MGMT 590 CAUSAL LEARNING IN BUSINESS – FINAL PROJECT

Team #2

Chun-Yi Chiang, Xiaoyu Guan, Chu-Yun Hsiao, Yi-Chun Huang, Yu-Hui Lin

1. Background and Problem Statement

Health is essential to all. As individuals, we care about understanding the relationship between personal habits and health conditions. The main topics we are interested in are diabetes and blood pressure. The project is to understand whether personal behavior, (specifically exercise, diet, and smoking) has causal relationships to these diseases.

The dataset [1] is sourced from National Health and Nutrition Examination Survey (NHANES) conducted by CDC[2], which is a program of studies designed to assess the health and nutritional status of adults and children in the United States.

The survey is unique in that it combines interviews and physical examinations. The interview part includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

To prevent the influence of COVID, we took the 2017-2018 dataset where 9,254 participants completed the interview and 8,704 were examined. The dataset contains numerous variables, and we have chosen to focus on the interested variables, including gender, income, diabetes, cardiovascular health (hypertension), alcohol use, dietary habits, sleep disorder, physical activity, smoking, and body measures.

As heart diseases and diabetes are constantly the leading causes of death in the United States, this project focuses on investigating the potential causes of them. In this study, we examine the following hypotheses:

1. Diabetes is related to being overweight and exercise habits.
2. Hypertension can be caused by overweight, smoking, and lack of sleep.

Due to the control variables including alcohol usage, this project is focusing on 21+ unpregnant adults. Other confounding effects might exist among demographic variables (e.g. income, gender, and education), and thus, demographic variables are taken as control variables as part of deconfounding measures.

2. Significance and Practical Impact

Under the predefined background, the project aims to provide information on the causality between behaviors and personal health, and, therefore, guidelines of a healthy lifestyle. These guidelines can be utilized as part of the preventive healthcare measures, and eventually release pressure on the national healthcare system for the country and the burden of medical care on household spending. This project offers invaluable insights into disease prevalence, risk factors, and healthcare utilization.

3. Operational and Analytic Goal

For this project, our goal is to perform causal analysis to prove the abovementioned assumptions, mapping out relationships among various factors and their influence on health conditions.

---

[1] https://wwwn.cdc.gov/Nchs/Nhanes/continuousnhanes/default.aspx?BeginYear=2017

[2] Data collected from U.S. Centers for Disease Control and Prevention (https://www.cdc.gov), 27 June, 2024.

We will conduct observation studies to estimate the treatment effects. That is, we plan to use the potential outcomes framework to estimate the average causal effect of the selected factors, like smoking, exercise frequency, and dietary, on the diseases. Moreover, conditional and grouped average treatment effects will also be estimated to compare the impact difference among people with different characteristics. The whole analytical process will be based on two important assumptions: unconfoundedness and overlap. Accordingly, we will try to include all potential confounders and ensure the validity of the overlap assumption by implementing relevant examination and adjustment.

To provide a solid conclusion, this project will utilize various methods, including Regression-Based method, Propensity-Score-Based methods, Double Robustness, and Partially Linear Model. The idea is to examine the outcomes of various methods and derive consistent inferences across, allowing us to provide suggestions with confidence. Our analytical process is shown in Figure 1.

Figure 1. Analytical Workflow of This Project

4. **Causal Analytics and Result**

In this section, we mainly focus on validating our hypotheses: the causality between chronic diseases and personal physical conditions and habits. Therefore, the ATE (average treatment effect) of conditions and habits on the risk of chronic diseases is of interest.

(A) Diabetes ~ Exercise

a. Variable Selection

- Dependent Variable: Diabetes is binary variable provided in Questionnaire database, where the participants answer whether they were diagnosed by doctor with diabetes (1 as yes and 0 as no.)
- Treatment Variable: exercise is binary variable which pre-processed with Questionnaire PAQ650 and PAQ665: participants answer whether they conduct vigorous or moderate recreational activities in a typical week respectively. If the participants answered yes to either of the mentioned questions, the exercise will be denoted as 1.
- Covariate: the covariates include dietary (‘unhealthy_condition’), alcohol consumption ('ALQ121’), family income (‘INDFMIN2’), gender (‘RIAGENDR’), age (‘RIDAGEYR_fixed’), BMI (‘BMXBMI’), smoking habit ('smoker_con')
- Observed units are 3,275.

b. Model and Estimates

- Regression-Based Model w/o Interaction: for the causality between diabetes and exercise, we tried Logistic Regression, Lasso (Logistic with L1 penalty), and Random Forest. Logistic and Lasso has similar AUC (around 0.79), while Random Forest with only 0.75, leading us to an estimated ATE -0.2549 and standard error 0.115.
- Regression-Based Model w/ Interaction: Based on previous, we further analyze with Logistic and Random Forest approaches. Logistic still outperform.
- Inverse Probability Weighted Estimator (IPWE): We applied Logistic approach in this model. We obtained an ATE estimate -0.0356, and standard error 0.014. The propensity score was obtained with logistic function and shows the covariate balance as below. The range of propensity score is between 0.11 to 0.82, aligning with the overlap assumption.

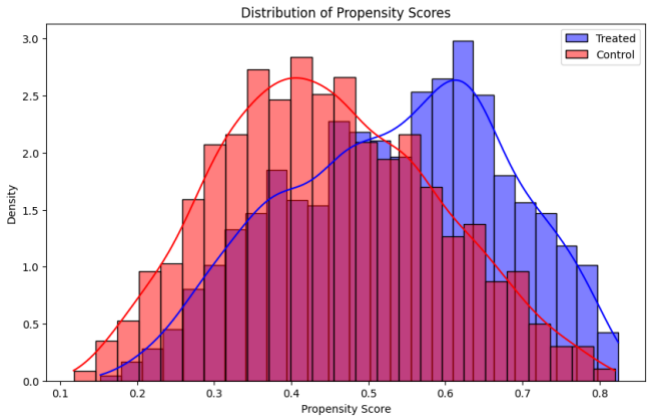

Figure 2. The Distribution of Estimated Propensity Scores (Diabetes ~ Exercise)

- Augmented Inverse Probability Weighted Estimator (AIPWE): Aligned with IPWE, AIPWE also applied Logistic function and obtained an ATE estimate -0.0309 and standard error 0.016.
- Doubly Robust Partially Linear Model method: For doubly robust, we applied both Logistic and Random Forest, and approached the outcome of an estimate ATE -0.0301 and standard error 0.009.

The overall outcome is shown as Table 2. We can see with all approaches that exercise is a significant negative factor to diabetes.

Table 1. Summary of Estimate Results (Diabetes ~ Exercise)

	Regression Based	*Regression Based (w/ Interaction)*	*IPWE*	*AIPWE*	*PLM*
ATE Estimate	-0.2549	-0.2319	-0.0356	-0.0309	-0.0301
Standard Error	0.115	0.146	0.014	0.016	0.009
t-statistic	-2.22	-1.59	-2.54	-1.93	-3.34

(B) Diabetes ~ Obesity

a. Variable Selection

- Dependent variable: Diabetes (DIQ010) indicates if a person is informed having diabetes by doctors, which is a binary variable selected from the Questionnaire dataset.
- As for the treatment variable**,** we created a new variable (BMI_above_30) as an indicator for people with obesity. The variable is binary. If a person's BMI is over 30, then the variable will be 1. (otherwise, 0)
- In order to analyze the causality between diabetes and obesity, we considered relevant covariates into our analysis for deconfounding. For **control variables**, we included unhealthy condition ('unhealthy_condition'), alcohol ('ALQ121'), household income ('INDFMIN2'), gender (RIAGENDR), age ('RIDAGEYR_fixed'), sleep condition (SLD012), blood pressure (PBXPULS, Sys_AVEBP), exercise habits, waist-hip circumference ratio ('WH_ratio'), smoking ('smoker_con')
- The total observed units: 2,640

b. Models and Estimates

- Regression-based model w/o interactions: We tried OLS, Lasso, random forest and decision tree. OLS performed the best with a higher AUC. With the 0.000 p-value lower than 0.05, the coefficient of BMI_above_30 estimated by OLS indicates that the positive relationship between Diabetes and BMI is significant.
- Regression-based model w/ interactions: We applied the same model and included the interactions terms into model. The positive effect between diabetes and BMI still statistically significantly exists.
- Inverse Probability Weighted Estimator (IPWE): By using the IPWE model, we can directly address confounding bias by weighting observations based on the inverse of the propensity score. This creates a pseudo-population where the distribution of covariates is balanced between treated and control groups, making the treatment assignment effectively random. The IPWE method with logistics regression model estimates a much smaller effect of 0.1033 units but with a higher precision (lower standard error). The t-statistic of 6.980 still indicates significance. The propensity score ranged from 0.138452 to 0.780745

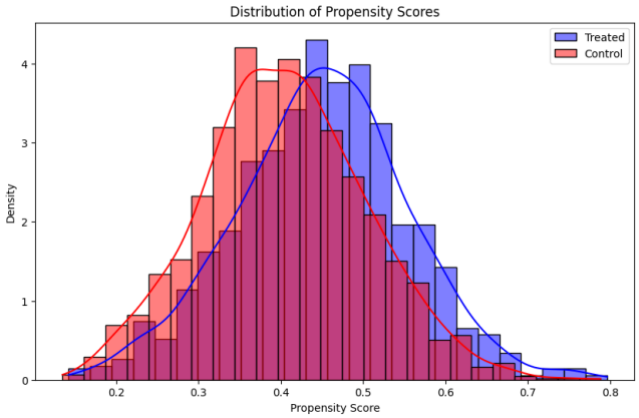

Figure 3. The Distribution of Estimated Propensity Scores (Diabetes ~ Obesity)

- Augmented Inverse Probability Weighted Estimator (AIPWE): By utilizing AIPWE model, we combine the strengths of IPWE and regression adjustment even if propensity score or outcome model is mis-specified, the estimator of ATE remains consistent. This provides additional robustness to model misspecification compared to using regression alone. By applying gradient boosting as outcome regression model, the estimated ATE is 0.0945. The standard error is slightly reduced compared to IPWE, and the t-statistic is higher at 7.269, indicating strong evidence for the effect.
- Doubly Robust Partial Linear Model: By considering the possibility of non-linear relationship between covariates and diabetes, we applied both logistics and random forest model to capture the non-linear relationships. The PLM provides an ATE estimate of 0.1032, with the lowest standard error among all models (0.0094) and the highest t-statistic (10.979). This suggests the PLM's estimate is highly statistically significant.

Table 2. Summary of Estimate Results (Diabetes ~ Obesity)

	Regression Based	*Regression Based (w/ Interaction)*	*IPWE*	*AIPWE*	*PLM*
ATE Estimate	0.7941	0.8016	0.1033	0.0945	0.1032
Standard Error	0.116	0.139	0.0148	0.0130	0.0094
t-statistic	6.865	5.766	6.980	7.269	10.979

(C) Hypertension ~ Obesity

a. Variable Selection:

- Dependent variable (LBDHDD): The hypothesis here is hypertension is associated with obesity, and HDL-C level, representing High-Density Lipoprotein Cholesterol, is used as the dependent variable in this causal analysis to assess the risk level for hypertension. HDL-C is known as the "good"

cholesterol because it helps remove other forms of cholesterol from the bloodstream. Higher levels of HDL-C are associated with a lower risk of heart disease and hypertension. For a healthy person, this value is normally larger than 60 mg/dL.

- Treatment (BMI_above_30): We use BMI index to evaluate obesity in this model. Following the normal standard, BMI value over 30 is treated as obesity.
- Control variables: to exclude the cofounding effects and keep the result consistency, we chose alcohol (ALQ121), blood pressure (PBXPULS, Sys_AVEBP), exercise (exercise), smoke (smoker_con), sleep (SLD012), waist-hip ratio (WH_ratio), dietary (unhealthy_condition), income (INDFMIN2), age (RIDAGEYR_fixed), and gender (RIAGENDR) as covariates. After excluding the null value, the number of observed units is 3,244.

b. Models and Estimates:

- Regression-based w/o interactions: We compared OLS, Lasso, Random Forest, Gradient Boosting, and ANN models here, the MSE value is around 168, 194, 178, 171, and 186 accordingly, lowest for OLS. With this method, the ATE estimate is -4.8640, and the result is significant under significant level 5% (p-value = 0.00).
- Regression-based w/ interactions: We applied OLS model here, the ATE estimate is -5.1344, and the result is significant (p-value = 0.00).
- Inverse Probability Weighted Estimator (IPWE): We fit logistic regression model here to estimate the propensity score, which is ranged from 0.056 to 0.910. Then using the IPWE method, we got the estimate ATE as -6.1644.

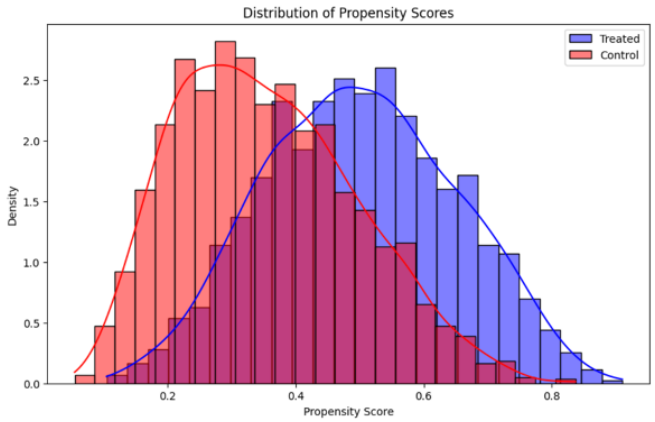

Figure 4. The Distribution of Estimated Propensity Scores (Hypertension ~ Obesity)

- Augmented Inverse Probability Weighted Estimator (AIPWE): We use logistic regression and gradient boosting models here to obtain the ATE estimate, which is -5.0379.
- Doubly Robust Partially Linear Model method (PLM): Based on the propensity score result we have got before, we calculate the ATE using doubly robust PLM, which is -4.9365.

Table 3. Summary of Estimate Results (Hypertension ~ Obesity)

	Regression Based	*Regression Based (w/ Interaction)*	*IPWE*	*AIPWE*	*PLM*
ATE Estimate	-4.8640	-5.1344	-6.1644	-5.0379	-4.9365
Standard Error	0.597	0.637	0.5632	0.4320	0.3298
t-statistic	-8.149	-8.06	-10.946	-11.6622	-14.9679

(D) Hypertension ~ Smoking

a. Variable Selection

- Dependent variable: This study hypothesizes that the risk of Hypertension can be increased by smoking behavior. In this section, we continue to use HDL-C (High-density lipoprotein cholesterol) level as a representative measure of the risk of Hypertension.
- Treatment: We selected the variable representing the amount of Cotinine from the laboratory test for the observed individuals and created a new variable by the rule: smoker_con = 1 if Cotinine >= 11 ng/mL; smoker_con = 1, otherwise. This rule is based on the fact that smokers generally have a higher amount of Cotinine in their body, and the scientific threshold is 11 ng/mL.
- Control variables: We also include covariates as control variables to exclude confounding effects, and the covariates include Alcohol consumption, BMI, Dietary, Sleep hours, Exercise habits, Family income, Gender, and Age. In the end, due to the exclusion of units with missing values, 3,817 out of the total units are analyzed in this section.

b. Models and Estimates: We applied different methods to estimate ATE and its standard error.

- Regression-based method w/o interactions: We compared OLS, Lasso, Random Forest, and Gradient Boosting models, and found that the Gradient Boosting model produced the lowest MSE on the test dataset. With the model, we obtained an ATE of -1.3546, and it is statistically significant under a significant level of 5%.
- Regression-based method w/ interactions: The same models are applied with interaction terms included. Again, we got a Gradient Boosting model with the lowest MSE. It estimated an ATE of -1.4727, which is significant.

- Inverse Probability Weighted Estimator (IPWE): To estimate the propensity score of each unit, we compared OLS Logistic regression, Lasso, Random Forest, and Gradient Boosting models and found the OLS Logistic regression model performed best in terms of the AUC matric. We then used the model to estimate the propensity scores, which range from 0.038 to 0.807(Fig. 2). To avoid the violation of the overlap assumption, we discarded about 20 units with a propensity score < 5%.

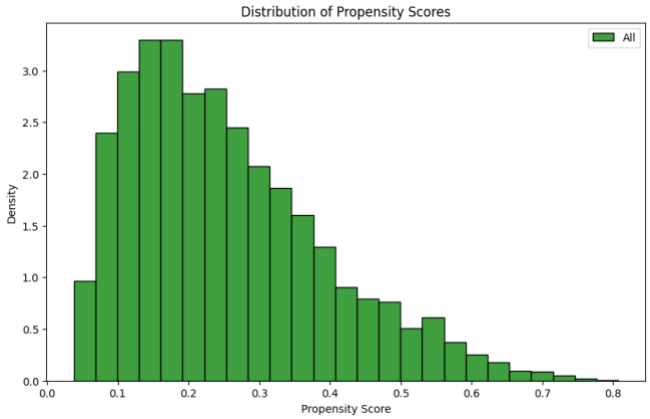

Figure. 5. The Distribution of Estimated Propensity Scores (Hypertension ~ Smoking)

We use the remaining samples and the estimated propensity scores to obtain the ATE of -2.0406 using IPWE, and it is significant.

- Augmented Inverse Probability Weighted Estimator (IPWE): In this method, we utilized both the Gradient Boosting model from the regression-based method and the Logistic Regression model from the propensity score method to produce a doubly robust estimator (AIPWE), which is -2.0403 and significant.
- Doubly Robust Partially Linear Model method: The same models are employed to obtain the outcome regression residuals and treatment regression residuals. From the residual-on-residual regression result, we obtained an ATE estimate of -1.4970, which is also significant.

Table 5 lists all the estimated results from 5 methods. No matter which method we use, we can produce a consistent negative treatment effect of smoking on the high-density lipoprotein cholesterol. Therefore, we can conclude that smoking can increase the risk of Hypertension.

Table 4. Summary of Estimate Results (Hypertension ~ Smoking)

	Regression Based	*Regression Based (w/ Interaction)*	*IPWE*	*AIPWE*	*PLM*
ATE Estimate	-1.3546	-1.4727	-2.0406	-2.0403	-1.4970
Standard Error	0.4957	0.5315	0.6530	0.5667	0.5130
t-statistic	-2.7327	-2.7706	-3.1249	-3.6001	-2.9160

(E) Hypertension ~ Sleep

a. Variable Selection

- Except for obesity and smoking, we're also interested in evaluating the relationship between hypertension and sleeping hours. Hence, we use HDL-C again as a representative measurement. In this case, we assumed the risk of Hypertension can be affected by sleeping hour.
- In terms of the treatment variable, we created a column named sleep_well followed by the rules: If the individual has less than 7 hours sleep on typical weekday will be viewed as not sleeping well. Sleep hours less then 7 hours would be 0; otherwise, 1. The intuition of setting this rule is based from the National Institutes of Health which states that adults who sleep less than 7 hours a night might have more health issue than those who sleep more than 7 hours.
- As for the covariates to exclude confounding effects, Alcohol consumption, BMI, Dietary condition, Exercise habits, Pulse, Family income, Gender, and Age are included. In the end, due to the exclusion of units with missing values, 3071 out of the total units are utilized.

b. Models and Estimates:

- Regression-based method w/o interactions: We use OLS model, and obtain an ATE of -0.9243 and it is not significant under a significant level of 5%.
- Regression-based method w/ interactions: The same model is applied, and we got an ATE of -0.8480. As a result, it is also not significant.
- Inverse Probability Weighted Estimator (IPWE): We chose Logistic approach again to estimate propensity score, which ranged from 0.447 to 0.8829. As a result, we obtained an ATE estimate -0.5613, and a standard error 0.6066. Still, the t-statistic revealed that there is no significant relationship between the variables being tested.

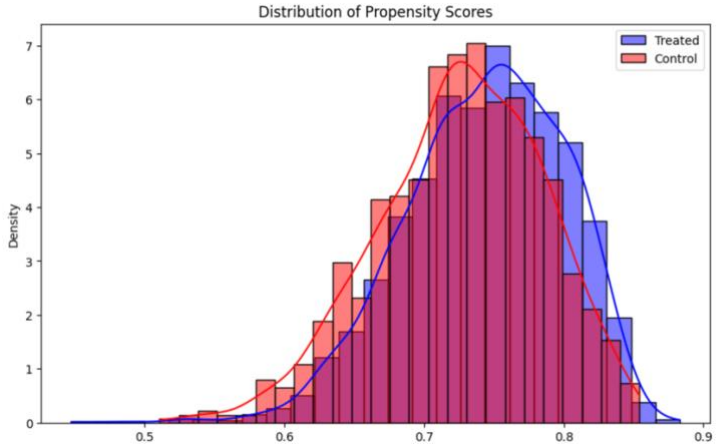

Figure 6. The Distribution of Estimated Propensity Scores (Hypertension ~ Sleep)

- Augmented Inverse Probability Weighted Estimator (IPWE): We also applied Logistic function and obtained ATE -0.5759 and standard error 0.5385. The t-statistics still indicate that there's no significance.
- Doubly Robust Partially Linear Model method: We applied both logistics and random forest model to capture the capturing non-linear relationships. We got an ATE with -0.3647, and standard error 0.3389. In this model, the t-statistic remains no significance.

Table 5. Summary of Estimate Results (Hypertension ~ Sleep)

	Regression Based	*Regression Based (w/ Interaction)*	*IPWE*	*AIPWE*	*PLM*
ATE Estimate	-0.9243	-0.8480	-0.5613	-0.5759	-0.3647
Standard Error	0.578	0.608	0.6066	0.5385	0.3389
t-statistic	-1.600	-1.395	-0.9253	-1.0694	-1.0761

5. Conclusion

The study confirms the significant impact of exercise and obesity on diabetes and the effects of obesity and smoking on hypertension. However, no significant relationship was found between sleep duration and hypertension. These findings highlight the importance of maintaining healthy habits to prevent chronic diseases, offering evidence-based guidelines for preventive healthcare. In short, you ate your way in, and you can walk your way out.

This observational study is limited by the inability to establish definitive causation due to potential unmeasured confounding factors. Furthermore, the data used in this study is from the 2017-2018 NHANES data. Expanding the study to include more recent data could provide more current insights. Additionally, future research could explore other potential factors that may contribute to chronic diseases, such as environmental factors. By addressing these limitations and expanding the scope of research, we can gain a more comprehensive understanding of the complex interactions between personal habits and health.

6. **Collaboration Plan:** All members equally participate in this project in writing, discussing, and data preparation.

Reference

Demographic database

Name	Category	Description
INDFMIN2	Continuous	Total family income (reported as a range value in dollars)
RIAGENDR	Binary	Gender of the participant (1=Female, 0=Male)
RIDAGEMN	Continuous	Age in months at screening - 0 to 24 months
RIDAGEYR	Continuous	Age in years at screening - 2 to 80 years
RIDAGEYR_fixed	Continuous	Adjust the unit of RIDAGEMN to year and combine data with RIDAGYR - 0 to 80 years

Dietary database

Name	Category	Description
DS1DS	Binary	Dietary supplement, 1 for Yes, 0 for No.
DS1AN	Binary	Antacids taken? 1 for Yes, 0 for No.
DBQ095Z	Binary	Used salt? 1 for ordinary salt, lite salt, and salt substitute. 0 for not using any salt.
DBD100	Binary	How often add salt to food? 1 for occasionally and very often. 0 for rarely and refused.
DRQSPREP	Binary	Salt used for preparation. 1 for occasionally and very often. 0 for never and rarely.
DR1_300	Binary	Compare food consumed yesterday to usual. 1 for much more than usual and usual. 0 for much less than usual and refused.
DR1TSFAT_criteria	Binary	Criteria for total saturated fatty acids ≤20. 1 stands for satisfied; otherwise, 0.
DR1TSUGR_criteria	Binary	Criteria for total sugars ≤50. 1 stands for satisfied; otherwise, 0.
DR1TSODI_criteria	Binary	Criteria for sodium ≤2000. 1 stands for satisfied; otherwise, 0.
DR1TFIBE_criteria	Binary	Criteria for dietary fiber ≥31. 1 stands for satisfied; otherwise, 0.
unhealthy_condition	Binary	Unhealthy condition indicator based on criteria. If those four criteria are satisfied no more than two, it stands for unhealthy; otherwise, healthy. 1 for unhealthy, 0 for healthy.

Laboratory database

Name	Category	Description

LBDHDD	Continuous	Direct HDL-Cholesterol (mg/dL), under 60 mg/dL indicates the risk of heart disease, stroke and hypertension.
LBXCOT	Continuous	Cotinine, Serum (ng/mL), can be used as markers for active smoking and as indices for secondhand smoke (SHS) exposure. Higher than 11 ng/mL can be recognized as smoker, typically around 0.1 ng/mL for non-smoker.
smoker_con	Binary	Derived from the 'LBXCOT' column, 1 for value larger or equal to 11, 0 for else.

Questionnaire database

Name	Category	Description
SLD012	Continuous	Average sleep hours on weekdays.
DIQ010	Binary	Doctor told you have diabetes. Yes is 1.
PAQ650	Binary	Vigorous recreational activities. Performing any during a typical week is 1.
PAQ665	Binary	Moderate recreational activities. Performing any during a typical week is 1.
sleep_well	Binary	If the individual has less than 7 hours sleep on typical weekday will be viewed as not sleeping well and denoted as 0.
exercise	Binary	If the individual has either PAQ650 or PAQ665 as 1, they will be viewed as doing exercise on the typical week, and denoted with 1.

Examination Database

Name	Category	Description
BPXPLS	Continuous	60 sec. pulse (30 sec. pulse * 2)
BPXPULS	Binary	Pulse regular or irregular? irregular is 1.
BPXSY1/2/3	Continuous	Systolic: Blood pressure (first/second/third reading) mm Hg
BPXDI 1/2/3	Continuous	Diastolic: Blood pressure (first/second/third reading) mm Hg
Sys_AVEBP	Continuous	Average Systolic Blood Pressure mm Hg
Dia_AVEBP	Continuous	Average Diastolic Blood Pressure mm Hg
BMXWT	Continuous	Weight (kg)
BMXHT	Continuous	Standing Height (cm)
BMXBMI	Continuous	Body Mass Index (kg/m**2)
BMXARMC	Continuous	Arm Circumference (cm)
BMXWAIST	Continuous	Waist Circumference (cm)
BMXHIP	Continuous	Hip Circumference (cm)
WH_ratio	Continuous	Waist Circumference/ Hip Circumference (ratio)