Risk Factors in Diabetes Diagnoses

Kyle Su Comorbidities, Health Factors, Odds Ratios, Some charts for parts 1 and 2; Step 3, parts of conclusion
Yiping Wang Lifestyles parts 1 and 2, Tests of Association; Coding differences analysis part 4
Nay Petrucelli Demographics, Medical Care, Some charts for parts 1 and 2; sex-based analysis in part 4, part of conclusion

1. Introduction

Diabetes is a widespread and growing health concern in the United States, affecting over 37 million people. Oftentimes, late-diagnosed diabetes can lead to severe complications, including cardiovascular disease, kidney failure, nerve damage, and blindness. Identifying individuals at high risk can help promote earlier intervention, behavior change, and better clinical outcomes. This report aims to build a reliable predictive model using health, lifestyle, comorbidities, and demographic data to understand which factors are most strongly associated with diabetes status. By identifying these predictors, we hope to contribute to early detection strategies and preventive efforts.

This report analyzes a subset of data from the Centers for Disease Control's Behavioral Risk Factor Surveillance System survey. Each record in the dataset includes an indicator of whether a respondent has diabetes and 21 health and demographic variables from the 2015 version of this annual survey. (Teboul 2022) Of the 21 predictors, 14 are binary and 7 take discrete numerical values. Of the 7 non-binary variables, 3 were recoded. The systems used for recoding these variables are below.

Mental Health (MentHlth) and Physical Health (PhysHlth) initially took values between 0 and 30. They indicate the number of days in a month respondents reported poor health. Most observations were concentrated at 0, and many levels had few or no observations. These variables were recoded with the following ordinal scale:

Table I: Mental Health and Physical Health Recoding

Value	Number of days reported poor health (in 1 month)
1	Rarely or never (0-3 days)
2	Less than half of the time (4-13 days)
3	Approximately half of the time (14-16 days)
4	More than half of the time (17-27 days)
5	Always or nearly always (28-30 days)

Body-Mass Index (*BMI*) is a ratio of body weight (in kilograms) to the square of body height (in meters) and is one simple metric to estimate body fat. (Cleveland Clinic 2022) The majority of the values fell between 20 and 40, with values spread as low as 13 and as high as 79. This variable was recoded using the clinical definitions for weight types.

Table II: BMI Recoding

Value	BMI Classification
1	Underweight
2	Optimal Range
3	Overweight
4	Class I Obesity
5	Class II Obesity
6	Class III Obesity

The appendix shows the distribution of these 3 variables before and after recoding.

2. Overview of Variables

Our response variable is a binary indicator of whether the study participant had a diagnosis of either diabetes or prediabetes. In our sample, the proportion of those who had this diagnosis was 0.495, or roughly half of the sample.

We classified the predictors in the dataset into 5 groups: demographic characteristics, lifestyle factors, health metrics, medical care, and comorbidities. Below, we consider descriptive statistics for each group of factors and their relationship to our response variable.

2.1 Demographic Characteristics

We considered age, sex, education, and income as the non-health-related demographic factors associated with each survey participant. Descriptive statistics show that just over half of the respondents are female, and the majority are older. The median age range is the low 60s, and the first quartile value corresponds to an age range in the low 50s. All 3 of the ordinal variables in this category had a strong left skew (graphs can be seen in the appendix). The education variable had a small number of values in the 2 categories below high school graduates and a high number of observations in the remaining three categories. The median value of 5 corresponds with some college attendance. For income, the median coded income level was 6, which corresponded with an income range of $35,000 to $49,999.

Demographic Variables (Binary)

Variable Short Name	Binary Variable Description	Proportion
Sex	Sex (Male = 1)	0.478

Demographic Variables (Ordinal)

Variable Short Name	Ordinal Variable Description
Age	Full scale 1-13 in codebook. 1= 18-24, 9=60-64, 13= 80 or more
Education	2= elementary 3= some high school 4= high school grad or GED 5= some college 6= college graduate
Income	Full scale 1-8 in codebook: 1= less than 10k , 5= less than 35k, 8= 75k or more

	Mean	Min.	Lower Quartile	Median	Upper Quartile	Max.
Age	8.714	1	7	9	11	13
Education	4.866	2	4	5	6	6
Income	5.577	1	4	6	8	8

2.2 Medical Care Treatment and Access

The variables included in this category relate to the interactions the survey respondent had with the medical industry. The large majority of respondents reported having had a cholesterol check, at least some health coverage, and did not report avoiding seeing a doctor due to cost.

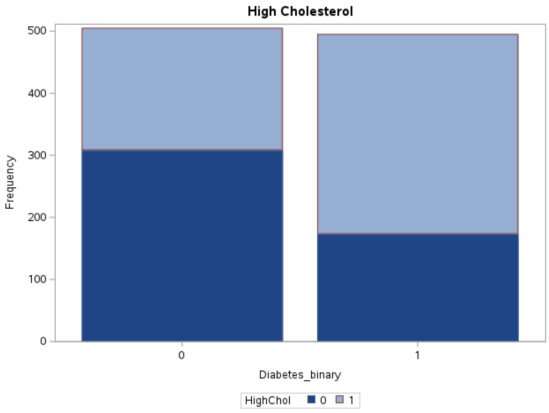
Variable Short Name	Binary Variable Description	Proportion
CholCheck	Had a cholesterol check in the past 5 years?	0.974
AnyHealthcare	Any healthcare coverage in the past year?	0.946
NoDocbcCost	Avoided doctor in past year due to cost?	0.105

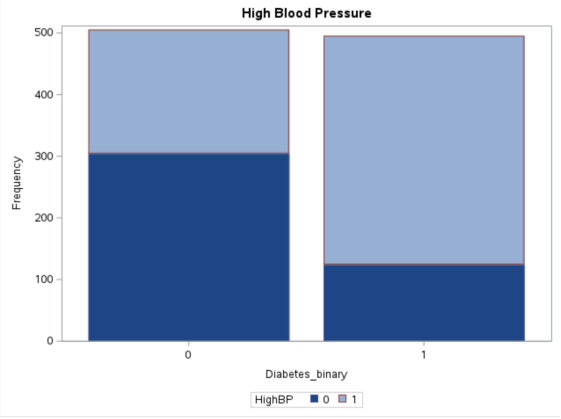
2.3 Comorbidities

Comorbidities refer to the presence of two or more medical conditions occurring alongside a primary illness. In the case of diabetes, common comorbidities include high blood pressure, high cholesterol, heart disease or heart attack, and stroke. In our dataset, high blood pressure (57%) and high cholesterol (51.7%) were the most prevalent, while heart disease (15%) and stroke (5.4%) were less common.

Variable Short Name	Binary Variable Description	Proportion
HighBP	High Blood Pressure?	0.57
HighChol	High Cholesterol?	0.517
HeartDiseaseorAttack	Have coronary heart disease or myocardial infarction?	0.15
Stroke	Ever had a stroke?	0.054

The Stacked bar plots below showcased that the proportion of individuals with high blood pressure or cholesterol was nearly twice as high among those with diabetes compared to non-diabetics.





Although less frequent overall, heart disease and stroke also appeared more commonly in the diabetic group. Thus, these consistent patterns support the idea that the presence of each condition is associated with a greater likelihood of having diabetes.

2.4 Health Metrics

In this study, Health metrics include BMI, general health (GenHlth), mental health (MentHlth), physical health (PhysHlth), and serious difficulty walking (DiffWalk). From the average BMI of 3.538 and the lower quartile of 3, we can notice that the majority of individuals fall into the overweight or obese categories.

Variable Short Name	Binary Variable Description	Proportion
DiffWalk	Experience serious difficulty walking?	0.265

Variable Short Name	Ordinal Variable Description
BMI	Body Mass Index (Scale in Table II)
GenHlth	General Health: 1 = Excellent 2 = Very Good 3 = Good 4 = Fair 5 = Poor
MentHlth	Number of days in past month MH poor (Scale in Table I)
PhysHlth	Number of days in past month PH poor (Scale in Table I)

	Mean	Min.	Lower Quartile	Median	Upper Quartile	Max.
BMI	3.538	1	3	3	4	6
GenHlth	2.912	1	2	3	4	5
MentHlth	1.49	1	1	1	1	5
PhysHlth	1.755	1	1	1	2	5

2.5 Lifestyle

In this study, lifestyle refers to the binary status of real-life activities reported by observed individuals, from which indicated physical activity, fruit consumption, and

veggie consumption statuses are more prevalent (proportion > 0.5), and smoking and heavy alcohol consumption are less common (proportion < 0.5).

Variable Short Name	Binary Variable Description	Proportion
Smoker	Smoked at least 100 cigarerres in lifetime?	0.478
PhysActivity	Phyical activity outside of work in past 30 days?	0.699
Fruits	Eat fruit at least once per day?	0.603
Veggies	Eat vegetables at least once per day?	0.78
HvyAlcoholConsump	More than 14 (men) or 7 (women) drinks per week?	0.039

3. Tests of Variable Associations

To understand the relationship between our predictor variables and our response variable, we conducted tests of association. We highlighted different tests in the narrative below, but in all cases, the likelihood ratio, score, and Wald metrics were in agreement. For all tests, we used the p-value significance threshold of 0.05 (or 95% confidence intervals). A summary of all associations is found in the table below, with those that are statistically significant highlighted in grey.

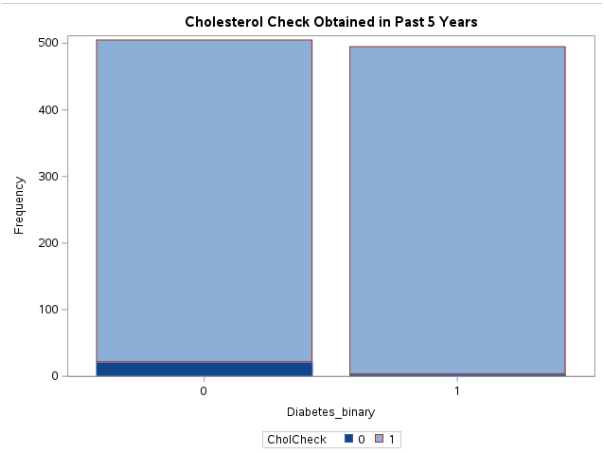
Category Name	Predictor Name	Chi-Square LR	Chi-Square Score	Chi-Square Wald	pvalue LR	pvalue Score	pvalue Wald	Consistent?
Comorbidities	HeartDiseaseorAttack	36.76	35.74	33.40	<.0001	<.0001	<.0001	TRUE
	HighBP	129.10	125.96	119.68	<.0001	<.0001	<.0001	TRUE
	HighChol	68.66	67.86	66.26	<.0001	<.0001	<.0001	TRUE
	Stroke	10.21	9.95	9.34	0.0014	0.0016	0.0022	TRUE
Demos	Age	79.95	76.84	71.28	<.0001	<.0001	<.0001	TRUE
	Education	42.69	41.91	40.39	<.0001	<.0001	<.0001	TRUE
	Income	63.49	62.11	59.24	<.0001	<.0001	<.0001	TRUE
	Sex	2.08	2.08	2.08	0.1492	0.1492	0.1497	TRUE
Health Metrics	BMI	87.17	79.55	71.46	<.0001	<.0001	<.0001	TRUE
	bmiRcd (*Recode)	99.51	95.44	87.51	<.0001	<.0001	<.0001	TRUE
	DiffWalk	100.80	97.29	89.14	<.0001	<.0001	<.0001	TRUE
	GenHlth	162.29	152.01	132.54	<.0001	<.0001	<.0001	TRUE
	MentHlth	2.38	2.37	2.36	0.1231	0.1235	0.1246	TRUE
	mhRcd (*Recode)	1.95	1.94	1.94	0.1628	0.1632	0.1642	TRUE
	PhysHlth	50.80	49.07	45.28	<.0001	<.0001	<.0001	TRUE
	phRcd (*Recode)	49.78	48.16	44.60	<.0001	<.0001	<.0001	TRUE
Lifestyle	Fruits	5.11	5.11	5.10	0.0237	0.0238	0.0239	TRUE
	HvyAlcoholComsump	4.33	4.24	4.09	0.0375	0.0394	0.0431	TRUE
	PhysActivity	33.85	33.55	32.86	<.0001	<.0001	<.0001	TRUE
	Smoker	0.47	0.47	0.47	0.4949	0.4949	0.495	TRUE
	Veggies	11.44	11.39	11.26	0.0007	0.0007	0.0008	TRUE
Medical Care	AnyHealthcare	1.43	1.43	1.42	0.2314	0.2321	0.234	TRUE
	CholCheck	13.68	12.43	9.89	0.0002	0.0004	0.0017	TRUE
	NoDocbcCost	0.69	0.69	0.69	0.4062	0.4063	0.4072	TRUE

3.1 Demographic Characteristics

In the demographic category age ($G^2 = 79.95$, p-value < 0.0001), education ($G^2 = 42.69$, p-value < 0.0001), and income ($G^2 = 63.49$, p-value < 0.0001) were all statistically significant, while sex was not. Increased age was associated with increased odds for diabetes, while increased education and income were associated with lower odds of diabetes. Interestingly, while the test for association was not statistically significant, sex does appear to be an important consideration for this study as discussed in the additional analyses section.

3.2 Medical Care Treatment and Access

The tests of association show only a statistically significant result for the variable cholesterol check ($G^2 = 13.68$, p-value 0.0002). From the stacked bar chart, we can see that although the proportions are visibly different between both the diabetes variable categories, the proportion of survey respondents who reported not having a cholesterol check was extremely low in both cases.



For the other two variables in this category, the large majority of respondents reported having access to healthcare and did not report avoiding the doctor due to cost. The proportions of those reporting reduced access to healthcare were slightly higher in those with diabetes in both variables, albeit not significant.

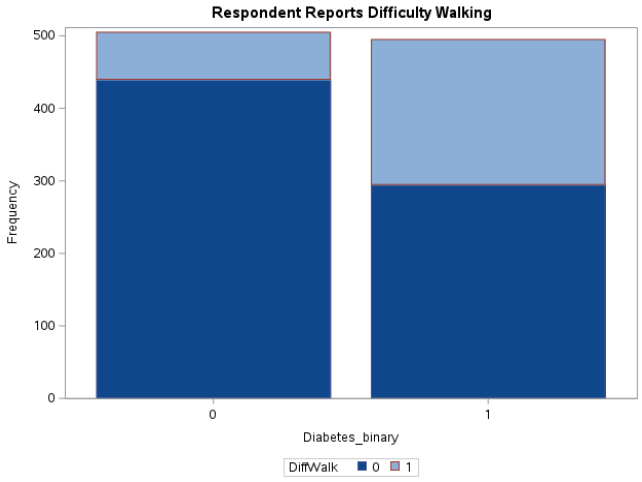
3.3 Comorbidities

To formally assess these relationships, we did Chi-square tests and fit separate logistic regression models with diabetes status as the outcome. All four comorbidities were found to be significantly associated with diabetes. These tests were consistent with the other tests of association.

Of these variables, high blood pressure was the strongest predictor (Wald $X^2 = 119.68$, $p < .0001$), followed by high cholesterol (Wald $X^2 = 66.26$, $p < .0001$), heart disease (Wald $X^2 = 33.40$, $p < .0001$), and stroke (Wald $X^2 = 9.34$, $p = .0022$). Additionally, the odds ratios and their 95% confidence intervals that are above 1 indicate that the presence of these comorbid conditions substantially increases the odds of diabetes. These findings reinforce the observed trends in the descriptive analysis and support the conclusion that comorbid conditions, especially cardiovascular-related ones, are related to the increased odds of having diabetes in this population.

3.4 Health Metrics

The stacked bar plot showed an interesting trend that the proportion of individuals with walking difficulty was substantially higher among those with diabetes.



Furthermore, the box plots showed that individuals with diabetes generally had higher BMI, poorer general health, and more days of poor physical health compared to non-diabetics. The only variable that does not seem to have a clear relationship with diabetes is Mental health. These observations suggest that worse physical condition and mobility limitations are most likely associated with a higher odds of diabetes.

Besides the graphic analysis, we also conducted Chi-square tests and fitted individual logistic regression models with diabetes status as the outcome. According to the table (Appendix 2), all variables except mental health showed statistically significant relationships with diabetes. General health was the strongest predictor, followed by difficulty walking, BMI, and physical health. In addition, it is also shown in the table from Appendix 4 that the odds ratios for the significant predictors were all above 1, and their 95% confidence intervals did not include 1. For example, the odds of diabetes were over 4.5 times higher for those reporting serious walking difficulty (OR = 4.589, CI:

3.345–6.297) and 2.2 times higher for those with worse general health (OR = 2.202, CI: 1.925–2.519). On the other hand, the 95% confidence interval for the odds ratio of mental health(0.968, 1.212) includes 1, meaning the relationship is statistically significant. These results confirm and strengthen the patterns observed in the descriptive plots, supporting the conclusion that poorer physical condition and reduced mobility are highly associated with increased likelihood of diabetes.

3.5 Lifestyle

In the test results from Chi-Square tests, the table demonstrates that all variables but smoking status had p-values smaller than 0.05, and all the test statistics provided consistent results. In the Chi-Square results between the diabetes response and smoking status, the p-values for the test statistics are all approximately 0.5, which is much larger than confidence level of 0.05, which indicate that there is not enough evidence to claim that there is statistically significant difference between the frequencies of diabetes between the binary categories of the smoking status. Furthermore, the results in the contingency table, the odds ratio table, and the stacked bar plots of the predictors, which also numerically and visually present the differences, corroborate the Chi-Square results that all variables but smoking status have significant differences in diabetes response between the categories. The odds ratio estimate of all predictors but smoking status had ratios lower than 1, indicating that the presence of the variable (predictor = 1) would indicate a lower odds of diabetes.

4. Final Model Selection and Results

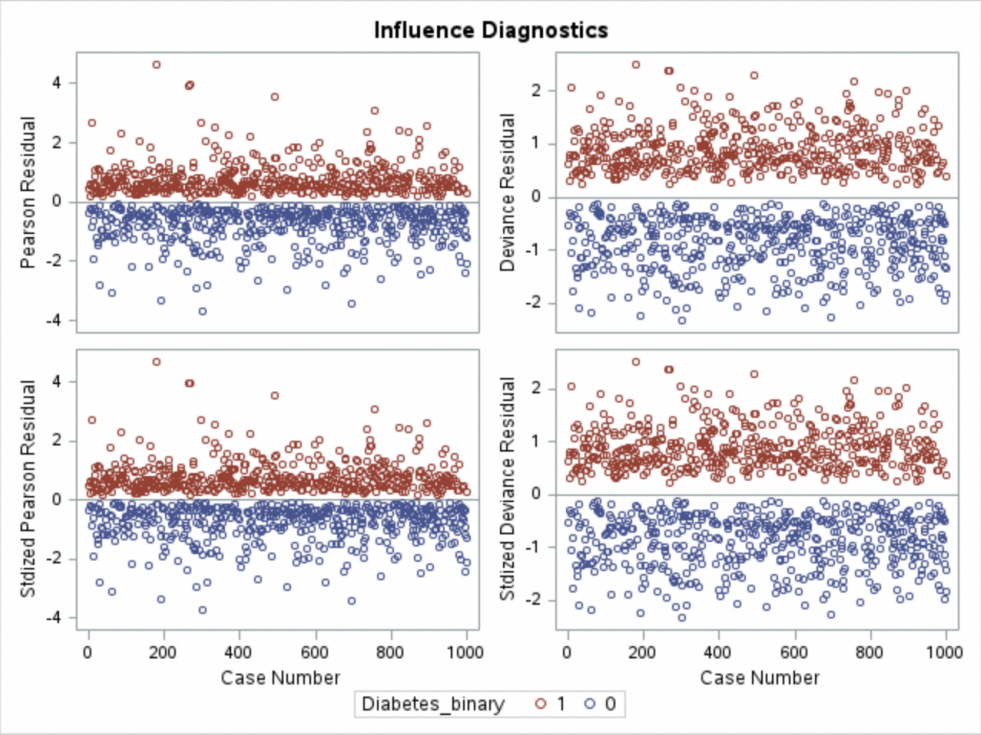
Based on the descriptive statistics in the exploratory analysis, we proceeded to build a predictive model to determine the predictors that best explain the outcome variable (Diabetes_binary). To begin with, we used a stepwise variable selection method in SAS with significance levels set at 0.05. This approach adds or removes predictors from the model based on their statistical contribution to predicting diabetes status. The selection was performed using all 21 variables. In this case, ordinal variables such as bmiRcd, GenHlth, and mhRcd were treated as categorical in the selection step to allow for more flexibility in capturing their potential non-linear effects on diabetes risk. This approach avoids imposing a linear trend across levels, which is likely not appropriate given the complexity of these health-related variables.

4.1 Final Model Selection

The stepwise procedure identified a final model with seven valid predictors: HighBP, HighChol, bmiRcd, GenHlth, DiffWalk, CholCheck, and Age. This model was then refit to the complete dataset to produce coefficient estimates, odds ratios, and diagnostic

plots. The inclusion of these variables aligns well with prior research and initial exploratory analyses, confirming their predictive relevance for diabetes diagnosis.

4.2 Model Diagnostics and Influence Assessment



To verify model assumptions and detect any influential observations, we examined several diagnostic plots and statistical summaries. As shown in the influence diagnostic plots and Appendix 5, we assessed likelihood residuals, deviance, along with leverage and CBAR values. The deviance residuals measure how much each individual observation deviates from the model's predicted value, while the Pearson residuals capture the standardized difference between observed and expected values. In this case, both deviance and Pearson residuals are symmetrically distributed and centered around zero, showing no notable outliers. Furthermore, the leverage values were also within standard thresholds, which indicates that no individual case exerted an extreme influence on the model. This supported our findings from the CBAR values, as none of the CBAR values exceeded 0.5.

Besides using Cbar and leverage, plots from Appendix 6 also displayed DFBETA plots for each predictor, illustrating how much each individual observation influenced the parameter estimates. For our model, all DFBETA values remained relatively close to zero, indicating that coefficient estimates are not overly sensitive to individual data

points. These diagnostics together indicate that the current model is already robust and not affected by outliers or influential data points.

4.3 Predictor Significance and Interpretation

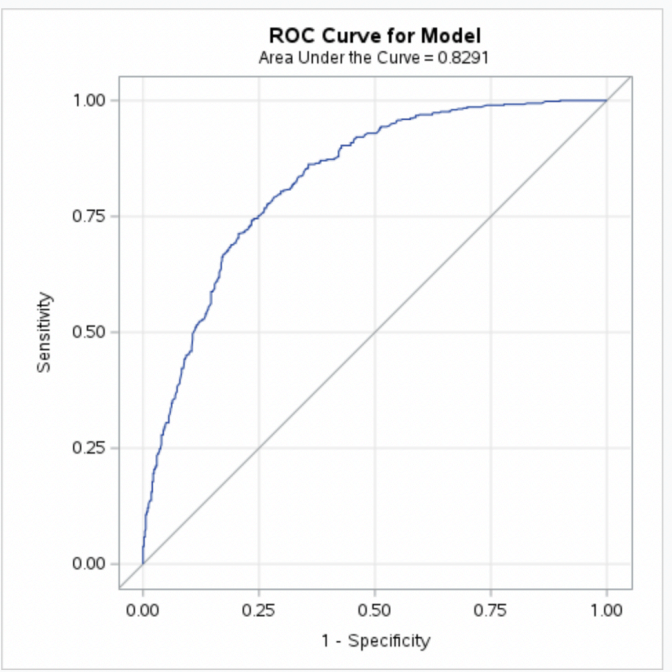
Type 3 Analysis of Effects for Final Model			
Effect	DF	Wald Chi-Square	Pr > ChiSq
HighBP	1	16.9446	<.0001
HighChol	1	12.2986	0.0005
bmiRcd	5	54.8654	<.0001
GenHlth	4	48.9365	<.0001
DiffWalk	1	8.6422	0.0033
CholCheck	1	4.0693	0.0437
Age	1	37.1296	<.0001

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
HighBP 1 vs 0	1	1.974	1.428	2.73
HighChol 1 vs 0	1	1.751	1.28	2.396
bmiRcd 2 vs 1	1	0.98	0.239	5.138
bmiRcd 3 vs 1	1	2.574	0.647	13.224
bmiRcd 4 vs 1	1	4.057	1.015	20.871
bmiRcd 5 vs 1	1	7.195	1.712	38.503
bmiRcd 6 vs 1	1	6.589	1.519	36.026
GenHlth 2 vs 1	1	1.828	0.961	3.63
GenHlth 3 vs 1	1	4.276	2.309	8.302
GenHlth 4 vs 1	1	5.784	2.931	11.897
GenHlth 5 vs 1	1	8.126	3.635	18.949
DiffWalk 1 vs 0	1	1.814	1.222	2.708
CholCheck 1 vs 0	1	3.65	1.125	14.581

The type 3 analysis of effects confirms that all included predictors are statistically significant at the 0.05 level using the Wald test. From the odds ratio estimate table, we can interpret the effects of each predictor. For instance, the HighBP variable shows that individuals with high blood pressure had nearly double the odds of having diabetes compared to those without (OR = 1.974). For bmiRcd, it shows that the risk of diabetes increased sharply with BMI. If individuals are in the highest category, they have an odds of having diabetes over 7 times higher (OR = 7.195) compared to the reference group. Lastly, the odds ratio for GenHlth showcases that poorer reported health was highly associated with substantially increased risk. Those in the worst health category had over 8 times the odds of being diabetic (OR = 8.126). Of note is that the odds Ratio 95

percent confidence interval for these variables all exclude 1, which is consistent with significance at the 0.05 level.

4.4 Model Performance and Predictive Power



After selecting our optimal model, we used multiple criteria to assess how well the model predicts the diabetes outcome. First, the Hosmer and Lemeshow Goodness-of-Fit Test has a p-value of 0.9336 (Appendix 7), which suggests that the model fits the data well, as we fail to reject the null hypothesis of good fit. Additionally, the global tests were all statistically significant (LR test statistic= 378.18, p-value <0.0001). This is also supported by the calibration curve in Appendix 7, as it shows that predicted probabilities align closely with actual outcomes across the full range of values, confirming good model calibration. Most importantly, the ROC curve shows an AUC of 0.8291, indicating strong discriminative ability. This proves that the model is effective at separating individuals with and without diabetes. Given the results above, we can conclude that our optimal model performs well in terms of discrimination (ROC) and calibration (calibration plot).

5. Additional Analyses

5.1 Impact of Variable Coding on Model Results

Decisions about variable coding have an impact on model results. Variable importance differed when we used the original categorical variables BMI, MentHlth, and PhysHlth to fit the logistic main-effects model in lieu of bmiRcd, mhRcd, and phRcd. First, in the

model with the original variable coding, CholCheck was no longer significant, and the AUC (area under the ROC curve) dropped from 82.9 to 82.1. In addition, for the model with the original coding, all GenHlth categories were statistically significant, but for the model with recoded variables, the second category of GenHlth had p-value = 0.0739 > 0.05, indicating that this level was not significant at alpha = 0.05. Another notable difference was that the BMI had a p-value less than 0.05 in the model with original variables, but the bmiRcd in the model with recoded variables had p-values greater than 0.05 for categories 2-4, indicating no significant difference of influence on the response variable compared to the reference level at confidence level alpha = 0.05. Last but not least, while the conclusions for Hosmer and Lemeshow Goodness-of-Fit test are the same for both models since both had p-values > 0.05, the p-value is significantly larger for the model with recoded variables.

To further illustrate this point, an identical model to our final model (containing the 3 recoded variables) was rerun with the small change of collapsing the BMI variable from 6 categories to 3 (1-2, 3-4, 5-6). In this model, CholCheck remained significant, but mental health was introduced as an additional significant predictor. This was on the cusp of significance in our selected model and was significant in the sex-specific model discussed below.

5.2 Sex-Based Differences in Diabetes Diagnostics

Existing literature suggests that the relative importance of risk factors for diabetes may differ between men and women. We followed the same model fitting and variable selection procedure used in the main analysis model to test whether variable significance would differ based on the subset of male survey participants compared with the subset of female participants. The global tests for fit were highly significant in both models. Variable selection indicated that there were some differences in significance between the sexes.

Variable Selection - Female

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	HighBP		1	1	98.0535		<.0001
2	GenHlth		4	2	69.5555		<.0001
3	bmiRcd		5	3	41.5724		<.0001
4	Age		1	4	22.9148		<.0001
5	mhRcd		4	5	10.7628		0.0294
6	HighChol		1	6	6.3914		0.0115

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	87.0	Somers' D	0.742
Percent Discordant	12.9	Gamma	0.742
Percent Tied	0.1	Tau-a	0.371
Pairs	67925	c	0.871

Variable Selection - Male

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	GenHlth		4	1	58.7125		<.0001
2	Age		1	2	32.1722		<.0001
3	bmiRcd		5	3	28.6238		<.0001
4	HighChol		1	4	10.8988		0.0010
5	DiffWalk		1	5	7.6666		0.0056

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.1	Somers' D	0.585
Percent Discordant	20.7	Gamma	0.586
Percent Tied	0.2	Tau-a	0.293
Pairs	57040	c	0.792

Four predictors appeared in both models: general health, age, high cholesterol, and BMI. High blood pressure and mental health were only additionally significant for women, and difficulty walking was only significant for men. Notably, the model performed better for predicting diabetes outcomes for women than for men.

Parameter Estimates - Female

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.2945	1.5006	23.6298	<.0001
HighBP	1	1	1.0228	0.2490	16.8682	<.0001
GenHlth	2	1	0.2081	0.4891	0.1810	0.6706
GenHlth	3	1	1.2005	0.4685	6.5655	0.0104
GenHlth	4	1	1.7047	0.5096	11.1916	0.0008
GenHlth	5	1	3.5230	0.6959	25.6283	<.0001
bmiRcd	2	1	2.0693	1.3704	2.2800	0.1311
bmiRcd	3	1	3.1400	1.3599	5.3310	0.0209
bmiRcd	4	1	3.7946	1.3632	7.7486	0.0054
bmiRcd	5	1	4.5209	1.4049	10.3559	0.0013
bmiRcd	6	1	4.3394	1.3961	9.6603	0.0019
Age		1	0.2139	0.0502	18.1396	<.0001
mhRcd	2	1	-0.3192	0.3689	0.7486	0.3869
mhRcd	3	1	-0.8333	0.7314	1.2982	0.2545
mhRcd	4	1	1.5340	0.8497	3.2592	0.0710
mhRcd	5	1	-1.2191	0.4679	6.7875	0.0092
HighChol	1	1	0.6181	0.2460	6.3111	0.0120

Parameter Estimates - Male

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.5532	1.4204	6.2578	0.0124
GenHlth	2	1	0.9523	0.4789	3.9536	0.0468
GenHlth	3	1	1.8084	0.4661	15.0556	0.0001
GenHlth	4	1	2.1620	0.5125	17.7968	<.0001
GenHlth	5	1	1.7273	0.5808	8.8452	0.0029
Age		1	0.2264	0.0448	25.5183	<.0001
bmiRcd	2	1	-1.4921	1.2809	1.3571	0.2440
bmiRcd	3	1	-0.5731	1.2585	0.2074	0.6488
bmiRcd	4	1	-0.1586	1.2617	0.0158	0.9000
bmiRcd	5	1	0.2656	1.2862	0.0426	0.8364
bmiRcd	6	1	0.4762	1.3250	0.1292	0.7193
HighChol	1	1	0.7273	0.2178	11.1529	0.0008
DiffWalk	1	1	0.8323	0.3043	7.4820	0.0062

In both models, age was significant at 0.0001 level or higher, and had coefficients of similar value in the expected direction (membership in a higher age group was associated with higher odds of a diabetes diagnosis). High cholesterol was also

statistically significant and in the expected direction for both models; however, the p-value was smaller and the odds ratio was higher for men compared with women ($e^{0.73} \approx 2.08$ and $e^{0.62} \approx 1.86$, respectively). General health in both models and BMI for only women performed consistently across levels, with levels further from the baseline generally having a more disparate odds ratio compared with the baseline (set as the lowest level) and also a stronger p-value. Worse self-reported general health (higher levels) is associated with a higher odds of having diabetes compared with the baseline. For women, a higher BMI is associated with a higher odds of having diabetes, however, this was not seen consistently with men. For men, only the highest BMI levels had an odds ratio greater than one, and although BMI was a significant predictor in the model, the parameter estimates were not statistically significant for any individual level.

Of the three variables that differed between the models, both high blood pressure (women) and difficulty walking (men) were significantly below the 0.01 level, with an affirmative value associated with increased odds of a diabetes diagnosis. The final variable, mental health (women), was significant in the model, however, the parameter estimates for the individual levels had coefficients that changed direction, and none were statistically significant. As discussed earlier in this analysis, different coding schemes can have a dramatic impact on these parameter estimates.

The results in this section are consistent with existing literature on sex differences in diabetes risk factors. Kautzky-Willer et al (2023) report that obesity was shown to be a higher risk factor for women than it is for men. Our measure of obesity was BMI, which showed a clear and statistically significant pattern of increasing odds of a diabetes diagnosis with increasing BMI levels. Additionally, the Kautzky-Willer article indicates high blood pressure as a risk factor for women, attributed at least in part to lower rates of adherence to prescribed medication regimens for cardiovascular trouble. For mental health considerations, a meta-analysis on diabetes-specific emotional distress found a correlation between the proportion of the study sample that was female and comorbid depressive symptoms, as well as diabetes distress symptoms. (Perrin et al 2017) The variable in our analysis, considering self-reported mental health, is more broadly defined, but it may have been partially capturing variation related to depression and distress.

6. Conclusions

Importantly, the final model includes predictors from four distinct conceptual categories: comorbidities (HighBP, HighChol), medical care (CholCheck), health metrics (bmiRcd, GenHlth, DiffWalk), and demographics (Age). This broad representation showcases that diabetes risk in this population is multifactorial, and is likely influenced not just by

underlying medical conditions, but also by physical health, demographic factors, and interactions with the healthcare system.

Despite our model's strong performance, there are still several limitations. A problem to note is that although bmiRcd was selected as a meaningful overall predictor, some of its subcategories had high p-values (0.97, 0.206, and 0.616), suggesting no statistically significant difference in diabetes risk for these groups (2, 3, 4) compared with the reference category. This is consistent with established clinical understanding that individuals who are underweight or at a healthy weight are generally at lower risk for diabetes. Given that we recorded BMI using clinical thresholds (see Table II in the Introduction), this lack of significance for lower-weight groups is not surprising.

One limitation of this analysis is the number of features available for each participant. Existing research discusses differences in diabetes outcomes based on cardiovascular risk and medication adherence for cardiovascular trouble (Kautzky-Willer et al 2023) as well as ethnicity (Xia et al 2025), among others. These factors have been shown to influence both diabetes prevalence and healthcare access, yet are not included in the current dataset. Other future models could also consider interaction effects in addition to the main effects considered in this model.

Section 4 addressed broader issues related to variable coding and the impact that even small changes can have on model results. There is an important ethical balance between developing a meaningful coding structure that is informed by the data and not using the coding structure to intentionally inflate the significance of results. Further analysis could look at existing literature on variable coding best practices and apply these strategies to model building.

Additionally, section 4 considered whether the variables that were significant for diabetes risk were the same for men and women. Consistent with existing literature, diabetes risk for women had a stronger relationship to BMI and was additionally associated with high cholesterol and mental health. Overly aggregated data can miss significant factors that are uniquely important for subsets of the population and, in the case of health studies, can lead to improved patient care and health outcomes. Further analysis can consider how other population subsets may have different influential factors for diabetes risk.

Bibliography

Alkhalaf, Arwa & Zumbo, Bruno. (2017). The Impact of Predictor Variable(s) with Skewed Cell Probabilities on Wald Tests in Binary Logistic Regression. Journal of Modern Applied Statistical Methods. 16. 40-80. 10.22237/jmasm/1509494640

CDC Codebook https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Cleveland Clinic. 2022. Body Mass Index (BMI)
https://my.clevelandclinic.org/health/articles/9464-body-mass-index-bmi

Cicek, Meryem, et al. “Characterizing Multimorbidity from Type 2 Diabetes: Insights from Clustering Approaches.” *Endocrinology and Metabolism Clinics of North America*, U.S. National Library of Medicine, Sept. 2021, pmc.ncbi.nlm.nih.gov/articles/PMC8383848/

Kautzky-Willer A, Leutner M, Harreiter J. “Sex differences in type 2 diabetes.” Diabetologia. 2023 Jun;66(6):986-1002. doi: 10.1007/s00125-023-05891-x. Epub 2023 Mar 10. Erratum in: Diabetologia. 2023 Jun;66(6):1165. doi: 10.1007/s00125-023-05913-8. PMID: 36897358; PMCID: PMC10163139. https://pmc.ncbi.nlm.nih.gov/articles/PMC10163139/

Perrin NE, Davies MJ, Robertson N, Snoek FJ, Khunti K. The prevalence of diabetes-specific emotional distress in people with Type 2 diabetes: a systematic review and meta-analysis. Diabet Med. 2017;34(11):1508–1520. doi: 10.1111/dme.13448
https://onlinelibrary.wiley.com/doi/10.1111/dme.13448

Teboul, Alex. 2022. Diabetes Health Indicators Dataset. Kaggle.
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Xia T, Nianogo RA, Yu Q, Horwich T, Srikanthan P, Inoue K, Allison M, Zhang Z, Watson KE, Chen L, “Racial Disparities of Type 2 Diabetes Through Exercise: The Multi-Ethnic Study of Atherosclerosis.” American Journal of Preventive Medicine, Volume 68, Issue 4, 2025, Pages 794-803, https://doi.org/10.1016/j.amepre.2025.01.009.
https://www.sciencedirect.com/science/article/pii/S074937972500008X

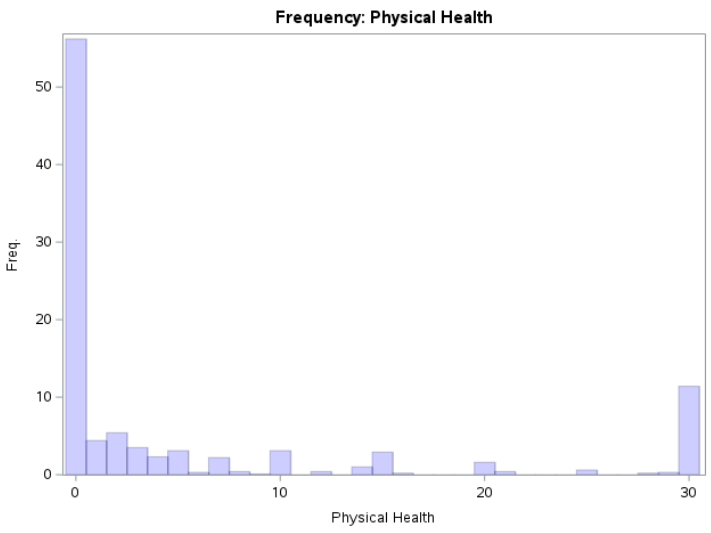
Appendices

Appendix 1: Recoded Variables

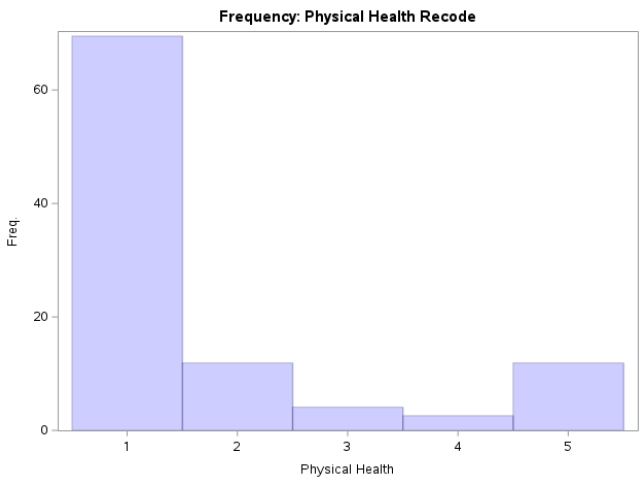
BMI, Physical Health, and Mental Health histograms before and after variable recoding.

Physical Health

Original

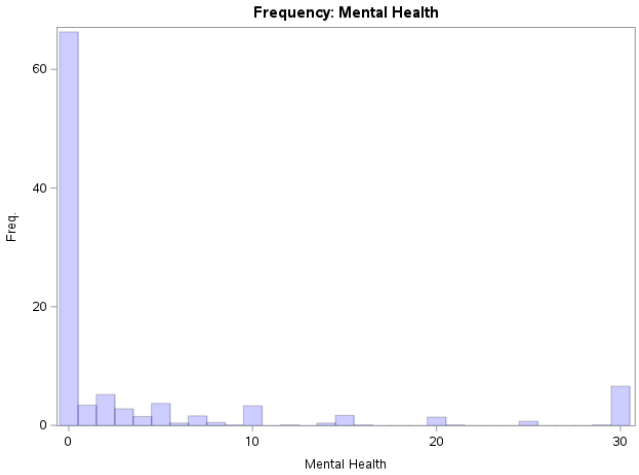


After Recode

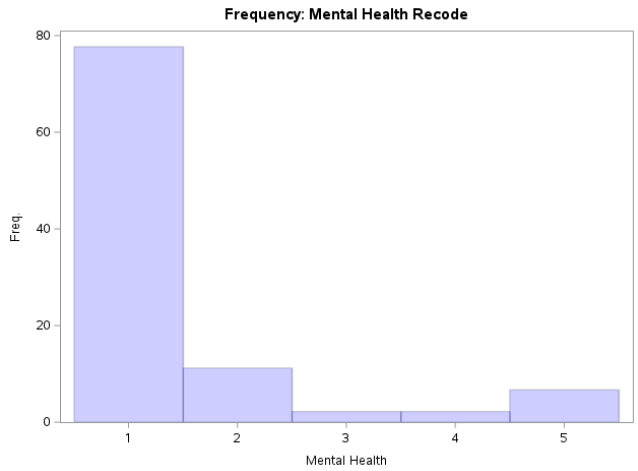


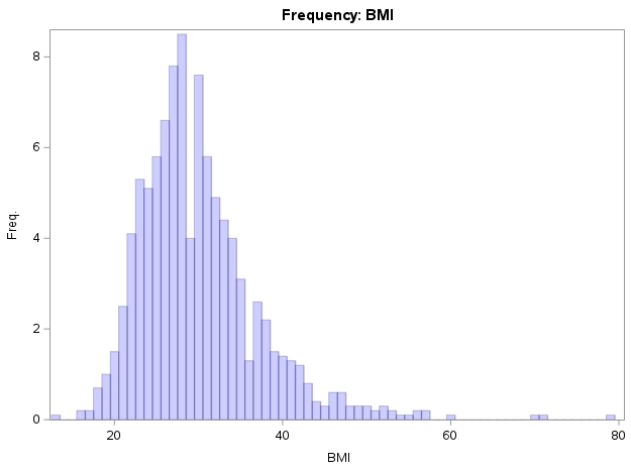
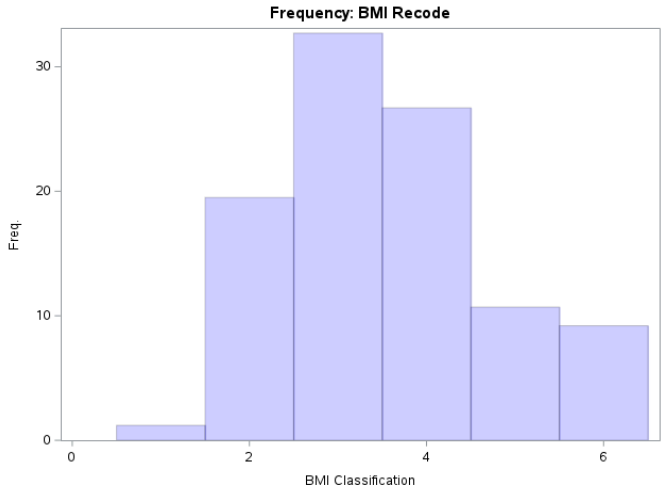
Mental Health

Original



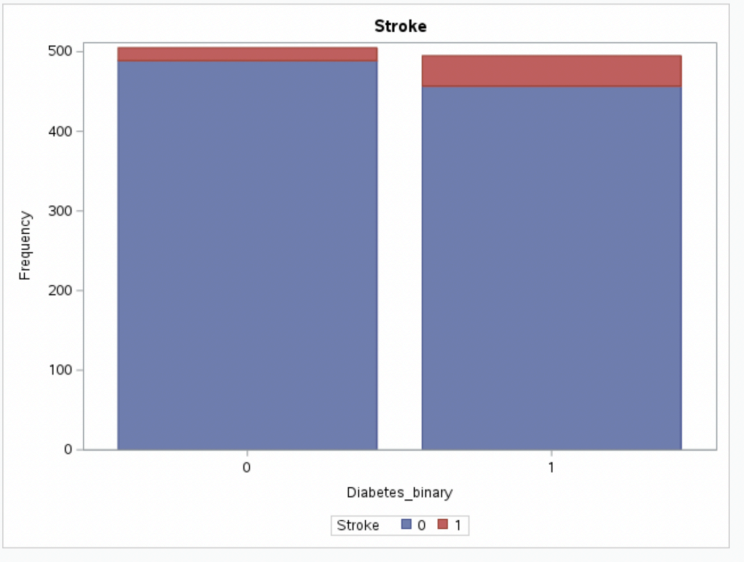
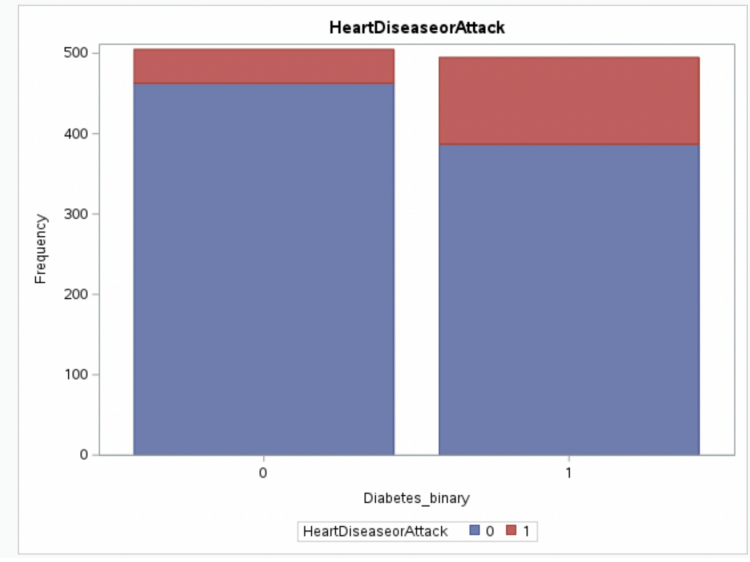
After Recode

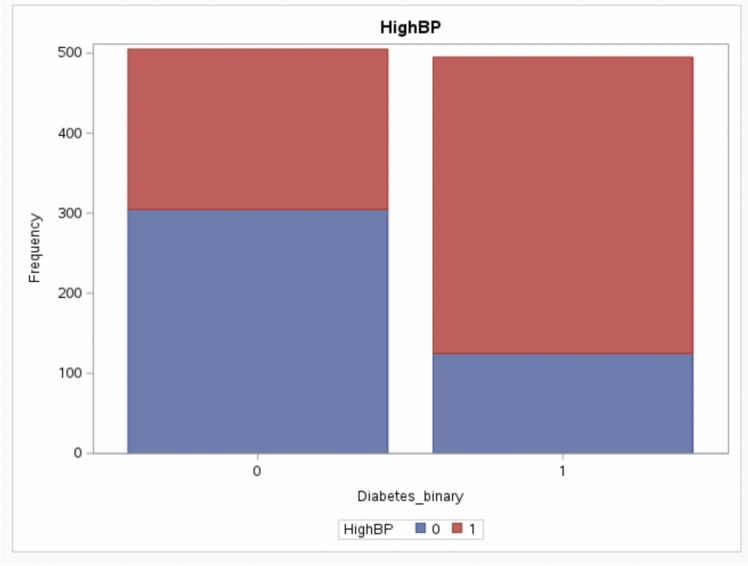


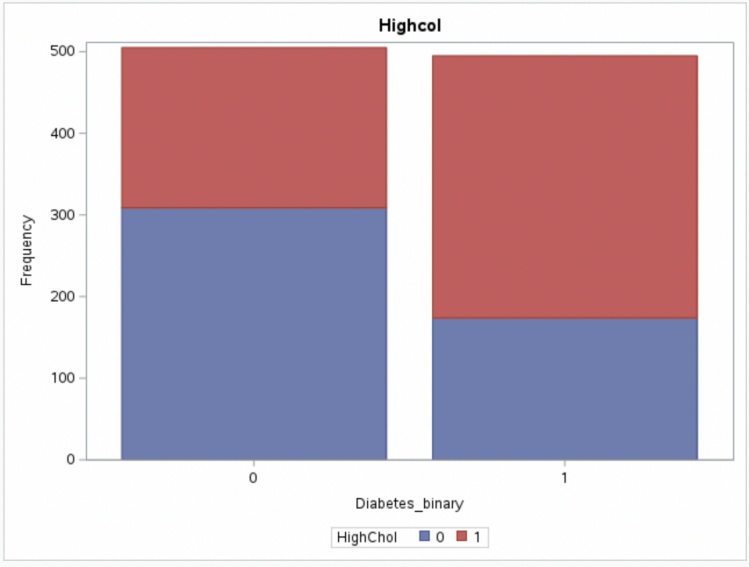
BMI	
Original	**After Recode**
	

Appendix 2: Associations with Diabetes Binary - Visualizations

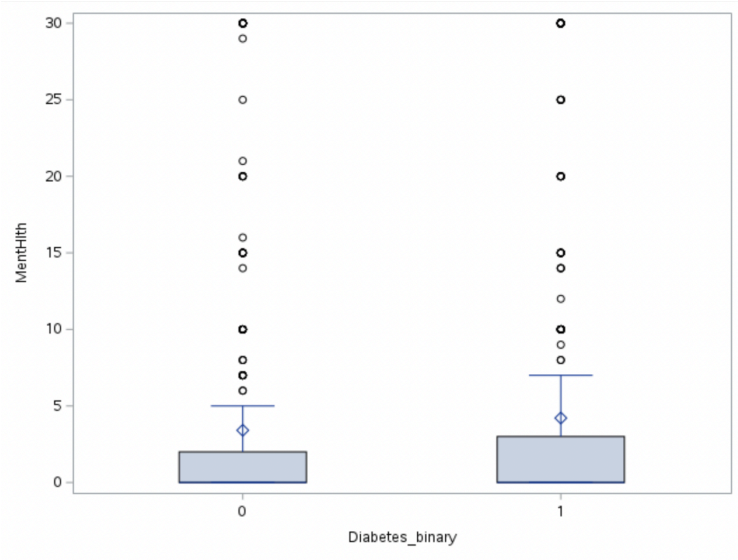
Stacked bar plots and boxplots that show the relationship between diabetes binary and the predictors from different groups (comorbidities, health metrics, medical care, lifestyle factors, and demographic factors).

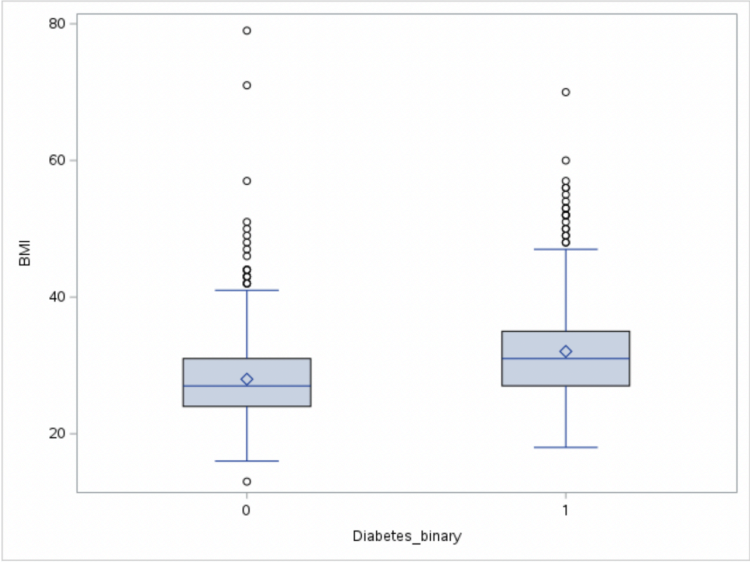
Comorbidities	
	

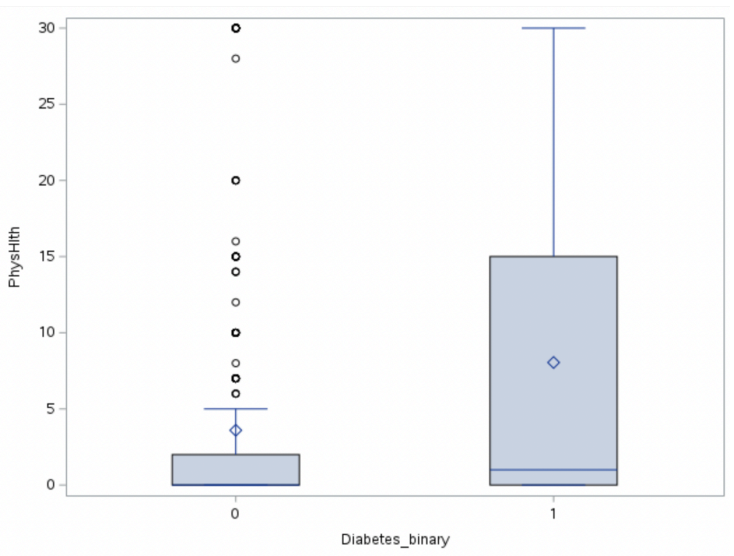


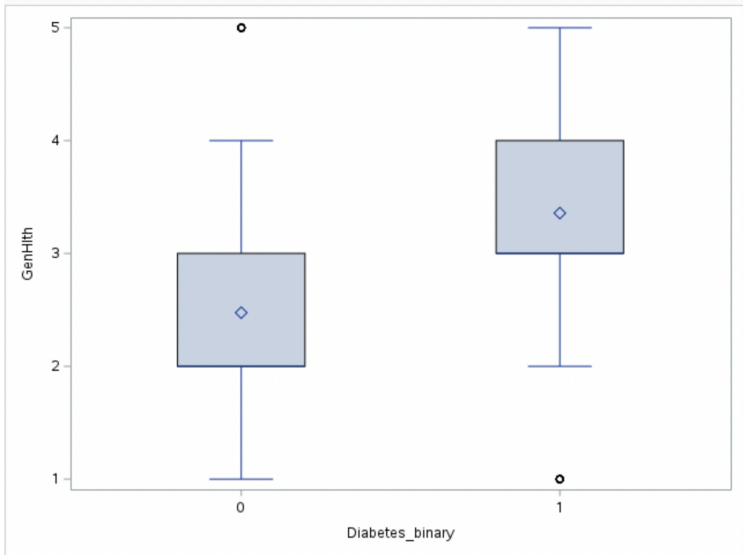


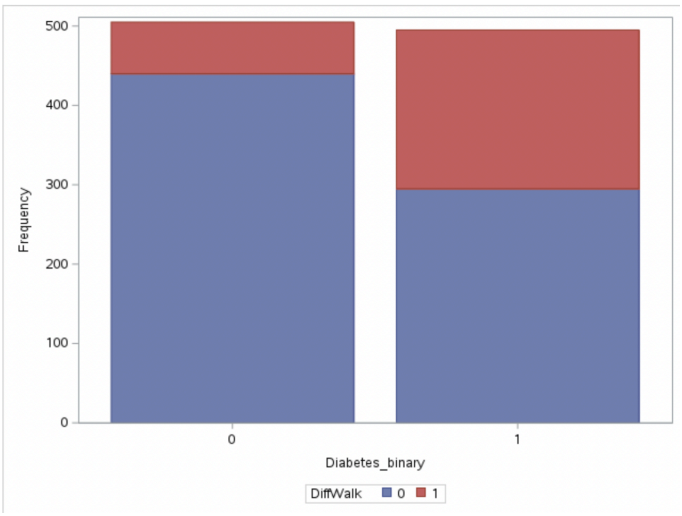
Health Metrics



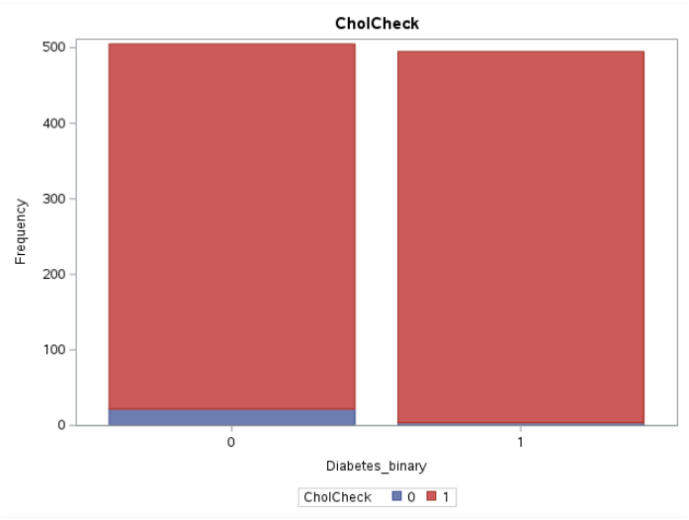


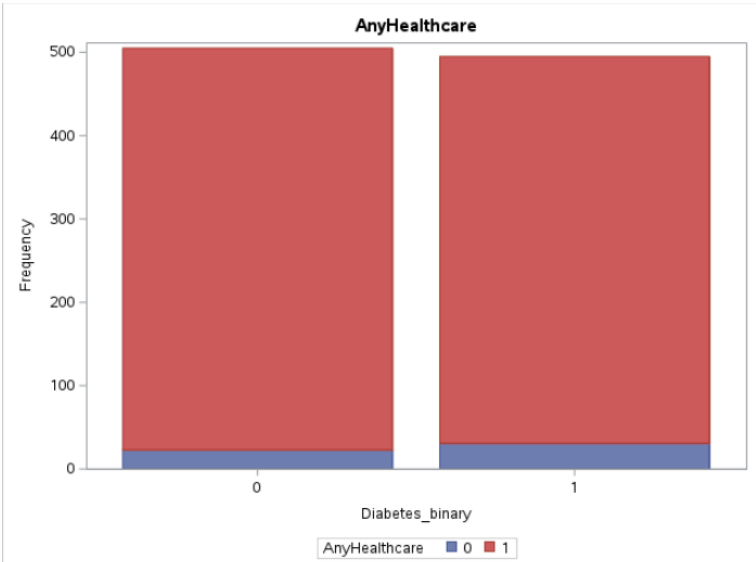

PhysHlth
30
25
20
15
10
5
0
0
1
Diabetes_binary

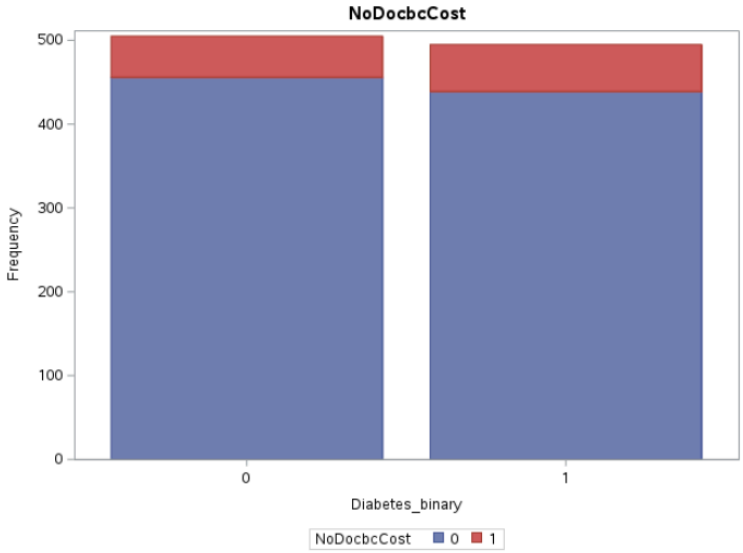

GenHlth
5
4
3
2
1
0
1
Diabetes_binary


Frequency
500
400
300
200
100
0
0
1
Diabetes_binary
DiffWalk 0 1

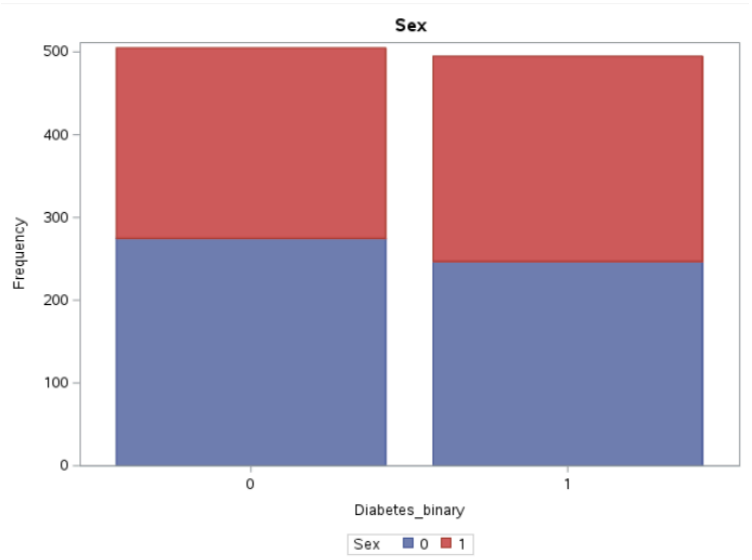
Medical Care - Treatment and Access

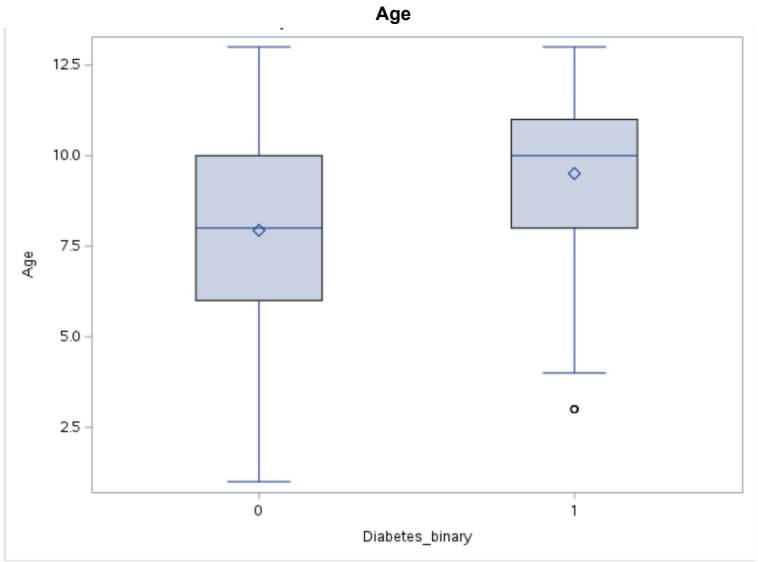


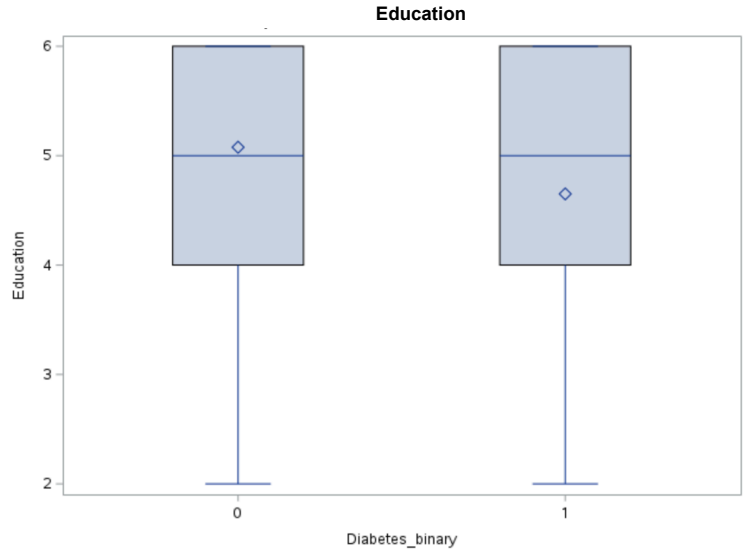


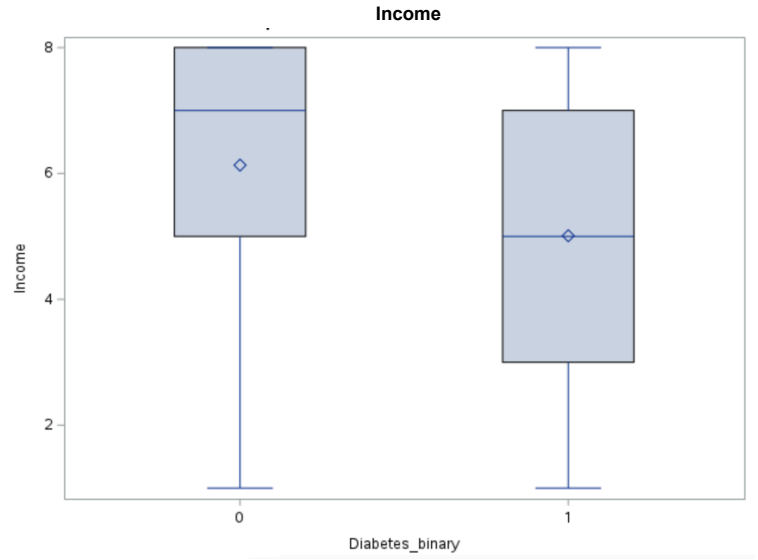


Demographic Factors

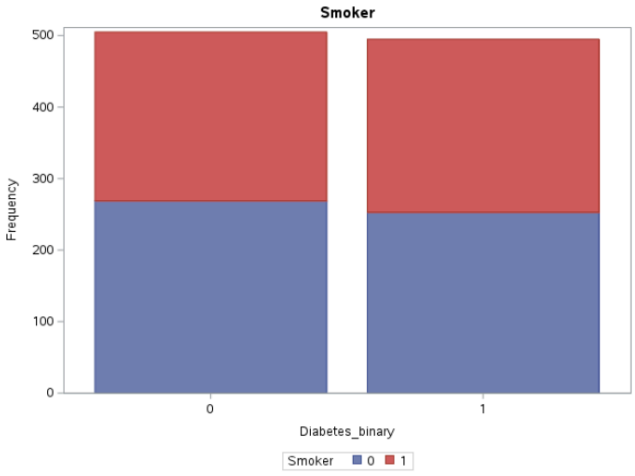


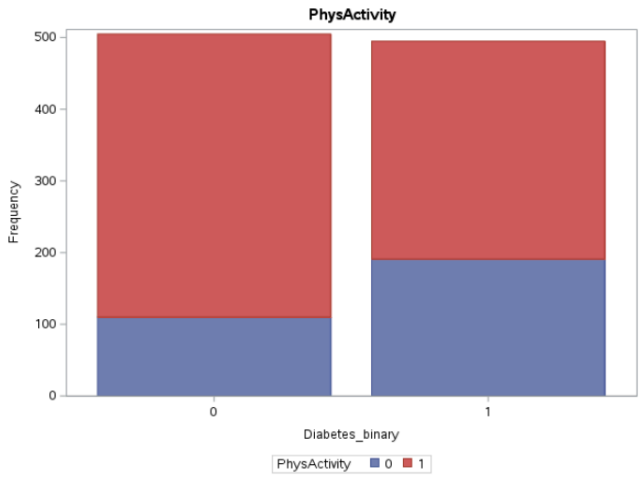


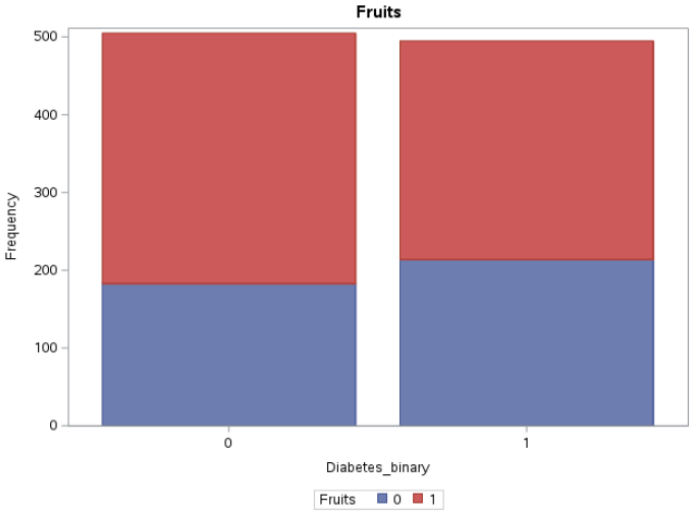


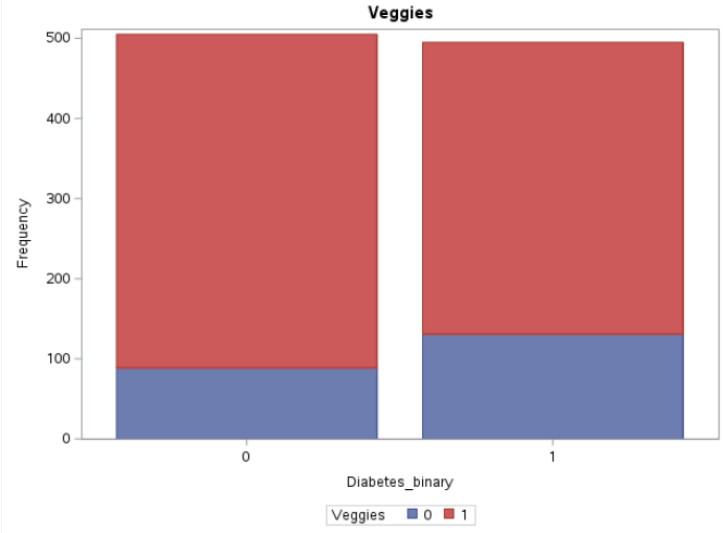


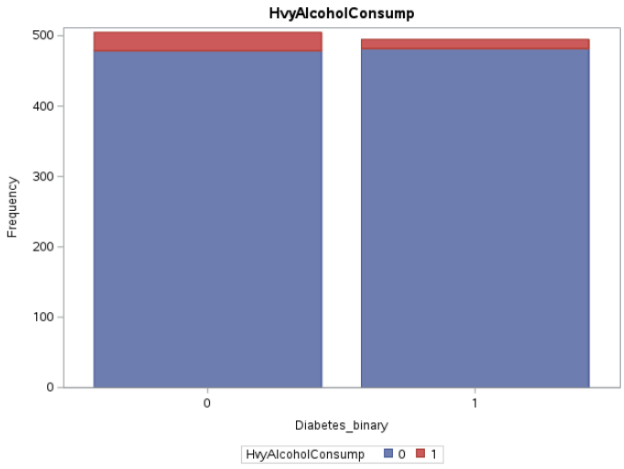
Lifestyle Factors











Appendix 3: Odds Ratio Estimates

Table with odds ratios and 95% confidence intervals for all variables separated by groups. Help determine the significance of variables and its relationship with diabetes.

Odds Ratio Estimates for Comorbidities			
Variables	Odds Ratio (Point Estimate)	95% CI – Lower	95% CI – Upper
HighBP	4.514	3.446	5.913
HighChol	2.908	2.249	3.761
HeartDiseaseorAttack	3.076	2.101	4.504
Stroke	2.54	1.397	4.618

Odds Ratio Estimates for Health Metrics			
Variables	Odds Ratio (Point Estimate)	95% CI – Lower	95% CI – Upper
BMI	1.74	1.55	1.954
GenHlth	2.202	1.925	2.519
MentHlth	1.083	0.968	1.212
PhysHlth	1.413	1.277	1.564
DiffWalk	4.589	3.345	6.297

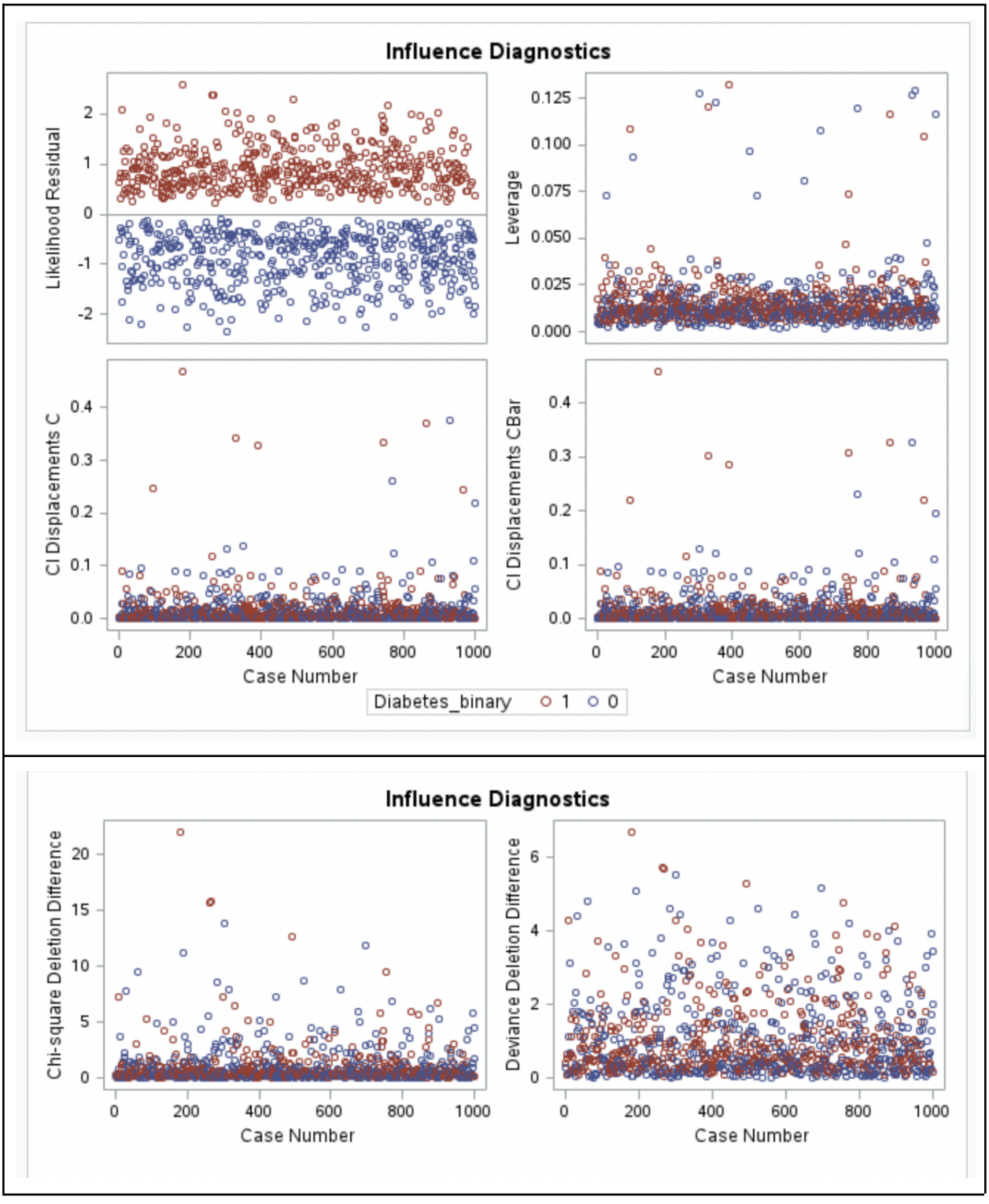
Odds Ratio Estimates for Medical Care			
Variables	Odds Ratio (Point Estimate)	95% CI – Lower	95% CI – Upper
CholCheck	5.589	1.912	16.338
AnyHealthcare	0.714	0.41	1.243
NoDocbcCost	1.187	0.791	1.78

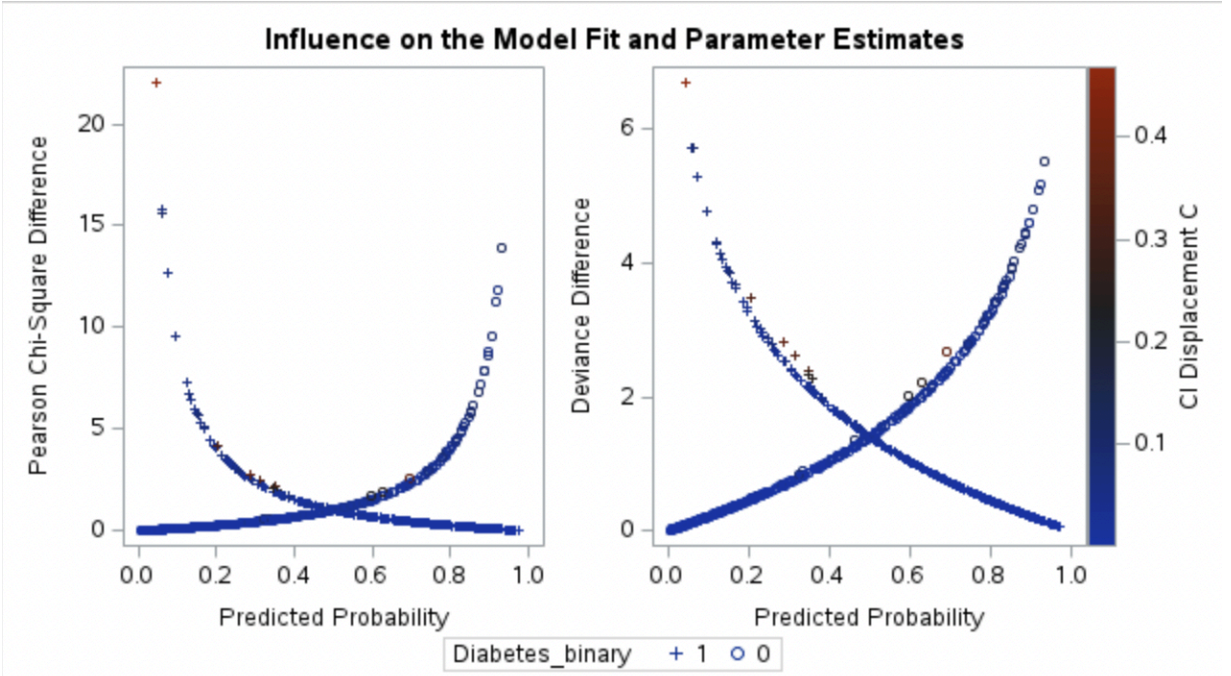
Odds Ratio Estimates for Lifestyle			
Variables	Odds Ratio (Point Estimate)	95% CI – Lower	95% CI – Upper
Smoker	1.09	0.851	1.397
PhysActivity	0.443	0.336	0.585
Fruits	0.746	0.579	0.962
Veggies	0.594	0.439	0.805
HvyAlcoholConsump	0.497	0.252	0.979

Odds Ratio Estimates for Demos			
Variables	Odds Ratio (Point Estimate)	95% CI – Lower	95% CI – Upper
Sex	1.2	0.936	1.539
Age	1.237	1.177	1.299
Education	0.665	0.587	0.754
Income	0.793	0.748	0.842

Appendix 4: Influence Diagnostics and Cbar (Descending Order)

Additional influence diagnostics, including likelihood residual, leverage, Pearson chi-square difference, and the first few rows for cbar values in descending order.

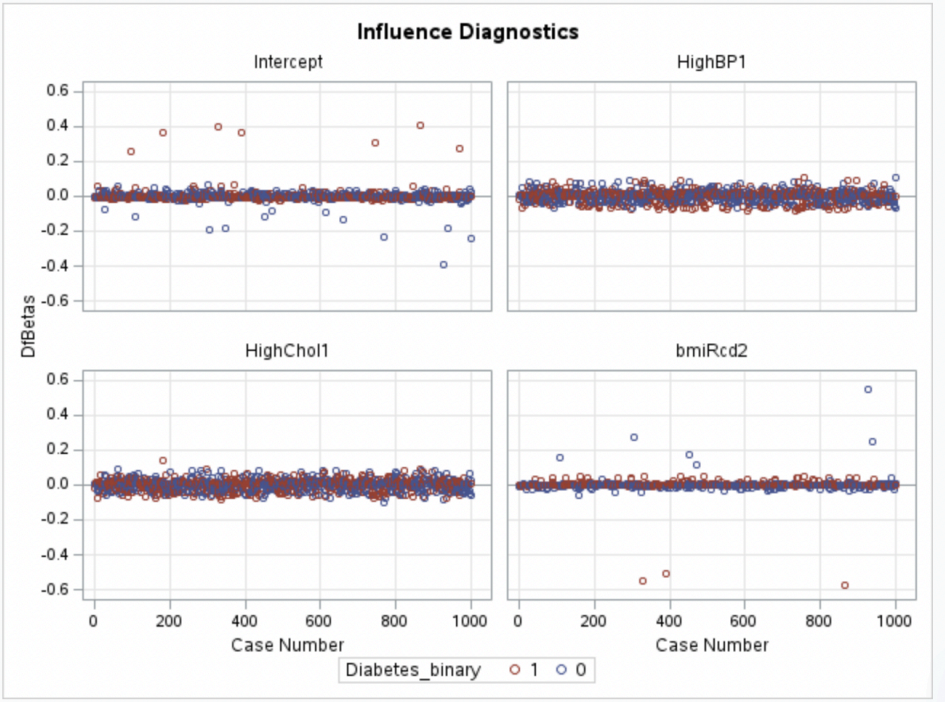


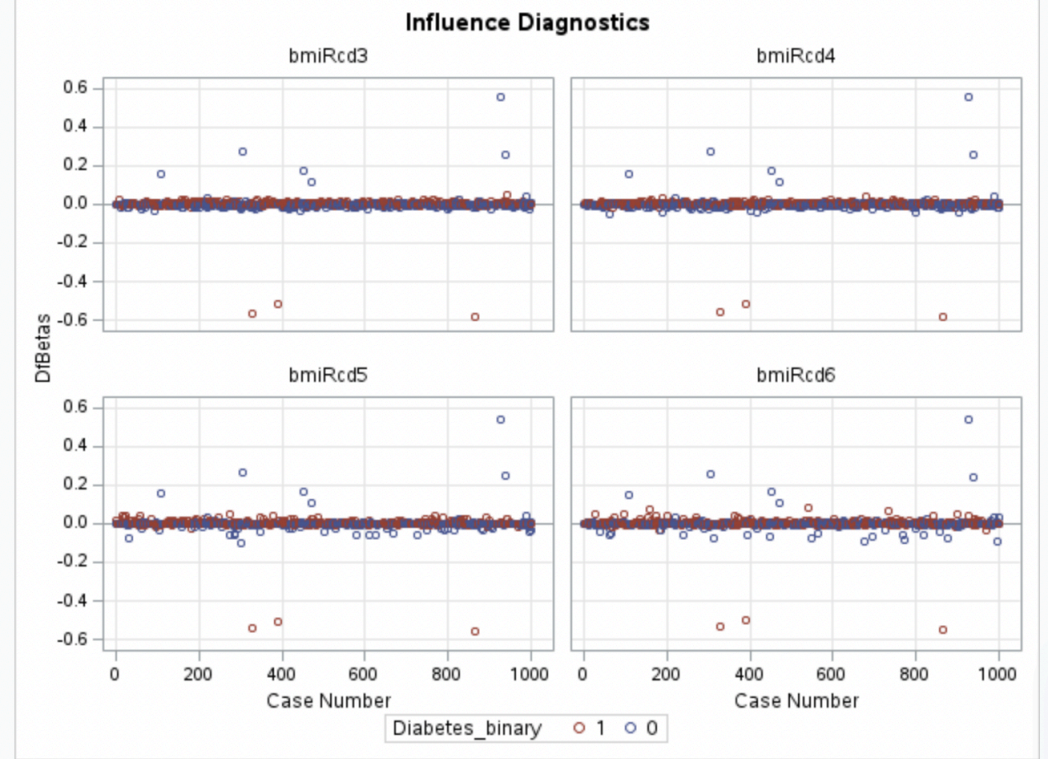


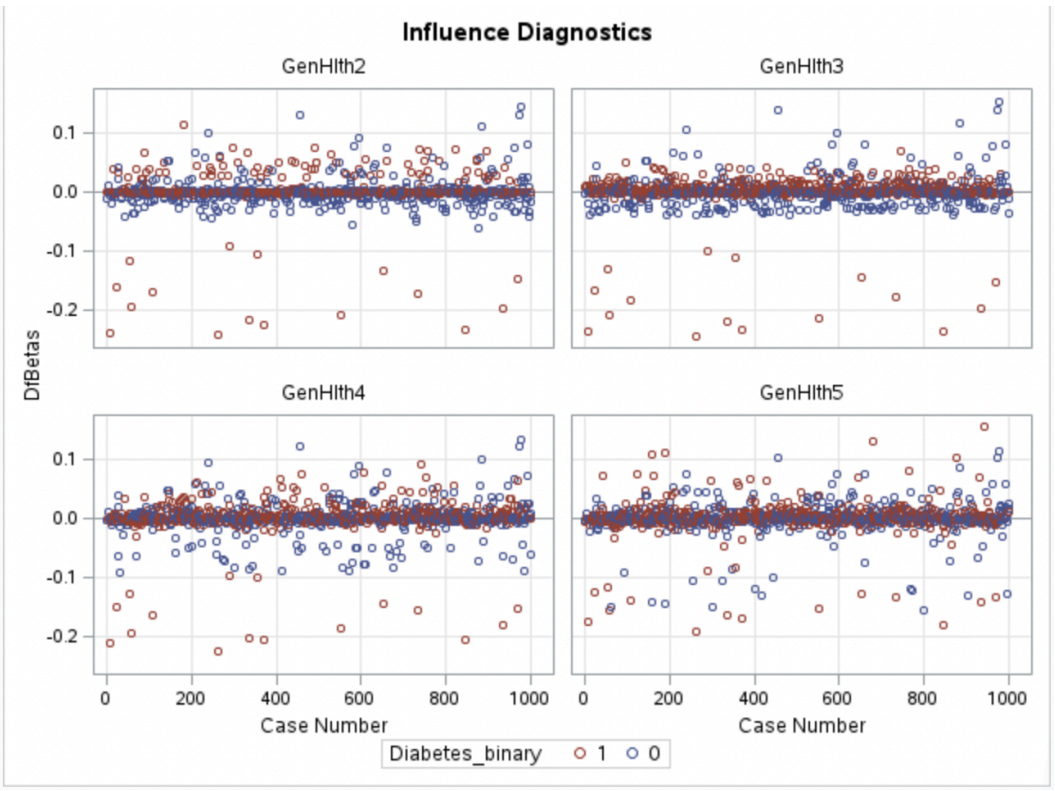
ObsID	_LEVEL_	predicted	leverage	cbar
180	1	0.04428	0.02082	0.45886
928	1	0.69268	0.12691	0.32762
865	1	0.28753	0.11647	0.32664
744	1	0.20393	0.07332	0.30889
330	1	0.31275	0.12063	0.30143
391	1	0.34851	0.13221	0.28480
767	1	0.62838	0.11985	0.23026
96	1	0.35609	0.10852	0.22013
967	1	0.34646	0.10419	0.21941
998	1	0.59709	0.11611	0.19467
302	1	0.93225	0.00936	0.12994
348	1	0.46387	0.12288	0.12121
772	1	0.87063	0.01763	0.12079
263	1	0.06051	0.00742	0.11605
997	1	0.85097	0.01870	0.10883
878	1	0.85813	0.01685	0.10365

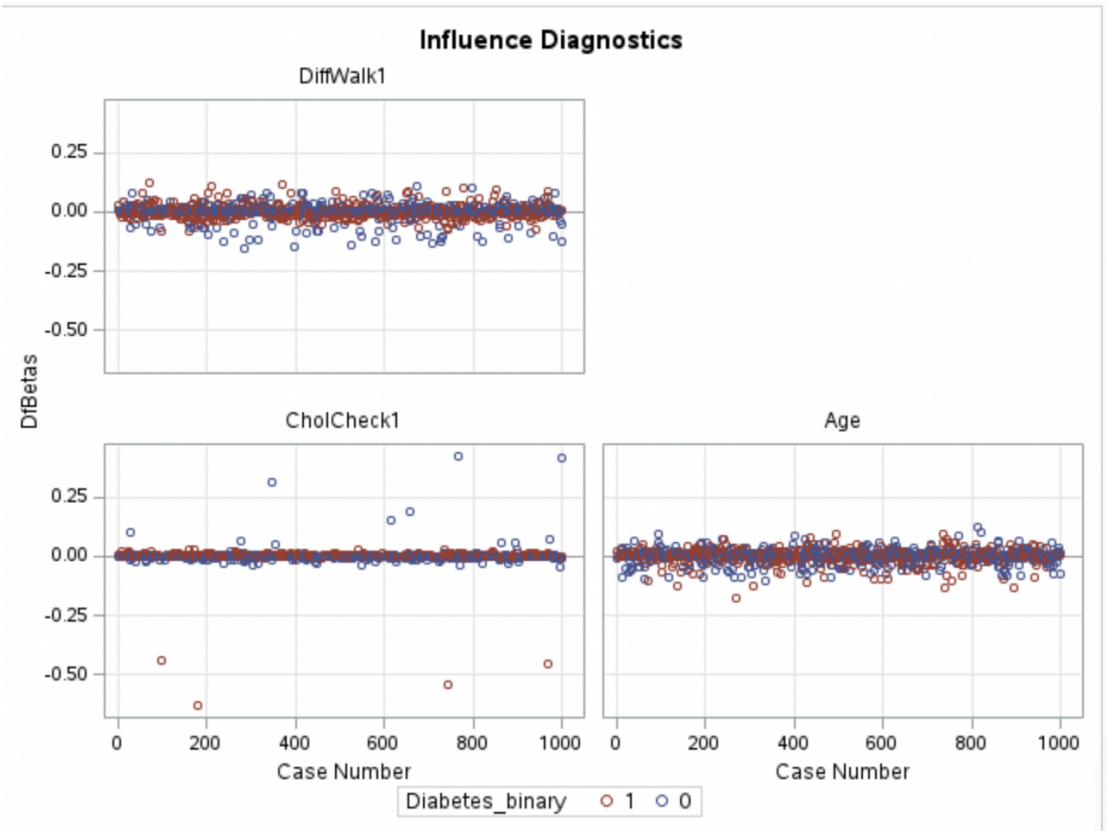
Appendix 5: DfBetas plots for each predictor in the Final Model

Dfbeta plots for all 7 variables included in the final model selected using the stepwise method. They help determine whether there are highly influential points.


Influence Diagnostics
Intercept
HighBP1
HighChol1
bmiRcd2
DfBetas
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
0
200
400
600
800
1000
Case Number
Diabetes_binary 1 0


Influence Diagnostics
bmiRcd3
bmiRcd4
bmiRcd5
bmiRcd6
DfBetas
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
0
200
400
600
800
1000
Case Number
Diabetes_binary 1 0


Influence Diagnostics
GenHlth2
GenHlth3
GenHlth4
GenHlth5
DfBetas
0.1
0.0
-0.1
-0.2
0
200
400
600
800
1000
Case Number
Diabetes_binary 1 0


Influence Diagnostics
DiffWalk1
CholCheck1
Age
DfBetas
0.25
0.00
-0.25
-0.50
0
200
400
600
800
1000
Case Number
Diabetes_binary 1 0

Appendix 6: Hosmer and Lemeshow Goodness-of-Fit Test

Hosmer and Lemeshow test shows strong agreement between observed and expected values. The p-value > 0.05 indicate that the model is a good fit.

Partition for the Hosmer and Lemeshow Test					
		Diabetes_binary = 1		Diabetes_binary = 0	
Group	Total	Observed	Expected	Observed	Expected
1	100	4	4.44	96	95.56
2	100	9	12.56	91	87.44
3	101	24	22.63	77	78.37
4	100	36	35.59	64	64.41
5	100	50	46.73	50	53.27
6	101	58	58.19	43	42.81
7	100	70	67.11	30	32.89
8	102	78	77.39	24	24.61
9	101	81	83.72	20	17.28
10	95	85	86.64	10	8.36

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
3.0114	8	0.9336

Appendix 7: Calibration for Diabetes_binary using Final Model

Calibration plot illustrating strong alignment between predicted probabilities and actual outcomes across the range of predictions.