BST 210 Final Project
Group Name: Bugs in the Matrix
Group Number: 8
Authors: Jurijs Alehins, Harshi Saha, Joyce Song
Date: 12/16/2024

The Association of Lifestyle Factors and Comorbidities with Heart Attack and BMI

Abstract

A variety of lifestyle factors and comorbidities are associated with cardiovascular disease, specifically heart attacks, which according to the CDC is the leading cause of death in America. Lifestyle factors such as smoking, alcohol use, and BMI, and comorbidities such as depression and stroke have been found to be associated with heart attack as well. This study aims to assess the strength of the associations between these and additional lifestyle factors, and these and additional comorbidities with the occurrence of heart attacks. We also investigate the relationship between these variables and BMI, to identify which variables, if any, are associated with both BMI and heart attack, and which are associated with BMI but not heart attack.

Survey data from the 2022 CDC Behavioral Risk Factor Surveillance System, focusing on U.S. adults[7][8] was used to assess these associations. Logistic regression models were used to investigate the association between variables of interest and the binary outcome of heart attack, and generalized ordinal logistic regression models were used to investigate the association between variables of interest and the ordered categorization of bmi into normal weight, overweight, and obese. Several lifestyle factors and comorbidities were identified as being associated with both heart attack and bmi category, including smoking, alcohol use, physical activity, kidney disease, arthritis, and chronic obstructive pulmonary disease. However, asthma, hours of sleep, and region were found to be associated with bmi category but not heart attack, and further investigation into these variables is warranted. Limitations of this study include constraints in analysis due to limited computational power, imputation methods, and imbalance in the heart attack outcome.

Introduction

Heart disease remains one of the main causes of death globally, with heart attacks contributing to mortality and morbidity rates significantly in the United States[1]. Understanding the risk factors associated with heart attacks is crucial for developing effective prevention strategies. Lifestyle factors such as smoking, alcohol consumption, and obesity have been widely implicated as modifiable risk factors for heart disease[2]. Additionally, comorbidities like depressive disorders, kidney disease, and asthma are also known to influence cardiovascular health[3][5][6]. Previous studies have explored the association between lifestyle factors, comorbidities, and heart attacks. However, many of these studies face limitations such as small sample sizes, lack of diversity in the study population, and potential confounding factors that may bias the results[2]. Moreover, few studies have directly compared the relative influence of lifestyle factors versus comorbidities on the occurrence of heart attacks within a large, representative sample of the U.S. population.

This study primarily aims to investigate the association between various lifestyle factors and non-communicable comorbidities with the occurrence of heart attacks among adults in the United States. Utilizing data from the 2022 CDC Behavioral Risk Factor Surveillance System (BRFSS)[7][8], which includes a large and diverse random sample of U.S. residents, this study will attempt to address previous limitations and provide a more comprehensive understanding of these associations and their strengths respectively. Additionally, this study investigates the categorical outcome of body mass index (BMI) and its associations with lifestyle factors and non-communicable comorbidities, independent of heart attack. Prior research indicates that sociodemographic factors and habits like smoking and alcohol consumption are closely linked to BMI and body composition, often differing between population subgroups based on age, sex, and lifestyle behaviors[4]. This analysis allows the identification of variables that are common predictors for both BMI and heart attack risk, as well as those uniquely associated with either outcome.

Research and Analysis Methods

Data Preprocessing:

The raw dataset contained 445,132 observations and 40 covariates, of which 15 covariates were excluded from the dataset in further analysis, as they did not align with the type of variables being assessed, that is, they were not lifestyle factors or non-communicable comorbidities. The remaining variables in the dataset, of which there were 25, were state, sex, generalhealth, physicalactivities, hadheartattack, hadangina, hadstroke, hadasthma, hadskincancer, hadcopd (chronic obstructive pulmonary disease), haddepressivedisorder, hadkidneydisease, hadarthritis, haddiabetes, smokerstatus, ecigaretteusage, raceethnicitycategory, agecategory, alcoholdrinkers, physicalhealthdays, mentalhealthdays, bmi, heightinmeters, weightinkilograms, and sleephours.

To categorize information contained in the state variable, a new variable called region was generated, using region definitions provided by the CDC[9]. 2238 observations were also identified where values for weightinkilograms and heightinmeters were provided, but not bmi. For these cases, the values for bmi were calculated as $\frac{weightinkilograms}{heighinmeters^2}$ and were filled into the corresponding observations, after which the columns weightinkilograms and heighinmeters were dropped. In addition, bmi was categorized in the variable bmi_category with the following thresholds from the CDC[10]: bmi < 18.5 (underweight), $18.5 \leq$ bmi < 25.0 (normal weight), $25.0 \leq$ bmi < 30.0 (overweight), $30.0 \leq$ bmi < 40.0 (obese), $40.0 \leq$ bmi (morbidly obese). All of retained variables were converted to appropriate types, and stored as factors with appropriate level order. Relevant non-binary categorical variables and their labels in the pre-processed dataset can be seen in [Table 1]. The primary variable of interest in this study is hadheartattack, and the secondary variable of interest is bmi_category.

Outliers:

Those with a bmi less than 18.5 represented $\approx 1.74\%$ of the dataset, while those with a bmi over 40 represented $\approx 5.67\%$ of the dataset. These thresholds were also visualized as reasonable for excluding outliers. Therefore observations for those with bmi less than 18.5 and above 40 were excluded, such that the remaining bmi_category values were normal weight, overweight, and obese, as the categories of underweight and morbidly obese were effectively dropped. Those with sleephours less than 3 represented $\approx 0.61\%$ of the dataset, while those with sleephours over 12 represented $\approx 0.35\%$ of the dataset. These thresholds were also visualized as reasonable for excluding outliers. Therefore observations for those with sleephours less than 3 and above 12 were excluded.

Missingness and Imputation:

The pre-processed dataset with added features and dropped outliers as discussed above was then filtered to keep observations with no missing values, and this dataset was used for fitting models as part of complete case analysis. The dimensions of the complete cases was 305,042, representing 68.35% of the original dataset of size 445,132. The dataset with added features and filtered outliers, but with missing values retained, was then assessed for missingness by each variable. The variables sex and state (and by extension region) did not have any missing values. The remaining variables all displayed missingness, but only the variables smokerstatus, ecigaretteusage, bmi, and alcoholdrinkers displayed missingness of above 5%, at 7.97%, 8.01%, 10.46%, and 10.46% respectively.

The missingness mechanism was determined as not being MCAR heuristically, as all of smokerstatus, ecigaretteusage, bmi, and alcoholdrinkers indicated a significant relationship between their corresponding missingness indicators and one or more other variables in the dataset. The frequency of missingness patterns of the four variables was visualized and the following patterns were the most common in order: only bmi missing (5.33%), all four variables missing (4.54%), all missing except bmi (2.76%), and alcoholdrinkers missing. The distributions of heightinmeters and weightinkilograms appeared similar by missingness of bmi, however, for those that do not report bmi but report at least one of heightinmeters or weightinkilograms, $\approx 80\%$ report heightinmeters but not weightinkilograms, likely indicating that bmi is MNAR.

The distribution of the missingness of smokerstatus by ecigaretteusage indicates that $\approx 92\%$ of those who did not report their smoking status did not report their e-cigarette use, compared to 0.74% for those who reported their smoking status. This indicates the possibility that smoking status is MNAR, as those with a high frequency of smoking might have been less likely to report smokingstatus, and these were likely mostly the same individuals who did not report ecigaretteusage due to high frequency use of e-cigarettes as well. Similar patterns were observed for the missingness of alcoholdrinkers for those where smokerstatus was and was not missing, at $\approx 92\%$ and 3% respectively.

The distribution of the missingness of ecigaretteusage by smokingstatus indicates that $\approx 91\%$ of those who did not report their e-cigarete use did not report their smoking frequency, compared to 0.68% for those who reported their e-cigarette use. This indicates the possibility that ecigaretteusage is MNAR, as those with a high frequency of e-cigarette use might have been less likely to report smokingstatus, and these were likely mostly the same individuals who did not report smokingstatus due to high frequency of smoking as well. Similar patterns were observed for the missingness of alcoholdrinkers for those where ecigaretteusage was and was not missing, at $\approx 96\%$ and 3% respectively.

The distribution of the missingness of alcoholdrinkers by smokingstatus indicates that $\approx 70\%$ of those who did not report their smoking frequency did not report their alcohol use, compared to 0.68% for those who reported their alcohol use. This indicates the possibility that alcoholdrinkers is MNAR, as those with a high frequency of alcohol use might have been less likely to report alcoholdrinkers, and these were likely mostly the same individuals who did not report smokingstatus due to high frequency of smoking as well. Similar patterns were observed for the missingness of ecigaretteusage for those where alcoholdrinkers was and was not missing, at $\approx 71\%$ and 0.34% respectively.

In order to account for the missingness and the effects of the missingness on conclusions, multiple imputation was conducted using the package mice in R. The dataset without the added categorical variables of region and bmi_category and without outliers of bmi and sleephours removed, was used to impute values for the variables with missingness above 5%, namely bmi, smokingstatus, ecigaretteusage, and alcoholdrinkers. The method used to impute values for all four variables, and only the four variables, was predictive mean matching, and 5 imputed datasets were generated using this approach. The regions and bmi_category variables were then added to each of the imputed datasets, with bmi values beyond the thresholds for outlyingess having bmi_category values as NA. In addition, the observations with sleephours beyond the thresholds for outlyingness were dropped from each imputed dataset. These processed imputed datasets were then combined into a single object for ease of pooling fitted models in further analysis. The imputed datasets had missingess of all variables less than 5%, and the dimensions of the complete cases in the 5 imputed datasets are 357,775, 357,809, 357,731, 357,761, and 357,756 in order, indicating that approximately 52,000 observations usable in model fitting were added via multiple imputation.

Analysis Methods:

The primary question investigated in this study is the association of lifestyle factors and non-communicable diseases with the outcome of heart attack, represented by the binary hadheartattack variable in the dataset. The form of model used to assess this association was logistic regression. The steps involved in deciding upon a model, for the complete case dataset, using which to make conclusions about associations were as follows: Fit a basic model using variables of interest, while excluding those identified as being likely confounders (agecategory, raceethnicitycategory, sex). Conduct covariate selection using LASSO regression and validate results for a lack of multicollinearity using VIF. Check for confounding by including likely confounders (agecategory, raceethnicitycategory, sex) one by one, and conduct LRT and compare AIC to check for suggested inclusion. Assess effect modification through inclusion of interaction terms of interest one by one and use the significance of the coefficients of the interaction terms and AIC to check for suggested presence of interaction effects. Check model performance of the final selected model using an ROC curve. The steps involved in deciding upon a logistic regression model, for the multiple imputed datasets, using which to make conclusions about associations were as follows: Retain the variables selected in complete case analysis after LASSO regression to fit a base model. Fit a model with the addition of suspected confounders (agecategory, raceethnicitycategory, sex) to the base covariates, checking the significance of coefficients of the added terms and comparing the AIC to the base model to assess confounding. Assess effect modification through the inclusion of interaction terms of interest one by one and use the significance of the coefficients of the interaction terms and AIC to check for suggested presence of interaction effects.

The secondary question investigated in this study is the differential association of lifestyle factors and non-communicable diseases with the outcome of bmi, represented by the ordinal categorical variable bmi_category, with levels normal weight, overweight, and obese, independent of heart attack, in the dataset. The form of model used to assess this association was generalized ordinal logistic regression, and this was chosen over multinomial due to natural ordering in bmi_category, and over ordinal to provide greater flexibility in modeling and to prevent the assumption of proportional odds, which might be violated in this dataset. The steps involved in deciding upon a model, for the complete case dataset, using which to make conclusions about associations were as follows: Fit a basic model using all variables of interest, while excluding those identified as being likely confounders (agecategory, raceethnicitycategory, sex). Check for confounding by including likely confounders (agecategory, raceethnicitycategory, sex) all together, and conduct

LRT and compare AIC to the model without these variables to check for suggested inclusion. Check which variables, if any, are not significant in predicting bmi_category, fit a model excluding those not statistically significant, and compare LRT and AIC values to the model fitted with their inclusion. Identify which variables, if any, are differentially associated with bmi_category and heart attack. These assessments and analyses were not repeated on the imputed datasets, and assessment of effect modification was not conducted, due to constraints in available computational resources.

Findings and Analysis:

Primary Investigation:

The primary investigation in this study was to determine associations between lifestyle factors and non-communicable comorbidities, and the outcome of heart attack. The confounders identified amongst variables of interest were age, race, and sex. A logistic regression model was fit for the binary outcome hadheartattack, using variables of interest excluding suspected confounders. Performing LASSO regression using the model above resulted in importance of covariates determined as in [Figure 4]. Based on these results, the following variables were excluded from further analysis: physicalhealthdays, mentalhealthdays, region, and sleephours. In addition to having low importance as per LASSO, the survey variables physicalhealthdays and mentalhealthdays were collected with non-specific questions that are highly subjective in responses as well, further supporting their exclusion in subsequent modeling [7].

The suspected confounders of age, race, and sex were individually added to the model with the covariates selected for after LASSO above, and the logistic regression models fitted indicated that all the terms had levels with statistically significant coefficients at the 0.05 level. Furthermore, the addition of the terms indicated in LRT support for their inclusion, and the AIC reduction was $\approx 4\%$, from 89474.6 to 85917.8. To assess effect modification, interaction terms for sex with race, sex with age, sex with bmi category, sex with smoker status, sex with diabetes, sex with general health, and sex with e-cigarette use were then each added to and assessed in comparison to the model with confounders. The inclusion of interaction terms indicated either no statistically significant results overall, or statistical significance for some but not all categories of the corresponding interaction terms. Furthermore, although a few terms slightly reduced AIC, based on the added complexity and the lack of significance in any or in some levels of multi-level interaction terms, no interaction terms were retained for the final selected model. No polynomial terms were assessed, as no continuous variables were included in the model selection process after filtering via LASSO regression.

Therefore, for the complete case dataset, the AIC of the final selected model was 85917.8, and this model had the equation as below:

$$\begin{aligned}
&\text{logit}(\Pr(\text{hadheartattack}_i = 1)) = \beta_0 + \beta_{1\ldots4} \cdot I_{\text{generalhealthcategory}_i} + \beta_5 \cdot \text{physicalactivities}_i \\
&+ \beta_6 \cdot \text{hadangina}_i + \beta_7 \cdot \text{hadstroke}_i + \beta_8 \cdot \text{hadasthma}_i + \beta_9 \cdot \text{hadskincancer}_i + \beta_{10} \cdot \text{hadcopd}_i \\
&+ \beta_{11} \cdot \text{haddepressivedisorder}_i + \beta_{12} \cdot \text{hadkidneydisease}_i + \beta_{13} \cdot \text{hadarthritis}_i \\
&+ \beta_{14\ldots16} \cdot I_{\text{haddiabetescategory}_i} + \beta_{17\ldots19} \cdot I_{\text{smokerstatuscategory}_i} + \beta_{20\ldots22} \cdot I_{\text{ecigaretteusagecategory}_i} \\
&+ \beta_{23} \cdot \text{alcoholdrinkers}_i + \beta_{24\ldots35} \cdot I_{\text{agecategory}_i} + \beta_{36} \cdot \text{sex}_i + \beta_{37\ldots40} \cdot I_{\text{raceethnicitycategory}_i} \\
&+ \beta_{41} \cdot I_{\text{bmi\_category}_i = overweight} + \beta_{42} \cdot I_{\text{bmi\_category}_i = obese}
\end{aligned}$$

The reference groups and the corresponding order of variable categories for each indicator group of β can be found in [Table 2.]

The only non-significant variable was asthma, but its removal resulted in a higher AIC, and it was selected for via LASSO, so this variable was retained in the final selected model. The final model chosen to explain associations was validated using AUC. The AUC for the complete case dataset was 0.886, although the class imbalance in our dataset for the hadheartattack variable may impact the confidence in these results.

For the 5 imputed datasets, an initial model was fitted using the same covariates selected in the complete case analysis after LASSO regression. The suspected confounders age, sex and race were then added all together to the model, and compared to the model without these covariates added. The coefficients of the confounder variables were all statistically significant at the 0.05 level, based on the model pooled on the imputed datasets. Comparison of the AIC values of the model with and without the confounders revealed that the AIC reduced by $\approx 6.75\%$ on average for all the 5 imputed datasets, from 113,637.5 to 105,966.8 upon inclusion of the suspected confounders. To assess effect modification, as in complete case analysis, interaction terms for sex with race, sex with age, sex with bmi category, sex with smoker status, sex with diabetes, sex with general health, and sex with e-cigarette use were then each added to and assessed in comparison to the model with confounders, and the results were similar to that in the complete case analysis. Therefore, interaction terms were not included in the final model, and no polynomial terms were assessed, as no continuous variables were included in the model selection process after filtering via LASSO regression.

Therefore, for the imputed datasets, the AIC of the final selected model was 105,966.8, and this model had the same equation as the model above for the complete case analysis. The summary of the fitted model, with the 95% CI of the exponentiated coefficients for each covariate after pooling on the imputed datasets can be seen in [Figure 2]. The final model chosen to explain associations was validated using AUC. The mean AUC for the imputed datasets was 0.885, nearly identical to that for complete case analysis, although the class imbalance in our dataset for the hadheartattack variable may impact the confidence in this result.

Based on complete case analysis and the imputed datasets, the lifestyle factors associated with heart attack are physical activities, smoker status, ecigarette use, and alcohol consumption, while the non-communicable comorbidities associated with heart attack are angina, stroke, skin cancer that is not melanoma, chronic obstructive pulmonary disease, depressive disorders, arthritis, and diabetes but not pre-diabetes or diabetes only during pregnancy. The variable asthma was not significant in both datasets. For complete case analysis and imputed dataset analysis, the odds of heart attack occurring for those in outlined levels compared to the reference or unexposed groups for each variable of interest, on average, according to these sample data, and holding all other covariates fixed, can be seen in [Figure 2].

In both complete case analysis and the imputed datasets, the associations found were similar. For associations with lifestyle factors: Engaging in physicalactivities shows a slight reduction in the odds of heart attack, with an odds ratio (OR) of approximately 0.92. The ORs for both smoker status and e-cigarette use are above 1, and the patterns of ORs in smoker status suggest an increase in the odds of heart attack in going from former smoking status, to those who currently smoke only some days, to those who are smoking everyday, when for e-cigarette use the odds of heart attack were not consistent with such an increasing pattern. Interestingly, the odds of heart attack for those who drink alcohol compared to those who do not drink alcohol is approximately 0.79, and this finding can likely be attributed to the imbalance in the outcome of heart attack as well as the non-specificity of the frequency of drinking captured by the binary alcoholdrinker variable. For associations with comorbidities: The largest OR was associated with the presence of angina, and the odds of heart attack for those that had an angina was around 12.34 times the odds of heart attack for those without. Additionally, the OR of heart attack for those who had stroke to those who had not had stroke was approximately 2.56. The odds of heart attack increased slightly for those with chronic obstructive pulmonary disease, depressive disorders, kidney disease, arthritis, and diabetes, while non-melanoma skin cancer was associated with a slightly lowered odds of heart attack.

Secondary Investigation:

The secondary investigation in this study was to determine differential associations between lifestyle factors and non-communicable comorbidities, and the outcome of bmi_category (normal weight, overweight, obese), independently of heart attack. The potential confounders identified as in the primary investigation were age, race, and sex. A generalized ordinal logistic regression model was fit for the ordered categorical outcome of bmi_category, using variables of interest excluding suspected confounders. In comparison to the initial model for heart attack in primary analysis, after LASSO regression and prior to the assessment of confounding, this model includes region and sleephours. All of the suspected confounders were then added to this model and compared to the model excluding these covariates. The generalized ordinal logistic regression model fitted indicated that all the terms corresponding to suspected confounders had statistically significant coefficients at the 0.05 level. The addition of the terms indicated in LRT support for their inclusion, and the AIC reduction was ≈ 2.43%, from 640,443.1 to 624,877.3.

Therefore, for the complete case dataset, the AIC of the final selected model was 624,877.3, and this model had the equation as below:

$$\log\left(\frac{P(Y_i \geq k)}{P(Y_i < k)}\right) = \beta_{k,0} + \beta_{k,1\ldots k,4} \cdot I_{\text{generalhealthcategory}_i} + \beta_{k,5} \cdot \text{physicalactivities}_i + \beta_{k,6} \cdot \text{hadangina}_i + \beta_{k,7} \cdot \text{hadstroke}_i + \beta_{k,8} \cdot \text{hadasthma}_i + +\beta_{k,9} \cdot \text{hadskincancer}_i + \beta_{k,10} \cdot \text{hadcopd}_i + \beta_{k,11} \cdot \text{haddepressivedisorder}_i + \beta_{k,12} \cdot \text{hadkidneydisease}_i + \beta_{k,13} \cdot \text{hadarthritis}_i + \beta_{k,14\ldots k,16} \cdot I_{\text{haddiabetescategory}_i} + \beta_{k,17\ldots k,19} \cdot I_{\text{smokerstatuscategory}_i} + \beta_{k,20\ldots k,22} \cdot I_{\text{ecigaretteusagecategory}_i} + \beta_{k,23} \cdot \text{alcoholdrinkers}_i + \beta_{k,24\ldots k,35} \cdot I_{\text{agecategory}_i} + \beta_{k,36} \cdot \text{sex}_i + \beta_{k,37\ldots k,40} \cdot I_{\text{raceethnicitycategory}_i} + \beta_{k,41\ldots k,44} \cdot I_{\text{regioncategory}_i} + \beta_{k,45} \cdot \text{sleephours}_i,$$

where Y_i represents *bmi_category*, with k = 1 corresponding to the normal weight category (reference), k = 2 corresponding to the overweight category, k = 3 corresponding to the obese category. The reference groups and the corresponding order of variable categories for each indicator group of β can be found in [Table 2.]

As the secondary analysis aimed to assess the differential associations between lifestyle factors and non-communicable comorbidities between the outcomes of heart attack and bmi_category, the results of fitting the model above were used to assess the presence or lack of association of covariates with bmi_category and not to interpret the associations or their strength. The lifestyle factors associated with bmi_category are physical activity, smoking status, e-cigarette use, and alcohol drinking, and sleephours, while the non-communicable comorbidities associated with bmi_category are angina, stroke, asthma, skin cancer that is not melanoma, chronic obstructive pulmonary disease, depressive disorder, kidney disease, arthritis, and diabetes. The primary investigation on both the completed and imputed datasets found that asthma was not significantly associated with heart attack, while the secondary analysis indicated that asthma was significantly associated with bmi_category at the 0.05 level. In addition, LASSO regression in the primary analysis indicated that sleephours and region were not associated with the outcome of heart attack, while analysis for association with bmi_category indicated that these variables were statistically significant. Therefore asthma, sleephours, and region were found to be associated with bmi_category, but not with heart attack, in this analysis.

Discussion:

Limitations:

The original dataset created by the CDC included over 300 variables, of which 40 were selected as being relevant to heart disease by the authors from which the dataset used in this study was sourced. However, it is possible that variables exist in the original CDC dataset that were not selected in the dataset used by us, that better explain the outcomes of heart attack and bmi. Further studies could investigate the variables in the original CDC dataset to select such potential variables.

The thresholds for the values of bmi were derived from those prescribed by the CDC, but potentially conducting literature review to formulate more meaningful bmi categories as relevant to heart attack may change the results of the assessments of the associations investigated here. Furthermore, the determination of thresholds for outlyingness and the methods of dealing with identified outliers were quite simplistic in this study, and further exploring outliers and patterns of associations with outliers may not only change conclusions but may also reveal interesting patterns not previously identified, which may be important in understanding the mechanisms of the associations between lifestyle factors and heart attack. This dataset also indicated imbalance in the outcome of hadheartattack, which was not addressed in this analysis, and could potentially influence the results outlined in this report.

The aspect of this study that has the greatest avenue for improvement is the assessment of missingness and the procedure of imputation. Assessment of missingness was only conducted on variables of interest, and more specifically on those that presented missingness of at least 5% in the dataset. Furthermore, only the method of predictive mean matching was used for multiple imputation, for the 4 variables that were determined to require multiple imputation prior to further analysis. This limit was decided upon in combination due to access to limited computational resources as well as the complexity of carrying out such methods. In addition to investigating potential missingness mechanisms for all variables with missingness over a threshold lower than 5%, it would also likely be beneficial to conduct multiple imputation for a greater number of variables with missingness using a variety of applicable methods of imputation in future attempts to generate datasets, using which to investigate the association between lifestyle factors and non-communicable comorbidities, and heart attack and bmi.

Additional limitations in this analysis were related to the availability of computational resources. Associations between bmi and lifestyle factors and non-communicable comorbidities were only assessed on complete cases, but not on the imputed datasets. Although associations of the same variables with heart attack were investigated on complete cases and imputed datasets, and resulted in similar conclusions for both, there is no measure of certainty for if this would have been the case for associations with bmi, which had relatively high missingness and was imputed for. The large size of the dataset, and the high number of covariates and levels in the non-binary categorical covariates, combined with the complexity of generalized ordinal logistic regression models, impacted the ability to conduct the secondary analysis on the imputed datasets, due to limitations in available computational resources. This limitation also impacted the ability to assess effect modification in the complete cases dataset. It is likely that upon further investigation into the associations for bmi using the imputed datasets, and effect modification in both complete case and imputed datasets, the variables determined as being differentially associated with bmi and heart attack are different. Due to the nature of the generalized ordinal logistic model as well, commonly used packages used to conduct covariate selection, such as stepwise selection and LASSO regression were not feasible, and further analysis in this regard as well might result in different findings.

Conclusions:

In investigation of the association between lifestyle factors and non-communicable comorbidities and the outcome heart attack, the following were found: The lifestyle factors associated with heart attack are physical activities, smoker status, ecigarette use, and alcohol consumption. The non-communicable comorbidities associated with heart attack are angina, stroke, skin cancer that is not melanoma, chronic obstructive pulmonary disease, depressive disorders, arthritis, and diabetes but not pre-diabetes or diabetes only during pregnancy. The variables sex, age category, race and ethnicity were found to confound the relationship between the covariates of interest listed above and heart attack, while no effect modification by sex was identified. Additionally, the final logistic regression model chosen to identify these associations was validated using AUC and this metric displayed a value of approximately 0.89, though imbalance in the heart attack variable may impact the confidence in these results.

In investigation of the association between lifestyle factors and non-communicable comorbidities and the outcome bmi_category, the following were found: The lifestyle factors associated with bmi_category are physical activities, smoker status, ecigarette use, alcohol consumption, and sleephours. The non-communicable comorbidities associated with bmi category are angina, stroke, skin cancer that is not melanoma, chronic obstructive pulmonary disease, depressive disorders, arthritis, and diabetes. The variables sex, age category, race and ethnicity were found to confound the relationship between the covariates of interest listed above and bmi_category as well. The comorbidity of asthma, the lifestyle factor of hours of sleep, and the demographic variable of region, were found to be associated with bmi category, but not with heart attack, in this analysis. Further investigations into these variables is warranted, as well as further analysis addressing limitations of the analysis identified in the study, as this could potentially uncover patterns of association relevant to the prognosis of heart attack.

Tables and Figures:

Table 1

Covariate	Levels
generalhealth	Poor, Fair, Good, Very good, Excellent
haddiabetes	No, No, pre-diabetes or borderline diabetes, Yes, but only during pregnancy (female), Yes
smokerstatus	Never smoked, Former smoker, Current smoker - now smokes some days, Current smoker - now smokes every day
ecigaretteusage	Never used e-cigarettes in my entire life, Not at all (right now), Use them some days, Use them every day
raceethnicitycategory	White only, Non-Hispanic, Black only, Non-Hispanic, Other race only, Non-Hispanic, Multiracial, Non-Hispanic, Hispanic
agecategory	Age 18 to 24, Age 25 to 29, Age 30 to 34, Age 35 to 39, Age 40 to 44, Age 45 to 49, Age 50 to 54, Age 55 to 59, Age 60 to 64, Age 65 to 69, Age 70 to 74, Age 75 to 79, Age 80 or older

Table 2: Indicator Functions and Reference Groups

Covariate	Reference Group and Coefficients
General Health	Reference group: Poor. β_1: Fair, β_2: Good, β_3: Very Good, β_4: Excellent
Had Diabetes	Reference group: No. β_{14}: Pre-diabetes or borderline diabetes, β_{15}: Yes, during pregnancy (female), β_{16}: Yes
Smoker Status	Reference group: Never Smoked. β_{17}: Former smoker, β_{18}: Current smoker (some days), β_{19}: Current smoker (every day)
E-Cigarette Usage	Reference group: Never used e-cigarettes in life. β_{20}: Not at all (right now), β_{21}: Use them some days, β_{22}: Use them every day
Age Category	Reference group: Age 18 to 24. β_{24} to β_{35}: Other age groups in order
Race/Ethnicity	Reference group: White only, Non-Hispanic. β_{37}: Black only, Non-Hispanic, β_{38}: Other race only, Non-Hispanic, β_{39}: Multiracial, Non-Hispanic, β_{40}: Hispanic
BMI Categories (For Primary Question)	Reference group: Normal Weight ($18.5 \leq \text{BMI} < 25.0$). β_{41}: Overweight ($25.0 \leq \text{BMI} < 30.0$), β_{42}: Obese ($30.0 \leq \text{BMI} < 40.0$)
Region Categories (For Secondary Question)	Reference group: Midwest. β_{41}: Northeast, β_{42}: Other, β_{43}: South, β_{44}: West

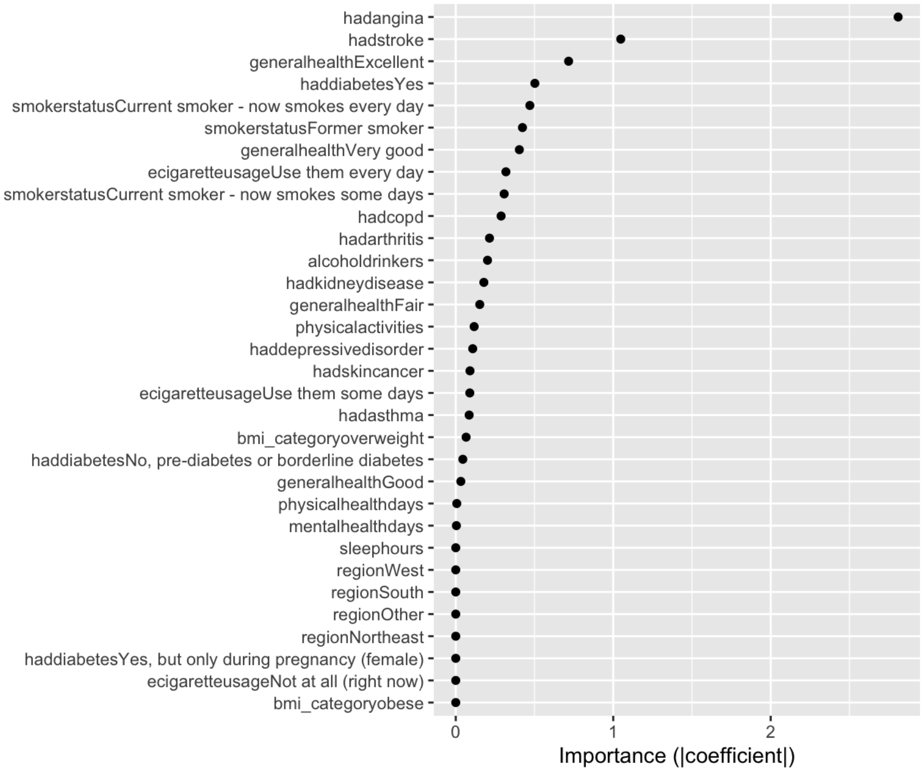
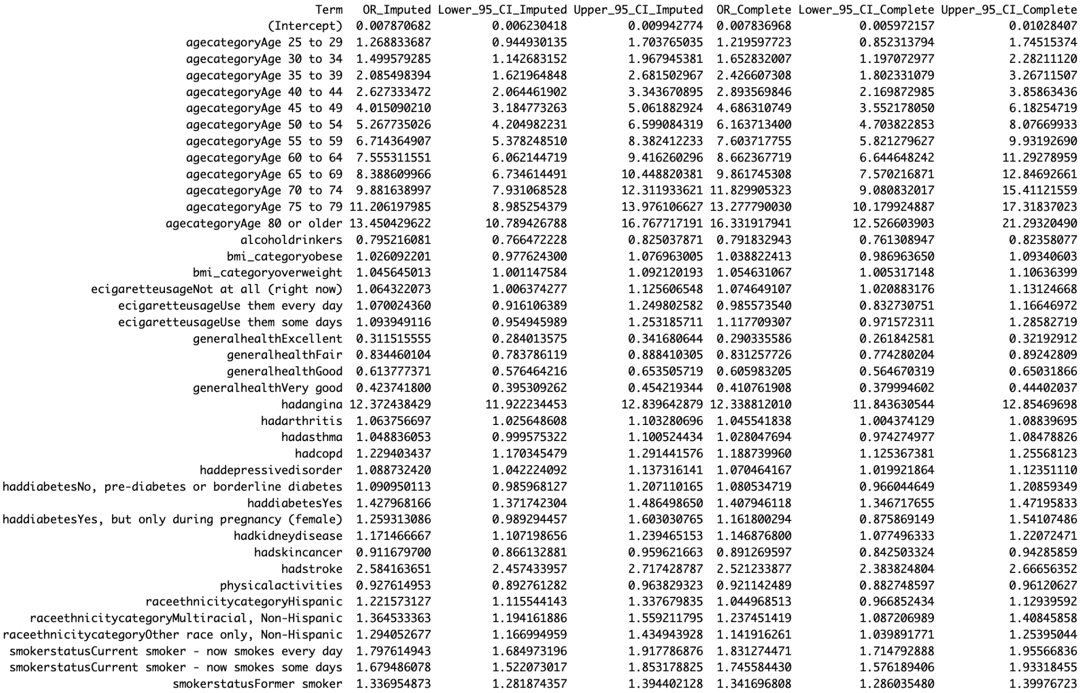


Figure 1: Primary Question LASSO Results for Complete Cases



```
                                              Term   OR_Imputed Lower_95_CI_Imputed Upper_95_CI_Imputed  OR_Complete Lower_95_CI_Complete Upper_95_CI_Complete
                                       (Intercept)  0.007870682         0.006230418         0.009942774  0.007836968          0.005972157           0.01028407
                           agecategoryAge 25 to 29  1.268833687         0.944930135         1.703765035  1.219597723          0.852313794           1.74515374
                           agecategoryAge 30 to 34  1.499579285         1.142683152         1.967945381  1.652832007          1.197072977           2.28211120
                           agecategoryAge 35 to 39  2.085498394         1.621964848         2.681502967  2.426607308          1.802331079           3.26711507
                           agecategoryAge 40 to 44  2.627333472         2.064461902         3.343670895  2.893569846          2.169872985           3.85863436
                           agecategoryAge 45 to 49  4.015090210         3.184773263         5.061882924  4.686310749          3.552178050           6.18254719
                           agecategoryAge 50 to 54  5.267735026         4.204982231         6.599084319  6.163713400          4.703822853           8.07669933
                           agecategoryAge 55 to 59  6.714364907         5.378248510         8.382412233  7.603717755          5.821279627           9.93192690
                           agecategoryAge 60 to 64  7.555311551         6.062144719         9.416260296  8.662367719          6.644648242          11.29278959
                           agecategoryAge 65 to 69  8.388609966         6.734614491        10.448820381  9.861745308          7.570216871          12.84692661
                           agecategoryAge 70 to 74  9.881638997         7.931068528        12.311933621 11.829905323          9.080832017          15.41121559
                           agecategoryAge 75 to 79 11.206197985         8.985254379        13.976106627 13.277790030         10.179924887          17.31837023
                        agecategoryAge 80 or older 13.450429622        10.789426788        16.767717191 16.331917941         12.526603903          21.29320490
                                   alcoholdrinkers  0.795216081         0.766472228         0.825037871  0.791832943          0.761308947           0.82358077
                                 bmi_categoryobese  1.026092201         0.977624300         1.076963005  1.038822413          0.986963650           1.09340603
                            bmi_categoryoverweight  1.045645013         1.001147584         1.092120193  1.054631067          1.005317148           1.10636399
             ecigaretteusageNot at all (right now)  1.064322073         1.006374277         1.125606548  1.074649107          1.020883176           1.13124668
                 ecigaretteusageUse them every day  1.070024360         0.916106389         1.249802582  0.985573540          0.832730751           1.16646972
                 ecigaretteusageUse them some days  1.093949116         0.954945989         1.253185711  1.117709307          0.971572311           1.28582719
                            generalhealthExcellent  0.311515555         0.284013575         0.341680644  0.290335586          0.261842581           0.32192912
                                 generalhealthFair  0.834460104         0.783786119         0.888410305  0.831257726          0.774280204           0.89242809
                                 generalhealthGood  0.613777371         0.576464216         0.653505719  0.605983205          0.564670319           0.65031866
                            generalhealthVery good  0.423741800         0.395309262         0.454219344  0.410761908          0.379994602           0.44402037
                                         hadangina 12.372438429        11.922234453        12.839642879 12.338812010         11.843630544          12.85469698
                                      hadarthritis  1.063756697         1.025648608         1.103280696  1.045541838          1.004374129           1.08839695
                                         hadasthma  1.048836053         0.999575322         1.100524434  1.028047694          0.974274977           1.08478826
                                           hadcopd  1.229403437         1.170345479         1.291441576  1.188739960          1.125367381           1.25568123
                             haddepressivedisorder  1.088732420         1.042224092         1.137316141  1.070464167          1.019921864           1.12351110
haddiabetesNo, pre-diabetes or borderline diabetes  1.090950113         0.985968127         1.207110165  1.080534719          0.966044649           1.20859349
                                    haddiabetesYes  1.427968166         1.371742304         1.486498650  1.407946118          1.346717655           1.47195833
haddiabetesYes, but only during pregnancy (female)  1.259313086         0.989294457         1.603030765  1.161800294          0.875869149           1.54107486
                                  hadkidneydisease  1.171466667         1.107198656         1.239465153  1.146876800          1.077496333           1.22072471
                                     hadskincancer  0.911679700         0.866132881         0.959621663  0.891269597          0.842503324           0.94285859
                                         hadstroke  2.584163651         2.457433957         2.717428787  2.521233877          2.383824804           2.66656352
                                physicalactivities  0.927614953         0.892761282         0.963829323  0.921142489          0.882748597           0.96120627
                     raceethnicitycategoryHispanic  1.221573127         1.115544143         1.337679835  1.044968513          0.966852434           1.12939592
    raceethnicitycategoryMultiracial, Non-Hispanic  1.364533363         1.194161886         1.559211795  1.237451419          1.087206989           1.40845858
raceethnicitycategoryOther race only, Non-Hispanic  1.294052677         1.166994959         1.434943928  1.141916261          1.039891771           1.25395044
 smokerstatusCurrent smoker - now smokes every day  1.797614943         1.684973196         1.917786876  1.831274471          1.714792888           1.95566836
 smokerstatusCurrent smoker - now smokes some days  1.679486078         1.522073017         1.853178825  1.745584430          1.576189406           1.93318455
                         smokerstatusFormer smoker  1.336954873         1.281874357         1.394402128  1.341696808          1.286035480           1.39976723
```

Figure 2: ORs for Complete and Imputed Data Analysis

References

2. Adams, B., Jacocks, L., & Guo, H. (2020, May 29). Higher BMI is linked to an increased risk of heart attacks in European adults: A mendelian randomisation study - *BMC Cardiovascular Disorders.* SpringerLink. https://link.springer.com/article/10.1186/s12872-020-01542-w

1. Brown, M. S., & Goldstein, J. L. (1996, May 3). Heart Attacks: Gone with the Century? *Science.* https://www.science.org/doi/10.1126/science.272.5262.655

10. Centers for Disease Control and Prevention. (n.d.). Adult BMI categories. *Centers for Disease Control and Prevention.* https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

9. Centers for Disease Control and Prevention. (2024, July 24). Geographic Division or region - health, United States. *Centers for Disease Control and Prevention.* https://www.cdc.gov/nchs/hus/sources-definitions/geographic-region.htm

3. Chiu, H.-Y., Huang, H.-L., Li, C.-H., Chen, H.-A., Yeh, C.-L., Chiu, S.-H., Lin, W.-C., Cheng, Y.-P., Tsai, T.-F., & Ho, S.-Y. (2015, September 25). Increased risk of chronic kidney disease in rheumatoid arthritis associated with cardiovascular complications – a national population-based cohort study. *PLOS ONE.* https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0136508

6. Iribarren, C., Tolstykh, I. V., & Eisner, M. D. (2004, May 4). Are patients with asthma at increased risk of coronary heart disease? *Oxford Academic.* https://academic.oup.com/ije/article-abstract/33/4/743/665470

7. Kamilpytlak. (n.d.). Data-science-projects/heart-disease-prediction/2022 at main · Kamilpytlak/Data-science-projects. *GitHub.* https://github.com/kamilpytlak/data-science-projects/tree/main/heart-disease-prediction/2022

5. Nemeroff, C. B., & Goldschmidt-Clermont, P. J. (2012, June 26). Heartache and heartbreak-the link between depression and cardiovascular disease. *Nature News.* https://www.nature.com/articles/nrcardio.2012.91

4. Ohlsson, B., & Manjer, J. (2020, January 7). Sociodemographic and Lifestyle Factors in relation to Overweight Defined by BMI and "Normal-Weight Obesity." *Wiley Online Library.* https://onlinelibrary.wiley.com/doi/full/10.1155/2020/6210201

8. Pytlak, K. (2023, October 12). Indicators of heart disease (2022 update). *Kaggle.* https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease