**Statistical Analysis**

Notes:

* Made subsets of data to retain records in models with fewer variables
* Due to difference in sample size, models need to be compared using BIC, not AIC
  + BIC accounts for different sample sizes, AIC is also more accurate with larger datasets than ours
  + “AICc” could be used with our dataset if the models are made to be same size
  + <https://link.springer.com/article/10.3758/s13428-018-1188-3#:~:text=However%2C%20since%20computation%20of%20the,groups%20and%20the%20group%20size>.
  + Smaller BIC is better. Indicates good model fit for dataset.
  + Calculate difference between model BIC. BIC penalizes the more complex dataset.
    - > 10 : very strong evidence
    - 6-10 : strong evidence
    - 2 – 6 : positive evidence
    - 0 – 2 : weak evidence
  + BIC = deviance + k \* ln(N)
    - K = number of parameters
    - N = sample size

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

**Multiple Logistic Regression Analysis**

To generate the statistical analysis components of this study R version 4.2.0 was used within Rstudio 2022.07.2 Build 576. After importing, mapping, and factorizing each variable accordingly several subsets of the data were prepared such that each subset only included the relevant features corresponding to six multiple logistic regression models to be used as starting points for the model evaluation and selection process. Logistic regression requires that the observations do not contain any missing values which is a significant challenge for this study due to the sparsity of NHANES datasets. By creating subsets of the available data, we can retain a larger proportion of observations after removing any records with missing values. The six base models’ features were determined by considering best practices regarding selecting groups of variables which share common characteristics. Table x describes the sample size and number of parameters in each of the base models. All six base models appear to have strong statistical significance against the empty model via Chi-Square.

An initial determination of model fit is evaluated by calculating the Bayesian Information Criterion (BIC). The resulting BIC values for the six models indicate that Model 3 may be the best fit model and Model 1 may be the second best, but the difference in BIC between Model 2 and Model 1 is much smaller than the difference between Model 3 and Model 2. BIC applies a penalty that favors models with fewer parameters which corresponds to the resulting BIC values observed from our models. Another key indicator used in this analysis to compare model fit is McFadden’s adjusted R2. By looking at this metric, Model 2’s value of 0.199 appears to have the best relative fit and is also the only model that gets extremely close to falling within the range of 0.2-0.4 which indicates very good model fit. This metric also applies a penalty that favors models with fewer parameters.

The resulting odds ratios and p-values for each of the variables in all three models are presented in Table x. The variables that appear to be significant are relatively consistent across the three models with slight differences that can be observed due to inclusion of different variables. Variables that present statistically significant levels in all three models are gender, age, number of healthcare visits, and mobility limitations. In Model 2, not having high blood pressure, not being a tobacco user, and not having a mobility limitation is found to be significant and suggests a protective effect against cardiovascular disease. Model 3 also indicates a significant protective effect in not having a mobility limit. In Model 1, being Hispanic was found to be significant and suggests a protective effect against cardiovascular disease compared to being White and being in a poverty index greater than 1.85 was significant and indicates a protective effect. In all models being a younger adult and being female indicates a protective effect, while being an older adult indicates a risk factor, though being a young adult in Model 3 was not statistically significant. All the models suggest a strong correlation between number of healthcare visits and history of cardiovascular disease indicating that high numbers of yearly healthcare visits are associated with higher likelihood of having had a cardiovascular emergency at some point in the past. Additionally, all three models indicate that higher RFM is associated with higher odds of cardiovascular disease, but this feature was only significant at a 95% confidence level in Models 2 and 1 while this feature is only significant at a 90% confidence level in Model 3.

The predictive capabilities of the models are evaluated by conducting out-of-sample testing and examining the sensitivity and specificity of the resulting estimations. Each of the models are trained using a random sample of 70% of the available records for that model, then the coefficient estimates are applied to generate probabilities using the remaining 30% of records using an initial threshold of 0.5 for classification. The resulting distribution of probabilities in Models 3 and 2 show an observable difference in the IQR and median values when grouping by the observed class of cardiovascular disease history which indicates a well fit model. The probabilities generated from Model 1 have a much more similar IQR, median, and spread of outliers indicating less optimal fit and suggests that the variables selected in Models 2 and 1 are important indicators for predicting cardiovascular disease.

All three models have an out-of-sample accuracy rate of close to 90% with Model 2 being the highest at 93% and Model 1 being the lowest at 89.6%. Each of the models have much higher sensitivity than specificity and are extremely effective at classifying the negative case where the participant reported not to have had a heart attack or stroke. The majority of probabilities generated by the models fall below 0.5 except for outliers, so the classification accuracy for the positive case could be improved by adjusting the classification threshold.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **N** | **Description of Variables** | **Independent Variables** |
| **Model 1f** | 1510 | Social Determinants – Females only |  |
| **Model 1m** | 1551 | Social Determinants – Males only |  |
| **Model 2f** | 719 | Social Determinants + health indicators – Females only |  |
| **Model 2m** | 705 | Social determinants + health indicators – Males only |  |
| **Model 3f** | 718 | Social determinants + health indicators + religious attendance – Females only |  |
| **Model 3m** | 703 | Social determinants + health indicators + religious attendance – Males only |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **Model 1f** | **Model 1m** | **Model 2f** | **Model 2m** | **Model 3f** | **Model 3m** |
| **N** | 1510 | 1511 | 719 | 705 | 718 | 703 |
| **Chi-Square Significance** | < 0.000 | < 0.000 | < 0.000 | < 0.000 | < 0.000 | < 0.000 |
| **BIC** | 891.8 | 1185.1 | 523.5 | 635.6 | 535.6 | 642.8 |
| **McFadden’s R^2** | 0.143 | 0.147 | 0.212 | 0.226 | 0.214 | 0.236 |
| **McFadden’s Adjusted R^2** | 0.091 | 0.110 | 0.075 | 0.126 | 0.067 | 0.128 |
| **Out-of-Sample test Sensitivity** | .81 | .73 | .87 | .95 | 0.50 | 0.17 |
| **Out-of-Sample test Specificity** | .67 | .69 | .58 | 0.51 | 0.92 | 0.93 |
| **Out-of-Sample test F1 Score** |  |  |  |  |  |  |
| **Out-of-Sample Test Balanced Accuracy** | .75 | .71 | .73 | 0.73 | 0.71 | 0.55 |