Is Physical Activity Associated with C-protein Level in US Population with Different Framingham Risk Scores?

https://github.com/YinsuW/Wang1Asfaw2Qian3_ENV872_EDA_FinalProject.git

Yinsu Wang, Yared Asfaw, April Qian

Contents

1 Rationale and Research Questions 5

2 Dataset Information 6

3 Exploratory Analysis 7

3.1 Table 1 Baseline Characteristics of Participants 7

3.2 Scatterplot of the relationship between PA and C-protein 8

3.3 Heatmap of PA and C-protein under different Framingham risk scores 8

4 Analysis 11

4.1 Question 1: Is physical activity associated with the level of C-protein? If so, what is the exact relationship? Is it dose-response? 11

4.2 Question 2: Are different levels of framingham risk scores the effect modifiers for this relationship between physical activity and C-protein? How do they modify the relationship? 11

5 Summary and Conclusions 14

6 References 15

List of Tables

List of Figures

1 The relationship between Physical Activity and C-protein Concentration 8

2 The relationship between Physical Activity and C-protein Concentration (Log-transformed) 9

3 C-protein Levels under Physical Activity Status and Framingham Risk Scores (Log-transformed) 10

4 C-protein Level under Different Physical Activity Status (Log-transformed) . . . 12

5 C-protein Level under Differen PA Status by Different Framingham Risk Scores (Log-tranformed) 13

1 Rationale and Research Questions

Cardiovascular disease (CVD) is the leading cause of death in the United States. About 659,000 people in the United States died from CVD each year (Centers for Disease Control and Prevention, 2022). The formulation of coagulation is the fundamental pathophysiological mechanisms behind CVD. Studies have shown that coagulation biomarkers such as protein-C may be regarded as potential CVD risk factors, thus its level might be related to the risk of CVD (Zakai et tal, 2018). The Framingham risk score is a gender-specific algorithm used to estimate the ten-year cardiovascular risk of a person. The higher the score is, the higher the risk will be. Therefore, it can be treated as a parameter reflecting the risk of CVD. Physical activity (PA) is a widely accepted approach both by European Society of Cardiology and American Heart Association guidelines on CVD treatment (Arnett et al, 2019; Visseren et al, 2021). However, little is known whether PA would reduce the coagulation process, therefore reducing the risk of CVD. Thus, in this study, we used NHANES dataset to explore the relationship between PA and protein-C, one of the biomarkers of coagulation, under different Framingham risk score, aiming to find whether PA would be related to reduced protein-C level in individuals with high Framingham risk scores.

- Question 1: Is physical activity associated with the level of C-protein? If so, what is the exact relationship? Is it dose-response?
- Question 2: Are different levels of framingham risk scores the effect modifiers for this relationship between physical activity and C-protein? How do they modify the relationship?

2 Dataset Information

The dataset used for this project is obtained from the National Health and Nutrition Examination Survey (NHNES) which were conducted on a periodic basis from 1999 to 2010 by the National Center for Health Statistics, Division of Health Examination Statistics, part of the Centers for Disease Control and Prevention. The survey conducted between the year 1999 to 2010 every year by interviewing on average 10,360 individuals of all ages in their homes (9,965 (1999-2000), 11,039 (2001- 2002), 10,122 (2003-2004), 10,348 (2005-2006), 10,149 (2007-2008), 10,537 (2009-2010) individuals of all ages). The data were collected between January of the beginning year and December of the ending year). The data is available for public at https://www.cdc.gov/nchs/nhanes.htm.

The major objectives of the survey as indicated in the CDC website are:

1. To estimate the number and percent of persons in the U.S. population and designated subgroups with selected diseases and risk factors;
2. To monitor trends in the prevalence, awareness, treatment and control of selected diseases;
3. To monitor trends in risk behaviors and environmental exposures;
4. To analyze risk factors for selected diseases;
5. To study the relationship between diet, nutrition and health;
6. To explore emerging public health issues and new technologies;
7. To establish a national probability sample of genetic material for future genetic research;
8. To establish and maintain a national probability sample of baseline information on health and nutritional status.

The survey has demographic, dietary, medical examination, laboratory, questionnaire, and limited access data categories. The target population for this survey is the civilian, noninstitutionalized U.S. population which includes over-sampling of low-income persons, adolescents 12-19 years, persons 60+ years of age, African Americans, and Mexican Americans. For this project the data categories that the team selected for the analysis are demographic data, examination data, laboratory data and questionnaire data. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests of participants. The dataset will be used to investigate the relationship between physical exercise or activity and C-protein level (to explain the behavior of reactive C-protein in response to physical activity) in patients that have different levels of Framingham scores.

Detail information on data collection procedures, household interview data collection procedures, sources of the questions, questionnaire target populations, health examination component, MEC operations, second day examinations and dietary interviews, home examinations, sample person demographics file and guidelines for data users is available at https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=1999.

In the process of data wrangling, we first included participants whose ages were equal or above 20 years using NHANES datasets from 1999 to 2010. We then excluded participants whose cholesterol (HDL, TC), blood pressure, diabetes, PA status, race, education level, family PIR were missing. Finally, a total of 1079 participants were included in the analyses.

3 Exploratory Analysis

The analysis takes into consideration of the main characteristics of the participants which might have an effect on their level of C-protein concentration in the presence and absence of physical activity. These participants characteristics considered in the analysis are age, gender, ethnicity, education levels, family PIR, systolic blood pressure, direct HDL-Cholesterol, total cholesterol, and body mass index. The proportion of participants in terms of those categories of characteristics/information is indicated in Table 1 below.

3.1 Table 1 Baseline Characteristics of Participants

	Total
	(N=1079)
Age	
Elder	523 (48.5%)
Middle Aged	435 (40.3%)
Young	121 (11.2%)
Gender	
Female	378 (35.0%)
Male	701 (65.0%)
Ethnicity	
Hispanic	194 (18.0%)
Non-Hispanic Black	231 (21.4%)
Non-Hispanic White	623 (57.7%)
Other	31 (2.9%)
Education	
Below high school	334 (31.0%)
College or above	474 (43.9%)
High school	271 (25.1%)
Family PIR	
0≤PIR<1	244 (22.6%)
1≤PIR<2	289 (26.8%)
2≤PIR<3	172 (15.9%)
PIR ≥3	374 (34.7%)
Systolic Blood Pressure (mmHg)	
Mean (SD)	131 (19.0)
Median [Min, Max]	128 [82.0, 204]
Direct HDL-Cholesterol (mmol/L)	
Mean (SD)	1.29 (0.447)
Median [Min, Max]	1.19 [0.340, 3.41]
Total Cholesterol (mmol/L)	
Mean (SD)	5.11 (1.18)
Median [Min, Max]	5.04 [2.56, 10.7]
Body Mass Index (kg/m2)	

	Total
Mean (SD)	30.6 (6.19)
Median [Min, Max]	30.0 [17.6, 65.4]

The project team conducted an explanatory analysis of the dataset by plotting a scatter plot to show the response of C-protein concentration at different physical activity levels (Figure 1). To meaningfully capture and investigate the relationship in the scatter plot, log transformation of the data made (Figure 2). Similarly, the team produced a heatmap of physical activity and C-protein concentrations to demonstrate the level of C-protein concentrations at different levels of physical activity under different Framingham Risk Scores (Figure 3).

3.2 Scatterplot of the relationship between PA and C-protein

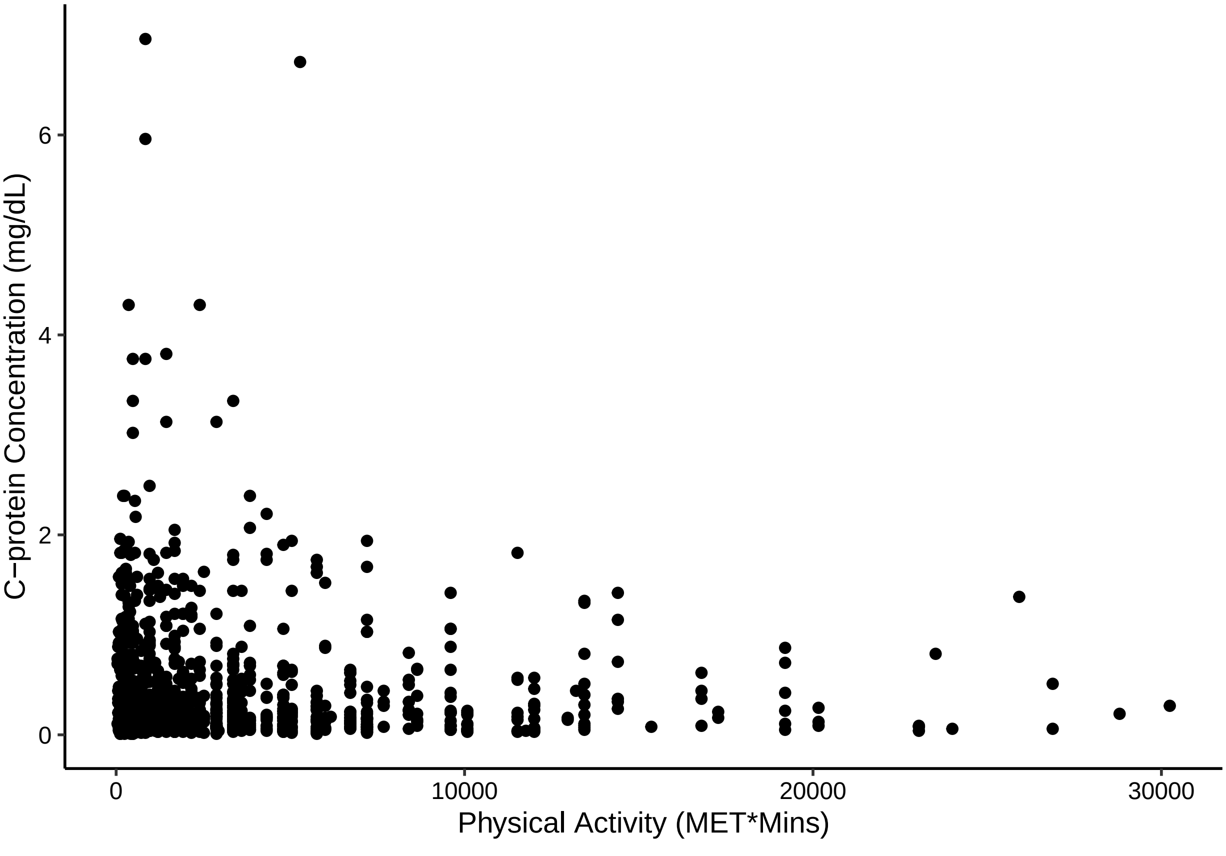

Figure 1: The relationship between Physical Activity and C-protein Concentration

3.3 Heatmap of PA and C-protein under different Framingham risk scores

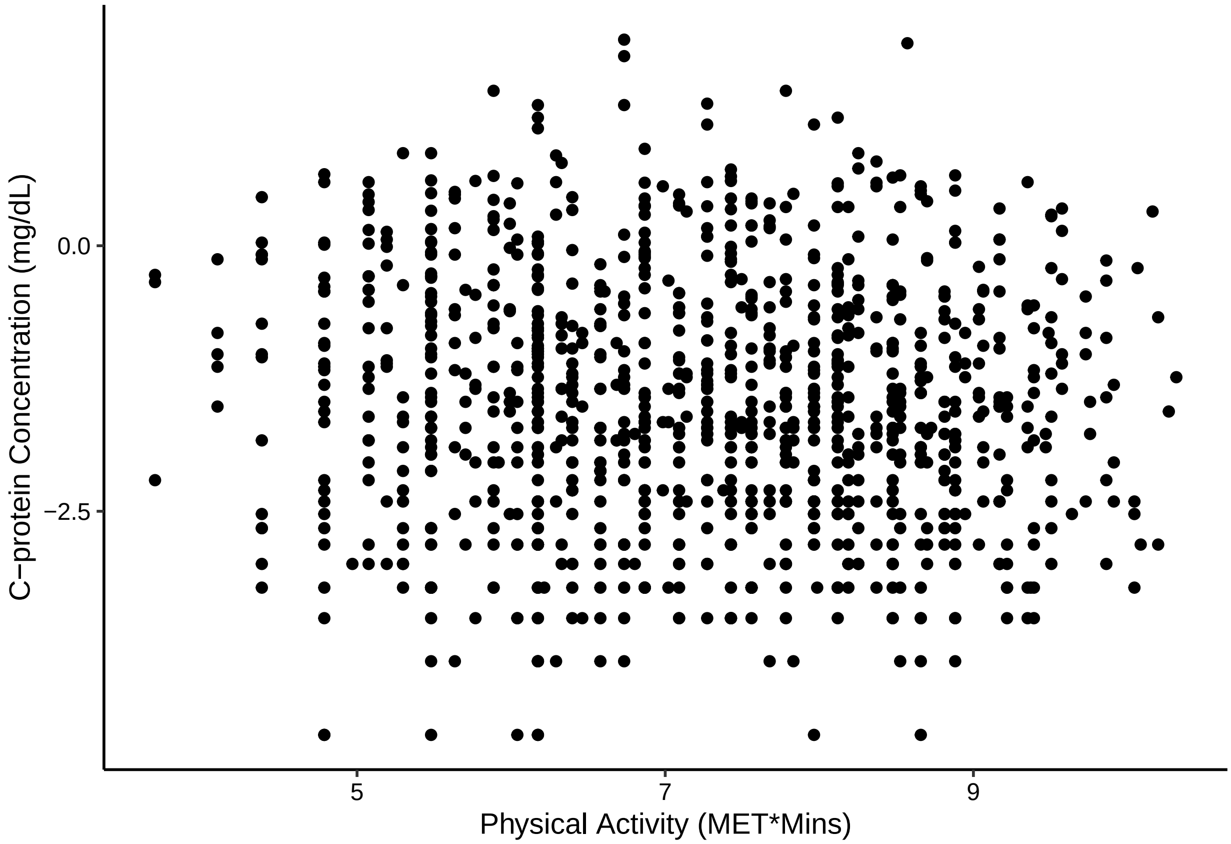

Figure 2: The relationship between Physical Activity and C-protein Concentration (Log-transformed)

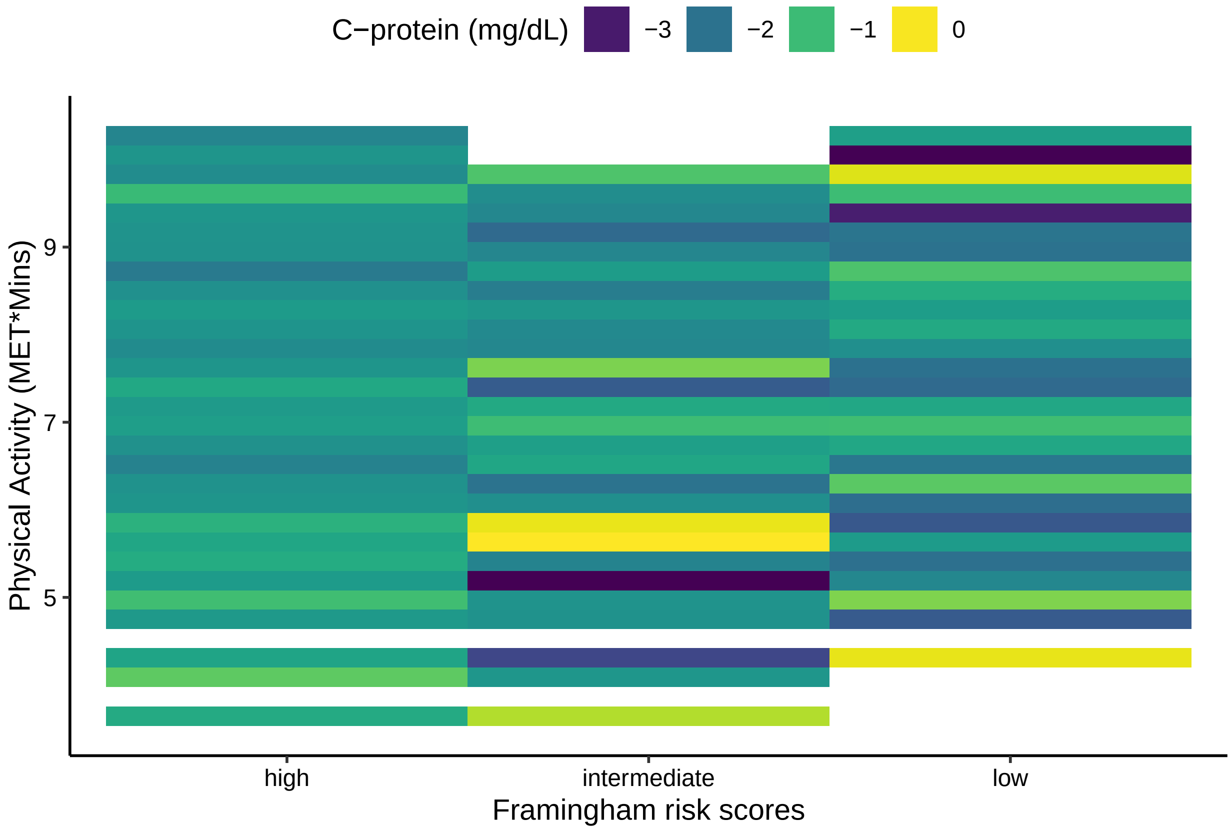

Figure 3: C-protein Levels under Physical Activity Status and Framingham Risk Scores (Log-transformed)

4 Analysis

4.1 Question 1: Is physical activity associated with the level of C-protein? If so, what is the exact relationship? Is it dose-response?

- Methods: We used the linear regression method to build up our model. First, after we plotted the relationship between PA and C-protein level with raw data, we found that most data positioned at the left bottom of the plot (Figure 1), suggesting that we need to take log transformation of our PA and C-protein variables. We, therefore, used the transferred data for further analysis. For the analysis, we first conducted a crude model using the "lm" function to see whether there was linear association between PA and C-protein level. We then adjusted gender and age to see whether this association changed or not. Finally, we conducted a linear model by adjusting multiple variables including gender, age, family PIR, education level, ethnicity, diabetes history, blood pressure, total cholesterol together with direct HDL-cholesterol to see whether there are significant association between the two variables.

- Results: In the crude model, PA shows no significant association with C-protein (F(1164)= 3.425; p-value: 0.06449, Figure 4). We then conducted the gender- and age-adjusted model for this relationship. In this adjusted model, we saw a relationship between gender and C-protein (F (1162) = 8.311; p-value: 1.802e-05) but not the association between PA and C-protein (p-value: 0.123). Finally, in the adjusted model of multiple variables, we still did not see a positive relationship between PA and C-protein (F(1066) = 24.24; p-value: 0.075685).

4.2 Question 2: Are different levels of framingham risk scores the effect modifiers for this relationship between physical activity and C-protein? How do they modify the relationship?

- Methods: We then conducted an analysis to identify whether different levels of Framingham risk scores modify the effect for the relationship between physical activity and C-protein by "lm" function. We first added an interaction term "PA*risk_level" to the previous crude linear model. We then added the same interaction term to the gender- and age-adjusted model.

- Results: The results shows that there no interaction between PA and Framingham risk scores (F-statistic: 1.175 on 5 and 1160 DF, p-value: 0.3191, Figure 5) in the crude model. In the adjusted model, we also found no interaction between PA and Framingham risk scores, suggesting that Framingham risk scores may not be an effect modifier for the relationship between PA and C-protein.

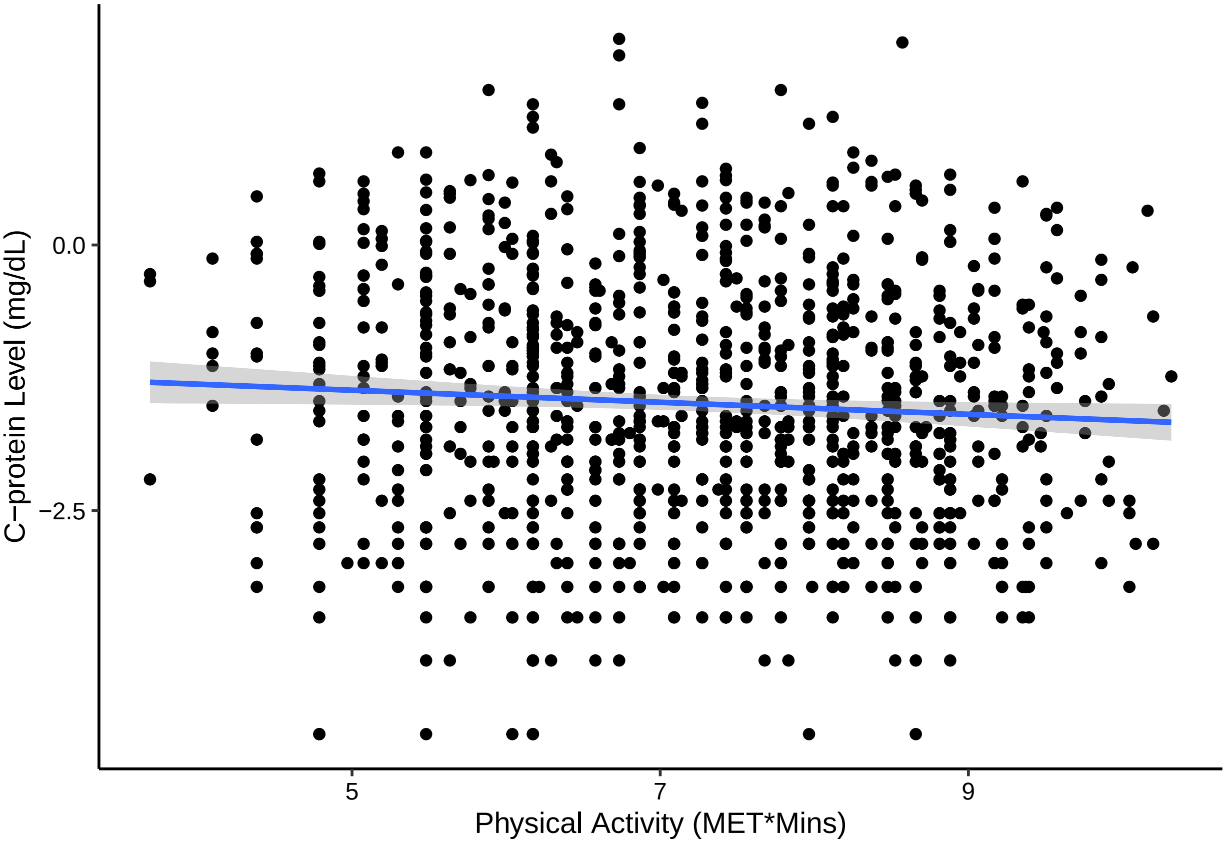

Figure 4: C-protein Level under Different Physical Activity Status (Log-transformed)

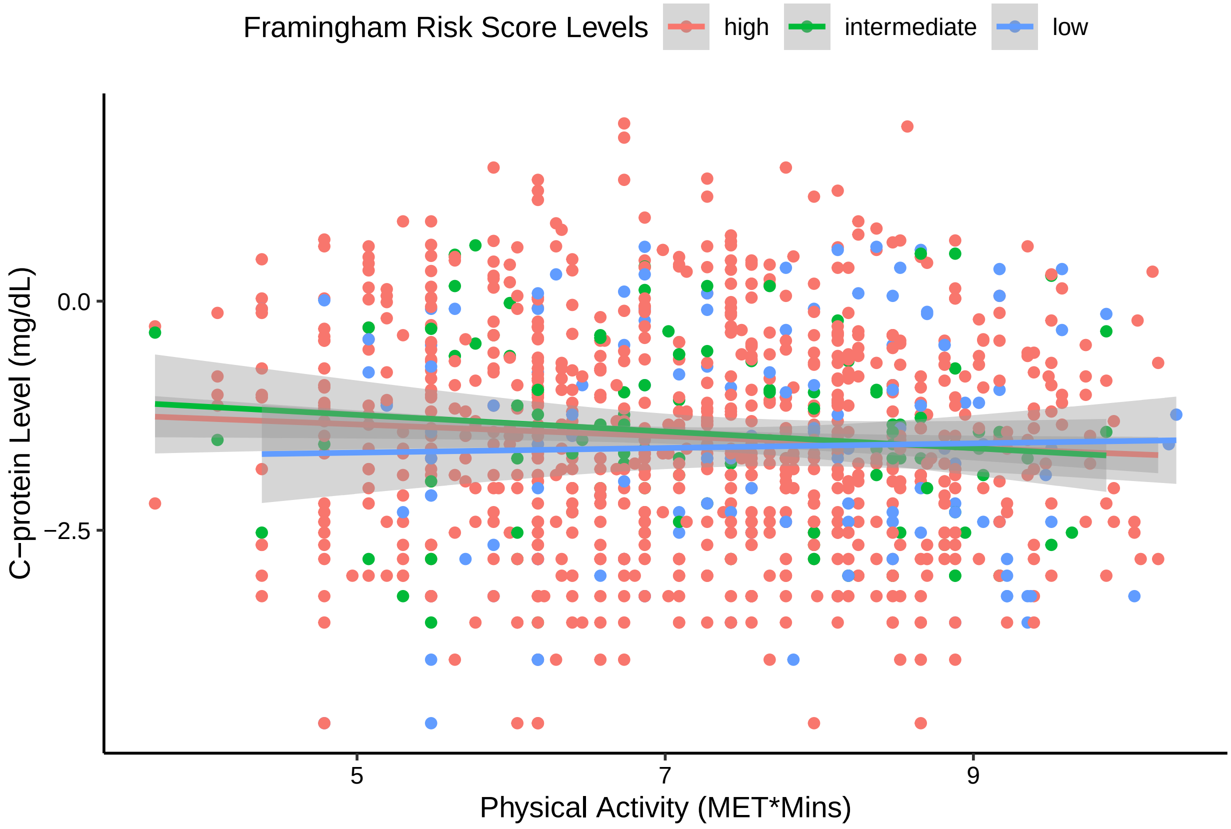

Figure 5: C-protein Level under Differen PA Status by Different Framingham Risk Scores (Log-tranformed)

5 Summary and Conclusions

Our analyses of the NHANES datasets suggest that there might not be a positive association between PA and C-protein level, with and without under different Framingham risk scores for population of the United States, even adjusted by gender and age, as well as other potential confounders such as ethnicity, education level, diabetes history, blood pressure and lipids. This might suggest that the relationship might not be exist. Previous studies demonstrated that low C-protein is a risk factor for venous thrombosis, but less is known whether it is also a risk factor for arterial disease such as CVD (Folsom et al, 2009). Previous studies also suggested that C-protein might not be an independent risk factor for CVD, in which case, PA might not be moderate the level of C-protein for people with high CVD level (high Framingham risk score) (Folsom et al, 2009). These results are congruent with ours. However, these analyses have their limitations. First, the processed data sample is relatively small, having only more than one thousand samples, which might cause random error of the results. Second, we used only linear model to conduct the analyses, potentially making this model not fit very well. The best value of R-squared is just more than 20% in the models, making it less confident to draw a strong conclusion using this linear model. Third, the study design is cross-sectional, which might not be validate. Further studies could be conducted using more complicated way of analyses such as using machine learning to establish a non-linear model to figure out the better model for this relationship between PA and C-protein. Studies such as cohort studies may also be required to seek for a potential positive result.

6 References

Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., Michos, E. D., Miedema, M. D., Muñoz, D., Smith, S. C., Jr, Virani, S. S., Williams, K. A., Sr, Yeboah, J., & Ziaeian, B. (2019). 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Journal of the American College of Cardiology, 74(10), e177–e232. https://doi.org/10.1016/j.jacc.2019.03.010

Folsom, A. R., Ohira, T., Yamagishi, K., & Cushman, M. (2009). Low protein C and incidence of ischemic stroke and coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study. Journal of thrombosis and haemostasis : JTH, 7(11), 1774–1778. https://doi.org/10.1111/j.1538-7836.2009.03577.x

Visseren, F., Mach, F., Smulders, Y. M., Carballo, D., Koskinas, K. C., Bäck, M., Benetos, A., Biffi, A., Boavida, J. M., Capodanno, D., Cosyns, B., Crawford, C., Davos, C. H., Desormais, I., Di Angelantonio, E., Franco, O. H., Halvorsen, S., Hobbs, F., Hollander, M., Jankowska, E. A., … ESC Scientific Document Group (2021). 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. European heart journal, 42(34), 3227–3337. https://doi.org/10.1093/eurheartj/ehab484

Zakai, N. A., Judd, S. E., Kissela, B., Howard, G., Safford, M. M., & Cushman, M. (2018). Factor VIII, Protein C and Cardiovascular Disease Risk: The REasons for Geographic and Racial Differences in Stroke Study (REGARDS). Thrombosis and haemostasis, 118(7), 1305–1315. https://doi.org/10.1055/s-0038-1655766