Risk of Osteoporosis Based on Dietary Vitamin D and Calcium (NHANES 2017-2018)

Anjana Renganathan & Yazmin Baldonado

04/28/2023

Introduction

Multiple logistic regression models are a common statistical methodology used when trying to model a response that is binary, or dichotomous. It aims to derive the model with the best fit and the highest efficiency that explains the relationship between the predictor and response variables.

Unlike typical regression models, a logistic regression transforms the model into a probability model for the outcome which can be interpreted as the probability that the response will occur based on the predictors in the model.

Logistic regression is ideal for modeling our question because it allows us to test both continuous and categorical predictors against a binary response.

For this project, we use a multiple logistic regression model to assess the risk of osteoporosis diagnosis based on a variety of predictors, including dietary and supplementary calcium and Vitamin D, as well as fractured wrists.

Osteoporosis is a chronic metabolic bone disease that is characterized by decreased bone mineral density and bone mass, resulting in weakened bone structure and strength. It presents in patients as wrist, hip and spine fractures, which are often caused from gentle stressors like falling from standing, lifting heavy objects or even coughing [13,14].

The impact of osteoporosis should not be understated considering that is it one of the most common chronic metabolic diseases in old age. In 2010, an estimated 10.2 million people above the age of 50 had osteoporosis in the United States [14]. The projected cost of bone fractures and osteoporosis in the United States is expected to reach 25.3 billion by 2025 [14].

Clinically significant risk factors for osteoporosis include sex, age, race, body and bone size, family history, diet and lifestyle factors. For diet, calcium and vitamin D levels are most important as they're both involved in the biochemical pathways involved in bone health. Lifestyle factors include levels of physical activity, alcohol use and smoking [13].

The objective of this analysis is to identify which dietary factors affect risk of osteoporosis diagnosis in older adults.

Methodology

Data Source

The data was sourced from the National Health and Nutrition Examination Survey (NHANES) run by the Center for Disease Control (CDC).The NHANES study has been run since the 1960's and provides population health data on adults and children in the United States [10].

The data is both publicly available and deidentified, therefore not requiring IRB approval. The survey includes demographic, socioeconomic, medical, dental, physiological, laboratory and dietary data [10,11,12].

NHANES uses survey weights to create a nationally representative dataset. This is achieved by oversampling minority populations, accounting for non-response and then making a post-stratification adjustment to match Census population counts [10]. This is described thoroughly in the NHANES survey documentation as well as the analytics modules created to assist researchers [12].

Dataset

For this project we selected the 2017-2018 NHANES cycle, which was completed and published prior to the COVID19 pandemic. The variables included self-reported age and gender from the demographic dataset. We took Dietary Calcium (mg), Dietary Vitamin D (Combined D2 + D3) (mcg), Calcium Supplementation (mg), Vitamin D Supplementation (Combined D2 + D3) (mcg) from the Day 1 Dietary Interview dataset. Finally, we took self-reported osteoporosis diagnosis, and ever broken or fractured wrist from the Osteoporosis dataset.

The dataset was restricted to participants with data collected for the response variable, self-reported osteoporosis diagnosis, and therefore was incidentally restricted to participants at or above the age of 50.

Data Management and Tools

Data management involved 'passing' over NA and missing data, rather than deleting them. This was necessary in order to preserve the weighting structure that makes the final calculations nationally representative. Other data management included renaming and recoding categorical variables and coalescing the WTDRD1 and WTMEC2YR weighting variables to create our final survey weights. The final survey sample included 3,069 participants.

Statistical analysis was completed in R Studio (Version 4.3.0), with heavy use of the packages 'RNHANES', 'survey', 'dplyr', 'tidyverse', 'jtools', 'ggplot2', 'ggeffects', 'WeightedROC', and 'ggsurvey' [1,2,3,4,5,6,7,8,9]. Version control was completed using Git and Github.

Analysis

Our initial exploratory analysis was done via a scatterplot matrix of the continuous variables (dietary calcium, calcium supplementation, dietary vitamin D and vitamin D supplementation) and weighted barplots for categorical variables (age and gender).

For descriptive statistics, we separated the dataset by self-reported osteoporosis diagnosis. A table of weighted population descriptive statistics was created, comparing the osteoporosis population to the osteoporosis-free population. Weighted means and standard deviations were calculated for continuous variables, while weighted percent populations were calculated for categorical ones.

We then tested for significance of the difference between the populations for each of the predictor variables across osteoporosis groups using either a weighted chi-square test or a weighted t-test.

Multivariate Logistic Regression

A multivariate logistic regression was then completed with all predictors included. The reference group was set to the osteoporosis-free, male population. After checking for significant predictors in the full model, we created our second model only including the significant predictors.

We then checked our model assumptions, using VIF to test multicollinearity and using another model to test if the continuous variables were significantly linearly related to the log odds. Finally, we produced plots of the predicted probabilities of significant predictors and created an AUC curve to assess model diagnostic ability.

Results

Exploratory Analysis

In order to visualize the data, we first created a scatterplot matrix of all the continuous variables.

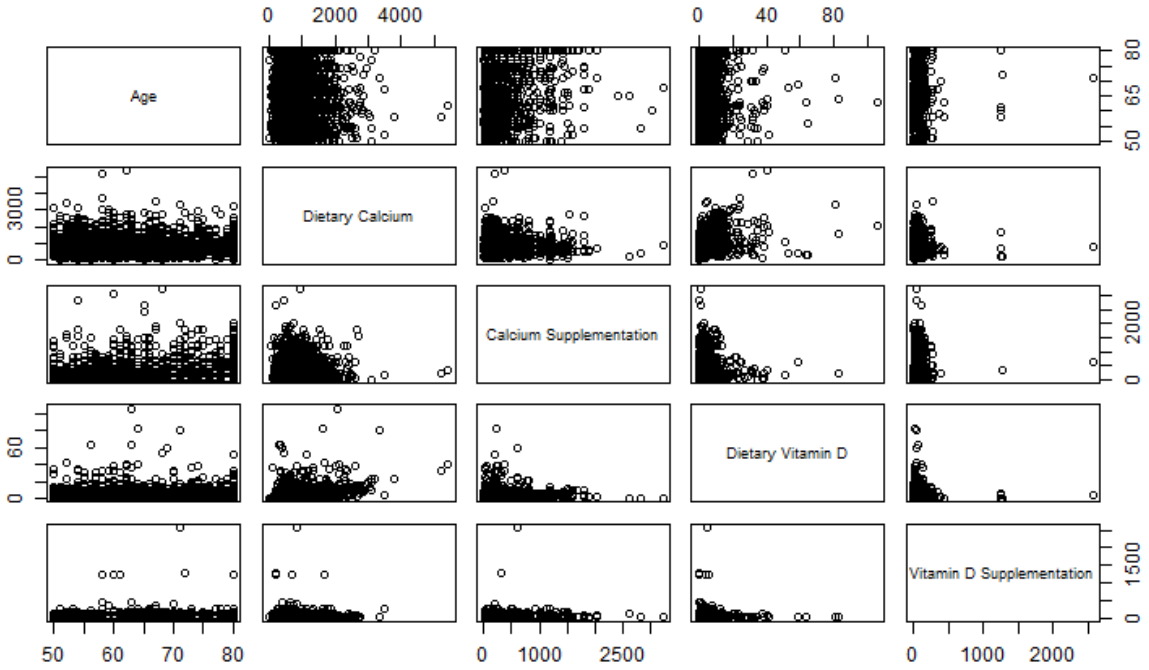

Figure 1: Scatterplot Matrix for continuous predictor variables.

Looking at the scatterplot matrix of continuous variables, it is evident there is no obvious linear relationship between the continuous predictors, which implies that correlation is not a problem for this analysis.

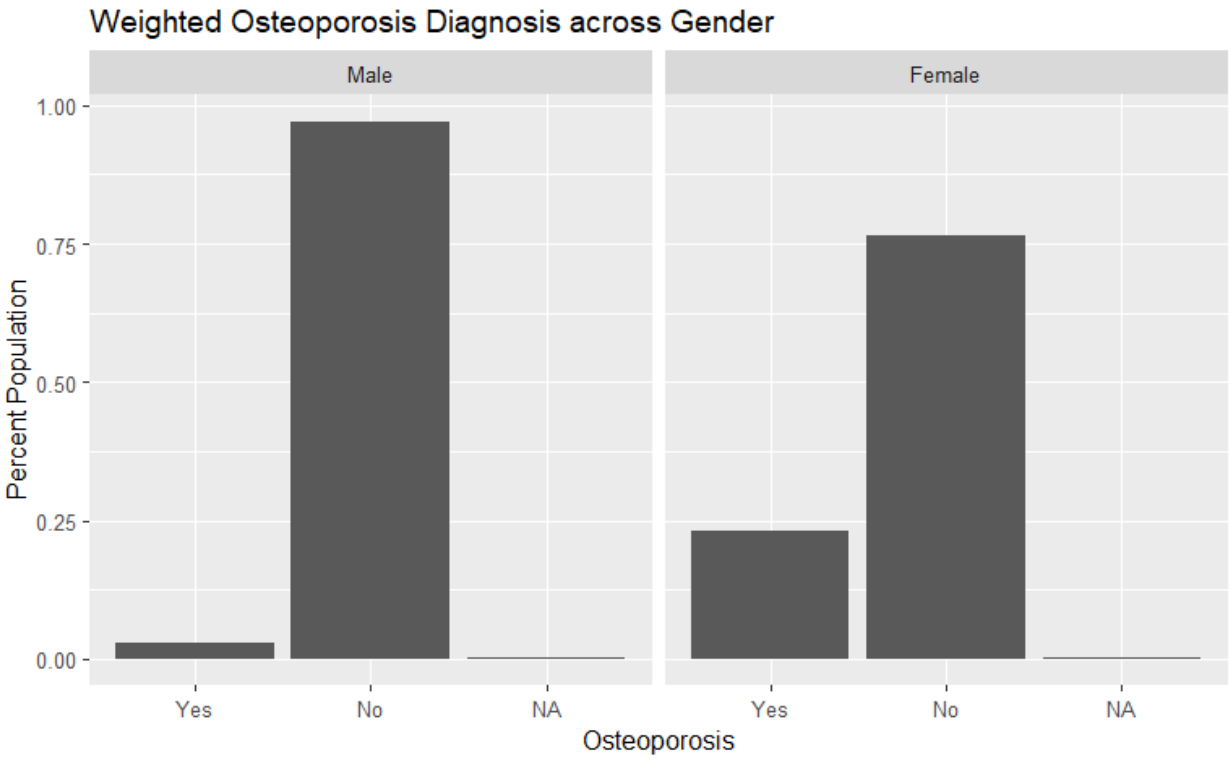

Figure 2: Weighted Osteoporosis Diagnosis across Gender

For the categorical variables, we created bar graphs for each variable to take a cursory glance at the data. Looking at the Gender variable, there is a much larger percent population of women with osteoporosis when compared to the percent population of men with osteoporosis (Figure 2). We can also tell that, in general, osteoporosis is a common disease, but does not exceed 25% in either of the bar plots.

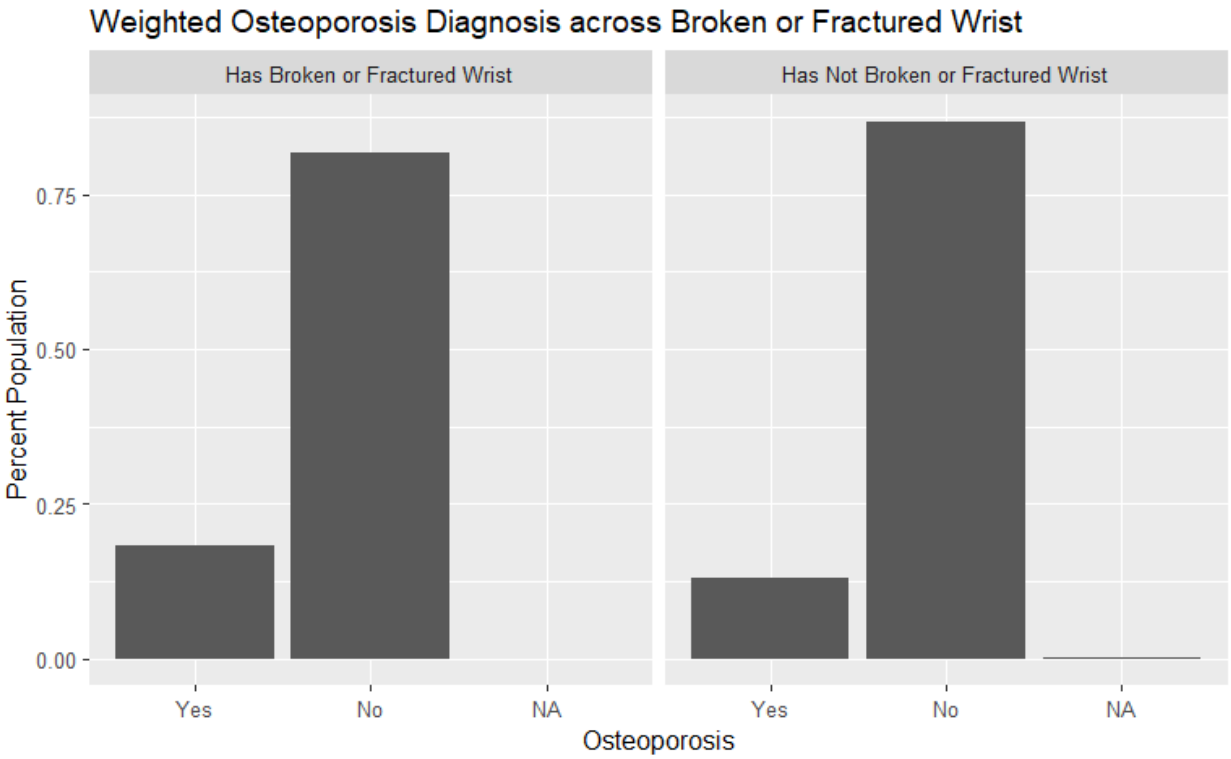

Figure 3: Weighted Osteoporosis Diagnosis across Broken or Fractured Wrist

When examining the Broken or Fractured Wrist variable, we can see that, within the population of people who have broken or fractured their wrist, the percent population of osteoporosis is higher than that which we see in the population of people who have not broken or fractured their wrist (Figure 3).

Descriptive Statistics

Table 1. Descriptive Statistics for the Osteoporosis and No Osteoporosis Populations. Weighted means and standard deviations calculated for continuous variables. Weighted percent populations calculated for categorical variables.

Weighted Means±Std.Dev or % Population	Osteoporosis	No Osteoporosis	Chi-Square Test P-Values
Age	68.94 ± 8.53	63.08 ± 8.832	<0.0001
Gender: Male	9.45%	52.70%	<0.0001
Gender: Female	90.55%	47.30%	
Dietary Calcium (mg)	823.78 ± 443.46	935.39 ± 540.27	<0.05
Dietary Vitamin D (Combined D2 + D3) (mcg)	4.31 ± 4.88	4.38 ± 6.24	
Calcium Supplementation (mg)	517.53 ± 418.88	317.66 ± 355.01	<0.001
Vitamin D Supplementation	63.43 ± 97.91	52.34 ± 97.09	
Broken or Fractured Wrist: Yes	19.16%	13.74%	
Broken or Fractured Wrist: No	80.84%	86.26%	

Comparing the Osteoporosis and No Osteoporosis population as seen in Table 1, we find that the weighted mean age of the Osteoporosis group is significantly greater than that of the No Osteoporosis group, with a chi-square value less than 0.0001.

Gender also significantly differs across the groups, with a chi-square value less than 0.0001. 90.55% of the Osteoporosis population is female, compared to 47.30% of the No Osteoporosis population.

Dietary Calcium (mg) was also found to be significantly different across groups (p-value<0.05), with a weighted mean of 823.78 mg in the Osteoporosis group compared to 935.39 mg in the No Osteoporosis group. Finally, Calcium Supplementation (mg) also differed across groups (p-value<0.001), with a weighted mean of 517.53 mg in the Osteoporosis population and a weighted mean of 317.66 mg in the No Osteoporosis population.

Full Model

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -8.3011264  1.3097256  -6.338 0.000223 ***
age               0.0749438  0.0141979   5.278 0.000748 ***
genderFemale      2.7986689  0.3239150   8.640  2.5e-05 ***
vitamin_d         0.0210421  0.0217499   0.967 0.361648
vitamind_supp     0.0006665  0.0013275   0.502 0.629153
calcium_intake   -0.0002028  0.0004197  -0.483 0.641880
calcium_supp      0.0002017  0.0002113   0.955 0.367646
fractured_wristNo -0.5382090  0.3725238  -1.445 0.186526
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: R output for full multivariate logistic regression model

From our model, the variables Age and Gender were found to be highly statistically significant with p-values well below 0.05. We see that the rest of the variables are not statistically significant. The coefficient estimates tell us the log odds of the variables relationship with our response.

In this case, for every one unit increase in age the log odds of having osteoporosis increases by 0.075. For gender, when compared to men, the log odds of osteoporosis for females increases by 2.8. This can also be interpreted as, for every additional year of age, the odds of having osteoporosis increase by 1.08, after controlling for gender and the other predictor variables (OR 1.08, 95% CI 1.05-1.11, p<0.001). After controlling for age and the other predictor variables, women had 16.42 times the odds of reporting osteoporosis compared to men (OR 16.42, 95% CI8.70-30.99, p<0.001).

Full Model Assumptions

In order to continue with our logistic regression analysis we need to ensure our model meets all the necessary assumptions.

Equation 1. Full multivariate logistic regression model with all predictors included.

$$\widehat{p} = \frac{exp(-8.30+0.075(Age)+2.80(Gender)-0.00(Calcium)+0.00(Suppl.Calcium)+0.02(VitaminD)+0.00(Suppl.VitaminD))}{1+exp(-8.30+0.075(Age)+2.80(Gender)-0.00(Calcium)+0.00(Suppl.Calcium)+0.02(VitaminD)+0.00(Suppl.VitaminD))}$$

The first assumption for any logistic regression model is having the response be binary, in our case this is Osteoporosis and it is binary with values of 0 (no risk) and 1 (there is a risk).

The variables also need to be independent of each other. Earlier, the scatterplot matrix confirmed that the continuous relationships were independent, however this assumption also refers to the inclusion of matched data (or matched pairs) and data with repeated measures. For our dataset we are not including either of the two and can conclude this assumption has also been met.

Table 2. Variance Inflation Factors for Full Model with all predictors included.

Predictors	Age	Gender	Calcium	Supplemental Calcium	Vitamin D	Supplemental Vitamin D	Broken or Fractured Wrist
VIF	1.88	2.59	1.68	2.08	2.26	1.75	3.78

With any model, there needs to be little to no multicollineary (or correlation) between our predictors as this can skew the results. We tested the correlation using the Variance Inflation Factor (VIF) test. As shown above, all our values were below the optimal cutoff of 5, which indicates we do not have multicollinearity. The sample size of our data set, excluding NAs, is 926 observations and is large enough for our analysis.

Table 3. Table of p-values for interaction between continuous variables and their log-odds for the Full Model.

Log(Predictor)	Age	Calcium	Supplemental Calcium	Vitamin D	Supplemental Vitamin D
P-value	0.60	0.44	0.33	0.39	0.96

Lastly, the independent continuous variables need to be linearly related to the logs odds. In other words, there needs to be a linear relationship between our predictors and their logs. To check this, we tested the interaction between each predictors and its log in our model. Each interaction resulted in a non-significant p-value, which tells us there is a linear relationship and our assumption has been validated. After checking all of our model assumptions we felt it was safe enough to continue analyzing this logistic regression model.

Final Model

We then created our final model with only the significant predictors, age and gender, included.

For this model, the significance of age and gender increased, with p-values less than 0.00001. For our final model, for every one unit increase in age the log odds of having osteoporosis increases by 0.075. For gender, when compared to men, the log odds of osteoporosis for females increases by 2.41.

In plain language, after controlling for other predictors, every additional year of age meant 1.08 higher odds of self-reporting osteoporosis (OR 1.08, 95% CI 1.06-1.10, p<0.00001). After controlling for other predictors, women had 11.20 times higher odds of reporting osteoporosis diagnosis when compared to men (OR 11.20, 95% CI 6.17-20.35, p<0.0001).

Final Model Assumptions

We again check all necessary assumptions for our final logistic regression model.

Equation 2. Final multivariate logistic regression model with only age and gender included.

$$\widehat{p} = \frac{exp(-8.52+0.075(Age)+2.41(Gender))}{1+exp(-8.52+0.075(Age)+2.41(Gender))}$$

We meet the first assumption of binary reponse variable, as we did in the full model.

The variables also need to be independent of each other. Earlier, the scatterplot matrix confirmed that the continuous relationships were independent, however this assumption also refers to the inclusion of matched

data (or matched pairs) and data with repeated measures. For our dataset we are not including either of the two and can conclude this assumption has also been met.

Table 4. Variance Inflation Factors for Final Model with only age and gender included.

Predictors	Age	Gender
VIF	1.89	2.59

We again test for multicollinearity using the VIF test. All values were far below the optimal cutoff of 5, which indicates we do not have multicollinearity. The sample size of our data set, excluding NAs, is 3,053 observations and is large enough for our analysis.

Table 5. Table of p-values for interaction between continuous variables and their log-odds for the Final Model.

Log(Predictor)	Age
P-value	0.51

Finally, age, as the only independent continuous variables, was linearly related to the logs odds, fulfilling the last requirement.

Predictive Probability Plots

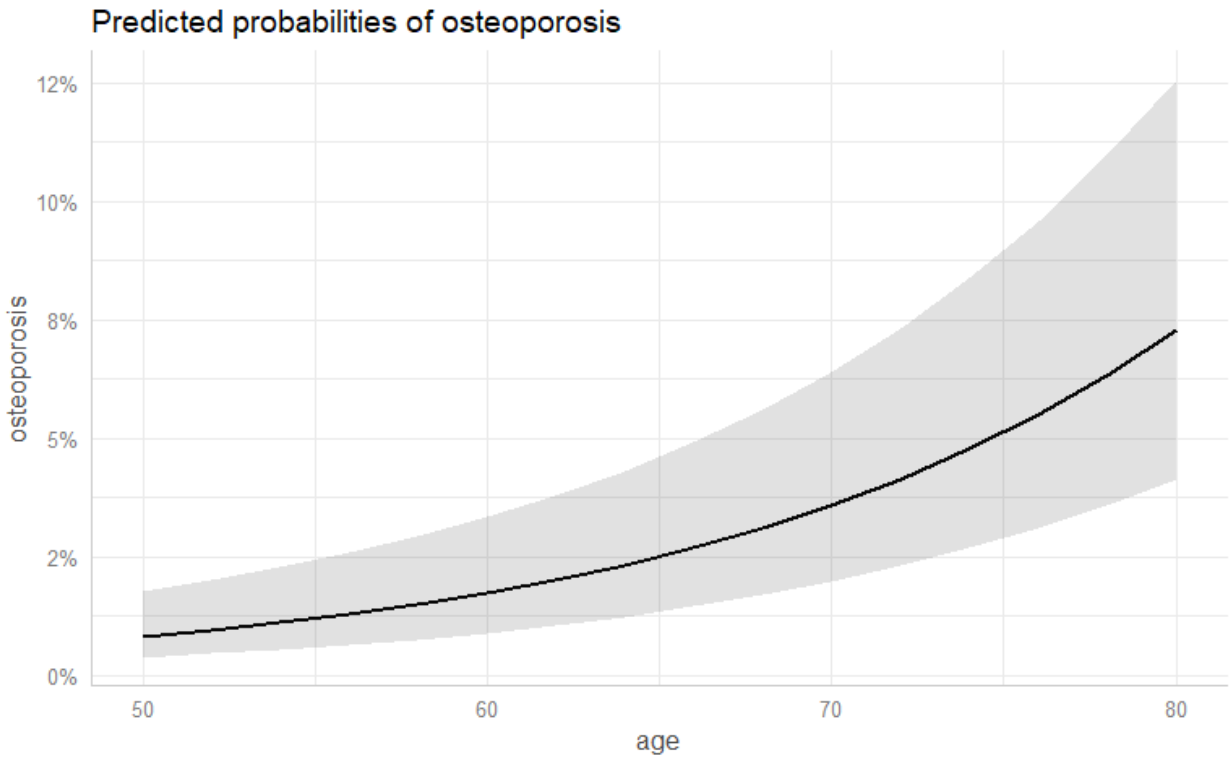

Figure 5: Predictive Probability Plot of Final Model for Age

Using the results of our final model, we plotted the predicted probabilities of Osteoporosis for each of our significant predictors. For Age, we see that as age increases the risk of a positive osteoporosis diagnosis increases up to 7%.

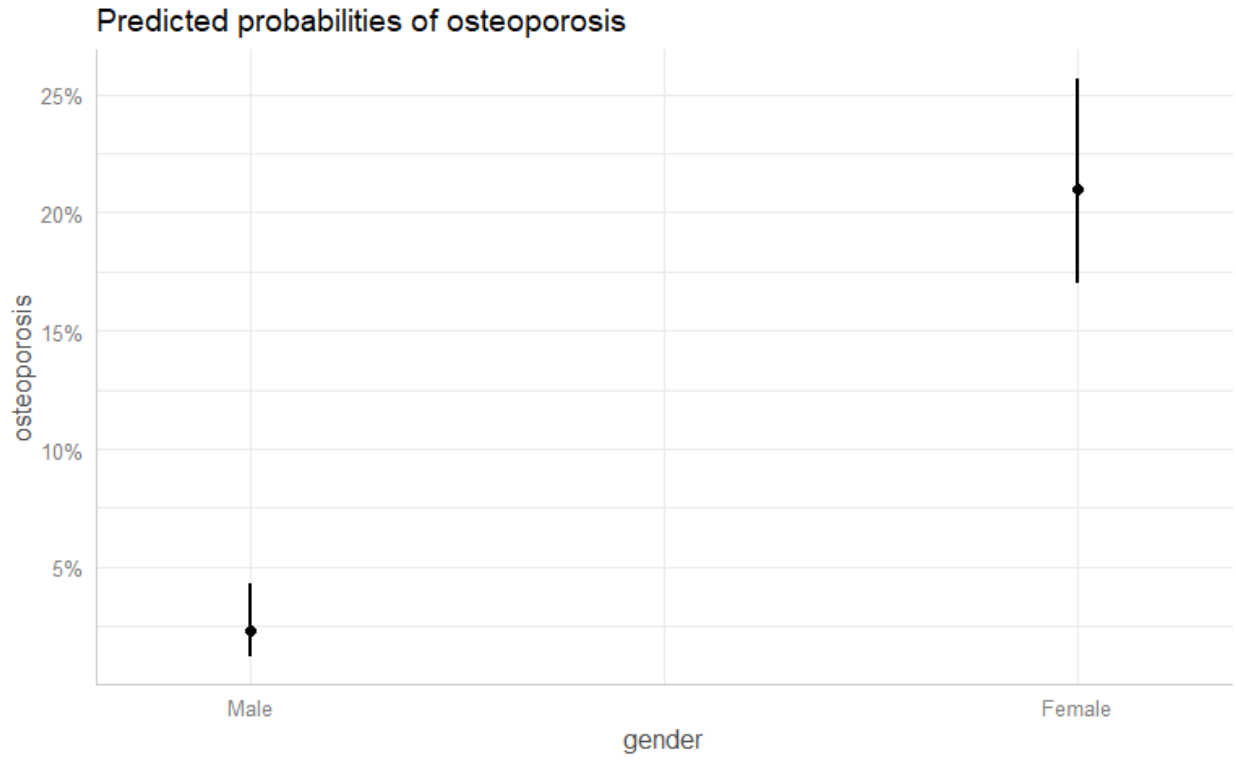

Figure 6: Predictive Probability Plot of Final Model for Gender

For gender, females are much more likely to receive a positive diagnosis than males, specifically 11.20 times more likely, The predicted probability for females to have osteoporosis is 21%, while the predicted probability of osteoporosis in males is 2%.

Model Test

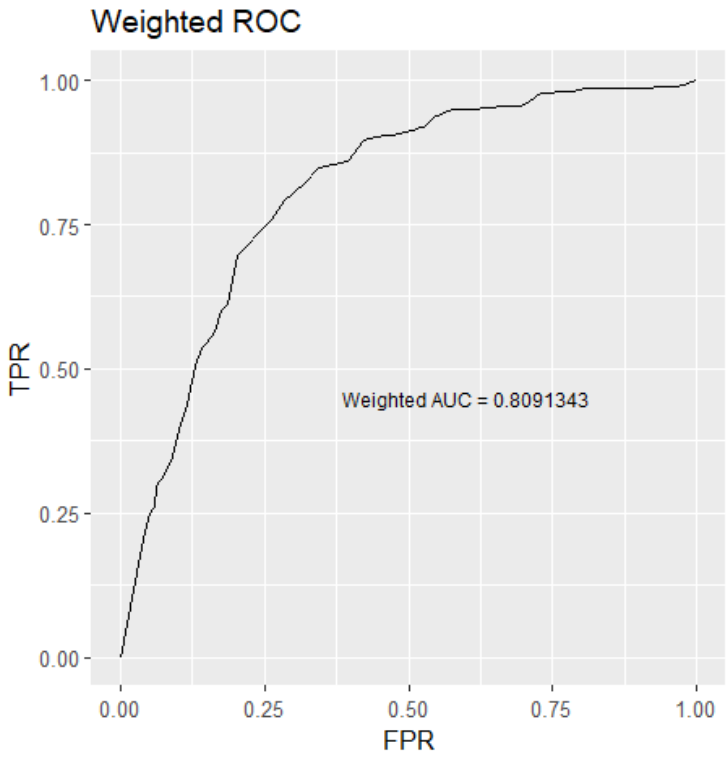

Figure 7: Receiver operating characteristic (ROC) Curve for Final Model

To test how well our model performs at correctly assessing the predictors effect on our outcome (Osteoporosis), we used a Receiver Operating Characteristic Curve that allows us to illustrate the diagnostic ability of our model. The curve plots the false positive rate (FPR) versus the true positive rate (TPR) of our model. To quantify our assessment, we assess the Area Under the Curve (AUC), which has a maximum value of 1. The goal is to have a high AUC as possible, as this will indicate a perfectly accurate test, or the better our model is at predicting 0 classes as 0 and 1 classes as 1. For our plot, the AUC is 0.809, or 0.81, since our value is above 0.80, this indicates our test is good.

Conclusions

In conclusion, we found that Age and Gender significantly impact the risk of self-resporting diagnosed osteoporosis. These are pre specified, clinically significant risk factors, so it is logical that they were highly significant. Initial analysis of dietary calcium and calcium supplementation did seem to indicate that, within the osteoporosis group, there was a significantly lower weighted mean of dietary calcium and a significantly higher weighted mean of calcium supplementation. This aligns with the clinical presentation and treatment of the disease, which often involved people with low calcium intake to supplement with additional calcium in order to prevent osteoporosis from occurring or worsening. Unfortunately, these dietary factors were not as significant in our logistic regression model.

Future Research

For future research, we could incorporate both Day 1 and Day 2 dietary and supplementary values for calcium and vitamin D. We did not initially incorporate both day values for Calcium and Vitamin D as

this would have complicated our survey design objects, analysis and models due to the addition of extra weights. We could also pursue further investigation of other risk factors including alcohol consumption, smoking status, activity level, weight and bone density. Additionally, working with a dataset specifically collected for osteoporosis studies, rather than population health studies, would also improve the precision of our model and definitions of our variables. It should also be noted that since NHANES is a survey, and none of our values came from the examination or laboratory data, all of our variables were self-reported and are therefore subjected to a number of biases, including recall bias and social desirability/conformity bias.

References

1. Alexander B (2022). *ggsurvey: Simplifying 'ggplot2' for Survey Data*. R package version 1.0.0, https://CRAN.R-project.org/package=ggsurvey.

2. Hocking TD (2020). *WeightedROC: Fast, Weighted ROC Curves*. R package version 2020.1.31, https://CRAN.R-project.org/package=WeightedROC.

3. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

4. Long JA (2022). *jtools: Analysis and Presentation of Social Scientific Data*. R package version 2.2.0, https://cran.r-project.org/package=jtools.

5. Lüdecke D (2018). "ggeffects: Tidy Data Frames of Marginal Effects from Regression Models." *Journal of Open Source Software*, *3*(26), 772. doi:10.21105/joss.00772 https://doi.org/10.21105/joss.00772.

6. Susmann H (2016). *RNHANES: Facilitates Analysis of CDC NHANES Data*. R package version 1.1.0, http://github.com/silentspringinstitute/RNHANES.

7. T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0. T. Lumley (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19 T. Lumley (2010) Complex Surveys: A Guide to Analysis Using R. John Wiley and Sons.

8. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

9. Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.2, https://CRAN.R-project.org/package=dplyr.

10. "Nhanes Tutorials." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, https://wwwn.cdc.gov/nchs/nhanes/tutorials/default.aspx.

11. "NHANES 2017-2018 Laboratory Data Overview." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewlab.aspx?BeginYear=2017.

12. "Nhanes Tutorials - Nhanes Dietary Analyses Module." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, https://wwwn.cdc.gov/nchs/nhanes/tutorials/DietaryAnalyses.aspx.

13. Burge, Russel, et al. "Incidence and economic burden of osteoporosis-related fractures in the United States, 2005–2025." Journal of bone and mineral research 22.3 (2007): 465-475.

14. Cooper, Cyrus, and L. Joseph Melton III. "Epidemiology of osteoporosis." Trends in Endocrinology & Metabolism 3.6 (1992): 224-229.

Appendix A

Code Repository can be found on Github at https://github.com/arenganathan28/StatAppProject.git