SMU Data Science Review

Volume 1 | Number 3 Article 12

2018

Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data

Alfeo Sabay
Southern Methodist University, asabay@smu.edu

Laurie Harris
Southern Methodist University, llharris@smu.edu

Vivek Bejugama
Southern Methodist University, vbejugama@smu.edu

Karen Jaceldo-Siegl
Loma Linda University, kjaceldo@llu.edu

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

 Part of the Cardiology Commons, Statistical Models Commons, and the Statistical Theory Commons

Recommended Citation

Sabay, Alfeo; Harris, Laurie; Bejugama, Vivek; and Jaceldo-Siegl, Karen (2018) "Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data," *SMU Data Science Review*: Vol. 1: No. 3, Article 12.
Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/12

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.

Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data

Alfeo Sabay[1], Laurie Harris[1], Vivek Bejugama[1], Karen Jaceldo-Siegl DrPH[2]

[1] Southern Methodist University (SMU), 6425 Boaz Lane, Dallas, TX 75205, USA
[2]Loma Linda Unversity, School of Public Health, 24951 North Circle Drive, Loma Linda, CA 92350, USA
{asabay,llharris,vbejugama}@smu.edu,kjaceldo@llu.edu

Abstract. In this paper, we present a heart disease prediction use case showing how synthetic data can be used to address privacy concerns and overcome constraints inherent in small medical research data sets. While advanced machine learning algorithms, such as neural networks models, can be implemented to improve prediction accuracy, these require very large data sets which are often not available in medical or clinical research. We examine the use of surrogate data sets comprised of synthetic observations for modeling heart disease prediction. We generate surrogate data, based on the characteristics of original observations, and compare prediction accuracy results achieved from traditional machine learning models using both the original observations and the synthetic data. We also use a large surrogate data set to build a neural network model (Perceptron) and compare the prediction results to the traditional machine learning algorithms (Logistic Regression, Decision Tree and Random Forest). Using traditional Machine Learning models with surrogate data, we achieved improved prediction stability within 2 percent variance at around 81 percent using ten fold validation. Using the neural network model with surrogate data we are able to improve the accuracy of heart disease prediction by nearly 16 percent to 96.7 percent while maintaining stability at 1 percent. We find the use of surrogate data to be a valuable tool, as a means to anonymize sensitive data and improve classification prediction.

1 Introduction

Traditional classification algorithms such as logistic regression and decision trees have historically been employed to design prediction models for medical data sets [1]. These approaches can provide very good accuracy of classification prediction. As machine learning algorithms become more popular and accessible, researchers may be tempted to apply neural network models to medical data sets in an attempt to improve classification prediction accuracy. However, medical data are often constricted by smaller sets of observations than what is usually preferred to allow for sufficient training and testing of models built using machine learning algorithms [2]. Without sufficiently sized data sets, it is very difficult to determine if a model is generalizable to previously unseen sets of data [3].

Using synthetic data to overcome constraints inherent in small medical research data sets could be a solution to protect patient privacy and allow for application of

machine learning algorithms. With tools such as the Synthpop package in R, researchers are able to efficiently generate extremely large data sets with the same characteristics of the original data to be used in machine learning algorithms. The larger data sets allow for sufficiently sized training and testing partitions which enable the machine learning algorithm to learn from experience by exposure to a large set of observations, and then to be tested upon another large set of observations that have not previously been introduced to the model.

In this paper, we examine the application of synthetic data generation to a heart disease prediction problem. Heart disease prediction is a well-studied classification problem and prior analyses serve as adequate baselines for our review. This analysis is completed in three stages. In stage one, we examine previously published results and replicate logistic regression, decision tree, and random forest models using the Cleveland Heart Disease data, as points of reference for our synthetic data research. In stage two, a surrogate data set of 50,000 observations is generated, based on the characteristics of the Cleveland data, using the Synthpop package in R. This surrogate data set serves as a stand in for the original observations. We carefully compare the surrogate data set to the original observations to determine that the original characteristics are maintained. We show that the distribution of the variables is consistent between the surrogate and original observations.

Using the synthetic data, we train and validate the Machine Learning Models then compare the prediction outcome accuracy to that using the original observations. Once satisfied with the consistency of classification prediction between the original data set and the surrogate data set, we generate an expanded surrogate data set in stage three. While based on the Cleveland data set, this expanded set contains previously unstudied attributes. This expanded data set is used to test and train a neural network model using the Keras API for Python, having partitioned the synthetic data into large testing and training subsets. We then compare the outcome of the prediction accuracy of the neural network model to the traditional logistic regression models. We find that using the expanded surrogate data set to build a neural network model results in the best classification prediction accuracy and stability. We pursue this model only after examining the synthetic data output and comparing it to the original observations. Finding that the characteristics of the original Cleveland data are maintained in the surrogate data sets, we find the use of surrogate data to be appropriate.

We conclude that the Synthpop package is a viable option to generate synthetic observations which can conceal sensitive data points and be used for deep learning. We are able to improve the accuracy of heart disease classification prediction by nearly 16 percent, using this approach. Based on these results, we consider this method to be a useful approach when analyzing relatively small clinical data sets.

The remainder of this paper is organized as follows. In section 2, we introduce the heart disease prediction problem, and describe the medical condition and risk factors. We also examine prior heart disease prediction analyses, noting prediction accuracy and modeling algorithms employed. In section 3, we describe the well-studied Cleveland heart disease data set, including the explanatory variables. Our design methods are presented in section 4 with a detailed explanation of the Synthpop package and discussion of our workflow process, describing each stage of our analysis. Our results are

presented in section 5, followed by the related analyses in section 6. In section 7, we offer a discussion of ethical concerns specific to medical data sets and the use of synthetic data. In section 8 we discuss potential research work to be done in the future using surrogate data generated from geographically diverse patient data sources. Finally, in section 9 we conclude on the work that is presented in this paper.

2 Coronary Heart Disease and Prediction

Heart Disease is a cardiovascular condition that affects 11.7 percent of American adults, with estimated costs, including healthcare, medications and lost productivity, of approximately 200 billion dollars [5]. Coronary Heart Disease (CHD), the most common type of heart disease, is a condition where the blood flow in the coronary arteries is impaired due to the narrowing effect of plaque buildup within the coronary arteries. There are well known causes to this condition including diet high in fat and cholesterol, inadequate physical activity, excess body weight, and tobacco and alcohol use. Without early detection and clinical intervention, studies show that almost half of the patients diagnosed with CHD will eventually die of the disease [7]. According to the National Center for Health Statistics report of 2016, heart disease was the leading cause of death causing 23.4 percent of the total reported deaths in 2015[8].

According to the American Heart Association[1], the terms Coronary Arterial Disease (CAD) and CHD are frequently used interchangeably. However, to be specific, a diagnosis of CAD is often a prequel to a diagnosis of the more general CHD[9]. In medical practice, CHD is a common term for the buildup of plaque in the arteries leading to and from the heart where restricted blood flow of the heart muscle, known as ischemia, will eventually lead to a heart attack.

2.1 Risk Factors for Heart Disease

There are multiple risk factors that can contribute to the development of CHD. A summary of the significant factors is shown in Table 1 below.

Although age, gender and family history are factors that cannot be changed or controlled, acknowledging them as risk factors can empower an individual to take initiative in monitoring for the controllable factors. Risk of CHD increases with age and positive family history of heart disease. In addition, men have more heart attacks than women and heart attacks in men occur earlier in age than women [10].

Many risk factors are routinely monitored by physicians. Blood pressure readings are measured at almost every clinic encounter. High blood pressure is one of the side effects of restricted blood flow. Arterial blockage which restricts blood flow in the blood vessels causes Hypertension due to increased resistance in blood flow when the heart is pumping. Blood pressure measurements higher than 140/90 are associated with increased heart disease risk.

Cholesterol readings are also frequently monitored for patients. Low Density Lipoprotein (LDL) Cholesterol, also known as the bad cholesterol, is carried in the blood stream

[1] Coronary Artery Disease - Coronary Heart Disease. http://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease

Table 1. Risk Factors Associated with Heart Disease

Risk Factor	Normal Range	Controllable
High LDL	<130 mg/dL depending on risk	Yes
Low HDL	>60 mg/dL, 40 for men, 50 for women	Yes
Hypertension	<140/90 mm Hg	Yes
Family History	N/A	No
Cigarette Smoking	N/A	Yes
Age	<45 to 55 (low risk)	No
Obesity	<25 Body Mass Index (BMI)	Yes
Gender	<45 for men, <55 for women (low risk)	No

and is indicated by the marker LDL-C. The amount of cholesterol carried in the bloodstream gives important information about the risk of developing CHD. Levels of LDL-C greater than 130 mg/dL are associated with increased risk of CHD. Another risk factor is low levels of the High Density Lipoprotein (HDL) cholesterol also known as the good cholesterol. Levels of HDL cholesterol under 60 mg/dL also indicate increased risk for CHD [6].

These clinical measurements can often be managed by a patient in order to reduce the risk of CHD. Hypertension, high LDL-C and low HDL-C can be controlled by diet, medication or physical activity. Regular physical activity has many benefits including stress reduction, weight control and lowering high blood pressure.

There are a few key risk factors that can be influenced by lifestyle choices, including cigarette smoking and obesity. Smoking is a risk factor and is highly correlated to high blood pressure. Smoking causes hardening of the arteries and causes blood pressure to rise. Avoidance of tobacco smoke, which may lower blood pressure, is yet another risk factor that can be improved by the patient. Additionally, the condition of obesity can be improved with diet and physical activity [10].

2.2 Studies in Heart Disease Prediction

CHD prediction algorithms and data mining techniques have been available since the 1960s. As with many medical concerns that potentially impact the health of a large population of individuals, prediction of heart disease is a topic that has been well studied.

In a 2010 publication, K. Srinivas et al. [11] applied data mining techniques to clinical documentation for arterial blockage measurements to predict diseased arteries. Their study brings attention to the vast amount of clinical data that are available within a patient medical record and considers, more broadly, how to utilize these data points to identify unseen patterns and unlock insights that can impact and improve patient health.

In their study, the percentage of arterial blockage was classified as healthy or non-healthy. For three of the arteries (Right Coronary Artery, Left Anterior Descending, and Left Circumflex) the cut-off values were set at 50 percent, with 70 percent or higher representing significant disease. For the Left Main Coronary Artery, the cutoffs were set at 30 percent and 50 percent, respectively. This adjustment was due to the presumption that blockage in this artery is likely to produce more disease than the others. Ultimately,

the authors identify a Naive Bayes model that can be used to predict heart disease with accuracy in the low to mid 80 percent range.

In a 2011 publication, Jabbar et al. utilized the Cleveland Heart Disease Data set to examine association rule techniques in predicting heart disease [12]. Their study transformed the measurements from the Cleveland data into binary classifications of variables. For example, where the data contained a specific cholesterol measurement for the subject (e.g., 200, 205, 262), the researchers transformed that value into either a 0 or 1 based on their established criteria. In their study, cholesterol values greater than 240 would have a binary value of 1 and those less than that threshold would have a value of 0.

This process of scoring clinical results enabled the researchers to apply clustering techniques and develop a Cluster Based Association Rule Mining Based on Sequence Number (CBARBSM) model. The model ultimately examined frequent item sets and results concluded that the association rule for heart disease includes the following: age greater than 45, systolic blood pressure greater than 120, maximum heart rate greater than 100 and old Peak (stress test depression) greater than 0 and Thallium (measurement of defect) greater than 3.

In 2012, Shouman et al., researchers built upon prior studies to combine decision tree techniques with k-means clustering to determine if heart disease prediction accuracy of existing models can be improved [13]. In the Shouman study, the researchers found that the best technique was to apply the k-means clustering to the age variable and use the remaining 13 attributes for the decision tree. Their results showed that their optimal design, using two clusters, produced approximately 83.9 percent accuracy of heart disease prediction. The researchers opined that their model be improved by adding more clusters or application to a larger data set.

A 2015 publication by El-Bialy et al. reinforces the motivation for the volume of research around the topic of cardiovascular disease, as it notes the fatality of the condition in that the illness causes over one million deaths each year and that almost one half of individuals with the condition will eventually die from the disease [7]. The study also discusses challenges associated with accumulating and using clinical data. Some of these include the sparse nature of the data, measurement errors, recording inaccuracies, and variations in professional interpretation, etc. As researchers, we also recognize this problem and acknowledge the delicate balance between the desire to utilize all the data that are available and simultaneously, build the most accurate model for use.

The authors concluded that applying the same machine learning techniques to different data sets can produce different results and noted that the accuracy of the models developed using C4.5 and fast decision trees were in the mid to high 70 percent range when applied to four distinct data sets. Table 2 below summaries the findings of these prior research studies.

In all of the above studies, the researchers' focus was mainly on the machine learning algorithms used in predicting heart disease. Machine learning algorithms such as logistic regression, decision tree, K-means clustering and fuzzy rule based models were used in their analysis. These models used the processed version of the Cleveland dataset and the models were fine tuned to maximize the accuracy metrics of the models [14]. These studies did not appear to consider techniques that could be applied to the data

Table 2. Summary of Prior Heart Disease Prediction Studies

Research Study	Year	Accuracy
K.Srinivas B.Kavihta Rani Dr. A.Govrdhan	2010	Low to mid 80%
Mai Shouman, Tim Turner, Rob Stocker	2012	83.90%
Randa El-Bialy, Mostafa A. Salamay, et al.	2015	Mid to high 70%

set used by the models in order to improve the accuracy and stability of the prediction accuracy.

2.3 Challenges Associated with Medical Data Sets

Application of today's Machine Learning techniques in heart disease prediction have challenges in data size and confidentiality matters. Processing times for statistical models that are core to modern day machine learning classification and Artificial Neural Network (ANN) algorithms have been greatly reduced thereby allowing for deeper analysis and enhanced validation techniques [7]. However, application of machine learning techniques often require very large data sets, much larger than what is traditionally available from medical research experiments, where observation sizes are constrained by cost, complexity and patient confidentiality compliance requirements. Such small datasets are typical in healthcare where they are adequate for human comprehension but insufficient in volume for machine learning models [2]. One reason for the small size of medical data sets is due to the lack of centralized medical databases. Hospitals, insurance companies, clinics and research organizations all maintain and protect their own patient databases. Patient privacy laws are the main reason for this fragmentation [22]. These data sets are categorized as small-data and often come in less than 500 observations and can be as small as 10 observations. Machine learning algorithms such as ANN are built through a process of model training where the machine is trained to learn by experience and exposure to data observations. The models are then tested using previously unseen data to assess model performance and ensure trained models can be generalized to new sets of data.

A novel solution for satisfying data volume requirements for machine learning models was developed by Torgyn Shaikhina and Natalia A. Khovanova in a 2016 publication titled Handling Limited Datasets with Neural Networks in Medical Applications: A Small-Data Approach [3]. In this research, a framework for generating surrogate data from small data sets (as small as ten observations) was developed and validated using neural network techniques. This technique utilizes multiple runs of 2000 neural networks in order to generate robust data sets that mimic the characteristics of the real data set and provides adequate data volumes to satisfy modern machine learning based prediction. The number of neurons required for this technique requires large computing resources therefore, in this experiment, we chose an alternative solution for surrogate data generation.

There are several tools available for surrogate data generation. Synthpop is an R language library that provides data synthesis and data comparison functions [26]. Simpop

is also an R language library that provides surrogate data synthesis utilizing S4 class implementation, but without the comparison tools provided by Synthpop to compare the seed data to the surrogate data [27]. There are numerous other methods that use the R "boot" package (bootstrap) to re-sample small data sets for surrogate data generation, but this was found to be a low level implementation and lacking built-in comparison tools like those provided by Synthpop [28]. Synthpop was chosen as the tool for surrogate data generation in this study because of the ease of use and the built-in tools that provide comparisons of the seed data to the surrogate data characteristics.

3 Heart Disease Data Set

Our analysis was performed using data from the Heart Disease Database available from the UCI Repository (Center for Machine Learning and Intelligent Systems) [14]. This data has been available since 1988, and is considered the gold standard in heart disease prediction research because of its availability and widespread use in research [13][24][25].

There are four available database categories as shown in Table 3 below. Of these four sources, the Hungarian, Switzerland, and Long Beach data sets have many missing values; therefore, the Cleveland dataset has been used the most due to its completeness of observations [15]. Although well documented in previous studies, we explored all of these data sets during the data preparation stage and have chosen to utilize the Cleveland data set because it has the least number (only six observations) of missing data.

Table 3. UCI Repository for Machine Learning Heart Disease Databases

Database	Donor	Author	Instances
Cleveland	Cleveland Clinic Foundation	Robert Detrano, M.D., Ph.D.	303
Hungarian	Hungarian Institute of Cardiology, Budapest	Andras Janosi, M.D.	294
Switzerland	University Hospital, Zurich, Switzerland	William Steinbrunn, M.D.	123
Long Beach	V.A. Medical Center, Long Beach	Robert Detrano, M.D., Ph.D.	200

The four Heart Disease data sets are available in raw and processed formats[2]. The raw format is space delimited text and where there is missing data, a "-9" was inserted. The Cleveland raw data set contains 76 attributes, but half (38) of the attributes are not usable due to missing or undefined data. The number of final usable records from the raw data set after the data cleanup was 282 records with 38 attributes. The processed data (the Cleveland 14 data set) is a comma delimited text file and has been reduced to 14 attributes as a result of past research [7]. In the processed databases, the missing data encoding (-9) still exists. Of the four processed databases, the Cleveland 14 data is the most complete with only 6 records of missing data making the total usable size 297 records with 14 attributes.

[2] Relevant Information (4): http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

Most of the recent research utilizes the processed Cleveland database containing 14 attributes (Table 4) [7]. The response variable (diag) indicates the presence or absence of heart disease. In the raw data, the values range from zero (0) to four (4). Zero indicates less than fifty percent arterial blockage and is classified as no disease, while one to four (1-4) is an indication of the degree of arterial blockage of over fifty (50) percent. For most studies the values of 1 through 4 are coded into a one (1) value so the prediction is based solely on whether heart disease is present or not. In addition to the response variable, there are 13 features used in this study from the Cleveland data set including age, sex, chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiography (ECG), maximum heart rate from Thallium test, exercise induced angina, ST depression (an ECG reading indicating blood flow), the slope of the ST depression, the number of blood vessels colored by fluoroscopy, and the heart defect status code. Table 4 lists the variables and descriptions. In this study, we refer to the processed Cleveland dataset as Cleveland 14.

All four datasets include variables which are patient readings collected from what is known as Stress or Exercise Test. Exercise Testing is performed by a Cardio Imaging Technician at the supervision of a Cardiologist where the patients heart rate is elevated by means of treadmill activity or by injection of a drug that makes the heart pump faster. The patients heart pumping activity is monitored both at rest and with elevated heart rate. Thallium, which is a radioactive material, is injected into the patients blood stream so that the medical imaging equipment can capture the patients blood flow and heart activity both at rest and with elevated heart rate [10]. Some Stress Test related metrics such as maximum heart rate, oldpeak and slope (electrocardiogram readings) are captured in the Cleveland 14 dataset.

Table 4. Cleveland 14 Data Set Variables

Attribute	Description
age	Age in years; continuous
sex	Gender; categorical
chest_pain	Scale of 1 to 4; categorical
resting_bp	Diastolic blood pressure; continous (mmHg)
cholesterol	Total Cholesterol; continuous (mg/dl)
fast_bsugar	Scale of 0 to 1; categorical
resting_ecg	Scale of 0 to 2; categorical
max_hrate	max heartrate from thalium test; continuous
exer_angina	Scale of 0 to 1; categorical
Oldpeak	ST depression relative to rest; continuous
Slope	Scale of 1 to 3; categorical
ca_mavessel	Scale of 0 to 4; categorical
heart_def_status	One of 3 values (per thalium test); categorical
diag	Response variable; 0 no disease, 1 disease; categorical

Management of the missing data was done in two ways. In the early stages of this study, a decision was made not to impute or substitute the missing data in the processed Cleveland data set in order to be consistent with the data used in past heart disease prediction studies [13]. In the final stage of this study, the raw Cleveland data (no component reduction) set was used to take advantage of the larger number of variables to train the Perceptron Neural Network Model. There were five (5) records with missing data and the missing data were filled in with imputed values (mean value of the attribute column). After data cleaning, the original 76 variables were reduced to 38 due to columns with all missing or unknown data. The final shape of the raw Cleveland data set after data cleaning was 38 x 282 (38 attributes and 282 samples).

The volume size of both the Cleveland 14 and the raw Cleveland data sets, which contain less than 300 records, are small and not ideal for stable performance of machine learning prediction models. This is because the training and validation steps for machine learning models involve splitting the data set into training and testing partitions. In a typical 10 fold cross validation process, the data set can be split where 80% of the records are used to train the models and 20% for testing. The training and test sets are further split into 10 partitions during the cross validation stage in order to observe the stability of the model. It is expected that the small volume size of both the Cleveland 14 and the raw Cleveland data sets will be problematic in the training and cross validation process and will result in unstable performance measurements. In order to introduce stability to the training and cross validation stages, we applied the use of data synthesis to produce surrogate data in this study.

4 Methods and Experiments

The usefulness of a good heart disease prediction model depends largely on its accuracy and stability. To achieve this, we divided this experiment into three (3) stages (Figure 1). In stage 1 of the experiment we established baseline models and their results. The main objective of Stage 1 was to validate and compare the accuracy and stability of the results of our own machine learning models with those of the results in prior research studies using the Cleveland 14 dataset [13][25][24]. In Stage 2, we introduced a variation of the Cleveland 14 dataset by generating synthetic data based on the Cleveland 14 data using the R Programming Language and the Synthpop Library. We used Synthpop to generate fifty thousand (50,000) records that mimic the characteristics of the Cleveland 14 data (containing 297 records). We then used this larger surrogate dataset to train and test the previous logistic regression, decision tree and random forest models. The purpose of Stage 2 was to see if there was an improvement in accuracy and stability of the same machine learning models by comparing the results to the baseline in Stage 1. In Stage 3 of this experiment, we generated sixty thousand records from the raw Cleveland dataset and used the surrogate data to train and test an ANN model of the Perceptron Forward and Back Propagation algorithm type. ANN models in general have the advantage of simpler feature selection because the models make no assumptions about the data input. We used the raw Cleveland dataset with 37 Explanatory Variables and 1 Response Variable. The Perceptron model performs well with large datasets, so a surrogate dataset of 60,000 records was generated from the raw Cleveland Dataset

(shape of 38 X 282). The purpose of Stage 3 was to compare the accuracy and stability of the Perceptron model to the Stage 1 and Stage 2 model accuracy and stability.

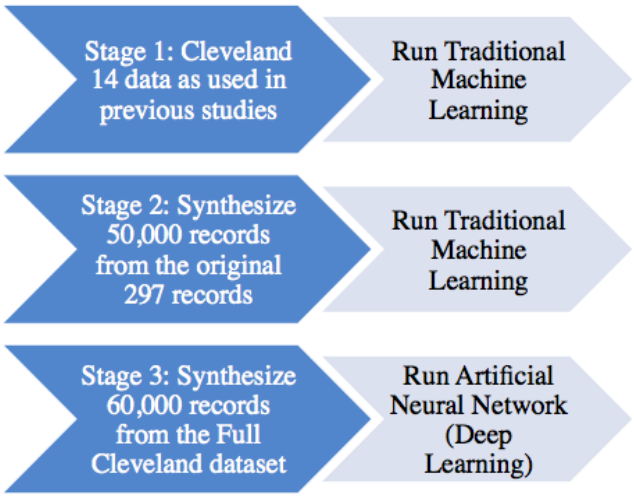


Fig. 1. Machine Learning Experiment Stages

4.1 Synthetic Data Generation using Synthpop

As previously mentioned, in stages 2 and 3, we expanded the volume of the data by generating a surrogate dataset from the original data. This was accomplished by a process of data synthesis using the R language Synthpop library. Synthpop generates synthetic data from the original dataset and produces a surrogate dataset that mimics the characteristics of the original dataset. The resulting surrogate dataset volume can be adjusted to a higher or lower volume to increase or decrease the total number of records in the surrogate dataset. The motivation behind the creation of Synthpop by its authors was to conceal confidential data by generating synthetic versions of the data without disclosing real data values while preserving the characteristics and the relationships between the variables [26].

Synthpop produces synthetic data using statistical techniques. The generation of the synthetic data is done via the function `syn()`. At its simplest usage, the function `syn()` only requires the argument `data` which is a data frame or matrix that contains the original data and can be supplied with the `k` argument that specifies the record size of the synthetic data. Other arguments specify random number `seed` and `drop.not.used` which if `TRUE`, variables not used in the synthesis are not saved in the synthesized data. Internally, the synthetic data is produced column by column based on the estimated distribution of the original data's column. The creation of the synthetic data is done by

projecting and fitting general linear/logistic models using the original data and by implementing Classification and Regression Trees (CART). The projection of the original data via the general linear/logistic models into surrogates of varied sizes allows for validation of the synthetic dataset by fitting the synthetic data via the general linear/logistic model and comparing the statistical characteristic to the original dataset via the function `compare()`. The `compare()` function plots a frequency plot of the percentage makeup of the original data versus the synthetic data. Depending on the argument passed to `compare()`, a comparison of the coefficient or Z confidence interval can also be plotted. The utility of the `compare()`function was used in stages 2 and 3 to demonstrate that the data characteristics of the Surrogate data set were preserved.

4.2 Model Evaluation Metrics

In all three stages of the experiment, three measures were used to assess the machine learning or classifier models' predictive capabilities. The first metric was the accuracy classification score. The accuracy score measured the fraction of the classifier's predictions that are correct and is expressed by the following formula:

$$accuracy(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y} = y_i)$$

The second metric is the precision or confidence. Precision is the proportion of predicted positive cases that are real positives and is expressed by the following formula:

$$precision = \frac{TruePositives}{TruePositives+FalsePositives}$$

The third metric is the recall or sensitivity. Recall measures the coverage of real positive cases as influenced by the number of false negative cases. The formula for recall is expressed as:

$$recall = \frac{TruePositives}{TruePositives+FalseNegatives}$$

All three metric values range from zero (0) to one (1). The closer the value is to one, the better the performance.

4.3 Stage 1

The objective of Stage 1 was to establish a baseline of the heart disease prediction metrics based on past research work. The baseline was established by using the Cleveland 14 data set the way it was used in some of the previous studies [13][24][25]. In the Shouman et al. study, a K-means clustering technique was used to pre-process the Cleveland 14 data using various centroid selection methods prior to using the Naive Bayes algorithm resulting in accuracy ranges from the high 70% to mid 80% range. We observed that although the results of the study showed improvement in the accuracy by applying the K-means preprocessing, the stability of the prediction metrics exhibited a wide range of results when selecting the number of clusters. In our study, the models used in stage 1 were from the Scikit-learn Python library. Three different algorithms were used from Scikit-learn. These models used the logistic regression, decision tree

and the random forest algorithms. In all three models, the model hyperparameters were fine tuned to the Cleveland 14 dataset using GridSearchCV. These models though single layer with no pre-processing, obtained similar results to the studies in the Shouman et al., Sumathi et al. and Marateb et al. studies shown in Table 2 with our own results as shown in Table 5. We were able to achieve comparable results with the logistic regression, decision tree and random forest models.

4.4 Stage 2

The objective in Stage 2 of the research was to improve the stability of the models used in Stage 1. The small number of observations (297 records) of the Cleveland 14 Dataset which is typical of real healthcare data does not meet volume requirements of Machine Learning models and will cause instability in the performance metrics results during 10 fold cross-validation on the Machine Learning models. Therefore, in Stage 2 we generated a 50,000 observation surrogate dataset based on subject characteristics from the Cleveland 14 dataset. We applied the Synthpop library in R Studio to generate the surrogate dataset and assessed the same prediction metrics using logistic regression, decision tree and random forest models.

The Synthpop library was used in R Studio to generate the surrogate dataset. To show that the surrogate dataset mimics the characteristics of the Cleveland dataset, all attributes of the synthetic data were compared on frequency plots to the Cleveland dataset (Figure 5). As an added comparison, the surrogate data were fitted to the general linear/logistic model used in the Synthpop data synthesis of the Cleveland 14 surrogate data and the resulting coefficients of the Cleveland 14 and the surrogate data set were compared and shown to overlap each other with a 95% confidence limit (Figure 2).

4.5 Stage 3

The objective of Stage 3 was to improve the accuracy, precision and recall results from Stage 2. We used a Perceptron Neural Network Model that was configured with 3 hidden layers and 3 dropout layers. Neural Network models, in general, produce higher accuracy measurements than traditional machine learning models; however, they have higher data volume training requirements [18]. To satisfy this, we generated a 60,000 record surrogate data set from the raw Cleveland dataset. The surrogate data set has 37 explanatory and one response variable and a shape of 38 X 60,000. To show that the Cleveland data set characteristics were preserved in the surrogate data set, we used the Synthpop compare function to plot frequency plots of all the data sets for comparison (Figure 5). In addition, the coefficient confidence interval was plotted to show that both the raw Cleveland dataset and the surrogate dataset coefficients have the same fit response to the general linear/logistic model used in Synthpop (Figure 3).

5 Results

In Stage 1, the results of our preliminary classification models (logistic regression, decision tree, and random forest) performed using the Cleveland 14 data set were consistent

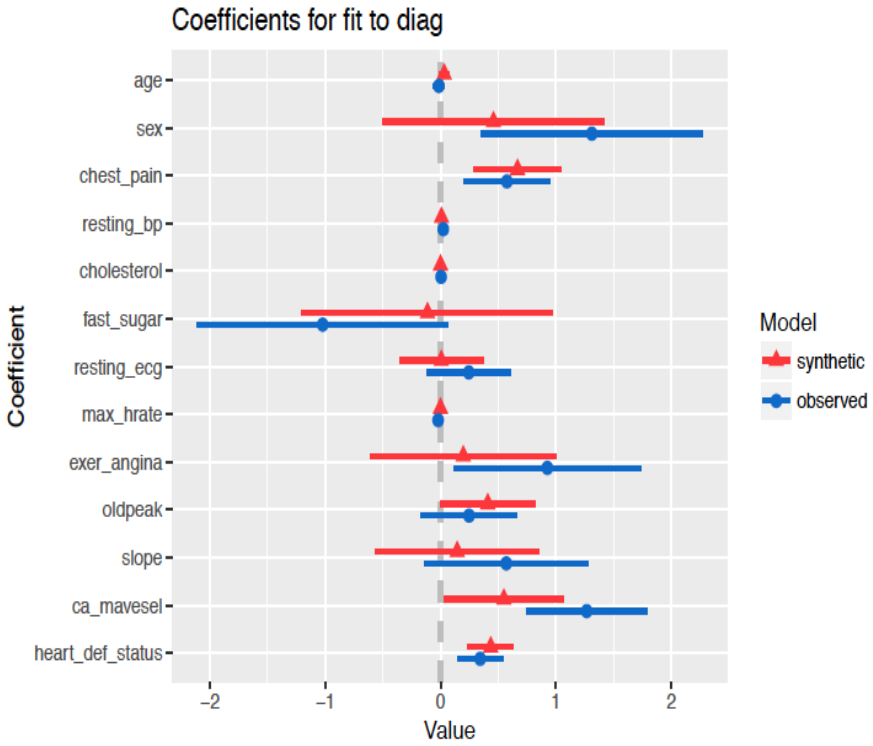


Fig. 2. Cleveland 14 Versus Surrogate Coefficient Confidence Interval Comparison

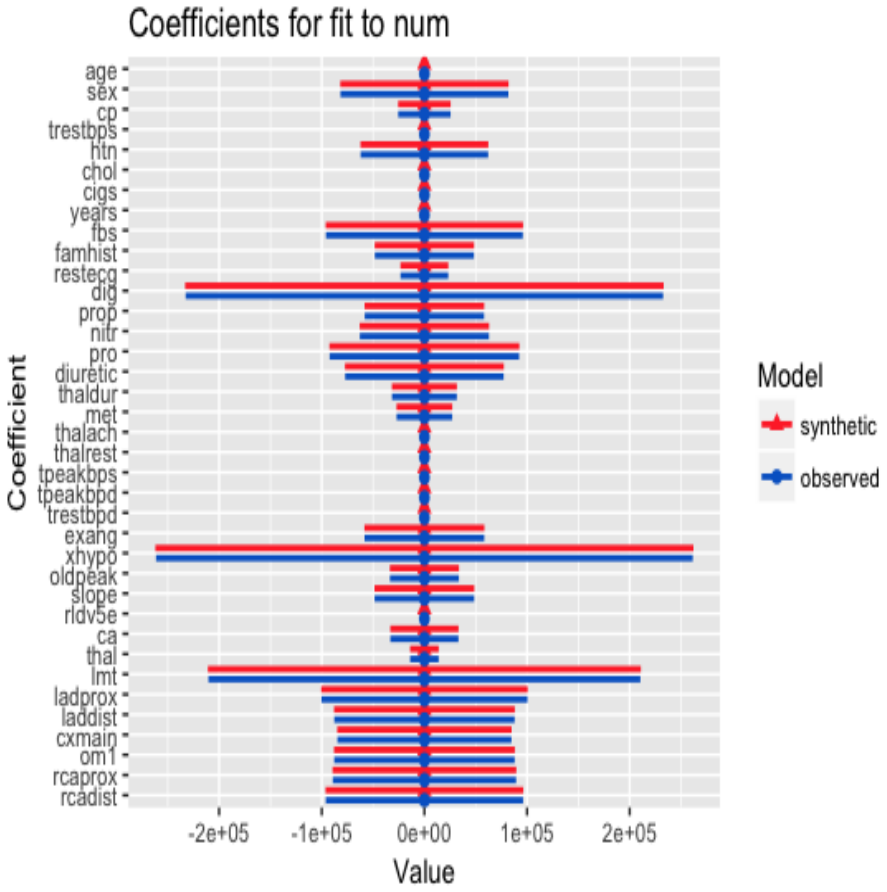


Fig. 3. Raw Cleveland Versus Surrogate Coefficient Confidence Interval Comparison

with results from previous research. Accuracy ranged from 83% - 88% using logistic regression, 74%-79% using a decision tree classifier, and 77%-81% using a random forest classifier.

Of the three, the logistic regression model produced superior accuracy prediction metrics - up to the 88 percent range. The recall metric score for logistic tegression has a 7 percent variation compared to that of the random forest model with a variation of 11 percent. The stability of the recall metric is of high interest since it is an indication of the false negative prediction rates. Figure 4 and Table 5 summarizes the accuracy, precision and recall results for the three models.

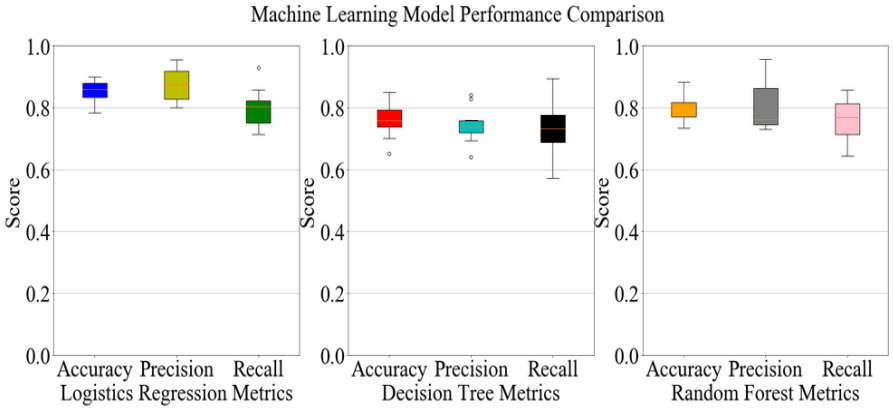


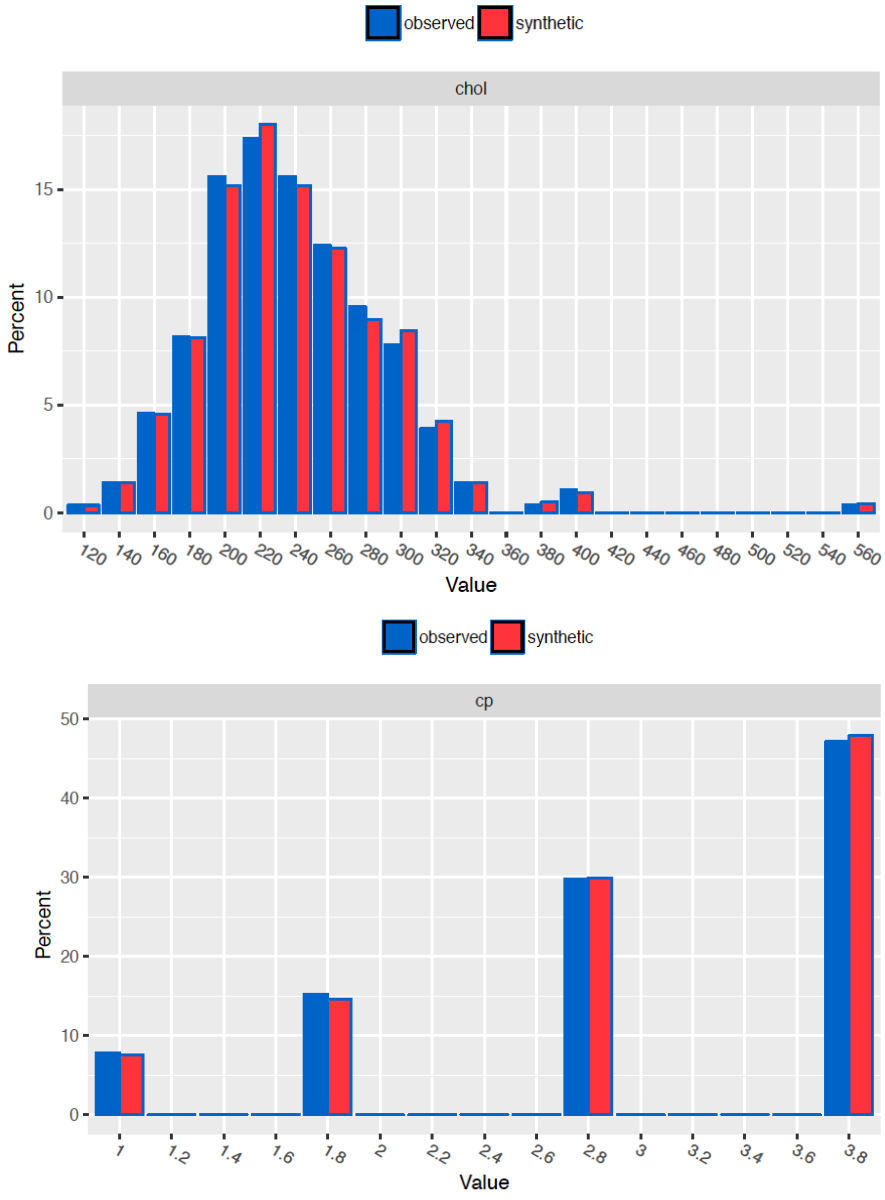
Fig. 4. Stage 1 Results: Range of accuracy, precision and recall values for 10-fold cross validation models for Logistic Regression, Decision Tree, and Random Forest.

Table 5. Stage 1 Machine Learning Metric Results

Machine Learning Model	Accuracy	Precision	Recall
Logistic Regression	83 - 88%	82 - 92%	75 - 82%
Decision Tree	74 - 79%	71 - 75%	69 - 77%
Random Forest	77 - 81%	74 - 84%	70 - 81%

The characteristics of the synthetic data that were generated using the original Cleveland 14 data set in stage 2 of our analysis shared very similar distributions compared to the original data. The distribution for each variable was reviewed and some examples are shown in Figure 5.

In stage 2, we applied the logistic regression model to the synthetic Cleveland 14 data in order to determine what the results would be using the same logistic regression model (as in stage 1) with the surrogate data. The results show that accuracy, precision


observed
synthetic
chol
Percent
0
5
10
15
120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500 520 540 560
Value
observed
synthetic
cp
Percent
0
10
20
30
40
50
1 1.2 1.4 1.6 1.8 2 2.2 2.4 2.6 2.8 3 3.2 3.4 3.6 3.8
Value

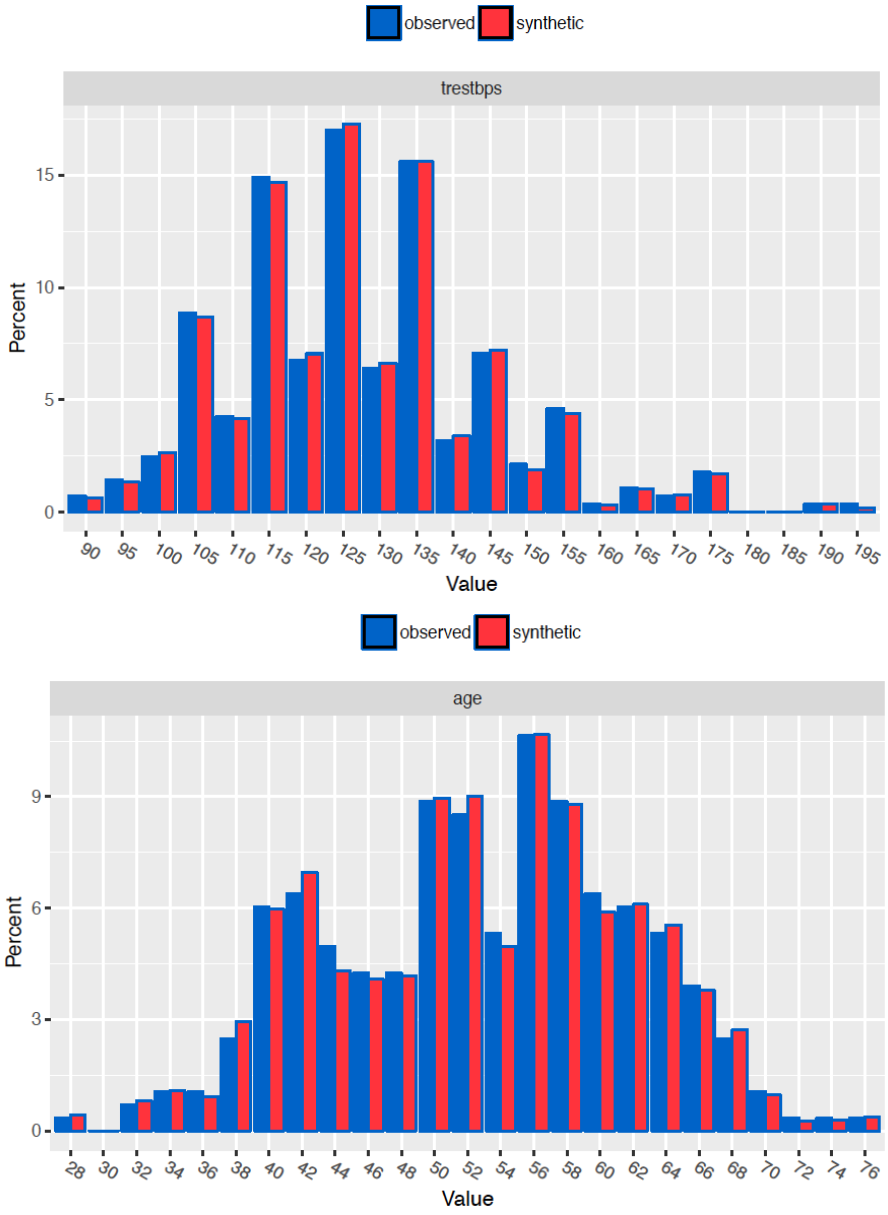


Fig. 5. Example Raw Cleveland to Surrogate Frequency Plots for Cholesterol, Chest Pain Type, Peak Exercise Blood Pressure and Age

and recall have similar levels between the original Cleveland 14 data in stage 1 versus the Cleveland 14 Surrogate data in Figure 6. The main difference is that the ranges are much narrower (within 2 percent) when using the surrogate data indicating a more stable performance of the logistic regression model when the data volume is significantly increased.

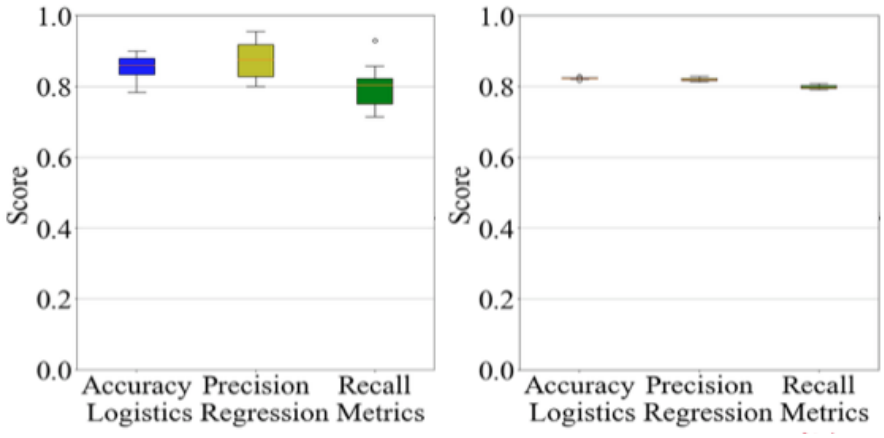


Fig. 6. Stage 2 results: Logistic Regression Model Accuracy, Precision and Recall Metrics for Stage 1 (left) compared to Stage 2 (right)

Finally in stage 3, after achieving improved stability with the application of the logistic regression model to the larger surrogate Cleveland 14 data set as compared to the original Cleveland 14 data set, we implemented a neural network model (Perceptron) using an expanded surrogate data set (60,000 samples) based on the raw Cleveland data set (38 attributes from 76, after data cleaning). Recall that the Cleveland 14 data set from the UCI Repository used in previous studies (also in stages 1 and 2) was also derived from the same raw Cleveland data set in this case. The performance of the Perceptron deep learning model shows the accuracy, precision and recall scores consistently at 96.7 percent with a 1 percent variation shown in Figure 7.

6 Analysis

As shown from the results of the traditional classification models, performance was consistent with what has been produced with prior studies. This outcome is consistent with what we expect from reviewing reproducible research and serves as a baseline for our additional analysis.

The primary focus of our analysis was to examine the synthetic data observations and compare them to the original observations from the Cleveland data set. We first looked at the response variable, which is the presence or absence of heart disease. For the presence of heart disease the Cleveland data set had 44.3 percent positive compared

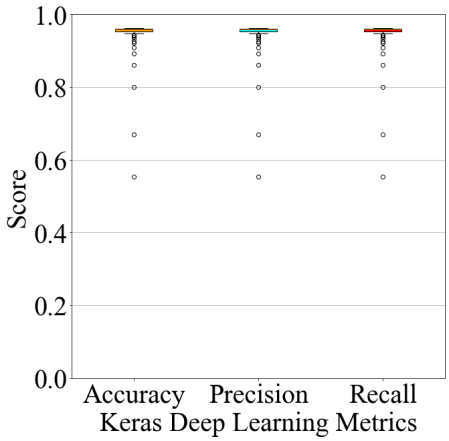


Fig. 7. Stage 3 results: Accuracy, Precision and Recall Metrics using Neural Network Perceptron Model with Surrogate Data.

to 55.7 percent negative. The surrogate data is similarly split with 43.8 percent positive compared to 56.2 percent negative.

In an effort to compare the mean distribution values, we calculated a 95 percent confidence interval for the proportion of individuals with heart disease and plotted the interval ranges in Figure 8. As expected, the variance for the synthetic data is very small due to the massive amount of observations. However, we find that the mean of the proportion for the surrogate data set is contained within the confidence interval for the proportion for the Cleveland data set.

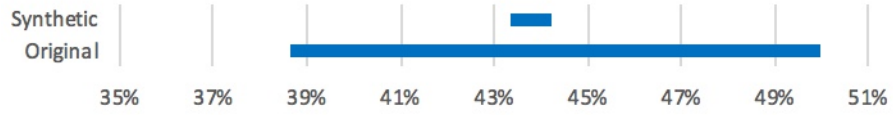


Fig. 8. 95 percent Confidence Interval of Mean Proportion of Heart Disease for Synthetic and Original Data Sets

Although we did not statistically test the values for the other variables we visually examine the distributions for each attribute. Those plots for a few of the more familiar attributes are shown in Figure 5. We are satisfied that the surrogate data set appropriately mimicked the original observations, therefore we used it to build a neural network model.

As shown in Figure 9, the neural network model built using the surrogate data set produces superior accuracy precision and recall compared to the traditional logistic regression models.

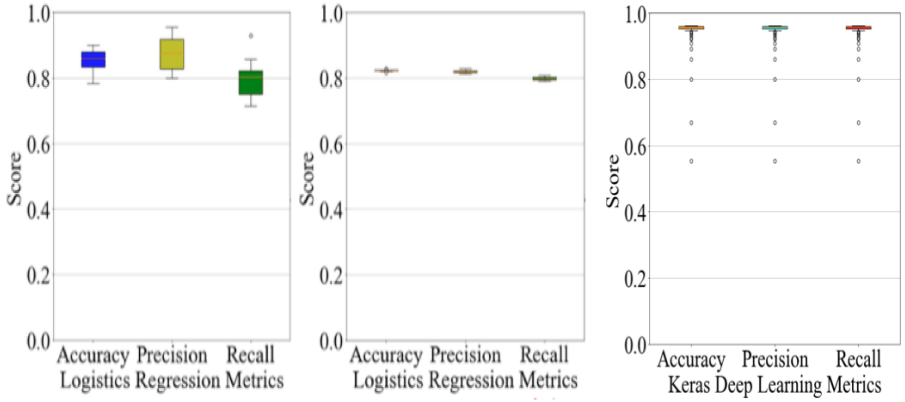


Fig. 9. Stage 1, 2 and 3 Results: Progression of improvement in Accuracy, Precision and Recall Metrics and Stability for Stage 1 (left), Stage 2 (middle) and Stage 3 (right). Stage 3 shows the highest scores for all 3 metrics at 96.7 percent with only 1 percent variation.

7 Ethics

Studies involving medical research require careful consideration of ethical concerns. Addressing those concerns often begins at the design stage of an experiment where a patient is required to give informed consent to participate in the study. During this stage, the patient is informed of the purpose of the study, the benefits and risks are explained and he is asked to voluntarily agree to participate.

However, even after a specific study is concluded, data gathered during the course of an experiment continues to live. Since much of the purpose of data mining is to identify previously unseen patterns, utilization of data collected in medical research may continue for many years and extend beyond original intents of a given study. Patients are not capable of giving specific consent for use of their information for unforeseen uses [19].

This motivates the need to protect patient privacy. Although many personally identifying elements are anonymized in medical research data sets, knowledge of the patient privacy protections allowed in the Health Insurance Portability and Accountability Act (HIPAA) of 1996 is critical to ensure compliance. HIPAA very explicitly provides for patient privacy protection by requiring de-identification of Protected Health Information (PHI). PHI describes any data element which could allow identification of the individual for which the health information pertains to. The following is the definition from the Code of Federal Regulations that defines health information subject to privacy protection [22].

> From 45 CFR 160.103
>
> Health information means any information, including genetic information, whether oral or recorded in any form or medium, that:

(1) Is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and

(2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual.

PHI must be removed from medical research data or masked to ensure patient privacy is protected under HIPAA. As stated in the description, this goes beyond anonymizing patient numbers and names. Any data element, or combination of data elements, that could reasonably be linked to an individual, including genetic data, is included in this measure.

To make data from medical research available, only the minimal necessary information should be disclosed. Although it can be difficult to determine the threshold for the minimally necessary information for each use, CFR 164 [23] contains an extensive listing of information elements that are potential identifiers and should be excluded from all data sets as shown in Table 6.

Table 6. Potential Identifying Elements

Potential Identifying Elements
Names
Postal Address
Telephone/Fax Numbers
Electronic Mail Addresses
Social Security Numbers
Medical Record Numbers
Health Plan Numbers
Patient Account Numbers
Certificate/License Numbers
Vehicle Identification / Serial Numbers
Device Identification Number
Internet Protocol Addresses
Biometric Identifiers
Full face photographic images
Web URLs

We view the opportunity to create surrogate data sets as a potential solution to minimize the risk of disclosing any potential individual identifying data related to medical research. Because the surrogate data is comprised of synthetic observations, there is no disclosure of any true patient identifying elements. The surrogate data sets would mimic the original observations. Depending upon the research purpose and analytic tools used, surrogate data sets of varying sizes can be generated. However, caution

should be taken to ensure that surrogate data sets are not used beyond research objectives and researchers should be transparent regarding the use of synthetic data.

8 Future Work

The use of surrogate data to train machine and deep learning models in this study appears to be an effective step towards improving heart disease prediction, but with limitations. The origin of the patient data used in this study was from the Cleveland, Ohio area and is not geographically diverse. From a global perspective, each geographical region has its own characteristic diet, lifestyle and availability of healthcare resources. According to the World Health Organization, in 2017, Turkmenistan had the highest death rate due to cardiovascular disease at 411.1 deaths per 100,000 while South Korea has the lowest at 30.76 [4]. Application of surrogate data in heart disease prediction using machine or deep learning techniques with patient data from geographically diverse sources would be interesting to explore in the future. By doing so, we can potentially discover data patterns that are not easily seen by using a geographically focused patient data set such as the Cleveland data set.

9 Conclusions

We find that the Synthpop package produces an adequate surrogate set of synthetic data which closely mimics the characteristics of the original observations. Our analysis shows, by comparison, that classification prediction outcomes from traditional machine learning models such as logistic regression, are reasonably similar whether the surrogate data set or the original observations are used. Variability of the prediction measurements are improved, in this case, due to the increased number of observations in the surrogate data set.

In addition, we improved heart disease prediction using surrogate data. Due to the large volume of synthetic observations that can be produced, the surrogate data are suitable for use with deep learning models such as ANN. Creating synthetic observations upon which the neural network can be trained and tested allows for an increase in classification prediction of nearly 16 percent.

These findings provide a basis for additional testing to be performed with other small clinical data sets. The use of surrogate data as a means to anonymize sensitive data and potential to improve classification prediction is a worthwhile area for additional exploration and research.

References

1. S. Aswal, N. J. Ahuja and Ritika, ”Experimental analysis of traditional classification algorithms on bio medical dtatasets,” 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2016, pp. 566-568.
2. Justin Collins, BS; Jordan Brown, BS; Christine Schammel, PhD; Kevin Hutson, PhD; and W. Jeffery Edenfield, MD: ”Meaningful Analysis of Small Data Sets: A Clinicians Guide”, Greenville Health System Proc.June 2017; 2 (1): 16-19

3. Shaikhina, Torgyn, and Natalia A. Khovanova. "Handling limited datasets with neural networks in medical applications: A small-data approach." Artificial intelligence in medicine 75 (2017): 51-63.
4. World health statistics 2017: Monitoring health for the SDGs, Sustainable Development Goals. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.
5. National Institute of Health. "Know the Differences: Cardiovascular Disease, Heart Disease, Coronary Heart Disease". Available at https://www.nhlbi.nih.gov
6. U.S. Department of Health and Human Services. "High Blood Cholesterol, What You Need To Know", NIH Publication No. 05-3290 Originally printed May 2001 Revised June 2005 Available at https://www.nhlbi.nih.gov/files/docs/public/heart/wyntk.pdf
7. Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, M. Essam Khalifa: Feature Analysis of Coronary Artery Heart Disease Data Sets, Procedia Computer Science, Vol. 65 (2015) pp459-468.
8. National Center for Health Statistics. Health, United States, 2016 With Chartbook on Long-term Trends in Health. Hyattsville, MD. 2017.
9. Gibbons, R. (2002). ACC/AHA 2002 Guideline Update for Exercise Testing: Summary Article: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1997 Exercise Testing Guidelines). Circulation, 106(14), pp.1883-1892.
10. Gibbons, R. (2003). ACC/AHA 2002 Guideline Update for the Management of Patients With Chronic Stable Angina–Summary Article: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on the Management of Patients With Chronic Stable Angina). Circulation, 107(1), pp.149-158.
11. Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) 2.02 (2010): 250-255
12. Jabbar, M. A., Priti Chandra, and B. L. Deekshatulu. "Cluster based association rule mining for heart attack prediction." Journal of Theoretical and Applied Information Technology 32.2 (2011): 196-201
13. Shouman, Mai, Tim Turner, and Rob Stocker. "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients." Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012
14. Andras Janosi, M.D., William Steinbrunn, M.D., Matthias Pfisterer, M.D., Robert Detrano, M.D., Ph.D. The UCI machine learning repository online. Available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease
15. Gadaras, I. and Mikhailov, L. (2009). An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artificial Intelligence in Medicine, 47(1), pp.25-41.
16. Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.
17. Nowok, Beata, Gillian M. Raab, and Chris Dibben. "synthpop: Bespoke creation of synthetic data in R." Journal of statistical software 74.11 (2016): 1-26.
18. Gulli, Antonio, and Sujit Pal. Deep Learning with Keras. Packt Publishing Ltd, 2017.
19. Mittelstadt, Brent Daniel, and Luciano Floridi. "The ethics of big data: current and foreseeable issues in biomedical contexts." Science and Engineering Ethics 22.2 (2016): 303-341.
20. Eggers, K., Ellenius, J., Dellborg, M., Groth, T., Oldgren, J., Swahn, E. and Lindahl, B. (2007). Artificial neural network algorithms for early diagnosis of acute myocardial infarction and prediction of infarct size in chest pain patients. International Journal of Cardiology, 114(3), pp.366-374.

21. Lee TH, Goldman L.: Evaluation of the Patient with Acute Chest Pain, New England Journal of Medicine. (2018).
22. HIPAA Privacy Rule. 45 CFR 160.103 2013.
23. Other requirements relating to uses and disclosures of protected health information. 45 CFR 164.514 2013.
24. R. Sumathi, E. Kirubakaran: "Enhanced Weighted K-means Clustering Based Risk Level Prediction For Coronary Heart Disease", ResearchGate, Volume 71 (2012), pp. 490-500. Available at https://www.researchgate.net
25. Hamid Reza Marateb, Sobhan Goudarzi: "A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system", ResearchGate, Volume 20 Journal of Research in Medical Sciences (2015), pp. 214-224. Available at https://www.researchgate.net
26. Beata Nowok, Gillian Raab and Chris Dibben: "Synthpop: Bespoke Creation of Synthetic Data in R", Journal of Statistical Software, Articles. Volume 74, pp. 1-26 (2016). Available at https://www.jstatsoft.org/v074/i11
27. Matthias Templ, Bernhard Meindl, Alexander Kowarik, Olivier Dupriez:"Simulation of Synthetic Complex Data: The R Package simPop", Journal of Statistical Software, Articles, Volume 79, pp. 1-38 (2017). Available at https://github.com/statistikat/simPop
28. A. C. Davison, D. V. Hinkley:"Bootstrap Methods and Their Applications", Cambridge University Press (1997) ISBN 0-521-57391-2. Available at http://statwww.epfl.ch/davison/BMA/