will use in the statistical analysis,
then we will turn some of the numeric variables to categorial variables.
*/
PROC FORMAT;
    VALUE Cgender 1='M' 2='F';
RUN;

```

```
PROC FORMAT;
    VALUE Crace 1='Mexican American' 2='Other Hispanic' 3='White' 4='Black'
        5='Other Race';
RUN;

PROC FORMAT;
    VALUE C_150OCD 1='Work' 4='Not Work';
RUN;

PROC FORMAT;
    VALUE C_670OCQ 1='Traditional' 2='Night' 3='Morning' 5='Variable';
RUN;

PROC FORMAT;
    VALUE YN 1='Yes' 2='No';
RUN;

DATA merged_data;
    SET merged_data;
    FORMAT RIAGENDR Cgender.
           RIDRETH1 Crace.
           OCD150 C_150OCD.
           OCQ670 C_670OCQ.
           PAQ605 YN.
           PAQ650 YN.;
RUN;
```

```
* Now we are gonna deal with the missing data;
* First, if an observation do not have phq-9 score, drop it.;

DATA xptread.analysis_data;
    SET merged_data;

    IF NOT MISSING(score);
RUN;

* Then we compute the numbers of missing value of each observation;

DATA MissingCounts;
    SET xptread.analysis_data;
    ARRAY var_array(*) RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR OCD150 OCQ180 OCQ670
        PAQ605 PAQ625 PAD630 PAQ640 PAQ650 PAQ670 PAD675 PAD680;
    Missing_Count=0;

    DO i=1 TO dim(var_array);

        IF missing(var_array(i)) THEN
            Missing_Count + 1;
    END;
    DROP i;
RUN;
```

```
PROC FREQ data=MissingCounts;
    TABLES Missing_Count/ nocum;
RUN;


* Compute the missing value for each variables;

PROC MEANS DATA=xptread.analysis_data NMISS;
    VAR RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR OCD150 OCQ180 OCQ670
        PAQ605 PAQ625 PAD630 PAQ640 PAQ650 PAQ670 PAD675 PAD680;
RUN;


* Try to visualize the pattern of missing data;

DATA missing_flag;
    SET xptread.analysis_data;
    IF INDFMPIR = . THEN INDFMPIR_FLAG = 1; ELSE INDFMPIR_FLAG = 0;
    IF PAQ625 = . THEN PAQ625_FLAG = 1; ELSE PAQ625_FLAG = 0;
    IF PAD630 = . THEN PAD630_FLAG = 1; ELSE PAD630_FLAG = 0;
    IF PAD680 = . THEN PAD680_FLAG = 1; ELSE PAD680_FLAG = 0;
RUN;

PROC MI DATA=missing_flag NIMPUTE=0;
    VAR INDFMPIR PAQ625 PAD630 PAD680;
    ODS SELECT MISSPATTERN;
```

```
RUN;


* Multiple Imputation;
/*
PROC MI DATA=xptread.analysis_data NIMPUTE=10 OUT=xptread.mi_data;
    CLASS OCD150 OCQ670 PAQ605;
    FCS PLOTS=TRACE(mean std);
    VAR RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR OCD150 OCQ180 OCQ670
        PAQ605 PAQ625 PAD630 PAQ640 PAQ650 PAQ670 PAD675 PAD680 score;
    FCS DISCRIM(OCD150 OCQ670 PAQ605 / classeffects=include) NBITER=100;
RUN;
*/

* We drop all the observations that contain missing value;


* Draw the histogram of the total score of phq-9;

ODS LISTING GPATH="/home/u63618391/Biostat203A/midterm/" image_dpi=300 style=Journal;
ODS GRAPHICS / IMAGENAME="Figure1" outputfmt=JPEG width=14cm height=7cm border=off;
PROC SGPLOT DATA = xptread.mi_data;
    WHERE _Imputation_ = 1;
    VBAR score / datalabel;
    XAXIS LABEL = 'Total Score';
    YAXIS LABEL = 'Frequency';
```

```
RUN;

* We seperate the observation into 2 group, No Depression (score<5) and Depression (score>=5);

DATA xptread.mi_data;
    SET xptread.mi_data;
    IF score>=5 THEN group='Yes';
    ELSE group='No';
RUN;

PROC EXPORT DATA=xptread.mi_data
        OUTFILE="/home/u63618391/Biostat203A/midterm/mi_data.xlsx" DBMS=XLSX
        REPLACE;
    SHEET="miData";
RUN;

* Now we compute the characteristics;
PROC FREQ DATA=xptread.mi_data;
    WHERE _Imputation_ = 1;
    TABLES RIAGENDR * group / nopercent norow;
RUN;

PROC FREQ DATA=xptread.mi_data;
    WHERE _Imputation_ = 1;
    TABLES RIDRETH1 * group / nopercent norow;
RUN;
```

```

* Now run some tests;

PROC IMPORT OUT= xptread.fdata
            DATAFILE= "/home/u63618391/Biostat203A/midterm/f.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

DATA xptread.fdata;
    SET xptread.fdata;
    FORMAT RIAGENDR Cgender.
        RIDRETH1 Crace.
        OCD150 C_150OCD.
        PAQ605 YN.
        PAQ650 YN.;
RUN;

PROC MEANS DATA=xptread.fdata MEAN std;
  VAR RIDAGEYR INDFMPIR OCQ180 PAQ625 PAD630 PAQ640 PAQ670 PAD675 PAD680;
  CLASS group;
RUN;


PROC TTEST DATA=xptread.fdata;
```

```
    CLASS group;
    VAR RIDAGEYR;
run;

PROC FREQ DATA=xptread.fdata;
   TABLE group * RIAGENDR / chisq NOCOL expected NOPERCENT;
RUN;

PROC FREQ DATA=xptread.fdata;
   TABLE group * RIDRETH1 / chisq NOCOL expected NOPERCENT;
RUN;

PROC FREQ DATA=xptread.fdata;
   TABLE group * OCD150 / chisq NOCOL expected NOPERCENT;
RUN;

PROC TTEST DATA=xptread.fdata;
    CLASS group;
    VAR OCQ180;
run;

PROC TTEST DATA=xptread.fdata;
    CLASS group;
    VAR INDFMPIR;
run;

```

```
PROC TTEST DATA=xptread.fdata;
    CLASS group;
    VAR PAQ625;
run;

PROC TTEST DATA=xptread.fdata;
    CLASS group;
    VAR PAD630;
run;

PROC TTEST DATA=xptread.fdata;
    CLASS group;
    VAR PAQ640;
run;

PROC TTEST DATA=xptread.fdata;
    CLASS group;
    VAR PAQ670;
run;

PROC TTEST DATA=xptread.fdata;
    CLASS group;
    VAR PAD675;
run;

PROC TTEST DATA=xptread.fdata;
```

```
    CLASS group;
    VAR PAD680;
run;

PROC FREQ DATA=xptread.fdata;
   TABLE group * PAQ605 / chisq NOCOL expected NOPERCENT;
RUN;

PROC FREQ DATA=xptread.fdata;
   TABLE group * PAQ650 / chisq NOCOL expected NOPERCENT;
RUN;

* Correlation of several variables;
PROC CORR DATA=xptread.fdata OUT=correlation_matrix;
  VAR INDFMPIR OCQ180 PAD630 PAD675 PAD680;
RUN;


/* Prepare the correlations coeff matrix: Pearson's r method */
%macro prepCorrData(in=,out=);
  /* Run corr matrix for input data, all numeric vars */
  proc corr data=&in. noprint
    pearson
    outp=work._tmpCorr
    vardef=df
  ;
  run;

  /* prep data for heat map */
data &out.;
  keep x y r;
  set work._tmpCorr(where=(_TYPE_="CORR"));
  array v{*} _numeric_;
  x = _NAME_;
  do i = dim(v) to 1 by -1;
    y = vlabel(v(i));
    r = v(i);
    /* creates a lower triangular matrix */
    if (i<_n_) then
      r=.;
    output;
  end;
run;

proc datasets lib=work nolist nowarn;
  delete _tmpcorr;
quit;
%mend;

proc template;
  define statgraph corrHeatmap;
   dynamic _Title;
```

```
    begingraph;
      entrytitle _Title;
      rangeattrmap name='map';
      /* select a series of colors that represent a "diverging"  */
      /* range of values: stronger on the ends, weaker in middle */
      /* Get ideas from http://colorbrewer.org                   */
      range -1 - 1 / rangecolormodel=(cxD8B365 cxF5F5F5 cx5AB4AC);
      endrangeattrmap;
      rangeattrvar var=r attrvar=r attrmap='map';
      layout overlay /
        xaxisopts=(display=(line ticks tickvalues))
        yaxisopts=(display=(line ticks tickvalues));
        heatmapparm x = x y = y colorresponse = r /
          xbinaxis=false ybinaxis=false
          name = "heatmap" display=all;
        continuouslegend "heatmap" /
          orient = vertical location = outside title="Pearson Correlation";
      endlayout;
    endgraph;
  end;
run;

ODS LISTING GPATH="/home/u63618391/Biostat203A/midterm/" image_dpi=300 style=Journal;
ODS GRAPHICS / IMAGENAME="Cor" outputfmt=JPEG width=10cm height=10cm border=off;

DATA try;
```

```
  SET XPTREAD.FDATA;
  WHERE group = "Yes";
  KEEP  RIDAGEYR INDFMPIR OCQ180 PAQ625
        PAD630 PAQ640 PAQ670 PAD675 PAD680;
  LABEL INDFMPIR = "Ratio of income to poverty"
        RIDAGEYR = "Age"
        PAQ625 = 'Number of days moderate work'
        PAQ640 = 'Number of days walk or ride'
        PAQ670 = 'Days moderate recreational activities'
        OCQ180 = 'Hours worked last week'
        PAD630 = 'Moderate work in a day'
        PAD675 = "Moderate recreation in a day"
        PAD680 = "Sedentary activity in a day";
RUN;

ODS LISTING GPATH="/home/u63618391/Biostat203A/midterm/" image_dpi=300 style=Journal;
ODS GRAPHICS / IMAGENAME="Cor" outputfmt=JPEG width=15cm height=10cm border=off;
%prepCorrData(in=try,out=result);
proc sgrender data=result template=corrHeatmap;
   dynamic _title= "Corr matrix for some variables";
run;

ods output ModelBuildingSummary=SUM;
ods output FitStatistics=FIT;
proc logistic data=xptread.fdata;
    class RIAGENDR RIDRETH1 OCD150 PAQ605 PAQ650;
   model group(event='Yes') = RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR OCD150 OCQ180 PAQ605 PAQ625
                                PAD630 PAQ640 PAQ650 PAQ670 PAD675 PAD680
                                / stepwise;
run;


ODS LISTING GPATH="/home/u63618391/Biostat203A/midterm/" image_dpi=300 style=Journal;
ODS GRAPHICS / IMAGENAME="roc" outputfmt=JPEG width=14cm height=7cm border=off;
proc logistic data=xptread.fdata plots(only)=roc;
    class RIAGENDR RIDRETH1 OCD150 PAQ605 PAQ650;
   LogisticModel: model group(event='Yes') = RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR OCD150 PAQ605 PAQ625 PAD630
                                 PAQ650 PAQ670 PAD680;
    output out=LogiOut predicted=LogiPred;
run;
```