5-2020

Characterizing Statin use among Prediabetic Patients with Predictive Analytics

Alexandra Gentile

Characterizing Statin Use among Prediabetic Patients with Predictive Analytics

A thesis submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Science in Industrial Engineering with Honors

Alexandra Gentile

University of Arkansas

Department of Industrial Engineering

Thesis Advisor: Shengfan Zhang, Ph.D.

Thesis Reader: Kelly Sullivan, Ph.D.

Acknowledgements

I would like to use this opportunity to thank those who have had a major influence over my college experience. First, I would like to thank the Department of Industrial Engineering for all they do for their students. Our department does such a great job of making students feel appreciated and setting them up for success. I definitely would not be where I am today without the support of our incredible faculty, staff, and students.

I would also like to thank the Honors College for supporting my research the past several years. The Honors College Research Grant I received helped support this research and support me while I was conducting this research. The Honors College Travel Grant I received allowed me to travel to the Institute for Operations Research and the Management Sciences (INFORMS) Annual Meeting 2019 in Seattle, Washington to present my work. I had a fantastic experience at my first major conference in Seattle, and it would not have been possible without support from the Honors College.

I would also like to thank Dr. Sullivan for all of his support. Outside of having him as a professor for five different courses, he also gave me the opportunity to gain some experience teaching in operations research under his guidance. Through the honors research program, he had a major role in encouraging me to start my research, apply for grants, and write my thesis. I am so thankful for his guidance and support through the years.

I would like to thank Dr. White for always pushing me to be better. He has challenged me more than any other professor, but I have always come out stronger on the other side. He has taught me many important life lessons and helped me realize how much I am truly cable of. I am very grateful for his investment in me the past several years.

Finally, I would like to thank Dr. Zhang for her investment in me and this research. She has always encouraged me and helped me believe I can accomplish whatever I set my mind to. I could not have accomplished this without her, and I will forever be thankful for her constant support.

Abstract

Diabetes is one of the leading causes of death in the United States and can cause severe impairments to those diagnosed. Prediabetes is a state when a patient has higher fasting plasma glucose levels than a non-diabetic person but is not quite high enough to be considered diabetes. Both diabetic and prediabetic patients are at higher risk for cardiovascular diseases (CVD), which is the leading cause of death in the United States. The primary form for prevention and treatment of CVD is through statin therapy. Statins are a class of medications used to treat and prevent CVD by limiting cholesterol production in the liver and stabilizing plaque in arteries. However, substantial research has found an association between statin use and the development of Type 2 diabetes. This is an important association to investigate because both statin use and diabetes are prevalent in the United States.

The association between statin use and the development of Type 2 diabetes poses a complicated risk for prediabetic patients. Because they are already at high risk for diabetes, taking a statin could further increase this risk. However, preventing CVD, which they are also at risk for, is critical as well. This research investigates the relationship between statin use and prediabetic subjects specifically.

An adult, prediabetic subpopulation was obtained from the National Health and Nutrition Examination Survey (NHANES), which is made publicly available through the Center for Disease Control and Prevention. Several random forest classifiers were built using this subpopulation to predict statin use among prediabetic patients. Analysis of the models found age, cholesterol levels, blood pressure levels, waist size, body mass index (BMI), and annual household income to be the best predictors of statin use in prediabetic subjects. Access to health insurance, gender, family history of heart attacks, and overall health rating were found to be the least impactful predictors of statin use among prediabetic subjects in all models. It appears the risk of CVD outweighs the risk of developing Type 2 diabetes, and doctors are continuing to prescribe statins for prediabetic patients despite the increased risk of developing diabetes.

Contents

1. Introduction

1.1 Background

Diabetes is the seventh-leading cause of death in the United States and affects over 30 million Americans [1]. Type 1 diabetes occurs when the body does not make enough insulin and accounts for 5% of cases of diabetes [2]. This form of diabetes cannot be prevented. Type 2 diabetes is much more common and is when the body cannot use insulin properly. Insulin is a hormone made by the pancreas used to regulate blood sugar levels, which can lead to health complications if not properly controlled [3]. People who have diabetes can suffer from complications such as blindness, kidney failure, heart disease, stroke, and loss of extremities [4].

Prediabetes is a state of latent impairment of carbohydrate metabolism in which the criteria for diabetes are not all satisfied [5]. According to the American Diabetes Association (ADA), there are approximately 3 million newly diagnosed prediabetes cases each year, and approximately 38% of adults in the U.S. have prediabetes [6] . Most patients with prediabetes have impaired fasting glucose, impaired glucose tolerance, metabolic syndrome, and high cholesterol levels. These conditions, along with lifestyle choices, cause prediabetic patients to have a high risk of cardiovascular disease, similar to patients with type 2 diabetes [7]. It is recommended pharmacological treatment be started to prevent progression into type 2 diabetes.

Prediabetic and diabetic patients are at high risk for cardiovascular disease (CVD). CVD, which is disease of the heart including diseased vessels, structural problems, and blood clots, is the leading cause of death in the United States [8]. The primary form for prevention and treatment of CVD is through statin therapy. Statins, one of the most widely prescribed types of medication in the United States, are a class of medications used to prevent CVD by limiting cholesterol production in the liver and stabilizing plaque in arteries [9]. Figure 1 shows the progression of cholesterol buildup progressively worsening from left to right. The image on the right of the figure shows a blood clot, which can be deadly. Statins can save lives by preventing this buildup.

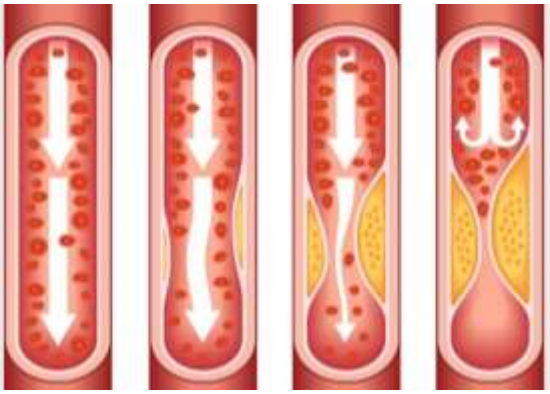

Figure 1. Healthy blood flow (left) progressing to a blood clot (right) as cholesterol builds up in arteries.

Unfortunately, substantial research has linked statin use with the development of Type 2 diabetes [10]. Figure 2 shows a clear difference between statin users (red) and non-statin users (blue) in the percent of Type 2 diabetes instances in the respective populations. The study concluded being a statin user was significantly associated with an increased risk of new onset diabetes [11]. Many other research studies have reached the same conclusion. This is an important association to investigate because both diabetes and statin use are so prevalent in the United States.

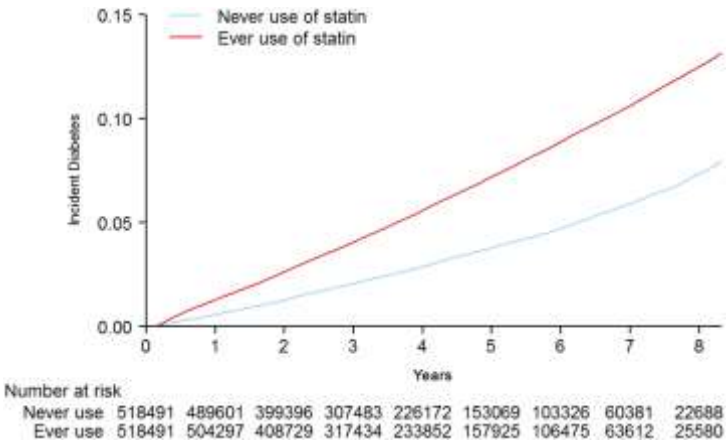

Figure 2. Comparison of percent of incident diabetes between statin users and non-users. [11]

The association between statin use and the development of Type 2 diabetes poses a complicated risk for prediabetic patients. Because they are already at high risk for diabetes, taking a statin could further increase this risk. However, preventing CVD, which they are also at risk for, is critical as well. There are some cases where taking a statin might be worth it, and others where it may not.

1.2 Literature Review

For years researchers have been investigating the association between statin use and new-onset diabetes mellitus (NODM), which is when a subject with no previous history of diabetes develops and is diagnosed with diabetes. A robust 2010 meta-analysis of over 90,000 subjects found statin therapy is correlated with NODM, especially in older subjects [12]. This analysis also found the longer a subject had been on statins, the more likely he or she was to develop NODM [12]. This study did not look at a prediabetic population specifically.

A 2012 study examined statin risk along with cardiovascular outcomes in the general population. The study found rates of diabetes were significantly higher in statin users ($p < .001$) over a median of 7.2 years [13]. However, major adverse cardiovascular events (MACE) and in-hospital mortality related to cardiovascular events were less [13]. Through risk-benefit analysis, it was determined that statin treatment was favorable in both high-risk and secondary treatment subjects in the general population [13]. Secondary treatment is defined as "screening to identify diseases in the earliest stages, before the onset of signs and symptoms, through measures such as mammography and regular blood pressure testing" [14].

A post-hoc analysis of a clinical trial found subjects with one or more risk factors for diabetes were at higher risk for NODM when subject to statin therapy than those with no major risk factors [15]. However, in individuals with one or more risk factors for diabetes, more adverse cardiovascular events and deaths were avoided than new-onset cases of diabetes [15]. This suggests prediabetic subjects may be at even higher risk for developing NODM than non-diabetic subjects. However, this risk may be worth it considering the reduction in adverse cardiovascular events.

In 2012, the U.S. Food and Drug Administration updated their guidelines to add a safety warning to statins indicating an increased risk for development of diabetes [16]. This change supports much of the research done in this area; however, little research has been done to analyze prediabetic subjects specifically. One study that did examine prediabetic subjects specifically found statin therapy is associated with an increased risk in development of NODM [17]. Despite the risk of NODM, prediabetic subjects would benefit from taking statins due to decreased risk of MACE and morbidity associated with cardiovascular events [17].

While many studies have addressed the link between statin use and the onset of Type 2 diabetes, research needs to be done specifically for high-risk patients like those with prediabetes. Little previous research analyzes how the medical community is currently handling this dilemma. While the long-term

research goal is to develop a guideline on statin initiation for prediabetes patients, this research analyzes the current practices with respect to statin use among prediabetes patients. We predict statin use among a prediabetic population and analyze the resulting model to determine how this dilemma is currently being handled.

The remainder of this thesis is organized as follows: Section 2 details the methodologies used to reach our conclusions, including data collection and processing, data validation, and model development. Section 3 states the results of the model, and Section 4 explains and draws conclusions from these results.

2. Methodologies

This section summarizes research methodologies. Subsection 3.1 gives an overview of the National Health and Nutrition Examination Survey (NHANES) data set, publicly available through the Center for Disease Control and Prevention (CDC), and how specific sections of the data were selected and processed for this research. Subsection 3.2 describes the extensive process of selecting and validating a subpopulation of prediabetic patients. Subsection 3.3 details the algorithms and methods used to build and train the model. Subsection 3.4 describes the methods used for model testing and validation.

2.1 Data Collection and Processing

NHANES is a survey of noninstitutionalized civilian residents of the United States population in their homes with a laboratory component taking place in mobile examination centers. The questionnaire portion of the survey ranges from weight history to smoking habits while the laboratory examination tests for diseases, pregnancy, and, most importantly for this research, levels of different measures in the body [18]. These measures include blood sugar levels, cholesterol levels, and insulin levels [18]. Participants are able to opt out of the laboratory examination if they choose to do so, and, for this reason, there are fewer responses for the laboratory examination portion. Respondent personal information, such as name and address, are not disclosed, and respondents are instead identified by a 5-digit sequence number. This sequence number allows us to link participants' responses from different sections of the survey and from the laboratory examination. We have chosen to use the 2015-2016 survey data because it was the most recent published survey at the time this research started.

Because NHANES does not classify or categorize respondents in any way, we had to classify prediabetic patients. There are various indicators commonly used by doctors to diagnose prediabetes and diabetes including fasting plasma glucose (FPG) levels, the oral glucose tolerance test (OGTT), and

hemoglobin A1C numbers. In the NHANES laboratory examination, FPG and OGTT were both taken. Because healthcare professionals most commonly use FPG to diagnose prediabetes and diabetes, we chose this test as our basis to classify respondents as “prediabetic”, “diabetic”, or “non-diabetic” [19]. Table 1, published by the American Diabetes Association (ADA), defines the criteria for prediabetes and diabetes based on FPG.

Result	Fasting Plasma Glucose (FPG)
Normal	less than 100 mg/dl
Prediabetes	100 mg/dl to 125 mg/dl
Diabetes	126 mg/dl or higher

Table 1. Summary table of prediabetes and diabetes classification standards [20].

The FPG test was completed by 2,972 respondents. Of these respondents, 1,227 were classified as prediabetic. In the diabetes section of the survey, there is a question asking if a respondent has ever been told they have diabetes by a doctor. Of those who responded “yes” to this question, 91 were a part of the initial prediabetic subpopulation. Because these respondents had previously been diagnosed with diabetes and were most likely treated for it, they were taken out of the subpopulation. Their treatment for diabetes could be the reason they are now prediabetic by FPG level. Removing these respondents brought the sample size down to 1,136. Finally, respondents under the age of 18 were removed. Pediatric and adult diabetes are treated differently by the medical community and should be analyzed separately. For this reason, we chose to focus on adult population. This subpopulation of 1,029 was used for subsequent analysis. The process for arriving at the subpopulation is illustrated in Figure 3.

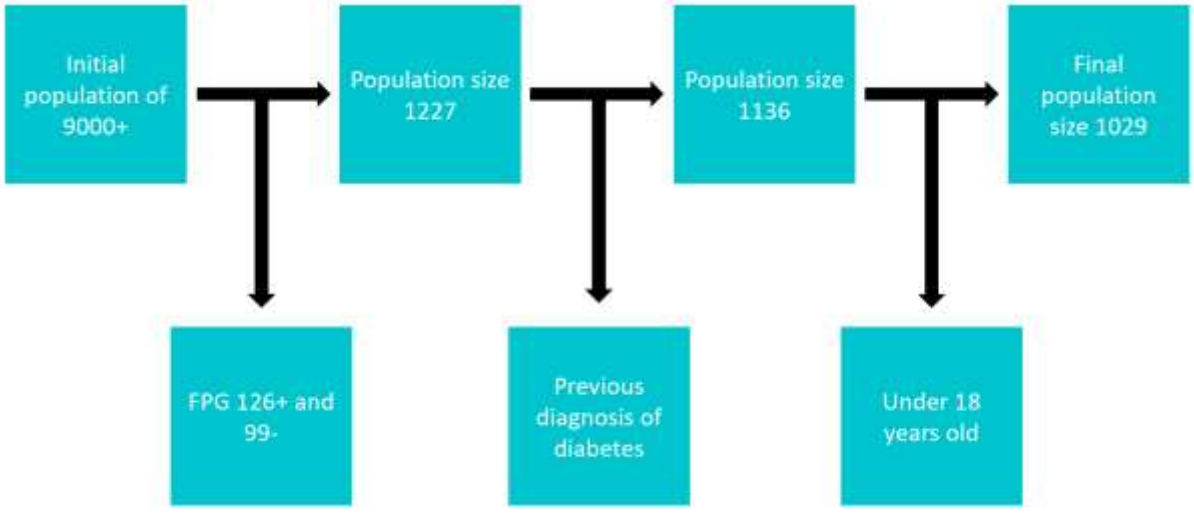

Figure 3. Process for arriving at final subpopulation for analysis.

There is a question on the NHANES survey asking respondents if they have ever been told by a doctor they have prediabetes. We chose not to use the answers to this question to define our subpopulation because of the high number of missing responses and because patients may be considered prediabetic even if they have not been diagnosed by a doctor. We did, however, use the responses to this question for data validation, which will be discussed further in Section 3.2. The demographic distributions of this final population are shown in Table 2.

Table 2. Distributions of demographics of final population.

Demographics Distributions										
	Male	**Female**								
Gender	55.00%	45.00%								
	18-29	**30-44**	**45-59**	**60-74**	**75+**					
Age	13.90%	24.39%	25.85%	25.36%	10.50%					
	Mexican American		**Other Hispanic**		**Non-Hispanic White**		**Non-Hispanic Black**		**Non-Hispanic Asian**	**Other**
Race/Ethnicity	18.76%		13.31%		36.93%		15.26%		12.05%	3.69%
	1	**2**	**3**	**4**	**5**	**6**	**7+**			
Household Size	14.38%	28.38%	17.59%	15.84%	10.59%	6.71%	6.51%			
	<20k	**(20k,45k]**	**(45k,65k]**	**(65k,75k]**	**(75k,100k]**	**>100k**				
Annual Income	23.05%	28.07%	16.22%	5.55%	10.99%	16.12%				
	Less than 9th grade		**9th-11th grade**		**High School grad/GED**		**Some college or AA**		**College grad or above**	
Education	13.08%		12.07%		22.33%		28.37%		24.14%	

The response variable for the model is whether a prediabetic respondent is prescribed statins. This response was determined using the “Prescription Medications” survey from the questionnaire section. A question in this survey asked respondents to list all prescription medications they were taking at the time of response. If one of the medications a respondent listed was a statin, they were given a “1” as their response value. If they did not list a statin, they were given a “0” as their response value. There was a question in the “Blood Pressure and Cholesterol” section asking if a respondent is prescribed medicine for cholesterol, which is often a statin. The answer to this question was not used as the response due to (1) a higher rate of missing values and (2) the “Prescription Medication” survey would have been used to verify this cholesterol medication was a statin. The response to this question was instead used for data validation, which will be discussed further in Section 3.2. 100% of the prediabetic subpopulation responded to the prescription medication survey, so there was no missing response data.

Eighteen predictors were chosen from various surveys in the questionnaire. Table 3 outlines which survey each predictor is from and whether the predictor is demographic, health-related, or behavioral. These categories will later be used to determine which types of factors are best at predicting statin use. Predictors were chosen based on their potential to have a relationship with diabetes or cholesterol levels. Age, gender, race, education, smoking, and family history of heart problems were used to predict statin use in a study examining adults in the United States [21]. Predictors were also chosen based on the quality of the data. Many survey questions had more than 25% of responses missing and were therefore eliminated from consideration.

Table 3. Categorization of predictors chosen from NHANES.

Survey	Predictor	Type
Blood Pressure & Cholesterol	Has a doctor ever told you that you have high blood pressure?	Health-related
	Has a doctor ever told you that you have high cholesterol?	Health-related
Body Measures	BMI	Health-related
	Waist size (cm)	Health-related
Consumer Behavior	Money spent on carryout food (past 30 days)	Behavioral
Demographics	Gender	Demographic
	Age (in years)	Demographic
	Household size	Demographic
	Annual family income	Demographic
	Education level	Demographic
	Race	Demographic
Diet and Nutrition	Diet health rating (self-rating)	Behavioral
	Number of fast food meals (past 30 days)	Behavioral
	Number of ready-to-eat foods (past 30 days)	Behavioral
General Health	Health rating (self-rating)	Health-related
Health Insurance	Covered by Insurance	Demographic
Medical Conditions	Close relative had a heart attack	Health-related
Smoking – Cigarette Use	Smoke at least 100 cigarettes in life	Health-related

2.2 Data Validation

To determine which analysis techniques were the most appropriate for the data set, further investigations were performed. A multicollinearity study was completed because many analysis techniques assume independence, and if multicollinearity was present, some options would not be appropriate. Each predictor was compared to the other 17 predictors to determine if there was a significant relationship between the variables. Due to the categorical nature of the data, the Pearson correlation coefficient could not be used to detect relationships between variables. Instead, a Chi-square test of independence was completed for each pair of predictors. A summary table of the results can be found in the appendix. Based on a level of significance of .05, about 57% of the predictor pairs had a significant relationship. In conclusion, the dataset exhibited strong multicollinearity. This makes analyses that assume independence of predictor variables, such as regression analysis, inappropriate for our model.

Most questions in the NHANES questionnaire were not completed by 100% of respondents. For this reason, we had to handle missing data. The question used for the response variable, asking respondents which prescriptions they were taking at the time of the questionnaire, was completed by all respondents in our subpopulation. Handling missing data is especially important for the response variable of a model, so it is fortunate this was the case. However, nine predictors, shown in Table 4, had at least one observation missing.

Factor	% Missing
Annual Income	2.24%
BMI	8.84%
Waist Size	13.61%
Health Rating	8.65%
Money on Carryout	4.28%
Number of FF Meals	22.93%
Number of Frozen Meals	0.10%
Education	3.40%
Family Heart Attacks	3.40%

Table 4. Predictors with at least one observation missing and their corresponding missing percentage.

Missing data was handled in three different ways: (1) treating missing as a category, (2) implementing a k-nearest neighbor imputation, and (3) excluding all missing observations. Because the

predictors are all categorical, it is appropriate to make missing its own category. This was done because the fact the observation was missing could hold predictive power. In one version of the dataset, all null values were replaced with a "-1" to indicate the variable was missing. Machine learning algorithms in python will not run with null values, so the null values had to be mapped to a numerical value. Negative one was chosen because no predictors take negative form.

The second approach to handling missing data was with a k-nearest neighbors (kNN) imputation. kNN is an algorithm that takes the k observations most similar to the observation with a missing value and calculates the median or mean, depending on the type of variable, for the missing category [22]. This mean or median replaces the missing value. The value k was chosen to be five because it is relatively low, which increases the prevalence of local effects, and odd, which prevents ties for binary predictors [22].

The third way of handling missing data was removing any observations with one or more values missing in any category. To ensure this was an appropriate measure to take, we verified the data was missing at random (MAR) [22]. We did this by comparing the distribution of the complete dataset with the distribution of the missing-removed dataset for each predictor. An example is shown in Figure 5. If the distributions look similar, it is fair to assume the data is missing at random. This was the case for all 18 predictors, making the missing-removed dataset valid.

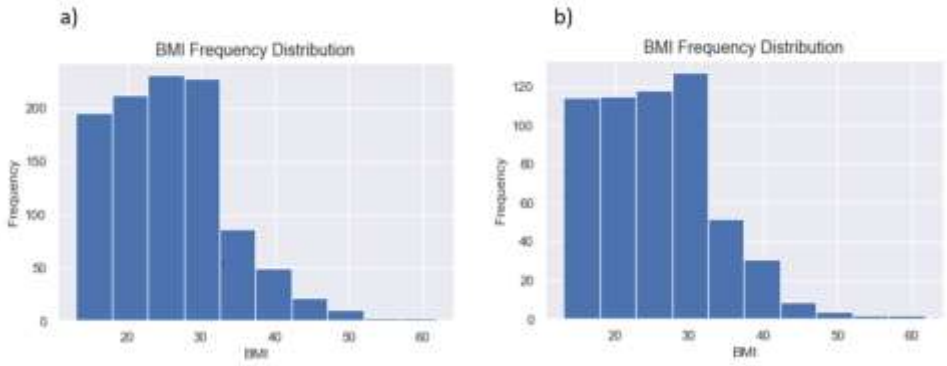

Figure 4. BMI distributions for a) the complete dataset and b) the missing-removed dataset

Because the response variable is so critical, in addition to comparing the distributions, a two-sample proportion test was completed. Because the response is binary, a proportion hypothesis test is ideal. At the level of significance of 0.1, no significant difference was found between the complete dataset and the missing-removed dataset. This also verifies the data is MAR because respondents who left questions blank were not more likely to be prescribed a statin.

2.3 Model Development

The machine learning algorithm chosen for this model was a Random Forest Classifier (RFC). Due to the presence of multicollinearity, discussed in Section 2.2, a model that assumes independence would be inappropriate. The categorical nature of most of the predictors and binary nature of the response variable were also factors in the selection of appropriate machine learning methods. For example, because we know for certain which category the response should fall into (either taking statin or not taking statin), a clustering algorithm such as k-means would not be appropriate. Algorithms requiring very large amounts of data, such as neural networks, would also not be appropriate because the final subpopulation included 1,029 observations. The ideal algorithm does not assume independence of predictors, does not require very large data sets, and works well for classification (non-continuous) responses. An RFC meets these requirements and therefore was used to create the models.

An RFC was created for each of the three ways of handling missing data, totaling in three models. All categorical predictors were mapped to integer values so the Python algorithm could run properly, and most were on an ordinal scale. This was the first step completed with each of the three datasets.

After being prepared, the datasets were read into Python as a Pandas data frame. They were then separated into two arrays: one for predictors and one for the response. These arrays were split into a training set and a testing set. The training set contained 70% of the data, and the testing set contained 30%. The training set is used to build and fit the model, while the testing set is used to evaluate how well the model is predicting. It is necessary to create both sets to properly evaluate a predictive model.

The RFC was built using a Random Forest Classifier method from the SciKit Learn library's "Ensemble" module [23]. The random state was set to 1, and 1,500 trees were used. Increasing the number of trees improved prediction performance without taking too much time. Because RFCs are not sensitive to overfitting, it was not a problem to use 17 predictors.

2.4 Model Testing

The testing set was used to evaluate model performance. For each algorithm built, the F1 score, percent accuracy, and confusion matrix were output. An F1 score is a common metric for evaluating classification algorithms that considers both precision and recall [24]. A confusion matrix is a 2x2 matrix containing the frequencies of true positives, false positives, false negatives, and true negatives. The F1 score and confusion matrix were computed using SciKit Learn's "Metrics" module. The percent accuracy

was computed using a method in SciKit Learn's "Ensemble" module. The results of these performance metrics for each model are shown in Table 5.

Model	F1 Score	Percent Accuracy	Confusion Matrix
Missing as category	0.495	83%	[[230 20] [33 26]]
kNN imputation	.519	83%	[[229 21] [31 28]]
Missing removed	0.51	86%	[[135 5] [20 13]]

Table 5. Model performance for each predictive model.

2.5 Subsequent Modeling

While RFCs can be great predictors, it is also common practice to use the importance information they output for subsequent modeling. We used this approach in the feature subsect selection process for further modeling. Cholesterol, blood pressure, age, waist size, and BMI were chosen to be used in additional models because they were consistently important features across all RFC models.

We wanted to continue using classification models, so we decided to use a logistic regression, which is a binary classification model. However, a simple logistic regression would not be sufficient on its own due to presence of multicollinearity in the data. For this reason, regularization techniques were used. Regularization techniques include penalty terms that penalize less important features so they do not affect the outcome of the model as much as other features [25].

There are three common types of regularization methods: ridge, lasso, and elastic net. Ridge regression can penalize features to make them less important, but they will never reach zero [25]. Lasso regression is similar to ridge regression but can penalize features to zero and get rid of their information completely [25]. Elastic net regression is a hybrid between ridge and lasso. It can perform similar to lasso regressions but may perform better when features are highly correlated [26].

Features from the original dataset were all standardized to a 0 to 1 scale, and dummy variables were created to represent the different age categories. Missing data was handled in only one way by imputing the missing values. The package glmnet was then used in R to create the regularized regressions [27]. The equation in Figure 5 shows how the penalty term is set up in this package. When the parameter α is set to 0, the lasso penalty disappears, and the model is a ridge regression. When α is

set to 1, the ridge penalty disappears, and the result is a lasso regression. An α between 0 and 1 results in an elastic net regression.

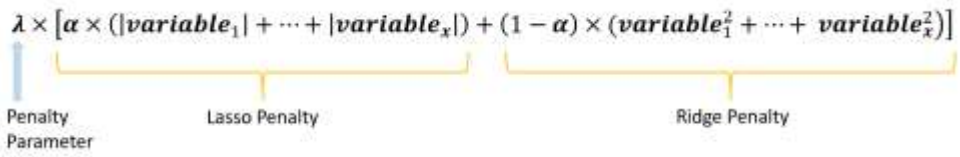

Figure 5. Demonstration of penalty term in glmnet package.

The confusion matrices and prediction accuracies for the three methods are shown in Figure 6. Of the values tested, α=.5 was the best-performing elastic net regression. The elastic-net regression performed the best for this data, and the ridge regression performed the worst. However, all models performed pretty similarly.

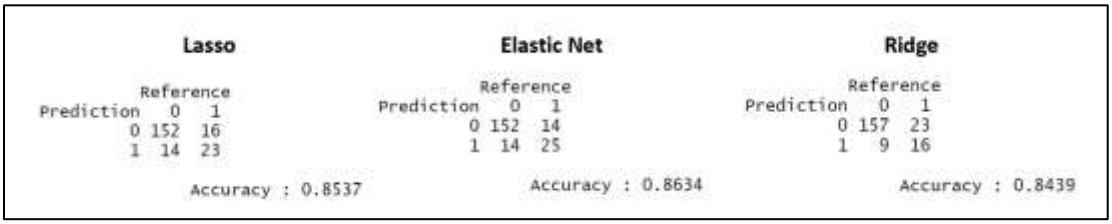

Figure 6. Confusion matrices and accuracy of predictions for each type of regularized logistic regression.

3. Results

An RFC is a black box machine learning algorithm, meaning it is difficult to see what is happening internally to cause the model to predict the way it is. For this reason, there are two ways to assess the effects of the predictors. There is a high-level conclusion, which shows the importance of predictors relative to each other. This is typically the more useful insight. There are also instance-level results, which show how one specific observation was affected by each predictor to reach the predicted response. This is not as useful because not as much insight can be gained from one specific instance. For this reason, we focused on the relative importance of predictors (referred to as feature importance).

The relative importance of the first model, where missing was a category, is shown in Figure 7. This model predicted with 83% accuracy. Generally, statin use was underpredicted. The model predicted roughly 13% statin use while the real number is close to 17%. This could be due to statin use being the minority response.

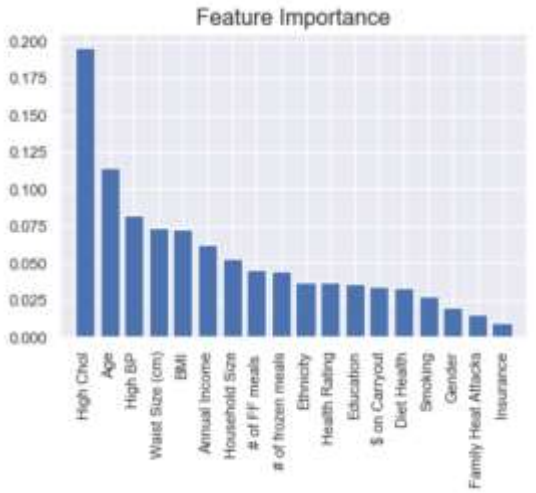

Figure 7. Feature importance obtained by the RFC where missing was a category

The relative importance of the second model, where missing observations were removed, is shown in Figure 8. This model predicted with 86% accuracy, making it more accurate than the previous model. This model also underpredicted statin use. While the most important features are similar to the previous model, they are not exactly the same. Age was a much more important feature when missing values are removed. High blood pressure is much less important, moving from the third most important feature to the eleventh. Waist size, BMI, and annual income are in similar positions. High cholesterol, as expected, is still a very important feature despite not being the most important in the second model. Whether a participant smokes changed from being one of the least important features in the previous model the seventh most important feature in this model.

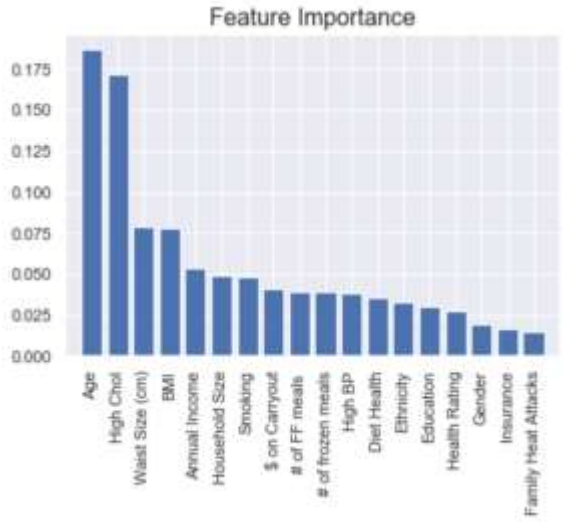

Figure 8. Feature importance obtained by the RFC when observations with missing values are removed.

The relative feature importance of the third model, where missing values were imputed, is shown in Figure 9. This model predicted with 83% accuracy, making it just as accurate as the first and less accurate than the second. The relative feature importance order in this model was very similar to the first. The top three features, high cholesterol, age, and high blood pressure, were the same. Waist size, BMI, and annual income were also very important in the first model but in a different order. Gender, insurance, and family heart attack history were the least important features in all three models.

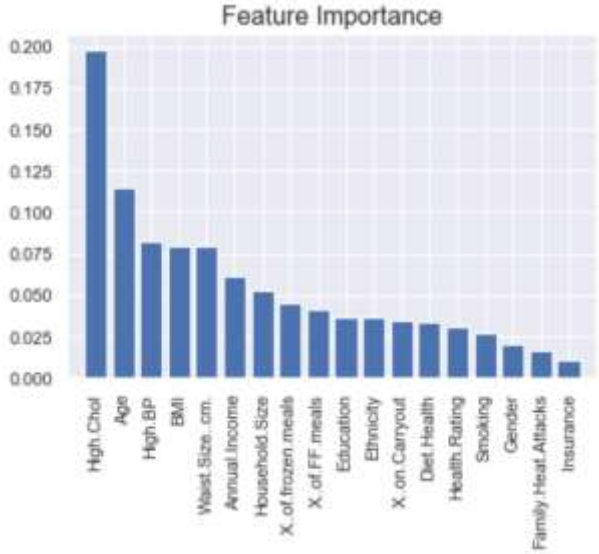

Figure 9. Feature importance obtained by the RFC when missing values were imputed.

4. Discussion and Conclusion

Among all three models, high cholesterol, age, BMI, waist size, annual income, and household size seemed to be the best predictors of statin use in prediabetic patients. High cholesterol being a very important feature in all models is intuitive because the U.S. Food and Drug Administration classifies statins as a class of drugs used to treat high cholesterol [28]. A 2006 study found a positive correlation between BMI and cholesterol in the general population; this could be a reason why BMI is fairly good predictor of statin use [29]. A 1998 study found larger households are less likely to be on a low-cholesterol diet, perhaps contributing to why household size was a relatively important feature [30].

Meanwhile, gender, family history of heart attacks, and whether a patient had insurance were the least impactful features. Aside from these three features, however, the order of the least impactful predictors varied between models. We found the fact that insurance is among the least impactful features in all three models surprising. Having medical insurance would seem to be a factor that would make accessing statins easier; however, this does not appear to be the case among prediabetic patients. One possible explanation for this is that statins are so important to patients they will ensure they have the medication whether insured or not.

Another result we found surprising was the low ranking of smoking. A 1991 study of adults found that cholesterol increased in both men and women for each cigarette smoked daily [31]. The U.S. Food and Drug Administration describes statins as a drug used to lower cholesterol [28]. Because smoking can raise cholesterol and statins can lower it, it would seem natural for them to have a relationship; however, that does not appear to be the case in these models.

In Table 3 in Section 2.1, the features of the model were categorized as “Demographic,” “Behavioral,” or “Health-related.” In general, health-related features such as blood pressure, cholesterol, BMI, and waist size were of the highest importance. Behavioral features tended to be in the

middle of the importance ranking. Demographic characteristics were generally less important with age and household size being exceptions.

Ideally, this research could be continued using data of the same patients over time to analyze the new onset of diabetes in statin-using prediabetic patients specifically. The results of the analysis could be used to determine which features are the most important in predicting whether a high-risk patient using statins will develop new-onset diabetes. If these features differ from our findings of which factors medical decision makers are currently taking into consideration when prescribing statins, this could be cause for these decision makers to re-evaluate the criteria they use to make these decisions. In this case we would have a baseline for current practices and an analysis determining which features should be used in practice.

For future research on current practices specifically, we recommend investigating different model types with different sets of features to try to improve model performance and including more health-related features. It could be impactful to see how or if comorbidities affect statin initiation in prediabetic patients. Oversampling methods could also be used in attempt to improve prediction accuracy because the model is currently under-predicting statin use.

5. Appendix

1) Results of Chi-square multicollinearity study

	Gender	Age	HH Size	Income	BMI	Waist Size	High BP	High Chol	Health Rating	$ on Carryout	Diet Health	# of FF meals	# Frozen Meals	Smoking	Education	Insurance	Ethnicity	Family Heart Attacks
Gender		0.162	0.297	0.298	0.269	0.309	0	0.081	0.189	0.108	0.086	0.241	0.506	0	0.029	0.231	0.062	0.009
Age			0	0.064	0.694	0.346	0	0	0.325	0.193	0	0.062	0.009	0	0	0	0	0
Household Size				0	0.806	0.113	0	0	0.0001	0	0	0.252	0.095	0	0	0	0	0
Annual Income					0.162	0.038	0.185	0.322	0	0	0.103	0.0002	0.11	0.002	0	0	0	0.048
BMI						0	0.829	0.925	0.302	0.0015	0.614	0.999	0.999	0.998	0.61	0.011	0.943	0.106
Waist Size (cm)							0.001	0.078	0.295	0.013	0.107	1	0	0	0.16	0	0.954	0.575
High BP								0	0	0.921	0	0.071	0	0	0	0.0002	0.004	0
High Chol									0.002	0.962	0	0.039	0	0	0	0	0.063	0
Health Rating										0.718	0	0.134	0	0.143	0	0.0501	0	0.002
$ on Carryout											0.203	0.0002	0.687	0.999	0.196	1	0.072	0.101
Diet Health												0	0	0	0	0.012	0	0
# of FF Meals													0	0.154	0	0.053	0.004	0
# of frozen meals														0	0	0.999	0.039	0
Smoking															0	0.431	0	0
Education																0	0	0
Insurance																	0	0.404
Ethnicity																		0.0001
Family HeatAttacks																		

References

[1] M. Heron, "Deaths: Leading Causes for 2017," Center for Disease Control and Prevention, 2019.

[2] National Institute of Diabetes and Digestive and Kidney Diseases, "Diabetes Overview," December 2016. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes.

[3] Center for Disease Control and Prevention, "Diabetes," August 2019. [Online]. Available: https://www.cdc.gov/diabetes/basics/diabetes.html.

[4] International Diabetes Federation, "Diabetes Complications," 1 January 2020. [Online]. Available: https://www.idf.org/aboutdiabetes/complications.html.

[5] Miller, Keane and O'Toole, Encyclopedia of Medicine, Nursing, and Allied Health, vol. 7, Elsevier, Inc, 2003.

[6] A. Menke, S. Casagrande, L. Geiss and C. Cowie, "Prevalnce of and Trends in Diabetes Among Adults in the United States," *Journal of the American Medical Association,* pp. 1021-1029, 2015.

[7] Sattar, Preiss and Murry et al., "Statins and Risk of Incident Diabetes: A Collaborative Meta-analysis of Randomized Statin Trials," *The Lancet,* vol. 375, no. 9716, pp. 735-742, 2010.

[8] Center for Disease Control and Prevention, "Heart Disease Facts," Center for Disease Control and Prevention, 2017. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm.

[9] U.S. Preventitive Services Task Force, "Statin Use for the Primary Prevention of Cardiovascular Disease in Adults," *JAMA,* 2016.

[10] A. A. Carter, T. Gomes, X. Camacho, D. N. Juurlink, B. R. Shah and M. M. Mamdani, "Risk of incident diabetes among patients treated with statins: population based study," *Bmj,* vol. 346, no. f2610, 2013.

[11] M. J. Ko, A. J. Jo, Y. J. Kim, S. H. Kang, S. Cho, C.-Y. Park, S.-C. Yun, W. J. Lee, D.-W. Park and S.-H. Jo, "Time- and Dose-Dependent Association of Statin Use With Risk of Clinically Relevant New-Onset Diabetes Mellitus in Primary Prevention: A Nationwide Observational Cohort Study," *Journal of the American Heart Association,* no. 8.8, 2019.

[12] e. a. Naveed Sattar, "Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials," *The Lancet,* vol. 375, no. 9716, pp. 735-742, 2010.

[13] e. a. Kang-Ling Wang M.D., "Statins, Risk of Diabetes, and Implications on Outcomes in the General Population," *Journal of the American College of Cardiology,* vol. 60, no. 14, pp. 1231-1238, 2012.

[14] Center for Disease Control and Prevention, "Picture of America Prevention".

[15] P. Ridker M.D., A. Pradhan M.D., J. G. MacFadyen, P. Libby M.D. and R. J. Glynn, "Cardiovascular benefits and diabetes risks of statin therapy in primary prevention: an analysis from the JUPITER trial," *The Lancet,* vol. 380, no. 9841, pp. 565-571, 2012.

[16] U.S. Food and Drug Administration, "FDA Drug Safety Communication: Important safety label changes to cholesterol-lowering statin drugs," 28 February 2012. [Online]. Available: https://www.fda.gov/drugs/drug-safety-and-availability/fda-drug-safety-communication-important-safety-label-changes-cholesterol-lowering-statin-drugs.

[17] e. a. Kang-Ling Wang M.D., "Risk of New-Onset Diabetes Mellitus Versus Reduction in Cardiovascular Events With Statin Therapy," *The American Journal of Cardiology,* vol. 113, no. 4, pp. 631-636, 2013.

[18] Center for Disease Control and Prevention, "National Health and Nutrition Examination Survey MEC Laboratory Procedures Manual," 2016.

[19] National Institute of Diabetes and Digestive and Kidney Diseases, "Diabetes Test and Diagnosis," U.S. Department of Health Services, December 2016. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/tests-diagnosis.

[20] American Diabetes Association, "Diagnosis," 2019. [Online]. Available: https://www.diabetes.org/diabetes.

[21] D. Adedinsewo, N. Taka, P. Agasthi, R. Sachdeva, G. Rust and A. Onwuanyi, "Prevalence and Factors Associated With Statin Use Among a Nationally Representative Sample of US Adults: National Health and Nutrition Examination Survey, 2011–2012," *Clinical Cardiology,* vol. 39, no. 9, 2016.

[22] Y. Obadia, "The use of kNN for Missing Values," Towards Data Science, January 2017. [Online]. Available: https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637.

[23] P. e. al., "Scikit-learn: Machine Learning in Python," *JMLR,* no. 12, pp. 2825-2830, 2011.

[24] R. Ng, "F1 Score," 23 April 2020. [Online]. Available: https://www.ritchieng.com/machinelearning-f1-score/.

[25] X. Liu, Ph.Dl, *Introduction to Modern Statistical Techniques for Industrial Applications Lecture 8,* Fayetteville, AR: University of Arkansas, 2020.

[26] MathWorks, "Lasso and Elastic Net," 2020. [Online]. Available: https://www.mathworks.com/help/stats/lasso-and-elastic-net.html.

[27] J. Freidman, T. Hastie and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software,* vol. 33, no. 1, pp. 1-22, 2010.

[28] U.S. Food and Drug Administration, "Statins," U.S. Food and Drug Administration, 16 December 2014. [Online]. Available: https://www.fda.gov/drugs/information-drug-class/statins. [Accessed 2020].

[29] H. M. Dashti, "Long Term Effects of Ketogenic Diet in Obese Subject with High Cholesterol," *Molecular and Cellular Biochemistry,* no. 286, pp. 1-2, 2006.

[30] R. M. Naya Jr., "Consumer characteristics associated with low-fat, low-cholesterol foods," *The International Food and Agribusiness Management Review,* pp. 41-49, 1998.

[31] J. E. Muscat, R. E. Harris, N. J. Haley and E. L. Wynder, "Cigarette Smoking and Plasma Cholesterol," *American Heart Journal,* pp. 141-147, 1991.