Inferring the Relationship between High-Density Lipoprotein (HDL) Cholesterol Levels and Alcohol Drinking Pattern

Abstract

In this study, a linear model is employed to infer the relationship between high-density lipoprotein cholesterol (HDL-C) and individuals drinking patterns. we utilize drinking days per year and heavy drinking as main study focus to assess the drink patterns. The results indicate that, under moderate alcohol consumption, an increase in drinking days per year is correlated with elevated HDL-C levels. Additionally, this effect diminishes with increasing age and intensifies with higher levels of poverty. For individuals defined as heavy drinkers in our study, it was observed that a higher poverty level is associated with a lower total HDL-C level. Our findings align with information from other studies.

Background

Epidemiological and clinical evidence has suggested high-density lipoprotein cholesterol (HDL-C) is beneficial for human health, particularly in diminishing the risk of heart failure, reducing inflammation, preventing the formation of blood blocks, and so on (Sirtori et al., 2019). There are many factors could impact the HDL-C level in the blood, from genetic to environmental and personal level (Weissglas-Volkov et al., 2010).

A prior investigation utilizing the NHANES II database reveals that a 1g alcohol consumption resulted in a 0.87mg/dl increase in the mean HDL-C level (Linn et al., 1993). Subsequently, another health professionals' follow-up study further validates this theory, and specifically pointed out that a drinking pattern for individuals matters more than the total amount of consumption when it comes to the impact on HDL-C levels (Mukamal et al., 2003). This implies that consistently maintaining a moderate drinking habit can raise HDL-C levels, thereby promoting positive effects on one's overall health condition. Notably, the term "moderate drink" is frequently mentioned, however, the determining standards vary across different studies. In our study, we focus on analyzing the drinking patterns of individuals. We adopt a standard definition from National Institute on Alcohol Abuse and Alcoholism to identify the heavy drinker group in our source dataset, and then we infer the impact of individual drinking patterns on HDL-C levels in the blood. Additionally, we explore the general interaction between other covariates from demographic parameters and socioeconomic factors.

Methods

Study Population

We initiate our study using the source database from NHANES, which includes 10000 observations and with 76 variables related to demographic information (Age, Sex, Race, etc.), Health indicators (Diabetes, Alcohol consumption, etc.) and socioeconomic factors (House ownership, Poverties.). NHANES strategically over-samples individuals aged 60 and older, African Americans, and Hispanics, we observe

duplicates resulting from this consideration. Removing those duplicate observations won't affect the inference result. Additionally, the target study group is drinkers, so people who don't consume any alcohol in a year are excluded from the source database. After eliminating all the missing data, our final study database involves 2672 observations.

Outcome variable

In our study, the variable of interest is direct HDL-C, commonly referred to as "good" cholesterol. The numerical range spans from 0.41 to 4.03. Modeling is based on direct HDL-C, with a median of 1.32, and a mean of 1.38. The distribution histogram of Direct HDL-C is right skewed. Therefore, we applied a log(Y) transformation, resulting in an overall distribution that appears more normal.

In figure 1, we compare the distribution, residuals, and QQ plots of the response variable with their log(Y) transformation. In the first column, the histogram distribution is right-skewed, and the residuals exhibit a trend of increasing variance with fitted values, indicating non-constant residual variance. The QQ plot reveals a departure from normality in the residuals. Conversely, in the second column, the residuals exhibit a more normal distribution. Residuals demonstrate a more constant variance with fitted values. The histogram distribution is symmetric and uniform.

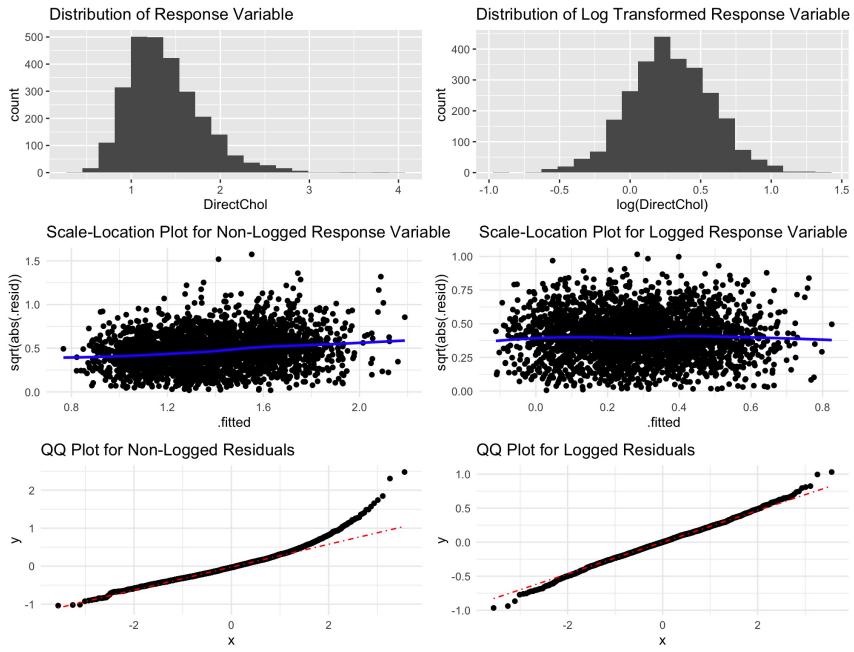

Figure 1: Comparison of Response Variable Pre/Post Log Transformation

Predictor of interest

In our study, we focus on two key variables related to drinking patterns: "drinking days per year" and "heavy drink". "Drinking days per year" denotes the estimated number of days in the past year during which participants consumed alcoholic beverages. We further calculated the "average drinking per day" based on the "drinking days per year." According to the research (Leaks K et al., 2023), individuals are classified as "heavy drinkers" if their average daily alcohol consumption exceeds 2 for males and 1 for females. Operationally, we define "heavy drink" as 1 and "moderate drink" as 0.

Covariates

We categorized our variables into three classes: demographics, socio-economic status, and health indicators. Demographic variables include age, gender, and race. Studies have emphasized the importance of adjusting for demographic variables when determining relationships among our variables of interest. The age range spans from 0 to 80 years, with an average of 35.5 and a median of 34; we centered it for interpretability. To understand the diverse racial representation, we classified individuals into Black, Hispanic, Mexican, and others, with "White" serving as our reference group for analytical purposes. Socio-economic status variables encompass poverty and education. We introduced a new poverty variable, ranging from 0 to 5, with higher values indicating increased poverty (0: very wealthy, 4: just reaching the poverty line, 5: very poor). Health indicator variables include BMI, diabetes, smoking, physical activities, and depression. To balance variables, we recoded depression as a binary variable, considering both "several depressed" and "most depressed" as depressed (0: non-depressed; 1: depressed).

Table 1 presents the descriptive statistics for all variables. Our covariates of interest indicate that 12% of participants are classified as heavy drinkers. From the GVIFˆ(1/(2Df)) values in the output, none of the values exceed sqrt (5), which is equivalent to VIF < 5, we do not observe any potential collinearity issues.

Table 1: Descriptive Statistic of Covariates

Variables	Description	Overall (N = 2672)
Direct HDL cholesterol (mmol/L)	HDL-C is a type of lipid considered beneficial for cardiovascular health and is commonly referred to as "good" cholesterol.	*
Drinking days per year (d)	Estimated number of days over the past year that participant drank alcoholic bever- ages.	*
Heavy drink (%)	1 = heavy drink, 0 = no heavy drink	323 (12.09)
Centered age	Centered at mean	*
Gender (%)	1 = male, 0 = female	1427 (53.41)
Race (%)	White	1733 (64.86)
	Black	355 (13.29)
	Hispanic	162 (6.06)
	Mexican	246 (9.21)
	Other	176 (6.59)
Poverty	Ratio of the gap between five times the poverty line and income to the poverty line itself.	*
Education (%)	8th Grade	121 (4.53)
	9-11th Grade	311 (11.64)
	High School	546 (20.43)
	Some College	846 (31.66)
	College Graduation	848 (31.74)

BMI (kg/m2) (%)	12 to 18.5	41 (1.53)
	18.5 to 24.9	758 (28.37)
	25.0 to 29.9	955 (35.74)
	30.0 to plus	918 (34.36)
Diabetes (%)	1 = yes, 0 = no	247 (9.24)
Smoking (%)	1 = smoked at least 100 cigarettes in their entire life, 0 = no	1271 (47.57)
Physic activities (%)	1 = does moderate or vigorous-intensity sports, fitness or recreational activities, 0 = no	1543 (57.75)
Depression (%)	1 = depressed, 0 = non-depressed	575 (21.52)

* Continuous variables, overall N = 2672

Interactions

Our study primarily identifies interactions based on a literature review. The combined impact of poverty, race, and gender on alcohol outcomes is multiplicative (Glass et al., 2017). Among drinkers, the influence of living near the poverty line on severe episodic drinking depends on race and gender. Compared to men and women from other racial/ethnic groups, Black men and Black women have a higher risk of severe episodic drinking when their income is close to the poverty line. Therefore, we have chosen gender, race, age, and poverty as key factors for interaction.

Statistical Analysis

In our statistical analysis, we focused on exploring the complex relationship between drinking patterns and direct HDL-C levels. Through careful data screening and processing, we aimed to extract more effective information. We removed duplicate IDs and specifically analyzed drinkers, resulting in a final observation count of 2672. We then honed in on two key variables: "drinking days per year" and "heavy drinking." Simultaneously, we addressed multicollinearity and balanced categorical variables. Drawing insights from the literature, we considered factors like race, gender, age, and poverty that may influence interactions. Following exploratory analysis and variable transformations, we established an initial model, evaluated its diagnostics, retained interactions involving age and poverty, and applied a log(Y) transformation. Detailed statistical results, effect sizes, and significance levels will be presented in subsequent sections.

Table shows the estimated regression coefficients and the p-values of the covariates in our final model. Adjusting for other variables, the drinking days per year associate with the total HDL-C level significantly: the more days a person drink per year, the higher the total HDL-C level will be (estimated coefficient = 0.060, 95% CI: (0.040, 0.080), p-value < 0.001), which means with every 100 drinking days increased per year, the total HDL-Cholesterol level is estimated to increase by 6.18%.

Model

The final model achieves an adjusted R^2 of 0.328, reflecting the extent to which the independent variables account for the variance in the dependent variable. Moreover, the model's significance is confirmed by the F-test, emphasizing its overall

effectiveness and statistical validity. Table 2 shows the estimated regression coefficients and the p-values of the covariates in our final model. Adjusting for other variables, the drinking days per year associate with the total HDL-Cholesterol level significantly: the more days a person drink per year, the higher the total HDL-C level will be (estimated coefficient = 0.060, 95%CI: (0.040, 0.080), p-value < 0.001), which means with every 100 drinking days increased per year, the total HDL-Cholesterol level is estimated to increase by 6.18%.

The table also displays the estimated coefficients for four interaction terms. Notably, for individuals classified as heavy drinkers, a higher poverty level is associated with a lower total HDL-C level (estimated coefficient = -0.020, 95% CI: (-0.041, 0.001), p-value = 0.064). To illustrate, a one-unit increase in the poverty level index among heavy drinkers corresponds to a 1.98% decrease in the total HDL-C level. Surprisingly, the estimated coefficient for the interaction between centered age and heavy drinking suggests that, as the age of heavy drinkers increases, their total HDL-C level also rises (estimated coefficient = 0.002, 95% CI: (0.000, 0.004), p-value = 0.051). In practical terms, a one-year increase in age among heavy drinkers is associated with a 0.2% increase in the total HDL-C level. Both p-values (> 0.05) indicate a lack of significance at the 0.05 level.

From the interaction term between age and drinking days per year, we can see that the benefits derived from an increase in drinking days decreases as age increases (estimated coefficient = -0.001, 95% CI: (-0.002, 0.001), p-value = 0.002), meaning that for each one-unit increase in age, an additional 100 drinking days per year will weaken the positive effect on total HDL-C by 0.1%. Finally, in the interaction between poverty and drinking days per year, the estimated coefficient suggests that with the increase of poverty level, the benefit brings by drinking days per year increases (estimated coefficient = 0.002, 95% CI: (0.000, 0.004), p-value = 0.039). Specifically, for every unit increase in the poverty level index, the positive impact of an additional 100 drinking days on the total HDL-C level increases by 0.2%.

Table 2: Model Summary (Adjusted $R^2 = 0.328$)

Variables	Estimated Coefficient	p-value
Intercept	0.461	<0.001***
Centered Age	0.003	<0.001***
Gender (male = 1, female = 0)	-0.198	<0.001***
Race_Black	0.084	<0.001***
Race_Hispanic	0.008	0.704
Race_Mexican	0.041	0.025 *
Poverty	-0.012	0.004 **
Education_9-11th Grade	0.037	0.165
Education_High School	0.049	0.087 .
Education_Some College	0.049	0.056 .
Education_College Grad	0.075	0.005 **
BMI_18.5-24.9	-0.075	0.057 .
BMI_25.0-29.9	-0.206	<0.001***
BMI_30.0-plus	-0.309	<0.001***

Diabetes	-0.048	0.004	**
Smoking	-0.029	0.004	**
Depression	-0.015	0.193	
Physic Activities	0.027	0.009	**
Drinking Days per Year [a]	0.060	<0.001	***
Heavy Drink	0.046	0.167	
Centered Age: Drinking Days per Year [a]	-0.001	0.002	**
Centered Age: Heavy Drink	0.002	0.051	.
Poverty: Drinking Days per Year [a]	0.007	0.039	*
Poverty: Heavy Drink	-0.020	0.064	.

[a] Coefficients represent the effect per 100 days

Model Diagnostics

We checked our final model against the fundamental assumptions of linear regression, known as the 'LINE' assumptions. To assess Linearity, we employed Partial Residual plots for non-categorical variables to examine their trend lines, and no significant deviations were found, confirming that our model adheres to the linearity assumption. We assessed Equal Variance with Scale-Location plots, revealing no noticeable pattern across the predicted range, supporting the assumption of constant variance. By applying a log transformation, the distribution became more normal, as shown by a well-fitting QQ-plot, ensuring our model meets the Normality Assumption. The Independence Assumption was examined with the Durbin-Watson Test, and the non-rejection of the null hypothesis suggests no apparent dependence in the model.

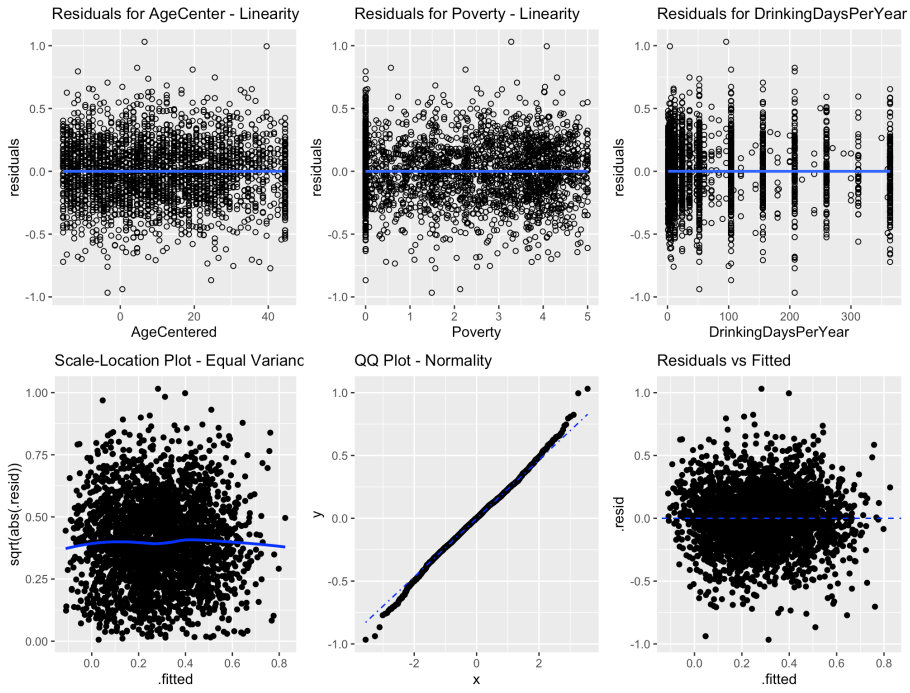

Figure 2: Comprehensive Model Diagnostics

Furthermore, we conducted Variance Inflation Factor (VIF) analysis to assess multicollinearity. The Generalized VIF (GVIF) values, calculated as GVIF^(1/(2*df)),

consistently stayed within the acceptable range of the square root of 5. This indicates that there are no notable issues of multicollinearity among the predictor variables.

Table 3: Multicollinearity Assessment using Generalized VIF

Variables	GVIF	df	GVIF^(1/(2*df))
Centered Age	2.097	1	1.448
Gender	1.108	1	1.053
Race	1.363	4	1.039
Poverty	2.109	1	1.452
Education	1.649	4	1.064
BMI	1.151	3	1.024
Diabetes	1.116	1	1.056
Smoking	1.146	1	1.071
Depression	1.075	1	1.037
Physic Activities	1.189	1	1.090
Drinking Days per Year	5.111	1	2.261
Heavy Drink	5.233	1	2.288
Centered Age: Drinking Days per Year	4.090	1	2.022
Centered Age: Heavy Drink	2.116	1	1.455
Poverty: Drinking Days per Year	4.057	1	2.014
Poverty: Heavy Drink	4.795	1	2.190

Lastly, we dealt with Outliers and High-leverage Points. Outliers were identified through studentized residuals, and high-leverage points were detected using the diagonal of the hat matrix. Following the removal of influential observations, model refitting revealed no noticeable differences, affirming the robustness of the model.

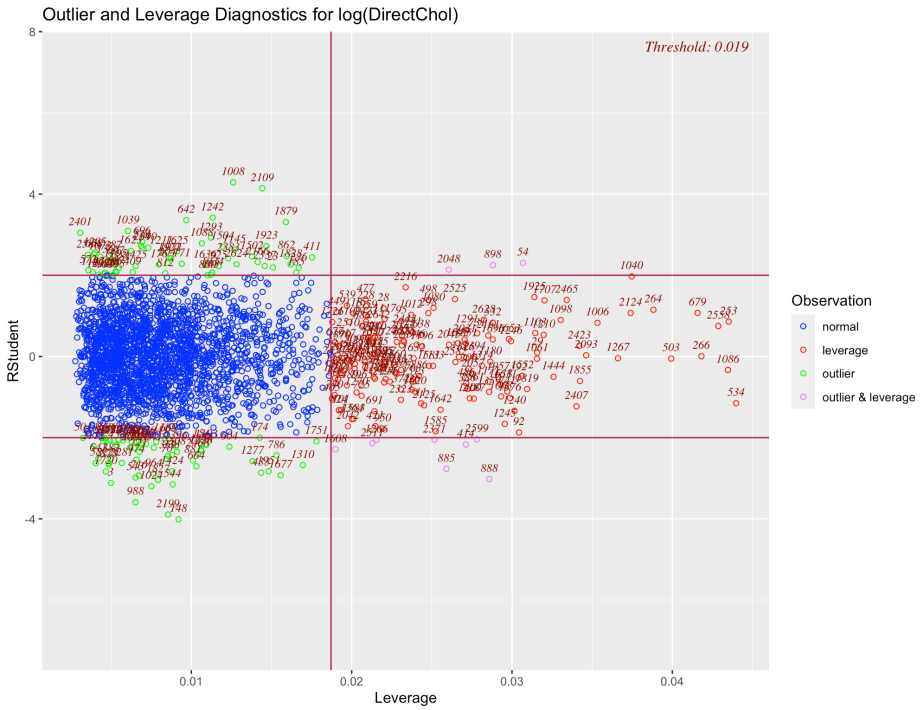

Figure 3: Identification of Influential Points

Conclusion and Limitations

In this study, we are interested in inferring how drinking patterns might affect HDL-C level. Our results show that if a person is consuming alcohol in a moderate rate with the above definition, every 100 drinking days per year increase will result in the increase in the HDL-C level by 6.18% significantly, adjusting for all other covariates, while whether a person consumes alcohol heavily or not does not have a significant influence on HDL-C level. Besides, although drinking days per year correlates with elevated HDL-C level under moderate alcohol consumption, this benefit diminishes over time.

From literature review, we also evaluated the interaction effects between heavy drink and poverty on HDL-C level. Specifically, the detrimental effects of heavy drinking on HDL-C level are exacerbated in the context of increased poverty. This suggests a synergistic negative impact where lower socioeconomic status may amplify the health risks associated with alcohol consumption. Conversely, we also observed that the number of alcohol-drinking days per year, typically perceived as beneficial for HDL-C level, seems to have a heightened positive effect in poorer populations. This finding invites further investigation into how socioeconomic factors modify the health impacts of alcohol consumption. Our analysis also highlights significant gender disparities. With the same frequency of alcohol consumption, males tend to have lower HDL-C level compared to females. This finding points to a possible biological or behavioral difference in how alcohol affects HDL-C level across genders, warranting further research to understand the underlying mechanisms. Last but not least, an upward trend in HDL-C level was observed with advancing age. This trend might reflect changes in metabolic processes, lifestyle factors, or even healthcare interventions that occur with aging, which can be due to different phenotypes and genotypes among people (Milman et al., 2014).

The limitation of our study is mainly due to the insufficiency of the dataset. Our findings indicate that the positive impact of alcohol-drinking days on HDL-C level diminishes with age. However, the dataset we used does not provide sufficient evidence to conclusively explain this trend. It is plausible that age-related physiological changes or lifestyle modifications could influence this relationship. A significant limitation of our study arises from the unclear definition of alcohol units in the NHANES database. The lack of standardization in the type and size of alcoholic beverages (e.g., bottles vs. glasses) introduces potential inaccuracies in our analysis and limits the precision of our conclusions. The dataset also did not include critical health history information that might significantly influence HDL-C level. Keys among these are liver diseases such as hepatitis and cirrhosis, which can directly impact cholesterol metabolism and alcohol's effect on the body (Osna et al., 2017). The absence of this data restricts our ability to fully understand the complex interplay between alcohol consumption, health status, and HDL-C level.

In reality, the relation between covariates and response might not be as linear as we assume. One possible substitute of model can be the Bayesian framework, and specifically, Bayesian Additive Regression Tree (BART). This is a nonparametric Bayesian regression technique, which has advantages that it does not assume linearity.

More importantly, BART is adept at automatically detecting and modeling interactions between variables. This is particularly useful in scenarios where interactions are not explicitly known or are too complex to model with traditional regression techniques, which can be very helpful in our case since the p-values of some interactions between covariates are just a little bit higher than our threshold value 0.05. Besides, BART includes built-in mechanisms to avoid overfitting, making it a robust choice for scenarios with large numbers of predictors, which can also be very helpful since we are dealing with a lot of covariates. Therefore, further research can be done in a methodological way to compare and analyze the difference in model outputs between traditional multiple linear regression and BART. That is, frequentist performance versus Bayesian performance.

Reference

[1] Sirtori, C. R., Ruscica, M., Calabresi, L., Chiesa, G., Giovannoni, R., & Badimon, J. J. (2019). HDL therapy today: From atherosclerosis, to stent compatibility to heart failure. *Annals of Medicine*, *51*(7-8), 345-359.

[2] Weissglas-Volkov, D., & Pajukanta, P. (2010). Genetic causes of high and low serum HDL-cholesterol. *Journal of Lipid Research, 51*(8), 2032–2057.

[3] Linn, S., Carroll, M., Johnson, C., Fulwood, R., Kalsbeek, W., & Briefel, R. (1993). High-density lipoprotein cholesterol and alcohol consumption in US white and black adults: data from NHANES II. *American Journal of Public Health*, 83(6), 811-816.

[4] Mukamal, K. J., Conigrave, K. M., Mittleman, M. A., Camargo Jr, C. A., Stampfer, M. J., Willett, W. C., & Rimm, E. B. (2003, January 9). Roles of drinking pattern and type of alcohol consumed in coronary heart disease in men. *New England Journal of Medicine, 348*(2), 109-118.

[5] Leaks, K., Norden-Krichmar, T., & Brody, J. P. (2023). Predicting moderate drinking behaviors in National Health and Nutrition Examination Survey participants using biochemical and demographical factors with machine learning. *Alcohol,*113 (2023), 1-10.

[6] Glass, J. E., Rathouz, P. J., Gattis, M., Joo, Y. S., Nelson, J. C., & Williams, E. C. (2017). Intersections of poverty, race/ethnicity, and sex: Alcohol consumption and adverse outcomes in the United States. *Social Psychiatry and Psychiatric Epidemiology*, 52(5), 515-524.

[7] Milman, S., Atzmon, G., Crandall, J., & Barzilai, N. (2014). Phenotypes and genotypes of high density lipoprotein cholesterol in exceptional longevity. *Current vascular pharmacology*, 12(5), 690-697.

[8] Osna, N. A., Donohue, T. M., Jr., & Kharbanda, K. K. (2017). Alcoholic Liver Disease: Pathogenesis and Current Management. *Alcohol Res*, 38(2), 147-161.

Contribution

Kexin Li: Initial data exploration; Model construction and diagnostics part of presentation and final report; plots and tables for the final report; format revision.

Xiaomeng Xu: Literature review; initial data exploration; Outcome Variable, Predictor of interest, Covariates, Interactions, and Statistical Analysis sections of the final report; plots and tables for the final report.

Hongjian Wang: Strategies for data cleaning; Crafting study designs, conducting literature reviews, and composing and reviewing final reports.

Yuan Lu: Literature review; initial data exploration; Conclusion and Limitations part of presentation and final report; potential methodological direction exploration; reference revision.

Gongyin Hong: Literature review; initial data exploration; Interpretation part of the presentation; Model part of the final report; tables for the final report.