01 Jul 2019

# Comparative Analysis Of Feature Selection Methods To Identify Biomarkers In A Stroke-Related Dataset

Thomas Clifford

Justin Bruce

Tayo Obafemi-Ajayi
*Missouri University of Science and Technology*, towd2@mst.edu

John Matta



## Recommended Citation

T. Clifford et al., "Comparative Analysis Of Feature Selection Methods To Identify Biomarkers In A Stroke-Related Dataset," *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2019*, article no. 8791457, Institute of Electrical and Electronics Engineers, Jul 2019.
The definitive version is available at https://doi.org/10.1109/CIBCB.2019.8791457

# Comparative Analysis of Feature Selection Methods to Identify Biomarkers in a Stroke-Related Dataset

Thomas Clifford
*Computer Science Dept.*
*Southern Illinois University Edwardsville*
Edwardsville, IL
tcliffo@siue.edu

Justin Bruce
*Computer Science Dept.*
*Southern Illinois University Edwardsville*
Edwardsville, IL
jubruce@siue.edu

Tayo Obafemi-Ajayi
*Engineering Program*
*Missouri State University*
Springfield, MO
tayoobafemiajayi@missouristate.edu

John Matta
*Computer Science Dept.*
*Southern Illinois University Edwardsville*
Edwardsville, IL
jmatta@siue.edu

***Abstract*—This paper applies machine learning feature selection techniques to the REGARDS stroke-related dataset to identify health-related biomarkers. A data-driven methodological framework is presented to evaluate multiple feature selection methods. In applying the framework, three classifiers are chosen in conjunction with two wrappers, and their performance with diverse classification targets such as *Current Smoker*, *Current Alcohol Use*, and *Deceased* is evaluated. The performance across logistic regression, random forest and naïve Bayes classifier methods, as quantified by the ROC Area Under Curve metric and selected features, was similar. However, significant differences were observed in running time. Performance of the selected features was also evaluated based on the accuracy of a prediction model generated using a multi-layer perceptron (MLP) classifier.**

***Index Terms*—machine learning, feature selection, classification**

## I. Introduction

The increasing availability of high-dimension medical data has made use of feature selection methods imperative in identifying relevant features that could be potential biomarkers. Diverse classifiers are available for feature selection, such as random forest, naïve Bayes, logistic regression, and k-nearest neighbors [1]. These classifiers are used with wrappers such as step forward, step backwards and exhaustive search [2] to identify a set of features useful for training a classifier to build an effective prediction model. The choice of feature selection algorithm as well as wrapper methods impacts running time. Some algorithms run quickly, such as naïve Bayes and logistic regression. Others are orders of magnitude slower, such as random forest and k-nearest neighbors.

Stroke is one of the leading causes of death and serious long-term disability in the United States [3]. REasons for Geographic and Racial Differences in Stroke (REGARDS) is a national, population-based, longitudinal study of 30,239 individuals $\geq$45 years old with data collected between 2003 and 2007 [4]. The REGARDS dataset is notable as it includes a vast amount of diverse information, including blood test results and other health indicators, as well as demographic information such as household income, education level, and smoking status. The REGARDS data was originally intended for determining factors that increase the risk of stroke in relation to racial differences and geographical location. Its large number of samples and varied attributes also make it attractive for studying other public health factors. It embodies a rich set of health information that can be mined with varied machine learning techniques to obtain new knowledge and address novel questions.

In this work, we utilize the REGARDS data as a testbed for analysis of various feature selection methods based on multiple classification target attributes such as stroke, death indicator, smoking status, drinking status, use of NSAIDs, and diabetes. We present a data-driven framework for comparison of the effectiveness of three classification methods used with two wrappers. We also examine the overall running time for each method and analyze the relevance of the features selected. The performance metric used for evaluation of features selected is the ROC area under curve (AUC) score along with the accuracy (as quantified by F-measure) of a prediction model based on the selected feature sets. A key application of feature selection on real datasets is that the discriminant features selected could be applied as potential biomarkers for the classification target attribute (or disease) in medical data analysis. Hence, in this work, we conduct a qualitative analysis of the features selected by each method to determine their usefulness as a potential biomarker for the related health factor (as specified by the target attribute).

The paper is organized as follows. In section II we review related work done on application of machine learning techniques to the REGARDS data. In Section III we describe the data-driven framework, while the results are presented in Section IV. Section V discusses the results and the conclusion is presented in Section VI.

978-1-7281-1462-0/19/$31.00 ©2019 IEEE

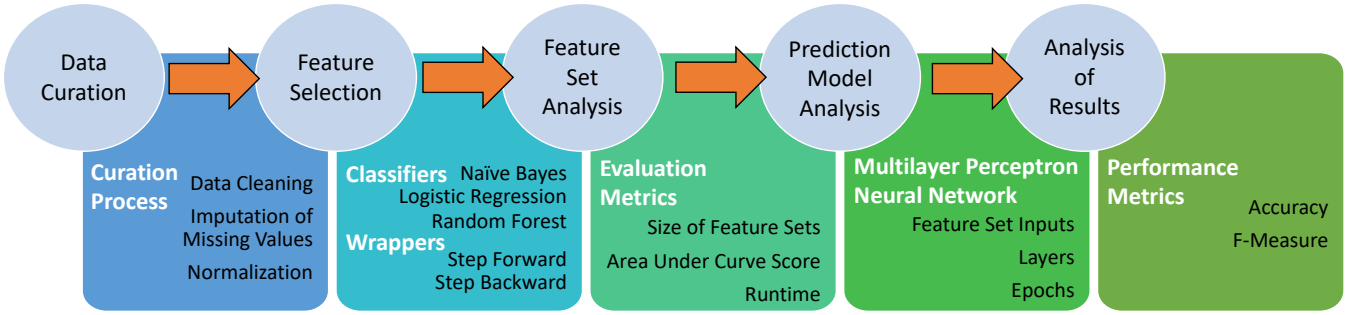


Fig. 1: Data-driven methodological framework to evaluate multiple feature selection methods.

## II. Related Work

Varied statistical analysis methods [5]–[9] have been applied to the REGARDS dataset to infer knowledge and association of multiple health factors related to stroke and other diseases. For example, Meshcia et al. [5] analyzed the data to determine whether African Americans are less likely than whites to be aware of having atrial fibrillation, or to be treated with the blood thinner Warfarin. Other studies examine the association between urinary albumin excretion and coronary heart disease in black vs white adults [6], and racial differences in the impact of elevated blood pressure on stroke risk [10]. Several studies examine kidney disease [7]–[9]. Lifestyle issues are also studied, such as caregiver strain [11], obesity management [12], and the risk of sepsis [13].

There are relatively few studies [14]–[16] in which machine learning techniques are used with the REGARDS data, as we present in this paper. Prineas et al. in [14] applied logistic regression to analyze the sensitivity of the various approaches utilized in the data to detect atrial fibrillation (AF). They concluded that the association of AF with residence in the Stroke Belt and black ethnicity was inversely related to the sensitivity of the method used to detect AF. Hence this could account for lower AF prevalence estimates in regions with higher stroke rates. Brown et al. in [15] applied logistic regression on the dataset to classify participants according to future risk of cardiovascular disease. O'Neal et al. in [16] also applied logistic regression to determine a link between environmental tobacco smoke exposure and AF.

In this work, we apply the REGARDS data to analyze multiple feature selection methods using multiple classification target attributes as well as quantitatively analyze resulting features to infer new knowledge.

## III. Methods

The objective of this work is to systematically evaluate feature selection classifiers with respect to both time to execute and accuracy of results. The framework for the approach is illustrated in Figure 1 and described below.

### A. Data Curation

The REGARDS dataset was collected according to rigorous standards and is widely regarded as being of high quality. The sample population consisted of 30,082 individuals and 79 features. There were missing values among some of the attributes. We applied data imputation techniques to resolve the missing values or discarded those samples or features all together. Among the participants, 105 ($< 0.4\%$) were missing blood-test data, which was deemed potentially important. These 105 samples were excluded from the analysis. Eighteen of the features were discarded due to a significant amount of missing data (30% or more). These included *White Blood Cell Count* with 10,267 missing and *Platelet Count* with 10,792 missing. The remaining features still had a few missing values, $\leq 1.7\%$ of the 29,977 remaining samples. Missing values for these features were imputed using the mean value for each feature across the sample population.

Regarding data type format, some of the features had nominal binary values, such as the self-reported attributes of *CAD Aneurysm*, *Diabetes*, *DVT (Deep Vein Thrombosis)*, and *Hypertension*. These attributes were left as *Yes-No* values. Six features containing multi-value attributes were processed using one-hot encoding. In one-hot encoding, categorical attributes that do not have an ordinal relationship are split into separate binary attributes. For example, the attribute *Alcohol Use* was split into three binary categories: *Current Alcohol Use*, *Alcohol Use in Past* and *Alcohol Use Never*. One-hot encoding added 15 new features, increasing the number of features to 76.

Another important phase of the data curation process is to verify the entries for correctness. For example, entries for the features *Falls Per Year* and *Heart Rate* were found to contain illogical (possibly placeholder) values of 888 or 999. A total of 130 possibly inaccurate entries ($<0.5\%$) were replaced by the mean value of the associated feature. All numeric attributes such as *Age*, *Height* and *Total Cholesterol* were normalized using min-max normalization.

Lastly, the features were pruned to discard redundant or irrelevant features by applying a pairwise correlation filter using Pearson coefficient. The features (5) with a correlation greater than 80% were discarded. These included *Smoke 100 Cigs*, *Weight Kg*, *Creatinine Serum*, *SR Hypertension*, and *LDL Cholesterol*.

### B. Feature Selection Methods

Feature selection is used to select a set of relevant or discriminant features for training a learning model with good performance that generalizes well [17]. This study compares

three classifiers: naïve Bayes [18], logistic regression [19], and random forest [20]. All are implemented as part of the Scikit-Learn Python package [21]. Naïve Bayes is a simple classification technique in which all attributes are assumed to be uncorrelated, and a probability model is used to assign instance probabilities to possible outcomes. Naïve Bayes has been proven effective in medical diagnosis applications [18]. Logistic Regression uses a linear model for binary classification. Instead of minimizing a linear cost function, as with linear regression, it minimizes a sigmoid function [22]. Random forest is a non-linear technique which constructs a classifier consisting of a collection of tree-structured classifiers where each tree casts a vote for the most popular class. Random forest classifiers and logistic regression models have been widely studied and applied on a variety of medical data [23], [24].

Two types of wrapper methods, the greedy step forward and step backwards algorithms [2], were used to create baseline performance data on 70 features, and then to select the $k$ features ($k \in [5, 10, 15, 20]$), per target, that yield the best performance for each classifier. In step forward, the performance of the classifier is evaluated with each feature, and the best performing feature is chosen as the first member of the feature set. This becomes the base feature, with which all other features are combined as a possible set. The feature that performs best in combination is added to the set, and the process continues until the feature set reaches the desired size. Step backwards is also a greedy algorithm. Unlike step forward, it is initialized with a set of all features. The performance of the classifier is evaluated by removing a single feature per iteration. The best performing set is kept, and features are eliminated until the feature set reaches the desired size. The wrapper methods are implemented as part of the Mlextend [25] Python package.

ROC area under curve (AUC) score is applied as the performance metric for the feature selection methods. Each algorithm was trained on the features that were chosen using each wrapper method, and the performance of the algorithm was evaluated on training (80%) and testing (20%) sets for each target feature. The sets were not stratified and were generated using the Scikit-Learn Python package [21]. Given that there were three primary classifiers, two wrapper methods, training and testing for sets of 5, 10, 15, 20 and 70 features over 6 possible output attributes implies that there are 180 different results obtained. The run-times of the algorithms (time to select a single feature) are shown in Figure 2.

TABLE I:
DISTRIBUTION OF CLASSIFICATION TARGET FEATURES

| Classification Target Feature | % Yes |
|---|---|
| Current Alcohol Use | 51.6 |
| Diabetic | 22.6 |
| Current Smoker | 14.6 |
| Regular Use of NSAIDs | 14.2 |
| Deceased | 13.5 |
| Self-Reported Stroke | 6.3 |

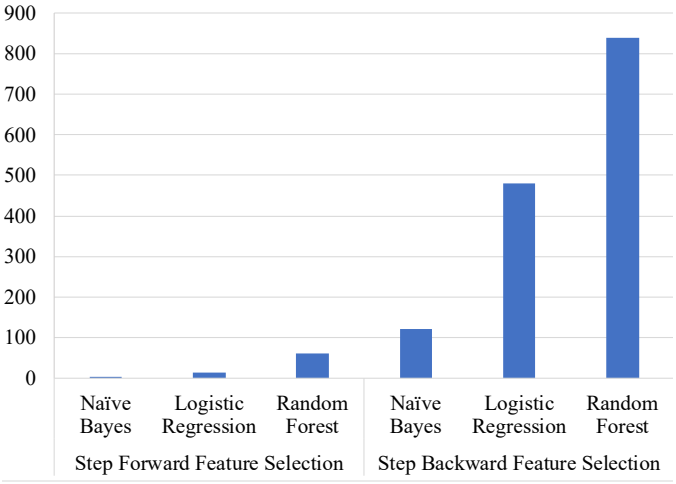


Fig. 2: Average Time to Select a Feature (Seconds)

### C. Classification Target Features

Given the goal of identifying potential biomarkers related to various health factors by mining the REGARDS data, six features were selected as classification target attributes. The selected features are binary (*Yes/No* responses) as illustrated in Table I. A classification model is implemented for each of the six target features: *Current Alcohol Use*, *Current Smoker*, *Self-Reported Diabetes*, *Regular Use of NSAIDs*, *Self-Reported Stroke*, and *Deceased*.

It is important to examine the balance of the dataset with respect to the target attribute as skewness of label could influence the outcome of the feature selection. *Current Alcohol Use* is the most balanced target, with roughly half the samples reporting *Yes* (Table I). Interestingly, though REGARDS is a Stroke targeted dataset, only 6.3% had a positive response (as denoted by Yes) for the *Self-Reported Stroke* attribute.

### D. Prediction Model Analysis

The top 5 features selected for each classification target were used to train a multi-layer perceptron (MLP) neural network classifier in Weka [26]. To reduce execution time of the MLP, continuous attributes were converted to nominal values as follows. *HDL Cholesterol*, *Urinary Albumin/Creatinine Ratio*, *Insulin*, and *Body Mass Index* were discretized into 10 bins by reducing the accuracy of the continuous variable. *Cholesterol*, *Heart Rate*, *Glucose*, and *Urinary Albumin* were discretized into low, medium and high based on their interpretation in established medical practice [27], [28]. *Cystatin C* was discretized according to guidelines given in [29].

The MLP model was implemented using a learning rate of 0.3, a momentum of 0.2, 1 hidden layer, in 10-fold cross-validation test mode with 500 epochs per fold. The performance of the model is demonstrated using F-measure (also called $F_1$ score). This metric is used because it takes the unbalanced nature of the classification targets into account. F-measure is defined, along with precision and recall, as follows. For a given data set of size $N$ and its corresponding classifi-

cation results, let $TP$ denote the true positives, $TN$ the true negatives, $FP$ the false positives, and $FN$ the false negatives. The measure *precision* defines the percentage of positives that are correctly classified i.e. $precision = \frac{TP}{TP+FP}$. *Recall* denotes the percentage of samples classified as true positives, given by $recall = \frac{TP}{TP+FN}$. F-measure is the harmonic mean of precision and recall: $F_1 = 2 \times \frac{precision \times recall}{precision + recall}$.

## IV. Results

### A. Feature Selection Outcome

The feature selection algorithms yielded sets of $k$ features for each classification target attribute. Overall, there was a great deal of overlap between features selected by the different methods. Table II illustrates the results for $k = 5$ by classification target attribute. For each target, all attributes returned by any of the algorithms are listed. Check marks indicate that the algorithm's results included the marked feature. Attributes are ordered by most-to-least selected. For example, for the target *Current Alcohol Use*, the features *Alcohol Use Never*, *Alcohol Use in Past* and *Number of Falls in Past Year* were chosen by all combinations of feature selection methods. For each target attribute there are a set of features that were consistently chosen regardless of the feature selection method and wrapper applied.

AUC scores for testing and training sets using all combinations of classifiers and wrappers are shown in Table III. For each target attribute, results are shown for sets of size $k$ = 5, 10, 15 and 20. The baseline performance, based on using all 70 features, is also included for comparison. In general the testing and training values were similar, which could imply that the model is not overfitting. The average difference between testing and training scores over all results is 0.4%, while the maximum is 2.72%. Only 26 of 144 comparisons had a difference of over 1%.

According to Table III, the AUC scores for *Current Alcohol Use* and *Current Smoker* for all cases was at least 99%. This suggests that some of the features could be directly correlated with the classification target attribute (outcome measure). Thus, the model might not be learning since it is biased by those features. Upon further analysis of the features for classification target attribute *Alcohol use*, the top features selected were *Alcohol Use Never*, *Alcohol Use in Past*, and *Alcoholic Beverages Per Week*. Intuitively, these features are almost identical with the outcome measure *Alcohol use*, hence the underlying reason for the superior performance of the classifier. For the target attribute *Current Smoker*, the related features were *Never Smoked*, *Smoked in Past*, and *Pack Years Smoked*. To eliminate the bias introduced by these features, the experiments were repeated with these features excluded (Table V, Current Smoker[2] results). The feature *Smoked 100 Cigarettes*, indicating that a person had smoked 100 cigarettes in their lifetime, had previously been removed during the correlation filter stage (see section III-A). Removal of the highly biased features resulted in this feature passing through the filter stage (unlike before), as shown in Table IV. It was

TABLE II:
Features Selected by Target Attribute

| | | Step Forward | | | Step Backward | | |
|---|---|---|---|---|---|---|---|
| | | Naïve Bayes | Logistic Regression | Random Forest | Naïve Bayes | Logistic Regression | Random Forest |
| Target | Selected Features | | | | | | |
| Current Alcohol Use | Alcohol Use Never | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Alcohol Use in Past | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Number of Falls in Past Year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Age | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Alcoholic Beverages Per Week | ✓ | ✓ | ✓ | ✓ | | |
| | SR* DVT | | | | | | ✓ |
| | Gender | | | | | | ✓ |
| | Race | | | | | ✓ | |
| Deceased | Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Cystatin C | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | SR General Health | ✓ | ✓ | ✓ | | ✓ | ✓ |
| | Gender | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Heart Rate | | | | | ✓ | ✓ |
| | Currently smoking | ✓ | ✓ | | | | |
| | AFib ECG | | | ✓ | | | |
| | Body Mass Index | | | | | | ✓ |
| | CAD Aneurysm | | | ✓ | | | |
| | PCS Health Score | | | | ✓ | | |
| | Never Smoked | | | | ✓ | | |
| Diabetic | Glucose | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | SR General Health | | ✓ | ✓ | ✓ | ✓ | |
| | Level of Insulin | ✓ | | ✓ | ✓ | | ✓ |
| | Use of Insulin | ✓ | ✓ | | ✓ | ✓ | |
| | Cholesterol | | ✓ | | ✓ | ✓ | |
| | SR Use of Hyper Meds | | ✓ | | | ✓ | |
| | Urinary Albumin/Creatinine Ratio | ✓ | | | | | |
| | C Reactive Protein | | | ✓ | | | |
| | Graduate College Education | | | ✓ | | | |
| | Undergrad College Education | ✓ | | | | | |
| | Heart Rate | | | | | | ✓ |
| | PCS Health Score | | | | | | ✓ |
| | Waist Measurement | | | | | | ✓ |
| Regular Use of NSAIDs | Race | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | PCS Health Score | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Gender | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Alcohol Use Never | ✓ | | ✓ | ✓ | | |
| | Body Mass Index | | ✓ | | | ✓ | ✓ |
| | Age | ✓ | | | ✓ | | |
| | Cystatin C | | ✓ | | | ✓ | |
| | Urinary Albumin/Creatinine Ratio | | | | | | ✓ |
| | SR General Health | | | ✓ | | | |
| | Sleep Interrupt | | | ✓ | | | |
| | Self-Reported Stroke | | | | | | ✓ |
| Current Smoker | Never Smoked | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Smoked in Past | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Income Refused | ✓ | ✓ | ✓ | ✓ | | |
| | Heart Rate | ✓ | | | ✓ | ✓ | |
| | Body Mass Index | ✓ | | | ✓ | | |
| | Age | | | | | | ✓ |
| | C Reactive Protein | | ✓ | | | | |
| | Graduate College Education | | | ✓ | | | |
| | Exercise 1 to 3 Times / Week | | | | | | ✓ |
| | Income Between 20k-34k | | | | | ✓ | |
| | Income Less Than 20k | | | | | ✓ | |
| | Insurance | | ✓ | | | | |
| | Regular Use of Aspirin | | | ✓ | | | |
| | Waist Measurement | | | | | | ✓ |
| Self-Reported Stroke | Deceased | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | SR Use of Hyper Meds | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Regular Use of Aspirin | | ✓ | ✓ | | ✓ | ✓ |
| | PCS Health Score | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | SR TIA | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | SR General Health | ✓ | | ✓ | ✓ | | |
| | Urinary Albumin | | | | | | ✓ |
| | Number of Falls in Past Year | | | | | | ✓ |

*SR indicates Self-Reported.

TABLE III:
PERFORMANCE QUANTIFIED BY ROC AREA UNDER CURVE (AUC) COMPARING TRAINING AND TESTING SETS FOR ALL COMBINATIONS OF CLASSIFIERS AND WRAPPERS.
$k$ INDICATES SIZE OF FEATURE SET. VALUES ARE PERCENTAGES.

| Target Feature | k | Step Forward Feature Selection | | | | | | Step Backward Feature Selection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Naïve Bayes | | Logistic Reg. | | Random Forest | | Naïve Bayes | | Logistic Reg. | | Random Forest | |
| | | train | test | train | test | train | test | train | test | train | test | train | test |
| Current Alcohol Use | 5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 10 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 15 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 20 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 70 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 |
| Deceased | 5 | 0.80 | 0.80 | 0.80 | 0.80 | 0.77 | 0.77 | 0.80 | 0.79 | 0.80 | 0.80 | 0.79 | 0.79 |
| | 10 | 0.81 | 0.81 | 0.81 | 0.82 | 0.78 | 0.79 | 0.81 | 0.80 | 0.82 | 0.82 | 0.80 | 0.80 |
| | 15 | 0.81 | 0.81 | 0.82 | 0.82 | 0.81 | 0.81 | 0.81 | 0.80 | 0.82 | 0.82 | 0.80 | 0.80 |
| | 20 | 0.81 | 0.81 | 0.83 | 0.82 | 0.80 | 0.81 | 0.81 | 0.80 | 0.83 | 0.82 | 0.81 | 0.80 |
| | 70 | 0.76 | 0.77 | 0.83 | 0.82 | 0.80 | 0.80 | 0.76 | 0.77 | 0.83 | 0.82 | 0.80 | 0.80 |
| Diabetic | 5 | 0.88 | 0.87 | 0.86 | 0.87 | 0.92 | 0.92 | 0.87 | 0.88 | 0.86 | 0.87 | 0.92 | 0.92 |
| | 10 | 0.87 | 0.87 | 0.88 | 0.88 | 0.92 | 0.92 | 0.88 | 0.88 | 0.88 | 0.88 | 0.92 | 0.92 |
| | 15 | 0.88 | 0.88 | 0.88 | 0.89 | 0.91 | 0.91 | 0.88 | 0.88 | 0.88 | 0.89 | 0.92 | 0.92 |
| | 20 | 0.88 | 0.88 | 0.89 | 0.89 | 0.91 | 0.90 | 0.88 | 0.88 | 0.89 | 0.89 | 0.92 | 0.91 |
| | 70 | 0.82 | 0.82 | 0.89 | 0.89 | 0.91 | 0.90 | 0.82 | 0.82 | 0.89 | 0.89 | 0.91 | 0.90 |
| Regular Use of NSAIDs | 5 | 0.65 | 0.64 | 0.66 | 0.65 | 0.62 | 0.60 | 0.65 | 0.64 | 0.66 | 0.65 | 0.66 | 0.65 |
| | 10 | 0.66 | 0.65 | 0.67 | 0.66 | 0.62 | 0.60 | 0.66 | 0.65 | 0.67 | 0.65 | 0.67 | 0.65 |
| | 15 | 0.66 | 0.65 | 0.68 | 0.66 | 0.62 | 0.60 | 0.66 | 0.65 | 0.68 | 0.66 | 0.67 | 0.65 |
| | 20 | 0.66 | 0.66 | 0.68 | 0.66 | 0.63 | 0.60 | 0.66 | 0.65 | 0.68 | 0.66 | 0.67 | 0.65 |
| | 70 | 0.63 | 0.62 | 0.68 | 0.66 | 0.67 | 0.66 | 0.63 | 0.62 | 0.68 | 0.66 | 0.67 | 0.66 |
| Current Smoker | 5 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | 10 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | 15 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | 20 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | 70 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Reported Stroke | 5 | 0.75 | 0.74 | 0.76 | 0.76 | 0.74 | 0.73 | 0.75 | 0.74 | 0.76 | 0.76 | 0.75 | 0.74 |
| | 10 | 0.78 | 0.77 | 0.78 | 0.78 | 0.75 | 0.74 | 0.78 | 0.77 | 0.78 | 0.78 | 0.76 | 0.75 |
| | 15 | 0.79 | 0.78 | 0.79 | 0.78 | 0.76 | 0.75 | 0.79 | 0.78 | 0.79 | 0.79 | 0.76 | 0.75 |
| | 20 | 0.79 | 0.78 | 0.80 | 0.79 | 0.76 | 0.75 | 0.79 | 0.78 | 0.80 | 0.79 | 0.77 | 0.75 |
| | 70 | 0.76 | 0.78 | 0.80 | 0.79 | 0.78 | 0.77 | 0.76 | 0.78 | 0.80 | 0.79 | 0.78 | 0.77 |

chosen by all classifier and wrapper combinations during the feature selection process. To eliminate the bias introduced by this feature, the feature selection was re-run a third time excluding this feature. Results are shown in Table V, Current Smoker[3].

For the *Current Alcohol Use* classification target, removal of alcohol-related attributes reduced the performance from 100% to an average of 70%. For the *Current Smoker* target, performance degraded from 100% to 80%, and removal of additional features further reduced performance to an average of 74%. A summary of the features obtained after pruning the feature space based on high degree of similarity to the target attributes is presented in Table IV. There is a greater degree of consistency among the features obtained across all the methods.

### *B. Outcome of Prediction Model Analysis using Selected Features*

The quality of the various feature sets obtained was further evaluated using a prediction model: a multi-layer perceptron classifier (see Section III-D). Results for feature sets of size $k = 5$ are shown in Table VI. Though the step backwards algorithm has a considerably longer runtime than step forward, the performance of the two algorithms, as quantified by the F-measure scores, is very similar (5% or less difference). In the two cases where F-measures varied most, the best accuracy occurred with the faster step forward algorithm.

The best AUC scores were for the targets *Diabetic* with 89.6%, and *Deceased* with 80.5%. *Diabetic* achieved the highest F-measure of any classification target with the MLP, averaging 0.86 F-measure with both step forward and step backwards. *Deceased* had slightly lower F-measures at 0.83

TABLE IV:
FEATURES SELECTED BY TARGET ATTRIBUTE USING ADJUSTED DATASETS

| | | Naïve Bayes | Logistic Regression | Random Forest | Naïve Bayes | Logistic Regression | Random Forest |
|---|---|---|---|---|---|---|---|
| Target | Selected Features | Step Forward | | | Step Backward | | |
| Current Alcohol Use[1] | Race | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Income Greater Than 75k | ✓ | ✓ | ✓ | ✓ | | ✓ |
| | Never Smoked | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Graduate College Education | ✓ | ✓ | ✓ | | ✓ | |
| | PCS Health Score | ✓ | ✓ | | ✓ | ✓ | |
| | HDL Cholesterol | | | | | ✓ | ✓ |
| | Age | | | | | | ✓ |
| | Less Than High School Education | | | | ✓ | | |
| | SR* General Health | | | ✓ | | | |
| | Gender | | | | | | ✓ |
| Current Smoker[2] | Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Smoked at Least 100 Cigarettes | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Heart Rate | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Body Mass Index | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Income Less Than 20k | | ✓ | ✓ | | ✓ | |
| | SR General Health | ✓ | | | ✓ | | |
| | Cystatin C | | | | | | ✓ |
| | Income Between 35k-74k | | | ✓ | | | |
| | Income Greater Than 75k | | | ✓ | | | |
| Current Smoker[3] | Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Heart Rate | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Body Mass Index | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Alcohol Use Never | ✓ | ✓ | ✓ | ✓ | | |
| | Graduate College Education | ✓ | | ✓ | | ✓ | ✓ |
| | SR General Health | | ✓ | | ✓ | ✓ | |
| | C Reactive Protein | | | | | | ✓ |
| | Income Between 20k-34k | | | ✓ | | | |
| | Income Less Than 20k | | | ✓ | | | |

[1]Removed Alcohol Use Never, Alcohol Use in Past, Alcoholic Bev. / Wk.
[2]Removed Smoke Never, Smoke Past, Pack Years.
[3]Removed Smoke Never, Smoke Past, Pack Years, Smoke 100 Cigs.
*SR indicates Self-Reported

for step forward and 0.81 average for step backwards.

In the case of *Self-Reported Stroke*, the AUC score averaged 77%, but the MLPs did not identify any *Yes* instances, except for the MLP created with features chosen by step backwards random forest. This led to undefined F-measure scores (shown as *n/a* in Table VI). *Regular use of NSAIDs* had the lowest AUC score with an average of 67%, but higher F-measure scores (average of 0.79). However, for the step forward random forest model, no instances were correctly classified as *True*, thus resulting in an undefined F-measure (n/a).

## V. DISCUSSION

Some interesting observations can be made from the results obtained. Intuitively, the presence of features that are closely related to the classification target biases the results to a large degree. For experiments with *Current Alcohol Use* as the classification target, removing closely related features such as *Alcoholic Beverages Per Week*, *No Alcohol Use Ever* and *Use of Alcohol in Past* resulted in a change of features selected. Initially, *SR DVT* and *Number of Falls In Past Year* along with the highly similar features listed above were selected. However, after pruning, new discriminant features selected included *Education*, *Income* and *PCS* health scores. For *Current Smoker*, only 14.6% of participants were smokers. After filtering out smoking-related attributes, the total union of features selected among all classifier types decreased from 14 to 9. The discriminant features selected included *Age* (chosen by all 6 models); *Heart Rate* and *Body Mass Index* (chosen by 5 models); and *Alcohol Use Never* and *Graduate College Education* (chosen by 4 models). There is a strong overlap among the features obtained across all the methods.

There is a significant difference in execution time among the different classifiers and wrappers. The step forward wrapper took only seconds for naïve Bayes, increasing to a minute per feature for random forest. Step backwards had the longest execution time, taking almost 15 minutes to select a feature using random forest.

The AUC scores for the chosen feature sets are highly consistent across the number of features chosen, $k$. The average change in AUC scores by increasing the selected feature set from $k = 5$ to 20 was only 1.44%. The largest change in scores was for the *Deceased* attribute training set, which went from 76.55% to 80.38%, an improvement of only 3.83%. Generally, the performance of the minimal 5-feature set compares favorably to the 70-feature baseline, with the largest AUC difference being 6% in the case of step forward random forest for *Regular Use of NSAIDs*. Although AUC scores were highly similar across classifiers, logistic regression seemed to be the best performer, with 81.4% average score using the step forward wrapper. Interestingly, logistic regression with the much slower step backward algorithm averaged a slightly worse score of 81.3%. By contrast the lowest AUC was for random forest step forward with an overall average of 79.2%.

One of the key methods used to evaluate the results of a feature selection process is to utilize the features to construct a classifier and then observe its performance as a prediction model using only those features. In this work, the classifier applied was the MLP neural network. The highest performing prediction model was obtained for the *Diabetic* classification target attribute. That MLP achieved an average F-measure of 0.86. This correlates well with the AUC scores obtained for this target using feature selection, which were also the highest of any classification target. AUC scores and MLP F-measures were second highest in both cases for *Deceased*.

The prediction model performed relatively well (0.80 F-measure) on the *Current Smoker* target attribute. The result was unanticipated, given the low average AUC score of 74% after excluding features that were highly similar to the target attribute. *Current Smoker* status was classified using features such as *Age*, *Heart Rate*, *BMI*, *Alcohol Use* and *College Education*. *Self-Reported Stroke* had AUC scores that were higher than those for *Current Smoker*. However, the prediction model analysis outcome was poor, as shown in section IV-B. This indicates that the features chosen were probably not discriminative.

TABLE V:
ROC-AUC SCORES AFTER REMOVING RELATED FEATURES
$k$ INDICATES SIZE OF FEATURE SET. VALUES ARE PERCENTAGES.

| | | Step Forward Feature Selection | | | | | | Step Backward Feature Selection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target Feature | k | Naïve Bayes | | Logistic Reg. | | Random Forest | | Naïve Bayes | | Logistic Reg. | | Random Forest | |
| | | train | test | train | test | train | test | train | test | train | test | train | test |
| Current Alcohol Use[1] | 5 | 0.68 | 0.67 | 0.69 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.69 | 0.68 | 0.67 | 0.66 |
| | 10 | 0.70 | 0.70 | 0.72 | 0.71 | 0.70 | 0.70 | 0.70 | 0.70 | 0.71 | 0.70 | 0.70 | 0.70 |
| | 15 | 0.71 | 0.70 | 0.73 | 0.72 | 0.70 | 0.69 | 0.70 | 0.70 | 0.72 | 0.72 | 0.70 | 0.70 |
| | 20 | 0.71 | 0.70 | 0.73 | 0.72 | 0.70 | 0.70 | 0.70 | 0.70 | 0.73 | 0.72 | 0.70 | 0.70 |
| | 67 | 0.68 | 0.68 | 0.73 | 0.72 | 0.70 | 0.70 | 0.68 | 0.68 | 0.73 | 0.72 | 0.70 | 0.70 |
| Current Smoker[2] | 5 | 0.86 | 0.87 | 0.87 | 0.88 | 0.85 | 0.86 | 0.86 | 0.87 | 0.87 | 0.88 | 0.86 | 0.87 |
| | 10 | 0.87 | 0.88 | 0.88 | 0.89 | 0.85 | 0.86 | 0.87 | 0.88 | 0.88 | 0.89 | 0.87 | 0.87 |
| | 15 | 0.87 | 0.88 | 0.89 | 0.90 | 0.85 | 0.85 | 0.87 | 0.88 | 0.89 | 0.90 | 0.87 | 0.87 |
| | 20 | 0.87 | 0.88 | 0.89 | 0.90 | 0.85 | 0.85 | 0.87 | 0.88 | 0.89 | 0.90 | 0.88 | 0.88 |
| | 67 | 0.84 | 0.83 | 0.89 | 0.89 | 0.87 | 0.87 | 0.84 | 0.84 | 0.89 | 0.89 | 0.87 | 0.87 |
| Current Smoker[3] | 5 | 0.71 | 0.72 | 0.73 | 0.74 | 0.70 | 0.70 | 0.72 | 0.72 | 0.72 | 0.74 | 0.72 | 0.72 |
| | 10 | 0.74 | 0.74 | 0.76 | 0.77 | 0.70 | 0.71 | 0.74 | 0.74 | 0.76 | 0.77 | 0.74 | 0.74 |
| | 15 | 0.74 | 0.75 | 0.77 | 0.78 | 0.71 | 0.72 | 0.74 | 0.75 | 0.77 | 0.77 | 0.75 | 0.74 |
| | 20 | 0.74 | 0.75 | 0.77 | 0.78 | 0.71 | 0.71 | 0.74 | 0.73 | 0.77 | 0.77 | 0.75 | 0.74 |
| | 66 | 0.70 | 0.69 | 0.78 | 0.78 | 0.75 | 0.74 | 0.70 | 0.69 | 0.78 | 0.78 | 0.75 | 0.74 |

[1]Removed Alcohol Use Never, Alcohol Use Past, Alcohol Drinks / Wk.
[2]Removed Smoke Never, Smoke Past, Pack Years.
[3]Removed Smoke Never, Smoke Past, Pack Years, Smoke 100 Cigs.

TABLE VI:
F-MEASURE SCORES OF FEATURE SETS OF $k = 5$ FROM PREDICTION MODEL ANALYSIS

| | Step Forward | | | Step Backwards | | |
|---|---|---|---|---|---|---|
| Target | NB | LR | RF | NB | LR | RF |
| Current Alcohol | 0.62 | 0.62 | 0.63 | 0.62 | 0.62 | 0.59 |
| Deceased | 0.83 | 0.83 | 0.82 | 0.82 | 0.79 | 0.82 |
| Diabetic | 0.87 | 0.81 | 0.90 | 0.89 | 0.81 | 0.90 |
| NSAID Use | 0.80 | 0.79 | n/a | 0.80 | 0.79 | 0.79 |
| Current Smoker | 0.80 | 0.81 | 0.79 | 0.80 | 0.80 | 0.80 |
| SR Stroke | n/a | n/a | n/a | n/a | n/a | 0.91 |

NB = Naïve Bayes, LR = Logistic Regression, RF = Random Forest

## VI. CONCLUSIONS

In this work, we have proposed a data-driven methodological framework to evaluate multiple feature selection methods. In the comparative analysis, it can be observed that AUC scores improved only slightly (less than 5%) as the number of features chosen $k$ increased from 5 to 20. Scores also varied only slightly across different classifiers and wrapper methods. The major difference was in running time, with step forward taking only 14 seconds per feature and step backward taking an order of magnitude longer at 480 seconds per feature. In analyzing the features selected, it was noted that AUC performance degraded quickly when closely related features were removed. However, the features chosen still produced high accuracy scores when used to create a MLP classifier. The results demonstrated similar performance for all three classifiers utilized for feature selection, using both step forward and step backwards wrappers. The step backwards resulted in run-times an order of magnitude higher. Future work will investigate multi-target classification models. Applying feature selection based on a combination of multiple target attributes that are related could shed light on the interaction of potential biomarkers.

## REFERENCES


[1] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[2] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[3] A. Towfighi and J. L. Saver, "Stroke declines from third to fourth leading cause of death in the united states: historical perspective and challenges ahead," *Stroke*, vol. 42, no. 8, pp. 2351–2355, 2011.

[4] V. J. Howard, M. Cushman, L. Pulley, C. R. Gomez, R. C. Go, R. J. Prineas, A. Graham, C. S. Moy, and G. Howard, "The reasons for geographic and racial differences in stroke study: objectives and design," *Neuroepidemiology*, vol. 25, no. 3, pp. 135–143, 2005.

[5] J. F. Meschia, P. Merrill, E. Z. Soliman, V. J. Howard, K. M. Barrett, N. A. Zakai, D. Kleindorfer, M. Safford, and G. Howard, "Racial disparities in awareness and treatment of atrial fibrillation: the reasons for geographic and racial differences in stroke (regards) study," *Stroke*, vol. 41, no. 4, pp. 581–587, 2010.

[6] O. M. Gutiérrez, Y. A. Khodneva, P. Muntner, D. V. Rizk, W. M. McClellan, M. Cushman, D. G. Warnock, and M. M. Safford, "Association between urinary albumin excretion and coronary heart disease in black vs white adults," *Jama*, vol. 310, no. 7, pp. 706–714, 2013.

[7] M. K. Tamura, V. Wadley, K. Yaffe, L. A. McClure, G. Howard, R. Go, R. M. Allman, D. G. Warnock, and W. McClellan, "Kidney function and cognitive impairment in us adults: the reasons for geographic and

racial differences in stroke (regards) study," *American Journal of Kidney Diseases*, vol. 52, no. 2, pp. 227–234, 2008.

[8] U. Baber, V. J. Howard, J. L. Halperin, E. Z. Soliman, X. Zhang, W. McClellan, D. G. Warnock, and P. Muntner, "Association of chronic kidney disease with atrial fibrillation among adults in the united states: Reasons for geographic and racial differences in stroke (regards) study," *Circulation: Arrhythmia and Electrophysiology*, vol. 4, no. 1, pp. 26–32, 2011.

[9] W. McClellan, D. G. Warnock, L. McClure, R. C. Campbell, B. B. Newsome, V. Howard, M. Cushman, and G. Howard, "Racial differences in the prevalence of chronic kidney disease among participants in the reasons for geographic and racial differences in stroke (regards) cohort study," *Journal of the American Society of Nephrology*, vol. 17, no. 6, pp. 1710–1715, 2006.

[10] G. Howard, D. T. Lackland, D. O. Kleindorfer, B. M. Kissela, C. S. Moy, S. E. Judd, M. M. Safford, M. Cushman, S. P. Glasser, and V. J. Howard, "Racial differences in the impact of elevated systolic blood pressure on stroke risk," *JAMA internal medicine*, vol. 173, no. 1, pp. 46–51, 2013.

[11] W. E. Haley, D. L. Roth, G. Howard, and M. M. Safford, "Caregiving strain and estimated risk for stroke and coronary heart disease among spouse caregivers: differential effects by race and sex," *Stroke*, vol. 41, no. 2, pp. 331–336, 2010.

[12] H. Kramer, D. Shoham, L. A. McClure, R. Durazo-Arvizu, G. Howard, S. Judd, P. Muntner, M. Safford, D. G. Warnock, and W. McClellan, "Association of waist circumference and body mass index with all-cause mortality in ckd: The regards (reasons for geographic and racial differences in stroke) study," *American Journal of Kidney Diseases*, vol. 58, no. 2, pp. 177–185, 2011.

[13] H. E. Wang, N. I. Shapiro, R. Griffin, M. M. Safford, S. Judd, and G. Howard, "Chronic medical conditions and risk of sepsis," *PLoS One*, vol. 7, no. 10, p. e48307, 2012.

[14] R. J. Prineas, E. Z. Soliman, G. Howard, V. J. Howard, M. Cushman, Z.-M. Zhang, and C. S. Moy, "The sensitivity of the method used to detect atrial fibrillation in population studies affects group-specific prevalence estimates: ethnic and regional distribution of atrial fibrillation in the regards study," *Journal of epidemiology*, pp. 0 906 190 091–0 906 190 091, 2009.

[15] T. M. Brown, J. H. Voeks, V. Bittner, and M. M. Safford, "Variations in prevalent cardiovascular disease and future risk by metabolic syndrome classification in the reasons for geographic and racial differences in stroke (regards) study," *American heart journal*, vol. 159, no. 3, pp. 385–391, 2010.

[16] W. T. O'Neal, W. T. Qureshi, S. E. Judd, L. A. McClure, M. Cushman, V. J. Howard, G. Howard, and E. Z. Soliman, "Environmental tobacco smoke and atrial fibrillation: the reasons for geographic and racial differences in stroke (regards) study," *Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine*, vol. 57, no. 11, p. 1154, 2015.

[17] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[18] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[19] S. Menard, *Applied logistic regression analysis*. Sage, 2002, vol. 106.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] S. Raschka, *Python machine learning*. Packt Publishing Ltd, 2015.

[23] C. R. Boyd, M. A. Tolson, and W. S. Copes, "Evaluating trauma care: the triss method. trauma score and the injury severity score." *The Journal of trauma*, vol. 27, no. 4, pp. 370–378, 1987.

[24] O. Pauly, "Random forests for medical applications," Ph.D. dissertation, Technische Universität München, 2012.

[25] S. Raschka, "Mlxtend: Providing machine learning and data science utilities and extensions to pythons scientific computing stack," *The Journal of Open Source Software*, vol. 3, no. 24, Apr. 2018. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00638

[26] F. Eibe, M. Hall, I. Witten, and J. Pal, "The weka workbench," *Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, vol. 4, 2016.

[27] C. Panzer, M. S. Lauer, A. Brieke, E. Blackstone, and B. Hoogwerf, "Association of fasting plasma glucose with heart rate recovery in healthy adults: a population-based study," *Diabetes*, vol. 51, no. 3, pp. 803–807, 2002.

[28] M. A. Williamson and L. M. Snyder, *Wallach's Interpretation of Diagnostic Tests: Pathways to Arriving at a Clinical Diagnosis*. Lippincott Williams & Wilkins, 2014.

[29] H. Finney, D. J. Newman, and C. P. Price, "Adult reference ranges for serum cystatin c, creatinine and predicted creatinine clearance," *Annals of clinical biochemistry*, vol. 37, no. 1, pp. 49–59, 2000.