# HEALTH CARE ANALYTICS: MODELING BEHAVIORAL RISK FACTORS ASSOCIATED WITH DISEASE

by

Elena Gritsenko Toth, B.S.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Health Information Management
with a Major in Health Information Management
May 2019

Committee Members:

Alexander McLeod, Chair

David Gibbs

Jacqueline Moczygemba

## FAIR USE AND AUTHOR'S PERMISSION STATEMENT

### Fair Use



### Duplication Permission

## DEDICATION

This thesis is dedicated to the memory of my dad, Henry, who passed away before I began my master's degree. He taught me that higher education and non-stop hard work will get you anywhere you want to go, no matter where you come from. I deeply valued his advice and encouragement and hope that he would be proud of my achievements.

I also dedicate this thesis to my mom, who always pushed me to reach my potential because she believed I was capable of great things. She pushed me even when I was discouraged or frustrated, which helped bring me to where I am today. I owe all my accomplishments to my parents.

## ACKNOWLEDGEMENTS

I would like to acknowledge all who played a part in the completion of this thesis. First, Dr. Alexander McLeod, my thesis chair. Dr. McLeod contributed many hours out of his busy schedule for over 10 months, helping me work and re-work many parts of this paper and providing careful guidance.

Secondly, my committee members, Dr. David Gibbs and Professor Jackie Moczygemba. I highly valued their detailed feedback and advice throughout this process.

Finally, my husband David, who has dealt with my busy schedule and never complained. While at times frustrating, working on this thesis was very rewarding, as research always is.

## TABLE OF CONTENTS

**Page**

ACKNOWLEDGEMENTS ..... v

LIST OF TABLES ..... viii

LIST OF FIGURES ..... ix

LIST OF ABBREVIATIONS ..... x

ABSTRACT ..... xi

CHAPTER

1.INTRODUCTION ..... 1

2.RESEARCH QUESTIONS ..... 4

3.LITERATURE REVIEW ..... 5

Big Data ..... 5
Disease Prediction ..... 6
Criticism of Analytics and Evidence-Based Practice ..... 13

4.METHODS ..... 17

Data and Collection ..... 17
Modeling Behavioral Factors ..... 19
Breast Cancer Risk Factors ..... 19
General Cancer Risk Factors ..... 21
Diabetes Risk Factors ..... 23
Coronary Heart Disease Risk Factors ..... 25

5.ANALYSIS ..... 29

Binary Logistic Regression Model ..... 30
Multinomial Logistic Regression Model ..... 31
Conditional Inference Classification Trees ..... 31

6.RESULTS ..... 34

Binary Logistic Model ..... 34
Post Hoc Analysis ..... 35
Multinomial Logistic Regression Model ..... 37
Decision Tree Models ..... 38

7.DISCUSSION ..... 44

Limitations ..... 49
Future Research ..... 50

8.CONCLUSION ..... 51

APPENDIX ..... 52

REFERENCES ..... 57

## LIST OF TABLES

**Table** **Page**

1. Analytical Methods From Prior Research ..... 15

2. Breast Cancer Risk Factors ..... 20

3. General Cancer Risk Factors ..... 22

4. Diabetes Risk Factors ..... 24

5. Coronary Heart Disease Risk Factors ..... 27

6. Results For Binary Logistic Regression For Breast Cancer ..... 35

## LIST OF FIGURES

**Figure** **Page**

1. Conditional Inference Decision Tree for Diabetes with Exercise, Stroke, and Depression ..... 39

2. Conditional Inference Tree for Diabetes with High Cholesterol and BMI ..... 40

3. Conditional Inference Tree for Diabetes, High Blood Pressure, and Cholesterol ..... 41

4. Conditional Inference Tree for Diabetes with High Blood Pressure and High Blood Cholesterol Created with Training Data Set ..... 42

## LIST OF ABBREVIATIONS

| Abbreviation | Description |
|---|---|
| HIM | Health Information Management |
| GDP | Gross Domestic Product |
| EHR | Electronic Health Record |
| BRFSS | Behavioral Risk Factor Surveillance System |
| ICU | Intensive Care Unit |
| GA | Genetic Algorithm |
| NET | Neuroendocrine Tumors |
| MRSA | Methicillin-resistant Staphylococcus aureus |
| BRCA | Breast Cancer Type 1&2 Susceptibility Protein |
| ICD | International Classification of Diseases |

## ABSTRACT

Currently, there is a lot of excitement in the healthcare field about using big data and healthcare analytics for disease risk prediction, clinical decision support, and overall support for personalized medicine. However, this excitement hasn't effectively translated to improved clinical outcomes due to knowledge gaps, a lack of behavioral risk models and resistance to evidence-based practice. Reportedly, only 10-20% of clinical decisions are known to be evidence-based (Moskowitz, McSparron, Stone, & Celi, 2015) and this problem is further highlighted by the fact that the U.S. spends more money on healthcare per person than any other nation, while still wrestling with poor health outcomes (Barrett, Humblet, Hiatt, & Adler, 2013). Critics say there are inadequate technological resources and analytical education for clinicians to make big data useful in the clinical world (Neff, 2013). Healthcare technology innovators often neglect important aspects of the reality of integrating clinical data into healthcare solutions (Neff, 2013). In response to these problems, this study examines big data and healthcare analytics for use in clinical applications and suggests HIM professionals develop behavioral risk factor prediction models to bridge the gap between data scientists and clinicians.

# 1. INTRODUCTION

Preventable chronic diseases are the most common cause of premature death in the U.S. population (Barrett et al., 2013). The U.S. continues to spend more money on healthcare per person compared to any other nation. Amazingly, about 5% of patients account for almost 50% of all healthcare spending (Bates, Saria, Ohno-Machado, Shah, & Escobar, 2014). In addition, chronic disease costs account for 86% of healthcare costs in the U.S. (Lin, Chen, Brown, Li, & Yang, 2017). This is partly due to a lack of intervention based on behavioral risk factors, such as obesity, and a large amount of funds being spent on high-cost disease interventions only after the disease has already developed.

The U.S. population continues to fall victim to diseases caused by preventable risk factors, and up to half of all U.S. deaths can be attributed to preventable behaviors related to things such as poor diet, inadequate exercise, or tobacco and alcohol use (Barrett et al., 2013). An even greater concern is an aging U.S. population, with 10,000 more people turning 65 every day between January 2011 and January 2030 (Lash & Escobedo, 2018). The health care industry is expected to consume up to a quarter of the country's Gross Domestic Product (GDP) in the near future (Fox, 2011). This poses an important question to healthcare stakeholders: how do we improve health outcomes while reducing healthcare costs?

It has been suggested that healthcare needs to move from a disease treatment centered approach toward a patient disease prevention centered approach in order to reduce costs prior to disease onset (Chawla & Davis, 2013). Through the use of big data

analytics, the U.S. healthcare sector could save more than $300 billion per year, with 2/3 of this value coming from reduced healthcare spending (Belle et al., 2015). Clinical treatment costs of $165 billion and research and development costs of $108 billion are the two largest areas for behavioral risk savings (Raghupathi & Raghupathi, 2014).

With so much behavioral risk data currently available in the healthcare field, there is an opportunity to incorporate big data and healthcare analytics into clinical decision making. Currently, most EHR data is an unused asset. A study done by the Medical Group Management Association reported that only 31 percent of healthcare providers currently use all analytics tools and capabilities offered in their EHR (Monica, 2017). Historically, clinical decisions have been mostly based on experience and intuition. A suggested improvement to this model is the use of evidence-based practices (Palaniappan & Awang, 2008). To do so, big data and healthcare analytics may consider behavioral risk prediction models to identify new modifiable factors in the population. Supporting clinicians with behavioral disease prevention models may help shift high risk patients from treatment to prevention saving healthcare dollars.

As for importance for HIM professionals, meaningful use standard (American Reinvestment & Recovery Act) stages 2 and 3 encourage the use of data to improve the health of patient populations and improve care coordination. Big data and healthcare analytics are essential in transitioning from an expert-based to an evidence-based practice, however those professionals most closely associated with data and information governance such as health information managers, may lack the tools to support clinicians (Mikalef, Krogstie, van de Wetering, Pappas, & Giannakos, 2018). One commonly missing tool for healthcare organizations is the lack of appropriate data management

software (Kent, 2018). Barriers to successful clinical integration may exist in terms of a lack of training for clinicians, a lack of appropriate processes and a lack of tools and modeling techniques (McLeod & Dolezel, 2018). For example, the International Data Corporation reported that more than 40 percent of healthcare organizations struggle to hire employees with the necessary analytics skills (Kent, 2018).

The goal of this research is to examine the use of big data and healthcare analytics for the creation of behavioral risk prediction models and clinical decision support in evidence-based practice from the perspective of Health Information Management (HIM). Using data from the Center for Disease Control and Prevention (CDC) 2017 Behavioral Risk Factor Surveillance System Telephone Survey (Centers for Disease Control and Prevention, 2016), the intent of this work is to provide exemplars of disease prediction models using behavioral risk factors so that healthcare organizations and health information management professionals understand how data analytics can aid in clinical decision making and make better use of the information-rich big data in evidenced-based practices.

## 2. RESEARCH QUESTIONS

The need to support transformational research begins with clinical healthcare providers and leadership (Aarons, 2006). Health information management professionals can play a pivotal role in creating suitable analytics for clinical researchers if they are knowledgeable about various modeling techniques and predictive analysis. Given the increased availability of clinical healthcare data from EHR systems and the need for additional health information management skills to support evidenced-based medicine, the following research questions were developed:

RQ1: What disease prediction models can be used to support evidence-based medicine?

RQ2: Can we provide a variety of disease prediction models and inform the health information management profession?

RQ3: How can predictive analytics be used by health information management professionals to create valid research models of disease prediction.

RQ4: Are behavioral risk factors associated with cancer and what models can be used to create predictive analytics cases?

RQ5: Can the association between behavioral risk factors and breast cancer be tested?

RQ6: How can conditional inference decision trees be used to compare classification models of various diseases?

# 3. LITERATURE REVIEW

Literature related to big data, analytics, disease risk prediction and statistics is relevant to the research questions created for this thesis. HIM professionals must be aware of the problems introduced and types of big data in order to manage datasets for prediction.

## Big Data

Understanding the limits of big data and analytics from a computing perspective is important. According to Viceconti, Hunter, and Hose (2015), big data can be defined by the "5V's": Volume (quantity of data), variety (different categories of data), velocity (quick generation of new data), veracity (quality of data), and value (within the data). In the Computer Science world, big data is defined as the amount of data which is slightly beyond our current capability to store, manage, or process efficiently, usually in volumes of exabytes (10^18). Big data is a moving target because data storage, processing, and management capabilities are constantly improving with technology. However, storage, management, and processing of such data are considered to be fundamental overarching issues in the area of big data today (Kaisler, Armour, Espinosa, & Money, 2013). This is partly due size of healthcare data. For example, Kaiser Permanente, a healthcare system in California with more than 9 million patients had between 24.5 and 44 petabytes of healthcare data from EHRs as of 2014 (Raghupathi & Raghupathi, 2014). Healthcare data is expected to grow extremely fast, with a compound annual growth rate of 36 percent by 2025 (Kent, 2018).

Big data is exploding in healthcare and there are exciting new sources of

healthcare data, such as passive sensors and crowdsourcing. Passive sensors, such as activity and sleep sensors (e.g., FitBit or Garmin), can be a source of detailed data on potential risk factors for an individual over a long period of time (Barrett et al., 2013). Crowdsourcing data, which has been important in predicting infectious disease outbreaks, comes from tracking online search queries, informal health data from social media websites, or Wiki-type websites such as WebMd discussion boards (Barrett et al., 2013). To create highly sensitive disease risk prediction models, we need large datasets that contain information about potential risk factors as well as disease outcomes.

## Disease Prediction

Historically, there have been successful implementations of large-scale risk assessment tools, such as disease onset prediction, hospital readmission models, and models predicting healthcare cost and utilization (Razavian et al., 2015). Big data analytics may address outcomes in areas such as length of hospital stay, complications, infections such as Methicillin-resistant Staphylococcus aureus (MRSA), disease progression, and causation of disease (Raghupathi & Raghupathi, 2014). Several large academic medical centers have begun to harness big data by applying it to clinical decision making. For example, the Mayo Clinic implemented a software package called the Ambient Warning and Response Evaluation (AWARE) system which supports clinical decision making in the ICU and operating room, as well as a syndromic surveillance system, which detects sepsis (Moskowitz et al., 2015). Columbia University Medical Center has created a prediction system that analyzes correlations of physiological data related to patients with brain injuries. This system is able to diagnose

serious complications 48 hours sooner than other clinical methods in patients who suffer a bleeding stroke from a ruptured brain aneurism (Raghupathi & Raghupathi, 2014).

Besides predicting disease risk, Bates et al. (2014) mentions other ways predictive systems can be used in healthcare, such as identifying high-cost patients, adverse events, decompensation (worsening condition of the patient), readmissions, triage, and optimizing treatment for diseases that affect multiple organ systems, such as lupus. Fox (2011) discusses using big data analytics for "intelligent case management," or creating predictive models that can identify how an intervention program is likely to impact the patient's health behavior. These models would be able to identify which patients are most likely to benefit from a disease management program, patients who are most likely to participate actively, the level of intervention that will be required, create data on patient adherence and compliance, and identify specific outreach and support that is most likely to impact that individual (Fox, 2011).

There are several analytics methods that can be used to turn big data into a disease risk prediction model. In reviewing other studies, common approaches to big data mining for prediction models include linear regression, decision trees, neural networks, and the Naïve Bayesian approach. These types of risk prediction models may allow clinicians to develop effective therapies or interventions more quickly, reducing the cost of care for the affected population (Steinberg, Church, McCall, Scott, & Kalis, 2014). Using existing risk factors that have been identified by previous studies, it is possible to develop disease risk prediction models using big data analytics.

Singh (2015) used a classification system termed the Genetic Algorithm (GA), a type of evolutionary computing, to create a risk prediction model using 17 breast cancer

causing genes as variables (such as BRCA 1 and 2), as well as several breast cancer risk factors. Based on the relationship of a patient's gene sequence, symptoms, and risk factors, the GA classifier could predict if the patient would be in the risky or safe category for breast cancer with an 87% sensitivity and 62% specificity (Singh, 2015). Classification systems are common modeling techniques.

Steinberg et al. (2014) used a big data analytic platform called Reverse Engineering and Forward Simulation (REFS) to create a risk prediction model for metabolic syndrome. The study conducted laboratory screenings of Aetna customers and calculated risk of metabolic syndrome, impact of incremental changes on risk factors, and impact of adherence to treatment plan. Two models using machine learning had good to excellent predictive ability (0.80 and 0.88 ROC/AUC). For example, the study could identify a man who had a 92% chance of developing metabolic syndrome in the next 12 months. As a result, Aetna piloted a metabolic syndrome intervention program, specifically focusing on reducing waist circumference, which was determined to be a strong risk factor (Steinberg et al., 2014)

In an early study, Wilson et al. (1998) developed a simplified algorithm using linear and logistic regression to predict heart disease using risk factor categories and longitudinal data from the Framingham Heart Study. The algorithm was adapted into simplified score sheets that could allow doctors to estimate a patient's heart disease risk based on continuous, categorical, and risk factor sum variables. The algorithm was built on previous models by integrating additional risk factors such as blood pressure and cholesterol and used continuous variables as well as categorical approaches. Although the predictive capability of this model was similar to existing models at the time, this study

was a predecessor to what, 20 years later, would be referred to as big data analytics providing clinical decision support.

Palaniappan and Awang (2008) developed a prototype termed the Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques such as decision trees, Naïve Bayes, and neural networks. The system used CRISP-DM methodology to create mining models using attributes from data such as serum cholesterol and resting electrographic results. Three models were developed using the three data mining methods. The results showed that the Naïve Bayes model was the most effective in predicting heart disease (86.53% positive prediction), followed by neural networks and decision trees. However, decision trees were the most effective at predicting patients without heart disease (89% positive prediction). The authors suggest that their model could be a valuable tool in training medical students or nurses to diagnose patients with heart disease, as well as providing clinical decision support for doctors to assist with diagnosis of heart disease (Palaniappan & Awang, 2008).

Mixymol (2017) used similar data mining techniques to create prediction models for dermatology, hepatitis, and heart disease. With the WEKA data mining tool, Mixymol used three different classification techniques, Naïve Bayesian, Nearest Neighbor (neural network), and Reptree (decision trees), which were selected because of their common use in similar disease prediction studies. Results showed that Nearest Neighbor had the greatest correct classification for dermatology and hepatitis. For heart disease, Naïve Bayesian had the highest correct classification, followed by Reptree, then Nearest Neighbor (Mixymol, 2017). This study shows that data for each individual disease must be tested using a variety of algorithms, as prediction accuracy is not the same across the

methods.

Razavian et al. (2015) created a risk prediction model for type 2 diabetes using insurance claims data. Using risk factors that were identified from other diabetes prediction models (such as cardiovascular disease history and diagnosis of obesity), a model was created using logistic regression and machine learning. The model could predict the risk of developing type 2 diabetes in three future time spans with a 21.6% accuracy, compared to 11.4% using traditional prediction methods (Razavian et al., 2015). Turnea and Ilea (2018) used the SimT2DMtutor software to create a predictive simulation for type 2 diabetes. The software generated decision trees using associative classification algorithms based on variables such as body mass index and diastolic blood pressure. The authors propose that such a predictive model may be used to diagnose type 2 diabetes before complications appear (Turnea & Ilea, 2018).

In their report, Viceconti et al. (2015) provide an example of a patient-specific predictive model for osteoporosis and fractures that went beyond assessing risk factors such as low bone mineral density. The model was informed by wearable sensors and medical imaging, allowing it to predict the relative risk of fracture of the hip and spine in a patient by simulating ten years of the patient's daily life. Using Monte Carlo simulations, the authors created a musculoskeletal model, neuromuscular control model, and an organ level model for prediction of spontaneous bone fractures (Viceconti et al., 2012). For example, 80-year old women with severe osteoporosis and variable degrees of neuromotor control and muscle sarcopenia had a 29% Actual Spontaneous Fracture Risk during a single level walking event. A later analysis showed that these models could increase the accuracy of prediction of fractures up to 80-85%, as compared to the

standard predictive measure of bone mineral density that gives an accuracy of only 60-65% (Viceconti et al., 2015).

Chawla and Davis (2013) strived to create a patient-focused model which delivered an individualized disease risk profile together with a management and wellness plan for the patient. They created the CARE (Collaborative Assessment and Recommendation Engine) system for individualized risk prediction. Theoretically, CARE should be able to provide the clinician with a short list of diseases for which a patient is high-risk. The model works off ICD-9-CM medical codes, but this needs to be updated to the newer ICD-10-CM. The authors mention patient empowerment as one benefit of the model, encouraging dialogue between the clinician and patient about prevention and early detection of disease (Chawla & Davis, 2013).

Prediction models can also be used in the lab to accurately identify pathology samples. Sarkar and Nag (2017) created a decision tree predictive model using the C4.5 algorithm to identify and diagnose breast cancer in at-risk patients. The study used a data set of pathology results from fine needle biopsies containing nine categories related to cell features and anomalies. The final decision tree model could identify breast cancer with an accuracy of 96.7% (Sarkar & Nag, 2017).

Besides creating predictive models for individual patients using existing risk factors, big data can also be used to identify which risk factors are associated with a disease. Pyo et al. (2016) used big data analysis to identify risk factors for rectal neuroendocrine tumors (NET), a rare type of cancer where risk factors are generally unclear. The authors selected 29 possible predictive factors from previous reports and conducted statistical analysis using a 3-step logistic regression to narrow down which risk

factors were the strongest determinants of rectal NETs. In this case, metabolic syndrome was determined to be the strongest determinant of rectal NETs and identified a total of four strong risk factors (Pyo et al., 2016).

One current gap in research is that most predictive models are specific to just one disease or condition (Bates et al., 2014). However, it is rare that a patient has, or is at risk for just one chronic condition. For example, diabetes, obesity, and heart disease may occur alongside each other. Even if studies have addressed multiple risks in one patient, they typically look at each clinical risk as an independent task, and fail to address the idea that risks are often correlated and dependent of each other (Lin et al., 2017). Therefore, there is a need to create predictive models that address multiple conditions, which is likely to have a larger impact on healthcare outcomes.

One of the first studies to recognize the need for multifaceted risk profiling was Lin et al. (2017), which used big data from EHR systems to develop a model to predict the risk of adverse health events in patients. The authors recognize that chronic care patients commonly face multiple clinical risks and the risks are correlated (such as stroke and heart attack). The authors propose a method in which multiple prediction models are correlated into a unified framework using a hierarchical Bayesian multitask learning environment. Choosing diabetes as their test case, the authors used three adverse health events to model simultaneously in diabetic patients (stroke, acute renal failure, and acute myocardial infarction). Study results showed that this multifaceted approach outperformed existing single-task prediction models and the Bayesian multitask learning platform outperformed existing multitask methods (Lin et al., 2017). This approach may be able to better support clinicians in identifying patients who are at high risk for multiple

clinical events and would otherwise not be given preventative treatment.

## Criticism of Analytics and Evidence-Based Practice

There are some critics that dispute the excitement surrounding big data analytics in healthcare saying there are areas where the research falls short. Neff (2013) argues that while the technology sectors of healthcare see big data as valuable, healthcare providers don't actually have the resources, expertise, or the time to utilize big data predictive analytics or quantified metrics for patient care. Technology innovators in healthcare tend to neglect important aspects of the reality of integrating data into healthcare solutions (also called social interoperability), which creates a big disconnect between data scientists and clinical practitioners. The author also argues that because a lot of resources are needed to make big data valuable, big data will not solve problems in clinical practice because it will never be free (Neff, 2013).

Belle et al. (2015) address potentials and challenges in three areas of use of big data in healthcare, including medical image analysis, genomic data processing, and physiological signal processing. For example, compression (reducing the volume of data while still maintaining relevant data) and preprocessing (reducing noise, artifacts, missing data, and contrast) are challenges with medical image analysis. Current approaches to signal processing, or creating alerts from physiological signals, typically relies on single sources of information and do not consider the patient's overall physical condition and if the signal is significant. This can create "alarm fatigue" in healthcare workers, as the alarms go off many times for no reason. Developing a signal processing approach that considers correlations and interactions among multimodal clinical signs is needed,

specifically because research has shown that humans are poor in processing changes in more than two signals (Belle et al., 2015).

There are also the complex issues of confidentiality, privacy, consent, patient access, and oversight of big data use (Barrett et al., 2013). Sometimes researchers lack the data needed to make an accurate prediction, because many predictive model outcomes come from low-risk groups or low risk data, which may not create an accurate model. For example, in creating models for diseases affecting multiple organ systems, there has been a big issue with lack or longitudinal data, and this data may take time to develop as EHR's become to be more widely used (Bates et al., 2014).

To have successful implementation of big data analytics into clinical practice, clinical staff will need to be educated on biostatistics as part of their medical school curriculum. Moskowitz et al. (2015) suggests that there needs to be promotion of ongoing, cross-disciplinary collaboration between clinical staff and data scientists, possibly even having data scientists participate in hospital rounds alongside clinicians to access data in real time and receive feedback on their input.

Krumholz (2014) discussed the importance of having better personalized prediction models in clinical science to make more informed decisions about prognosis and treatment response. Medical researchers and clinicians will need to begin utilizing machine learning, data mining, and other advanced analytic techniques, which will require new training in data science. This transition will also require new methods of disease classification, beyond using clinical diagnostic labels, but with increased complexity that is required for customizable interventions (Krumholz, 2014).

Summarizing the various analytical methods found in previous research provides

an overview of potential tools and techniques available for HIM professionals to use when collaborating with clinicians.

Collaborative Filtering, Bayesian, Neural Networks, Decision Trees, Linear Regression, Logistic Regression, and Classification Algorithms have been used in clinical studies modeling a variety of diagnostic and disease process predictive research shown in Table 1. There is a need to apply predictive models to multiple conditions as well as utilize machine learning, data mining, and other advanced analytic techniques. Because healthcare providers often lack the resources, time and skills to operationalize these models in the clinical setting, HIM professionals may fill the analytical gap by creating clinically relevant models for practitioners.

**Table 1. Analytical Methods from Prior Research**

| Authors A-Z (Year) | Title | Analytical Method |
|---|---|---|
| Chawla and Davis (2013) | Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework | Collaborative Filtering |
| Lin et al. (2017) | Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach | Bayesian |
| Mixymol (2017) | Disease Prediction and Risk Analysis using Classification Algorithms | Neural Networks, Bayesian, Decision Trees |
| Palaniappan and Awang (2008) | Intelligent Heart Disease Prediction System Using Data Mining Techniques | Neural Networks, Bayesian, Decision Trees |
| Pyo et al. (2016) | Evaluation of the risk factors associated with rectal neuroendocrine tumors: a big data analytic study from a health screening center | Logistic Regression |
| Razavian et al. (2015) | Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors | Logistic Regression |
| Sarkar and Nag (2017) | Identifying patient at risk of breast cancer through decision trees | Decision Trees |

| | | |
|---|---|---|
| Singh (2015) | Prediction of Breast Cancer using Rule Based Classification | Classification Algorithms |
| Steinberg et al. (2014) | Novel Predictive Models for Metabolic Syndrome Risk: A "Big Data" Analytic Approach | Reverse Engineering and Forward Simulation |
| Turnea and Ilea (2018) | Predictive Simulation for Type II Diabetes Using Data Mining Strategies Applied to Big Data | Decision Trees, Classification Algorithms |
| Viceconti et al. (2012) | Are spontaneous fractures possible? An example of clinical application for personalized, multiscale neuro-musculoskeletal modelling | Monte Carlo Method |
| Wilson et al. (1998) | Prediction of Coronary Heart Disease Using Risk Factor Categories | Linear Regression, Logistic Regression |

# 4. METHODS

This study used the Behavioral Risk Factor Surveillance System (BRFSS) behavioral data to create four different disease risk prediction models: Multinomial regression model for general cancer type, Binary Logistic regression model for breast cancer, and a Rule-based classification decision tree model for diabetes. Another way to model disease risk is creating a chart identifying risk factors. This was done for coronary heart disease in this study. All data analysis was done using IBM SPSS and R statistical software.

## Data and Collection

Data for the predictive modeling portion of this study came from the 2017 Behavioral Risk Factor Surveillance System (BRFSS) Telephone Survey conducted by the Centers for Disease Control and Prevention (CDC). This was a combined landline and cell phone data set which included data for the 50 states, the District of Columbia, Guam, and Puerto Rico. The objective of this survey was to collect uniform state-specific data on health risk behaviors, chronic diseases, access to healthcare, and the use of preventative health services related to the leading causes of death in the United States. Thus the BRFSS data is a good example of big data based on volume (quantity of data), variety (different categories of data), velocity (quick generation of new data), veracity (quality of data), and value (within the data). The BRFSS data can be effectively used to demonstrate big data analytics techniques that can be applied to data generated by EHRs.

The 2017 BRFSS data set contained 450,016 records (Centers for Disease Control and Prevention, 2018b). Data collection is managed by state health departments

following protocols established by the CDC. States and US territories collect data for each of the 12 calendar months, submitting the data to the CDC at the end of each month. The CDC begins processing the data for the survey year as soon as states submit their data for each month (Centers for Disease Control and Prevention, 2018b). More information on the background, design, data collection and processing for this survey can be found at https://www.cdc.gov/brfss/annual_data/annual_2017.html.

Several studies have been done to assess the reliability and validity of the BRFSS data set. According to Pierannunzi, Hu, and Balluz (2013), who conducted a systematic review of related studies, the BRFSS data are reliable and valid because prevalence rates correspond well with other national surveys which relied on self-reports. Prevalence estimates from the data set also correspond well with findings from surveys based on face-to-face interviews such as the National Health Interview Study and the National Health and Nutrition Examination Survey (C. Li et al., 2012).

The 2017 BRFSS data set contains variables that were created naturally by the questions asked, as well as calculated variables, which are computed from the responses to other questions in the survey. There are two types of calculated variables included in the data set: Intermediate variables, and variables used to categorize or classify respondents (Centers for Disease Control and Prevention, 2018a). Intermediate variables are taken from a question response and are used to calculate another variable or risk factor. An example is the Body Mass Index (_BMI5) variable being calculated from individual computed weight and height variables WTKG3 (Computed Weight in Kilograms) and HTM4 (Computed Height in Meters), with WTKG3 originally being calculated from the variable WEIGHT2 (Reported Weight in Pounds). The other type of

calculated variable is used to classify or categorize respondents for simplifying analysis or identifying risk of specific injury or illness. Some of these computed variables will group continuous variables, such as weight or age, into categories, while others regroup categorical variables. The CDC provides the SAS code that was used to calculate each variable. (Centers for Disease Control and Prevention, 2018a).

## Modeling Behavioral Factors

To answer the research questions posed in this thesis, several models were needed to test hypotheses related to disease prediction. To model variables associated with disease, this work looked to the medical literature. Historically, heart disease, cancer and diabetes have been in the top 10 causes of death in the U.S. (Nichols, 2018) and these diseases were chosen to determine behavioral factor associations. The first disease selected to model and test was breast cancer.

## Breast Cancer Risk Factors

The American Cancer Society (2017b) provides guidance on potential risk factors related to the occurrence of breast cancer and these variables were organized in Table 2. The breast cancer risk factors identified by the American Cancer Society (2017b) included influencers related to age, race, alcohol consumption, body mass index and physical activity. A review of the BRFSS dataset, produced the variables seen in column 3 of Table 2, including - respondents sex (SEX), reported age in five-year categories calculated variable (_AGEG5YR), imputed age collapsed above 80 (_AGE80), imputed age in six groups (_AGE_G), computed race-ethnicity grouping (_RACE), computed five level race/ethnicity category (_RACEGR3), drink any alcoholic beverages in past 30 days

(DRNKANY5), computed number or drinks of alcohol beverages per week (_DRNKWEK), heavy alcohol consumption calculated variable (_RFDRHV5), days in

| **Table 2. Breast Cancer Risk Factors (American Cancer Society, 2017b)** | | |
|---|---|---|
| **Risk Factor** | **Risky Class** | **BRFSS Variables** |
| Gender/Sex | Female | Respondents Sex (SEX) |
| Age | 55+ | Reported age in five-year categories calculated variable (_AGEG5YR), Imputed age collapsed above 80 (_AGE80), Imputed age in six groups (_AGE_G). |
| Race | White over 45 yrs., African-American under 45 yrs. | Computed Race-Ethnicity Grouping (_RACE), Computed Five level race/ethnicity category (_RACEGR3). |
| Alcohol Consumption | Consumes Alcohol; 1 drink/day=small risk, 2-3 drinks=20% increase in risk | Drink any alcoholic beverages in past 30 days (DRNKANY5), Computed number or drinks of alcohol beverages per week (_DRNKWEK), Heavy Alcohol Consumption Calculated Variable (_RFDRHV5), Days in past 30 had alcoholic beverage (ALCDAY5), Average alcoholic drinks per day in past 30 (AVEDRNK2) |
| BMI | Overweight or Obese | Computed BMI (_BMI5), Computed BMI Categories (_BMI5CAT), Overweight or Obese calculated variable (_RFBMI5) |
| Physical Activity | Not Physically Active | Exercise in Past 30 Days (EXERANY2), Leisure Time Physical Activity Calculated Variable (_TOTINDA), (EXRACT11), (EXRACT21) |

past 30 had alcoholic beverage (ALCDAY5), average alcoholic drinks per day in past 30 (AVEDRNK2), computed body mass index (_BMI5), computed body mass index categories (_BMI5CAT), overweight or obesity calculated variable (_RFBMI5) exercise in past 30 Days (EXERANY2), leisure time physical activity calculated variables (_TOTINDA), (EXRACT11), (EXRACT21). Within these breast cancer variables, the risky class includes females over 55 years of age, who are either White or African-

American, consume alcohol, are overweight or obese and not physically active.

## General Cancer Risk Factors

The next model crafted included the general cancer risk factors, excluding skin cancer. The National Cancer Institute (2015) provides a list of cancer disease risk factors and these were mapped to the CDC BRFSS variables. These included reported age in five-year categories as a calculated variable (_AGEG5YR), imputed age collapsed above 80 (_AGE80), imputed age in six groups (_AGE_G) drinking of any alcoholic beverages in past 30 days (DRNKANY5), computed number of drinks of alcohol beverages per week (_DRNKWEK), heavy alcohol consumption as a calculated variable (_RFDRHV5), days in past 30 having alcoholic beverages (ALCDAY5), average number of alcoholic drinks per day in the past 30 days (AVEDRNK2), drinking of regular soda or pop that contains sugar (SSBSUGR2), drinking of sugar-sweetened drinks (SSBFRUT3), eating potatoes (POTATOE1), eating French fries or fried potatoes (FRENCHFI), eating dark green vegetables (FVGREEN1), eating fruit (FRUIT2), consuming vegetables one or more times per day (_VEGLT1A), consuming fruit one or more times per day (_FRTLT1A), being involved in high risk situations for HIV (HIVRISK5), ever getting tested for HIV (HIVTST6), ever been tested for HIV as a calculated variable (_AIDSTST3), computed body mass index (_BMI5), computed body mass index categories (_BMI5CAT), being overweight or obese as a calculated variable (_RFBMI5), being a current smoker as a calculated variable (_RFSMOK3), (SMOKER3), (COPDSMOK). For these behavioral risk factors, being older in age, drinking alcohol, having an unhealthy diet that includes sweeteners, having or being exposed to hepatitis,

HIV/AIDS, HPV Epstein-Barr virus, or H. Pylori, being overweight or obese and being a current or previous smoker increases chances for cancer in general. Table 3 shows the various general cancer risk factors as discerned from the National Cancer Institute.

**Table 3. General Cancer Risk Factors (National Cancer Institute, 2015)**

| Risk Factor | Risky Class | BRFSS Variables |
|---|---|---|
| Age | Older in Age | Reported age in five-year categories calculated variable (_AGEG5YR), Imputed age collapsed above 80 (_AGE80), Imputed age in six groups (_AGE_G). |
| Alcohol | Drinking alcohol | Drink any alcoholic beverages in past 30 days (DRNKANY5), Computed number or drinks of alcohol beverages per week (_DRNKWEK), Heavy Alcohol Consumption Calculated Variable (_RFDRHV5), Days in past 30 had alcoholic beverage (ALCDAY5), Avg alcoholic drinks per day in past 30 (AVEDRNK2) |
| Diet | Unhealthy diet, including sweeteners | How often did you drink regular soda or pop that contains sugar (SSBSUGR2), How often did you drink sugar-sweetened drinks (SSBFRUT3), How often do you eat potatoes (POTATOE1), How often do you eat French fries or fried potatoes (FRENCHFI), How many times a day do you eat dark green vegetables (FVGREEN1), How many times did you eat fruit (FRUIT2), Consume vegetables 1 or more times per day (_VEGLT1A), Consume fruit 1 or more times per day (_FRTLT1A) |
| Infections | Hepatitis, HIV/AIDS, HPV Epstein-Barr virus, H. Pylori, | HIVRISK5 (Do any high-risk situations apply), Ever tested HIV (HIVTST6), Ever been tested for HIV calculated variable (_AIDSTST3) |
| BMI | Being overweight or obese | Computed BMI (_BMI5), Computed BMI Categories (_BMI5CAT), Overweight or Obese calculated variable (_RFBMI5) |
| Smoking | Current or previous tobacco use | Current Smoking Calculated Variable (_RFSMOK3), (SMOKER3), (COPDSMOK) |

## Diabetes Risk Factors

Diabetes is the 7th leading cause of death in the U.S. and so the next model created considered the risk factors for Type II diabetes as delineated by National Institute of Diabetes and Digestive and Kidney Diseases (2016). Table 4 shows the BRFSS variables mapped to the recognized diabetes risk factors.

For the Diabetes model, the variables associated with the disease prediction model were reported age in five-year categories calculated variable (_AGEG5YR), imputed age collapsed above 80 (_AGE80), imputed age in six groups (_AGE_G), computed race-ethnicity grouping (_RACE), computed five level race/ethnicity category (_RACEGR3), computed BMI (_BMI5), computed body mass index categories (_BMI5CAT), being overweight or obese calculated variable (_RFBMI5), exercising in past 30 days (EXERANY2), leisure time physical activity as a calculated variable (_TOTINDA), ever told blood cholesterol high (TOLDHI2), currently taking medicine for high cholesterol (CHOLMED1), high cholesterol as a calculated variable(_RFCHOL1), ever told blood pressure high (BPHIGH4), currently taking blood pressure medication (BPMEDS),ever told you had a depressive disorder (ADDEPEV2), ever diagnosed with a stoke (CVDSTRK3), ever diagnosed with angina or coronary heart disease (CVDCRHD4), and ever had coronary heart disease or myocardial infarction (_MICHD).

The risk factors related to these variables include age, gender, race, body mass index, smoking, blood pressure, coronary heart disease, cholesterol, physical activity, alcohol consumption, diabetes and prediabetes, stress, and diet. Risky Classes in these factors were being 45+ years of age, African-American, Alaska Native, American-Indian, Asian American, Hispanic, Native Hawaiian, Pacific Islander, being overweight or obese,

not participating in physical activities, having high cholesterol, having high blood pressure, having a history of depression, having a history of stroke, and having a history of heart disease.

**Table 4. Diabetes Risk Factors (National Institute of Diabetes and Digestive and Kidney Diseases, 2016)**

| Risk Factor | Risky Class | BRFSS Category |
|---|---|---|
| Age | 45+ Years of Age | Reported age in five-year categories calculated variable (_AGEG5YR), Imputed age collapsed above 80 (_AGE80), Imputed age in six groups (_AGE_G). |
| Race | African-American, Alaska Native, American-Indian, Asian American, Hispanic, Native Hawaiian, Pacific Islander | Computed Race-Ethnicity Grouping (_RACE), Computed Five level race/ethnicity category (_RACEGR3). |
| BMI | Overweight or Obese | Computed BMI (_BMI5), Computed BMI Categories (_BMI5CAT), Overweight or Obese calculated variable (_RFBMI5) |
| Physical Activity | No Physical Activity | Exercise in Past 30 Days (EXERANY2), Leisure Time Physical Activity Calculated Variable (_TOTINDA) |
| Cholesterol | High cholesterol | Ever Told Blood Cholesterol High (TOLDHI2), Currently taking medicine for high cholesterol (CHOLMED1), High cholesterol calculated variable(_RFCHOL1) |
| Blood Pressure | High blood pressure | Ever told blood pressure high (BPHIGH4), Currently taking blood pressure medication (BPMEDS) |
| Depression | History of Depression | Ever told you had a depressive disorder (ADDEPEV2) |
| Stroke | History of Stroke | Ever diagnosed with a stoke (CVDSTRK3) |
| Heart Disease | History of Heart Disease | Ever diagnosed with angina or coronary heart disease (CVDCRHD4), Ever had CHD or MI (_MICHD) |

## Coronary Heart Disease Risk Factors

In considering the next disease model, Coronary Heart Disease, risky classes included men over 45+, women over 55+, African-American, Hispanic, American Indian, Native Hawaiian, Asian-American, being overweight or obese, being a current smoker, not being physically active, consuming more than 2 drinks per day for men, more than 1 drink per day for women, presence of diabetes or prediabetes, presence of stress or anxiety, and consuming an unhealthy diet and added sugars. These classes were associated with risk factors related to coronary heart disease. The risk factors were age, gender, race, body mass index, smoking, blood pressure, cholesterol, physical activity, alcohol consumption, diabetes and prediabetes, stress, and diet.

These risk factors were then associated with the BRFSS variables, reported age collapsed above 80 (_AGE80), imputed age collapsed above 80 (_AGE80), imputed age in six groups (_AGE_G), respondents sex (SEX), computed race-ethnicity grouping (_RACEGR3), computed body mass index (_BMI5), computed body mass index categories (_BMI5CAT), being overweight or obese as a calculated variable (_RFBMI5), current smoker calculated variable (_RFSMOK3), computed smoking status (SMOKER3), number of years smoking tobacco products (COPDSMOK), ever told blood pressure high (BPHIGH4), currently taking blood pressure medication (BPMEDS), ever diagnosed with heart attack (CVDINF4), taking aspirin daily or every other day (CVDASPRN), ever had congestive heart disease or myocardial infarction (_MICHD), taking aspirin to reduce chance of heart attack (RDUCHART), change in eating habits to improve blood pressure(BPEATHBT), ever diagnosed with angina or coronary heart disease (CVDCRHD4), ever told blood cholesterol high (TOLDHI2), currently taking

medicine for high cholesterol (CHOLMED1), high cholesterol as a calculated variable (_RFCHOL1), any exercise in past 30 days (EXERANY2), type of physical activity (EXRACT11), other type of physical activity giving most exercise during past month (EXRACT21), days in past 30 had alcoholic beverage (ALCDAY5), average alcoholic drinks per day in past 30 (AVEDRNK2), drinking any alcoholic beverages in past 30 days (DRNKANY5), computed number of drinks of alcohol beverages per week (@_DRNKWEK), ever been told by a doctor or other health professional that you have pre-diabetes or borderline diabetes (PREDIAB1), ever told you have diabetes (DIABETE3), ever taken class in managing diabetes (DIABEDU), had a test for high blood sugar or diabetes in past three years (PDIABPST), now taking insulin (INSULIN), how often check blood for glucose (BLDSUGAR), times seeing health professional for diabetes (DOCTDIAB), times checked for glycosylated hemoglobin (CHKHEMO3), number of days mental health not good, including stress (MENTHLTH), how often have you felt this kind of stress (SDHSTRES), computed mental health status (_MENT14D), satisfaction with life (LSATISFY), how often get emotional support needed (EMTSUPRT), how often did you drink regular soda or pop that contains sugar (SSBSUGR2), how often did you drink sugar-sweetened drinks (SSBFRUT3), how often do you eat potatoes (POTATOE1), how often do you eat French fries or fried potatoes (FRENCHFI), how many times a day do you eat dark green vegetables (FVGREEN1), how many times did you eat fruit (FRUIT2), and consume vegetables 1 or more times per day as seen in Table 5.

| Table 5. Coronary Heart Disease Risk Factors (American Heart Association, 2014) | | |
|---|---|---|
| **Risk Factor** | **Risky Class** | **BRFSS Variables** |
| Age | Men (45+), Women (55+) | Reported age collapsed above 80 (_AGE80), Imputed age collapsed above 80 (_AGE80), Imputed age in six groups (_AGE_G). |
| Gender | Male | Respondents Sex (SEX) |
| Race | African-American, Hispanic, American Indian, Native Hawaiian, Asian-American | Computed Race-Ethnicity Grouping (_RACEGR3) |
| BMI | Overweight or Obese | Computed BMI (_BMI5), Computed BMI Categories (_BMI5CAT), Overweight or Obese calculated variable (_RFBMI5) |
| Smoking | Current smoker | Current Smoking Calculated Variable (_RFSMOK3), Computed smoking status (SMOKER3), How many years have you smoked tobacco products (COPDSMOK) |
| Blood Pressure | High levels increase risk | Ever told blood pressure high (BPHIGH4), Currently taking blood pressure medication (BPMEDS) |
| Coronary Heart Disease | | Ever diagnosed with heart attack (CVDINF4), Take aspirin daily or every other day (CVDASPRN), Ever had CHD or MI (_MICHD), Take aspirin to reduce chance of heart attack (RDUCHART), Change eating habits for BP (BPEATHBT), Ever diagnosed with angina or coronary heart disease (CVDCRHD4) |
| Cholesterol | High levels increase risk | Ever Told Blood Cholesterol High (TOLDHI2), Currently taking medicine for high cholesterol (CHOLMED1), High cholesterol calculated variable(_RFCHOL1) |
| Physical Activity | Not Physically active | Exercise in Past 30 Days (EXERANY2), Type of physical activity (EXRACT11), Other type of physical activity giving most exercise during past month (EXRACT21) |
| Alcohol Consumption | More than 2 drinks/day for men, more than 1 drink/day for women | Days in past 30 had alcoholic beverage (ALCDAY5), Avg alcoholic drinks per day in past 30 (AVEDRNK2), Drink any alcoholic beverages in past 30 days (DRNKANY5), Computed number of drinks of alcohol beverages per week (@_DRNKWEK) |
| Diabetes and Prediabetes | Presence of Diabetes or | Ever been told by a doctor or other health professional that you have pre-diabetes or |

| | | |
|---|---|---|
| | Prediabetes | borderline diabetes (PREDIAB1), Ever told you have diabetes (DIABETE3), Ever taken class in managing diabetes (DIABEDU), Had a test for high blood sugar or diabetes in past three years (PDIABPST), Now taking insulin (INSULIN), How often check blood for glucose (BLDSUGAR), Times seen health professional for diabetes (DOCTDIAB), Time checked for glycosylated hemoglobin (CHKHEMO3) |
| Stress | Presence of stress or anxiety | Number of days mental health not good, including stress (MENTHLTH), How often have you felt this kind of stress (SDHSTRES), Computed mental health status (_MENT14D), Satisfaction with life (LSATISFY), How often get emotional support needed (EMTSUPRT) |
| Diet | Unhealthy diet and added sugars | How often did you drink regular soda or pop that contains sugar (SSBSUGR2), How often did you drink sugar-sweetened drinks (SSBFRUT3), How often do you eat potatoes (POTATOE1), How often do you eat French fries or fried potatoes (FRENCHFI), How many times a day do you eat dark green vegetables (FVGREEN1), How many times did you eat fruit (FRUIT2), Consume vegetables 1 or more times per day |

# 5. ANALYSIS

Before beginning analysis, some of the data was transformed. A missing value analysis was conducted using IBM SPSS and variables that were favored for inclusion in the analysis showed less than 20% missing data. Garson (2015) recommends a conservative cutoff of 20% missing values. All survey variables required some transformation and recoding using SPSS. Models using binary logistic regression require a dependent variable which is binary, yes-no, true-false, male-female. The dependent variable was coded as binary. For multinomial regression and classification trees, the variables were recoded so that they would be equivalent in code/response to each other, as many questions were coded differently. On categorical yes/no variables where "Yes"=1, "No"=2 (or vice versa), "Don't know/not sure" (usually coded as 7) and "Refused/missing"(usually coded as a value of 9) responses were recoded into "No" categories (in some questions these two were combined into a common code of 9, or similar, which was recoded into "No"). For example, on variable "Drink any alcoholic beverages in the past 30 days", 1="Yes," 2= "No," 7= "Don't know/not sure," 9= "Refused/Missing." 7 and 9 were recoded into 2's, signifying people who did not answer "Yes".

For discrete variables, respondent's answers were kept the same, only recoding the "Don't know/Not sure/Refused/Missing" category, which can be signified by various codes depending on the question, but usually 9. For example, on variable "Computed number of drinks of alcohol beverages per week", 0=Did not drink, 1-98999=Number of drinks per week specified by respondent, and 99900=Don't know/Not sure/Refused/Missing. The 99900 category was recoded into 0, as these respondents

didn't specify they drank any alcoholic drinks. Following transformation, the data was examined for outliers, mis-specification and error.

## Binary Logistic Regression Model

Binary logistic regression is similar to multiple linear regression, however, the response variable is binomial (Sperandei, 2014). Logistic regression is used to get an odds ratio when there is a presence of more than one explanatory variable. The result shows the impact of each variable using the odds ratio of the studied event and analyzes the association of all variables in the model together. If multiple explanatory variables are to be analyzed independently, we disregard the covariance among variables and may end up with cofounding effects (Sperandei, 2014).

A binary logistic regression result produces the odds ratio, which is then evaluated for significance using t-test and subsequent p value. For example, when using logistic regression to predict risk factors for rectal neuroendocrine tumors, Pyo et al. (2016) got an odds ratio of 1.768 with a confidence interval of 95%. This meant that when looking at the variable "presence of metabolic syndrome", people with metabolic syndrome were 1.768 times more likely to develop a rectal neuroendocrine tumor than those without metabolic syndrome.

For the binary logistic regression breast cancer model, the Type of Cancer (CNCRTYP1) variable was used as dependent and was transformed into a binary dependent variable to signify breast cancer patients and survivors (Breast cancer response was transformed into 1, all other respondents transformed into 0). The subsequent binary logistic model was formed as Breast Cancer = Gender + Age + Race + Alcohol

Consumption + BMI + Physical Activity.

## Multinomial Logistic Regression Model

It is possible to create a simple disease prediction algorithm using multinomial logistic regression. Multinomial logistic regression is an extension of binary logistic regression. This method is used when the categorical dependent variable has more than two categories (Chan, 2005). Instead of predicting only two groups, for example normal weight and overweight, we may predict four groups, underweight, normal weight, overweight, and obese. For the multinomial regression model, we selected Type of Cancer (CNCRTYP1) as the dependent variable. Independent variables for each risk factor category were analyzed for significance and a multinomial model was created. The variables were noted as significant with a selection value of $p<0.05$ or 95% confidence interval.

## Conditional Inference Classification Trees

Classification is a technique in data mining and machine learning that has been used in many real-world applications by data scientists. In order to build a classifier, the researcher first needs to collect a data set with previously defined cases that can be used as training examples (X. Li & Liu, 2014). A predetermined classification algorithm can then be applied to the training data to assign the previously defined classes to test current instances for evaluation (X. Li & Liu, 2014). There are many classification techniques, but here we will focus on rule-based classification.

There are advantages to using and teaching rule-based classification to non-data scientist professionals, since the rules are easy to explain, and can be understood by many

different types of practitioners improving interpretation (X. Li & Liu, 2014). Rules are typically represented in logic form as IF-THEN statements. For example, in using rule-based classification for predicting breast cancer, Singh (2015) used IF statements such as Gender=Female, Age>=60, and Gene Mutation=BRCA2. Therefore, if a woman was found to have a breast cancer risk factor, such as the BRCA2 mutation, then she would be classified into the "Risky Class" by the algorithm (Singh, 2015).

Another method of mining big data to create disease prediction models is decision trees. General uses of decision trees include segmentation (identifying categories), stratification (assigning into categories), prediction (creating rules and predicting future events), data reduction and variable screening, interaction identification (identifying relationships), and category merging or banding continuous variables (IBM, 2012).

A decision tree model allows us to create a classification system that can predict or classify future cases based on a set of rules (IBM, 2012). The rule induction process uses existing big data, such as disease risk factors and outcomes, to build a set of rules to classify future cases. There are several tree-building algorithms available for classifying and segmenting data. For example, Sarkar and Nag (2017) used the C4.5 algorithm, which builds either a rule set or a decision tree (IBM, 2012). Designing the decision tree can be a difficult process, and for those without a computer science background new commercial tree building software (such as TreeAge Pro) has made the process easier (Bae, 2014). Decision tree analysis is also part of many business intelligence tools, such as IBM SPSS.

Conditional inference classification decision tree models were made with R "party package" using the "ctree" algorithm. Diabetes and Coronary Heart Disease were

selected as dependent variables. Independent variables were selected using risk factors identified in previous literature and were tested for significance using binary logistic regression. Variables which were significant were selected for inclusion in the decision tree models. First, the entire dataset was used to create a tree without creating separate testing and training data sets. The algorithm selects the most important variable as the first split, second important variable as the next split, and so on. Three conditional inference decision trees were created for diabetes using different dependent variables.

# 6. RESULTS

## Binary Logistic Model

For the binary logistic regression model, the Type of Cancer variable was transformed into a binary dependent variable to signify breast cancer patients and survivors. Breast cancer response was transformed into a 1, and all other respondents transformed into 0). The final model had a Nagelkerke R Square value of 0.141, and a Cox & Snell R Square value of .005. For both R Square values, a value of 1 would mean the model perfectly predicts the outcome. With these R Square values, the model fit is fairly weak.

The following variables were significant with a 95% confidence interval ($p<.05$): Respondents Sex (SEX)(ExpB=86.104), Computed Body Mass Index Categories (_BMI5CAT) (ExpB=1.225), Reported Age in Five-Year Categories (_AGEG5YR)(ExpB=1.329), Computed Five Level Race/Ethnicity Categories (_RACEGR3)(ExpB=0.768), and Heavy Alcohol Consumption Calculated Variable (_RFDRHV5)(ExpB=0.527) ($p=0.000$ for all variables). See Table 6. All variables analyzed for physical activity in this model were not significant.

| Table 6. Results for Binary Logistic Regression for Breast Cancer | | | | | | | |
|---|---|---|---|---|---|---|---|
| **BRFSS Variable Name** | **BRFSS Variable Code** | **B** | **S.E.** | **Wald** | **df** | **P value** | **Exp(B)** |
| Respondents Sex | SEX | 4.456 | 0.318 | 196.669 | 1 | 0.000 | 86.104 |
| Computed Body Mass Index Categories | _BMI5CAT | 0.203 | 0.035 | 32.890 | 1 | 0.000 | 1.225 |
| Reported Age In Five-Year Categories | _AGEG5YR | 0.284 | 0.012 | 524.962 | 1 | 0.000 | 1.329 |
| Computed Five Level Race/Ethnicity Categories | _RACEGR3 | -0.265 | 0.038 | 48.774 | 1 | 0.000 | 0.768 |
| Heavy Alcohol Consumption Calculated Variable | _RFDRHV5 | -0.640 | 0.144 | 19.770 | 1 | 0.000 | 0.527 |
| | Constant | -16.442 | 0.680 | 585.054 | 1 | 0.000 | 0.000 |

## Post Hoc Analysis

A post hoc analysis was done using binary logistic regression on BMI, Race, and Age. When looking at age, a bimodal distribution of significance was seen. Women ages 30-34 (p=.001), 35-39 (p=.001), and 40-44 (p=.003), compared the younger significant group. Women ages 45-49 (p=.305), 50-54 (p=.621), 55-59 (p=.887), 60-64 (p=.501) were not significant. Women ages 65-69 (p=.055) and 70-74 (p=.060) were marginally significant. Women ages 75-79 (p=.016) and 80+ (p=.006) made up the older significant group. Exp (B) odds ratio values were higher for older women, with women 80+ being 3.497 times more likely to have had breast cancer. Table 7 presents these results.

| Table 7. Post Hoc Analyses of Age | | |
|---|---|---|
| **Age** | **P value** | **Exp (B)** |
| 30-34 | 0.001 | 0.029 |
| 35-39 | 0.001 | 0.127 |
| 40-44 | 0.003 | 0.175 |
| 65-69 | 0.055 | 2.384 |
| 70-74 | 0.060 | 2.350 |
| 75-79 | 0.016 | 2.998 |
| 80+ | 0.006 | 3.497 |

When looking at weight, women who were underweight (p=.006), normal weight (p=.000) and obese (p=.000) were significant. The overweight group was not significant. Women who were obese had the highest Exp (B) odds ratio, being 1.324 times more likely to have had breast cancer. Table 8 shows these results.

| Table 8. Post Hoc Analysis of Body Mass Index | | |
|---|---|---|
| **BMI** | **P Value** | **Exp(B)** |
| Underweight (less than 18.5) | 0.006 | 0.504 |
| Normal weight (18.5-25) | 0.000 | 0.667 |
| Obese (30+) | 0.000 | 1.324 |

When looking at race, White (p=.002) and Other (p=.011) races among women were predictive. Black (p=.069) and Multi (p=.084) race women were only marginally significant. Being a Hispanic woman was not significant. White women had the highest odds ratio, being 7.043 times more likely to have had breast cancer (Table 9).

| Table 9. Post Hoc Analysis of Race | | |
|---|---|---|
| **Race** | **P Value** | **Exp(B)** |
| White | 0.002 | 7.043 |
| Black | 0.069 | 4.367 |
| Other | 0.011 | 4.778 |
| Multi | 0.084 | 4.720 |

## Multinomial Logistic Regression Model

In the analysis of the multinomial regression model, Type of Cancer (CNCRTYP1) was selected as the dependent variable. Specifically, individuals who had breast, lung, or colon cancer were selected for analysis. Independent variables for each risk factor category were analyzed for significance and the multinomial model yielded a pseudo R-Square value of 0.080 (Nagelkerke). These values mean that the model only explains 8% of total variance in cancer type, respectively.

**Table 10. Results for Multinomial Regression for Type of Cancer (CNCRTYP1)**

| BRFSS Variable Code | BRFSS Variable Name | -2 log Likelihood of Reduced Model | Chi-Square | df | P value |
|---|---|---|---|---|---|
| | Intercept | 4566.83 | .000 | 0 | |
| _AGEG5YR | Reported age in five-year categories | 5727.13 | 1160.30 | 39 | 0.00 |
| _RACE | Computed Race-ethnicity grouping | 4603.30 | 36.47 | 9 | 0.00 |
| _SMOKER3 | Computed Smoking Status | 4687.77 | 120.94 | 6 | 0.00 |
| DRNKANY5 | Drink any alcoholic beverages in past 30 days | 4609.93 | 43.09 | 9 | 0.00 |
| _VEGLT1A | Consume Vegetable 1 or more times per day | 4577.96 | 11.12 | 6 | 0.085 |
| _FRTLT1A | Consume Fruit 1 or more times per day | 4595.91 | 29.07 | 6 | 0.00 |
| _RFBMI5 | Overweight or obese calculated variable | 4890.43 | 23.60 | 6 | 0.00 |
| HIVRISK5 | Do any high-risk situations apply | 4580.35 | 13.51 | 9 | 0.141 |

The following variables were significant with a p value of <0.05 or 95% confidence interval: Reported Age in Five Year Categories (p=.00), Computed race/ethnicity grouping (p=.00), Computed smoking status (_SMOKER3) (p=.00), Drink any alcoholic beverages in past 30 days (DRNKANY5) (p=.00), Consume fruit 1 or more times per day (_FRTLT1A) (p=0.00) and Overweight or obese calculated variable (_RFBMI5) (p=.000). Consume vegetables 1 or more times per day (_VEGLT1A) (p=0.085) was marginally significant. Do any high-risk situations apply (HIVRISK5) was not significant (p=0.141) (Table 10). Overall, each risk factor category had at least one significant variable except for various related infections. Individual parameter estimates for each variable can be seen in the appendix.

## Decision Tree Models

Four conditional inference decision trees were created for diabetes using different independent variables. The first tree captured the relationship between Exercise (EXERANY2), Stroke (CVDSTRK3), and Depression (ADDEPEV2) and Diabetes (DIABETE3).

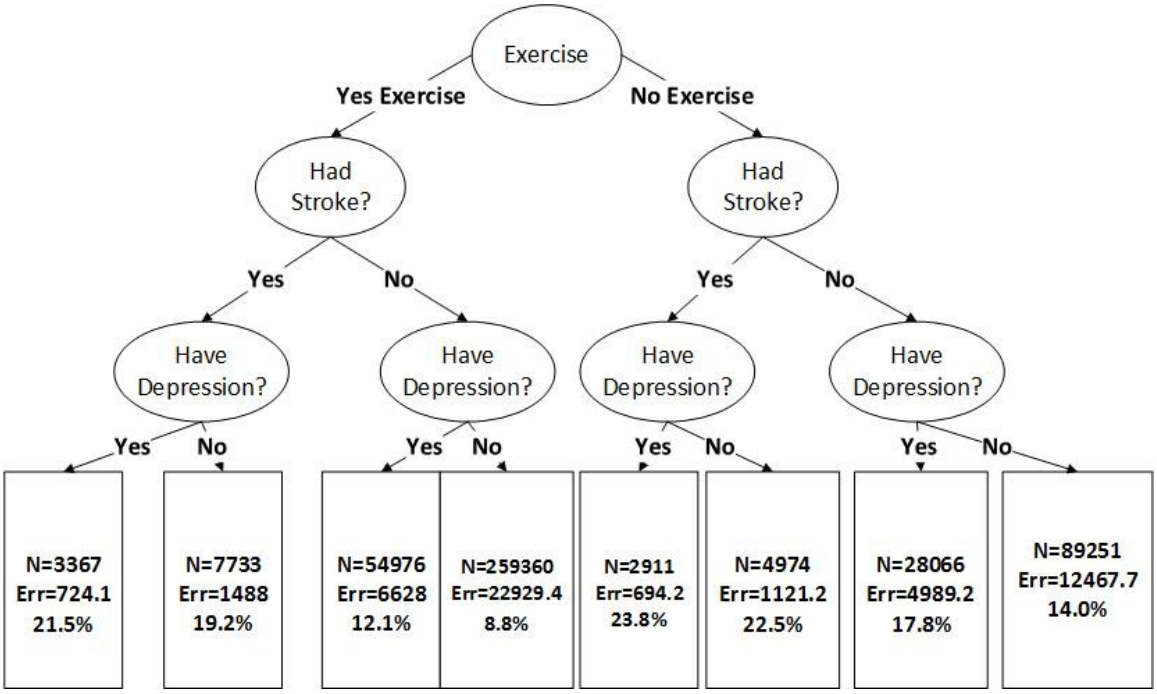


**Figure 1. Conditional Inference Decision Tree for Diabetes with Exercise, Stroke, and Depression**

Figure 1 shows that this tree had the highest classification error rates ranging from 8.8% - 23.8%, with an average of 17.46%, meaning this tree was the least accurate at predicting diabetes. The second tree examined the relationship between High Blood Cholesterol (TOLDHI2) and BMI (_BMI5CAT) and Diabetes (DIABETE3). Figure 2 shows that this decision tree had the lowest classification error rates ranging from 3.4% to 22.2%, with an average of 10.64%, meaning it was the most accurate at predicting diabetes.

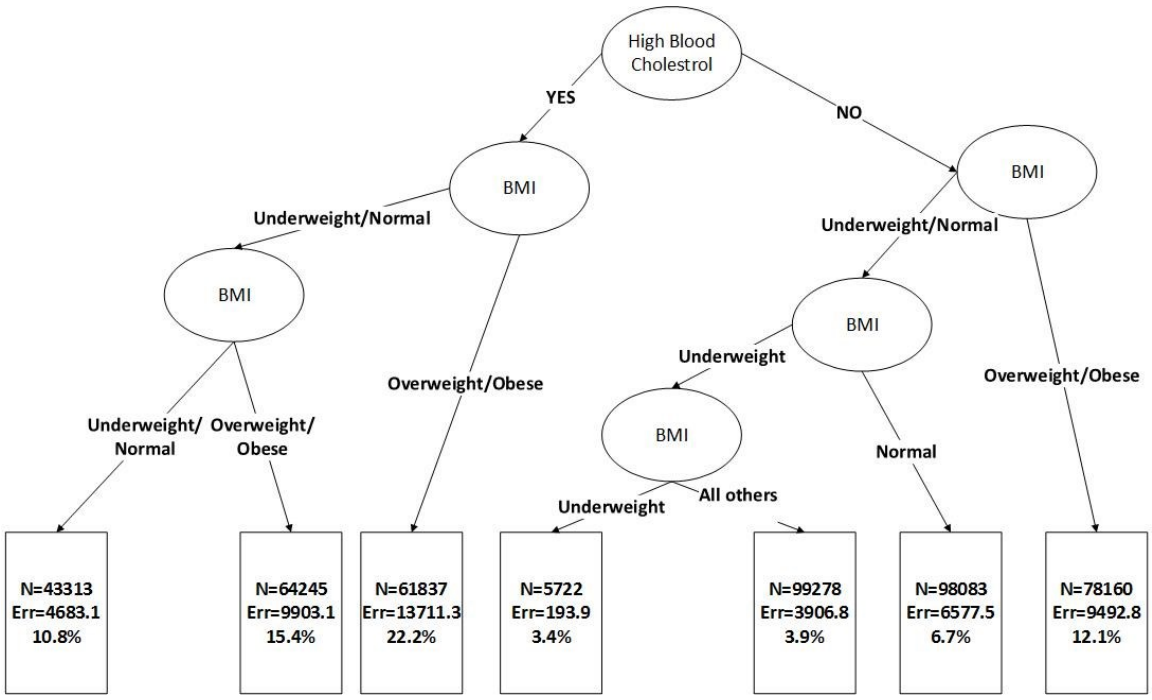


**Figure 2. Conditional Inference Tree for Diabetes with High Cholesterol and BMI**

The third tree showed the relationship between High Blood Cholesterol (TOLDHI2) and High Blood Pressure (BPHIGH4) and Diabetes (DIABETE3). Figure 3 shows that this decision tree had error rates ranging from 3.93% - 21.5%. The average error rate for this tree was 11.01%, meaning that its classification value was slightly poorer compared when compared to the second tree.

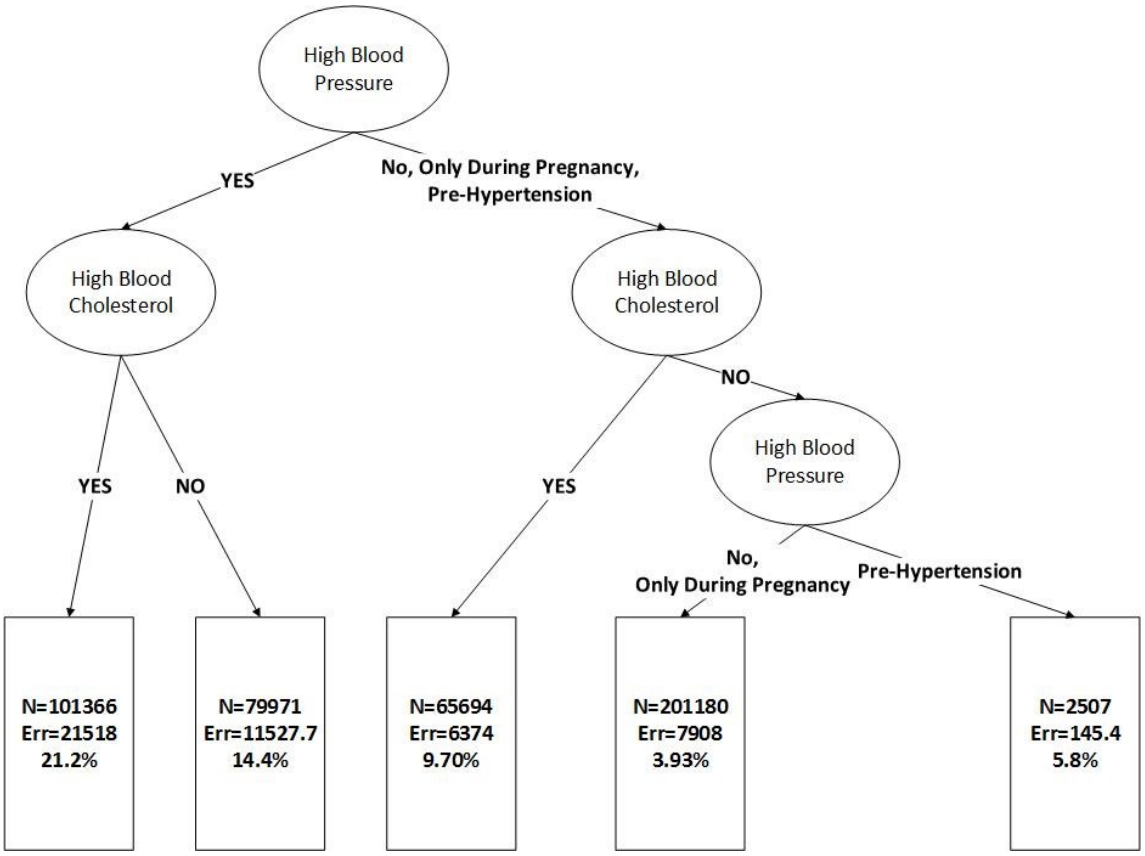


**Figure 3. Conditional Inference Tree for Diabetes, High Blood Pressure, and Cholesterol**

A fourth tree was created showing the relationship between High Blood Cholesterol (TOLDHI2) and High Blood Pressure (BPHIGH4) and Diabetes (DIABETE3) but created with training and validation data sets. Figure 4 shows these results. The data was split into training (80%) and validation (20%) data sets. The tree was then created using the training data set. A prediction using the predict (tree) function in R was then determined using the training data set. The prediction was then tested again using the validation data set. With the initial classification, there was an average error rate of 17.6%. When testing the prediction, there was an average error rate of 17.9%.

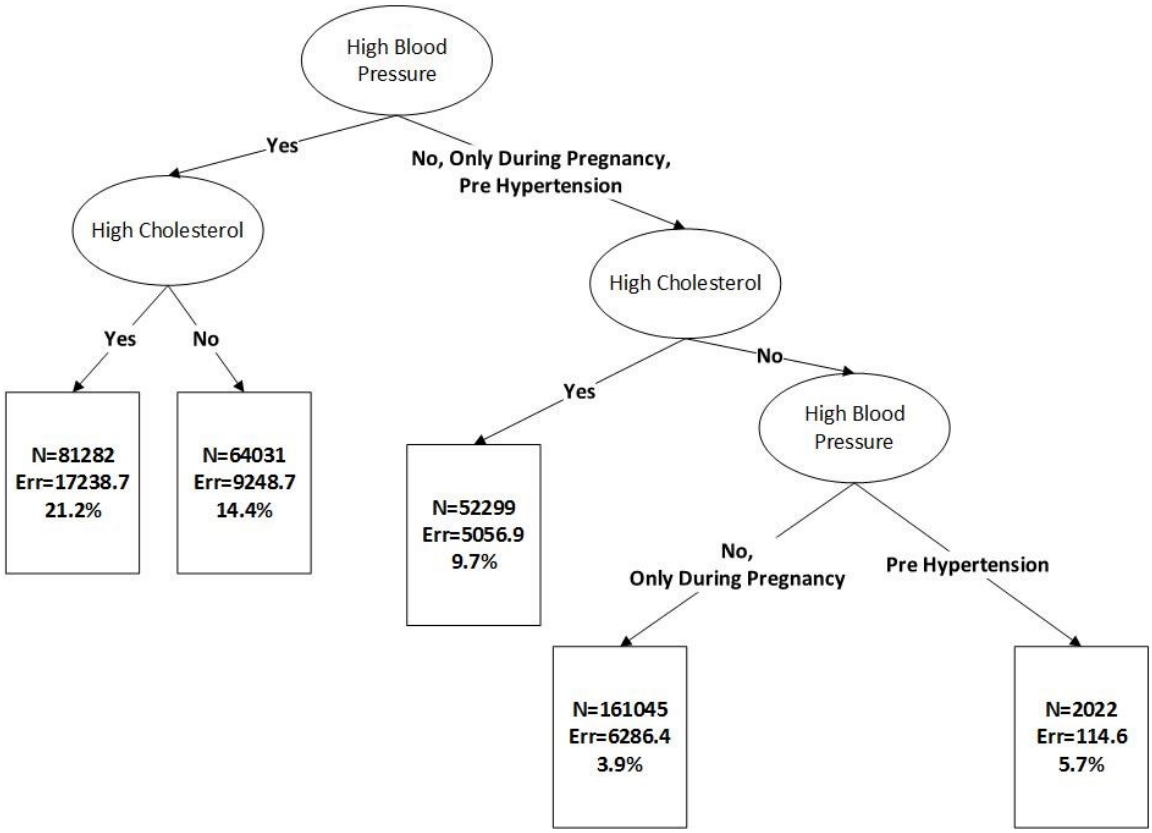


**Figure 4. Conditional Inference Tree for Diabetes with High Blood Pressure and High Blood Cholesterol Created with Training Data Set**

When compared to the tree using the entire data set, this model had a slightly poorer predictive value. However, this tree has the advantage of being able to make a prediction of disease risk versus just classifying individuals. Error rates for individual nodes are noted in Table 11.

**Table 11. Results for Conditional Inference Tree Prediction Using Validation Data**

| Tree 4 High Blood Pressure and High Cholesterol, Prediction with Training Data | | | |
|---|---|---|---|
| Node | N | Error | Pct Error |
| 1 | 56468 | 24814 | 43.94% |
| 2 | 52819 | 11212 | 21.23% |
| 3 | 46627 | 5672 | 12.16% |
| 4 | 1900 | 122 | 6.42% |
| 5 | 154492 | 6553 | 4.24% |
| | | Average | 17.60% |
| Tree 4 High Blood Pressure and High Cholesterol, Prediction Tested with Validation Data | | | |
| Node | N | Error | Pct Error |
| 1 | 13911 | 6199 | 44.56% |
| 2 | 13176 | 2738 | 20.78% |
| 3 | 11833 | 1477 | 12.48% |
| 4 | 454 | 33 | 7.27% |
| 5 | 38440 | 1698 | 4.42% |
| | | Average | 17.90% |

# 7. DISCUSSION

To predict disease associated with risk factors, healthcare literature was searched, variables were identified and matched with a corresponding BRFSS variables from CDC research data. Research models were then crafted to test prediction or inference classification. Statistical tests were calculated, and models tested using various analysis methods.

The binary logistic regression model showed that sex, body mass index, age, race, and alcohol consumption were all significant risk factors in predicting breast cancer. Women were 86.104 times more likely to develop breast cancer than men, which is consistent with previous literature. The American Cancer Society (2017b) reports that breast cancer is 100 times more common in women than in men. There was a bimodal distribution when it came to breast cancer and age and there was some significance seen in women who were younger, ages 30-44 years. Older women, however, were still more likely to have had breast cancer, with women 80+ having the highest odds ratio. This is somewhat consistent with previous research, as the American Cancer Society (2017b) identifies women 55+ being at higher risk. The model did identify a younger age group that would typically not be considered at higher risk. This may be the effect of the specific survey sample, or other unknown risk factors not captured in this study, such as the BRCA1 or BRCA2 genetic mutation which puts younger women at higher risk but is only responsible for 5-10% of all breast cancers (American Cancer Society, 2017b).

In considering weight, women who were underweight, normal weight, and obese were significant groups to consider when predicting breast cancer. Previous research identifies women who are overweight or obese to be at the highest risk, so the

insignificant result for the overweight variable was not consistent. However, the American Cancer Society (2017b) reports that the link between body weight and breast cancer risk is complex. For example, where a person carries their weight (waist vs. hips and thighs) or when weight was gained (adulthood vs. childhood) are important considerations. Weight also has different effects on different types of breast cancer. For example, women who are overweight before menopause have a higher risk of triple negative breast cancer. Triple negative breast cancer is when breast cancer cells don't have estrogen or progesterone receptors and don't have too much of the HER2 protein. These breast cancers are more aggressive than most other breast cancers and cannot be treated with hormone therapy or targeted cancer drugs (American Cancer Society, 2017a). These weight complexities may have influenced the result, since the patients were randomly sampled and were not identified by breast cancer sub-types.

When looking at race, White women and other races were predictive factors for breast cancer. Black and multi race women were only marginally significant. White women had the highest odds, being 7.043 times more likely to have had breast cancer. This is somewhat consistent with literature, as White women are slightly more likely to develop breast cancer compared to Black women. However, in women under age 45, Black women are more likely to develop breast cancer and more likely to die from it at any age (American Cancer Society, 2017b). The result is consistent that Hispanic women have a lower risk of developing or dying from breast cancer and being Hispanic was not significant in our analysis. The multi and other race categories could be significant due to self-identification of race by respondents and being mixes of higher and lower risk race groups, such as Asian and Black.

The multinomial logistic regression model showed that age, race, weight, smoking status, alcohol consumption, and healthy diet were all significant in predicting cancer. Infections, such as HIV/AIDS, were not significant in the multinomial regression model. Race was significant, although not identified as a risk factor by the National Cancer Institute (2015). These results show that our data is consistent with previous research and that it could be used to accurately identify and classify individuals with behavioral risk factors.

Four conditional inference decision trees were created, three to classify people with regards to diabetes, and one to predict the disease. The first three trees were created using the entire sample without splitting the data into training and validation sets. The first classification tree looked at the relationship between exercise, stroke, depression, and diabetes. This tree had an average error rate of 17.46%, meaning it was correct in classifying diabetes or no diabetes in individuals 82.54% of the time. This tree had the highest error rate, which could indicate that these particular risk factors are weaker in predicting diabetes.

The second classification tree looked at the relationship between high blood cholesterol, BMI, and Diabetes. This tree had an average error rate of 10.64%, meaning it was correct in classifying diabetes or no diabetes in an individual 89.36% of the time. This tree had the best classification capability and the lowest error rate, which could indicate that these risk factors are more strongly associated with diabetes.

The third classification tree looked at the relationship between high blood cholesterol, high blood pressure, and diabetes. This tree had an average error rate of 11.01%, meaning it was correct in classifying diabetes or no diabetes in individuals

88.99% of the time, making it slightly poorer in classification compared to the second tree. This could indicate that these risk factors are good classifiers of diabetes.

The fourth conditional inference tree was created with the purpose of predicting which individuals would have diabetes using machine learning. The tree looked at the relationship between high blood cholesterol, high blood pressure, and diabetes, similar to the third tree. This tree had an average error rate of 17.6% when evaluated for classification strength. When the prediction was tested using the validation data set, there was an error rate of 17.9%, meaning it was correct at predicting whether an individual had or did not have diabetes 82.1% of the time. Overall this tree performed slightly poorer compared to the third tree when it came to predictive capability. This could be a result of splitting the data into two sets, instead of using the whole data set with more values.

HIM professionals should consider using decision trees for classification purposes and predictive analysis for disease outcomes. We must also consider the value of machine learning using decision trees and being able to predict who may develop a disease in the future. Another consideration is the combination of these risk factors and their combined effect. Different combinations of risk factors could affect prediction and classification results, and many combinations could be tested to identify the strongest predictive values for each disease. Risks are correlated and dependent on each other, and so predictive models need to address multiple simultaneous conditions, requiring examination of correlations among interactions of multimodal clinical signs and risk factors (Belle et al., 2015).

Healthcare providers need resources, expertise, and available time to utilize big

data predictive analytics. However many report that incomplete data and insufficient technology are the biggest obstacles in implementing predictive analytics (Society of Actuaries, 2016). Hospitals are more likely to lack sufficient technology required to take advantage of predictive analytics. Staff also need to be educated on biostatistics and predictive analytics. However, medical groups and clinics are twice as likely to lack employees who are skilled in predictive analytics (Society of Actuaries, 2016).

The healthcare industry has historically made decisions differently than other business sectors (Society of Actuaries, 2016). In the 90s, there was a push for evidence-based medicine, which can help doctors provide the optimum disease management for their patients. There are several primary ideas in the use of evidence-based medicine such as clinical decisions needing to be based on best available scientific evidence. The clinical issue-rather than habit or protocol- should determine medical intervention. The best evidence often includes epidemiological and biostatistical ways of thinking. Information from critical evidence is only useful if it's put into action in making clinical decisions, and we should be constantly monitoring performance (Davidoff, Haynes, Sackett, & Smith, 1995).

To transition to evidence-based practice, medical authorities will have to adapt a new way of thinking about research, including switching from the primary use of deductive reasoning to inductive reasoning and pattern recognition (Krumholz, 2014). Medical researchers and clinicians will also need to begin utilizing machine learning, data mining, and other advanced analytic techniques, which will require more resources and new training in data science.

However, doctors are notoriously busy and may not have time to read countless

medical journal articles or run statistical analysis on their electronic health record data. Some in the medical community think that the evidence-based medicine movement is in crisis. The sheer volume of evidence has become unmanageable, and evidence-based guidelines often translate poorly to complex medical problems (Greenhalgh, Howick, & Maskrey, 2014). Clinicians must learn to sift through an unfathomable amount of data and clinical guidelines to find marginal benefits in clinical practice.

However, in a country that continues to be ravaged by chronic disease, there is still tremendous value for using evidence-based medicine in clinical decisions. To be effective, evidence-based medicine must be individualized to the patient. The clinician must not be merely bound by rules and guidelines but be taught to apply those rules in the context of each patient. A recent campaign in the United Kingdom, "Too Much Medicine," led by academics, clinicians, and patients is hoping to reduce over screening, overdiagnosis, and overtreatment and increase the use of personalized medicine (Greenhalgh et al., 2014).

## Limitations

There were some limitations regarding the nature of self-reported survey results. One limitation is the clumping of data around the whole number. For example, when asked their weight, respondents would be more likely to report 150 than 151 lbs. Data smoothing to account for this effect can be done but is complex and out of scope for this paper. There are also other limitations with self-reported data, such as people underestimating their tobacco/alcohol usage, misreporting their age, overestimating frequency of seeking healthcare and following medical advice.

Another limitation concerns cancer risk factors. The patients in this survey have already been diagnosed with cancer and may have altered their lifestyle due to the disease. For example, a smoker may have quit upon their diagnosis and reported that they do not currently smoke. Or a breast cancer patient who was overweight or obese prior to diagnosis may have lost weight because of cancer treatment or changing their lifestyle. The final limitation is that not all disease risk factors that were identified in literature had corresponding BRFSS variables and had to be excluded from analysis. It is possible these other risk factors had effect on the disease state but could not be analyzed.

## Future Research

Looking forward, the next step might be standardization of disease risk prediction models for clinical use. The Society of Actuaries conducted a survey analyzing the state of predictive analytics in healthcare. The survey identified that within the U.S. healthcare industry, fewer than half (43%) of healthcare organizations are currently using predictive analytics (Society of Actuaries, 2016). While most payers in healthcare are using predictive analytics (80%), only 39% of medical groups/clinics, and only 36% of hospitals are using these tools. For those who are using predictive analytics, the most common use is predicting hospital readmissions and costs. Medical groups and clinics, which is where predictive analytics could be used to predict chronic disease, were more likely to predict adverse events (Society of Actuaries, 2016).

Future research would include implementing the predictive and classification models created in this study into the clinical setting. This would include using models to create clinical decision support tools and evaluating their usefulness in medical practice.

# 8. CONCLUSION

Although disease prevention awareness campaigns have become more prevalent, the United States continues to be ravaged by chronic disease, in both mortality and cost. In their most recent report, the Partnership to Fight Chronic Disease (2016) estimates the projected total cost of chronic disease in America to reach $42 trillion between 2016-2030. The number of people with three or more chronic diseases in the United States is expected to reach 83.4 million by 2030, compared to 30.8 million in 2015. With behavioral changes, new interventions, and treatment advances, 16 million lives could be saved in the next 15 years (Partnership to Fight Chronic Disease, 2016).

Although many state of the art diagnostic and disease classification tools exist, the healthcare field is still lacking comprehensive predictive models to prevent chronic disease and plan interventions. Using big data predictive analytics could provide a definitive risk profile for each individual patient and help personalize interventions. Healthcare providers currently say that clinical outcomes and costs are the most valuable data to predict (Society of Actuaries, 2016). By focusing on prediction of disease risk we can improve clinical outcomes and reduce costs. As HIM professionals, we are responsible for assisting clinicians and health care organizations in utilizing big data predictive analytics, maintaining clean data, and bridging the gap between clinicians and data scientists.

# APPENDIX

## Codebook for BRFSS 2017 Data:

https://www.cdc.gov/brfss/annual_data/2017/pdf/codebook17_llcp-v2-508.pdf

## Multinomial regression parameter estimates for individual variables:

**Parameter Estimates**

| GENCNCR[a] | | B | Std. Error | Wald | df | Sig. | Exp(B) | 95% Confidence Interval for Exp (B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower Bound | Upper Bound |
| 0 | Intercept | 34.312 | 453.282 | .006 | 1 | .940 | | | |
| | [@_AGEG5YR=1] | 11.592 | 149.615 | .006 | 1 | .938 | 108233.365 | 4.808E-123 | 2.436E+132 |
| | [@_AGEG5YR=2] | 11.616 | 162.271 | .005 | 1 | .943 | 110905.525 | 8.308E-134 | 1.480E+143 |
| | [@_AGEG5YR=3] | .899 | 1.230 | .534 | 1 | .465 | 2.457 | .220 | 27.384 |
| | [@_AGEG5YR=4] | .296 | 1.124 | .069 | 1 | .792 | 1.344 | .149 | 12.158 |
| | [@_AGEG5YR=5] | .299 | 1.124 | .071 | 1 | .790 | 1.348 | .149 | 12.197 |
| | [@_AGEG5YR=6] | .813 | 1.160 | .491 | 1 | .483 | 2.255 | .232 | 21.911 |
| | [@_AGEG5YR=7] | -.037 | 1.060 | .001 | 1 | .972 | .963 | .121 | 7.690 |
| | [@_AGEG5YR=8] | -.942 | 1.024 | .846 | 1 | .358 | .390 | .052 | 2.900 |
| | [@_AGEG5YR=9] | -1.157 | 1.018 | 1.291 | 1 | .256 | .314 | .043 | 2.314 |
| | [@_AGEG5YR=10] | -1.259 | 1.017 | 1.532 | 1 | .216 | .284 | .039 | 2.085 |
| | [@_AGEG5YR=11] | -1.521 | 1.016 | 2.241 | 1 | .134 | .218 | .030 | 1.601 |
| | [@_AGEG5YR=12] | -1.784 | 1.017 | 3.076 | 1 | .079 | .168 | .023 | 1.233 |
| | [@_AGEG5YR=13] | -2.167 | 1.013 | 4.575 | 1 | .032 | .115 | .016 | .834 |
| | [@_AGEG5YR=14] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACE=1] | -1.034 | .711 | 2.112 | 1 | .146 | .356 | .088 | 1.434 |
| | [@_RACE=2] | -.879 | .743 | 1.398 | 1 | .237 | .415 | .097 | 1.783 |
| | [@_RACE=3] | -1.849 | .771 | 5.749 | 1 | .017 | .157 | .035 | .714 |
| | [@_RACE=4] | 10.956 | 200.513 | .003 | 1 | .956 | 57301.243 | 1.205E-166 | 2.724E+175 |
| | [@_RACE=5] | 11.044 | 543.442 | .000 | 1 | .984 | 62595.610 | .000 | .[c] |
| | [@_RACE=6] | 11.470 | 393.784 | .001 | 1 | .977 | 95798.198 | .000 | .[c] |
| | [@_RACE=7] | -.099 | 1.001 | .010 | 1 | .921 | .905 | .127 | 6.444 |
| | [@_RACE=8] | .676 | .915 | .546 | 1 | .460 | 1.966 | .327 | 11.816 |
| | [@_RACE=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_SMOKER3=1] | -.622 | 1.026 | .368 | 1 | .544 | .537 | .072 | 4.013 |
| | [@_SMOKER3=2] | -.329 | 1.083 | .092 | 1 | .761 | .719 | .086 | 6.006 |
| | [@_SMOKER3=3] | -.732 | 1.006 | .530 | 1 | .467 | .481 | .067 | 3.453 |
| | [@_SMOKER3=4] | -.685 | 1.004 | .465 | 1 | .495 | .504 | .070 | 3.609 |
| | [@_SMOKER3=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RFSMOK3=1] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RFSMOK3=2] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RFSMOK3=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACEGR3=1] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACEGR3=2] | 0[b] | . | . | 0 | . | . | . | . |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | [@_RACEGR3=2] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACEGR3=3] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACEGR3=4] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACEGR3=5] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACEGR3=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [DRNKANY5=1] | -11.864 | 453.279 | .001 | 1 | .979 | 7.043E-6 | .000 | .[c] |
| | [DRNKANY5=2] | -12.167 | 453.279 | .001 | 1 | .979 | 5.200E-6 | .000 | .[c] |
| | [DRNKANY5=7] | .175 | 552.153 | .000 | 1 | 1.000 | 1.192 | .000 | .[c] |
| | [DRNKANY5=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_VEGLT1A=1] | -.158 | .308 | .265 | 1 | .606 | .853 | .467 | 1.559 |
| | [@_VEGLT1A=2] | -.391 | .330 | 1.400 | 1 | .237 | .677 | .354 | 1.292 |
| | [@_VEGLT1A=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_FRTLT1A=1] | -.245 | .373 | .431 | 1 | .511 | .783 | .377 | 1.626 |
| | [@_FRTLT1A=2] | -.084 | .382 | .048 | 1 | .826 | .919 | .434 | 1.946 |
| | [@_FRTLT1A=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RFBMI5=1] | .263 | .296 | .790 | 1 | .374 | 1.301 | .728 | 2.326 |
| | [@_RFBMI5=2] | -.154 | .278 | .307 | 1 | .579 | .857 | .497 | 1.479 |
| | [@_RFBMI5=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [HIVRISK5=1] | -12.175 | .873 | 194.464 | 1 | .000 | 5.156E-6 | 9.314E-7 | 2.854E-5 |
| | [HIVRISK5=2] | -12.023 | .710 | 286.883 | 1 | .000 | 6.007E-6 | 1.494E-6 | 2.415E-5 |
| | [HIVRISK5=7] | .067 | 1114.868 | .000 | 1 | 1.000 | 1.069 | .000 | .[c] |
| | [HIVRISK5=9] | 0[b] | . | . | 0 | . | . | . | . |
| 1 | Intercept | 24.415 | 453.283 | .003 | 1 | .957 | | | |
| | [@_AGEG5YR=1] | -.404 | 168.347 | .000 | 1 | .998 | .667 | 3.367E-144 | 1.323E+143 |
| | [@_AGEG5YR=2] | -.385 | 182.611 | .000 | 1 | .998 | .681 | 2.479E-156 | 1.868E+155 |
| | [@_AGEG5YR=3] | -2.392 | 1.638 | 2.133 | 1 | .144 | .091 | .004 | 2.266 |
| | [@_AGEG5YR=4] | -1.477 | 1.278 | 1.337 | 1 | .247 | .228 | .019 | 2.791 |
| | [@_AGEG5YR=5] | -1.136 | 1.255 | .820 | 1 | .365 | .321 | .027 | 3.757 |
| | [@_AGEG5YR=6] | .640 | 1.244 | .265 | 1 | .607 | 1.897 | .166 | 21.733 |
| | [@_AGEG5YR=7] | .063 | 1.145 | .003 | 1 | .956 | 1.065 | .113 | 10.058 |
| | [@_AGEG5YR=8] | -.533 | 1.109 | .232 | 1 | .630 | .587 | .067 | 5.151 |
| | [@_AGEG5YR=9] | -.486 | 1.102 | .195 | 1 | .659 | .615 | .071 | 5.328 |
| | [@_AGEG5YR=10] | -.003 | 1.099 | .000 | 1 | .998 | .997 | .116 | 8.594 |
| | [@_AGEG5YR=11] | -.254 | 1.099 | .053 | 1 | .817 | .776 | .090 | 6.685 |
| | [@_AGEG5YR=12] | -.218 | 1.100 | .039 | 1 | .843 | .804 | .093 | 6.943 |
| | [@_AGEG5YR=13] | -.423 | 1.095 | .149 | 1 | .699 | .655 | .077 | 5.603 |
| | [@_AGEG5YR=14] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACE=1] | -.014 | .795 | .000 | 1 | .986 | .986 | .207 | 4.684 |
| | [@_RACE=2] | -.071 | .832 | .007 | 1 | .932 | .931 | .182 | 4.755 |
| | [@_RACE=3] | -.496 | .871 | .324 | 1 | .569 | .609 | .110 | 3.357 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [@_AGEG5YR=13] | -.423 | 1.095 | .149 | 1 | .699 | .655 | .077 | 5.603 |
| [@_AGEG5YR=14] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RACE=1] | -.014 | .795 | .000 | 1 | .986 | .986 | .207 | 4.684 |
| [@_RACE=2] | -.071 | .832 | .007 | 1 | .932 | .931 | .182 | 4.755 |
| [@_RACE=3] | -.496 | .871 | .324 | 1 | .569 | .609 | .110 | 3.357 |
| [@_RACE=4] | 11.199 | 200.514 | .003 | 1 | .955 | 73032.441 | 1.534E-166 | 3.477E+175 |
| [@_RACE=5] | .016 | 608.863 | .000 | 1 | 1.000 | 1.016 | .000 | .[c] |
| [@_RACE=6] | 11.701 | 393.785 | .001 | 1 | .976 | 120639.650 | .000 | .[c] |
| [@_RACE=7] | .578 | 1.095 | .278 | 1 | .598 | 1.782 | .208 | 15.244 |
| [@_RACE=8] | -.231 | 1.036 | .050 | 1 | .824 | .794 | .104 | 6.053 |
| [@_RACE=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_SMOKER3=1] | -.108 | 1.125 | .009 | 1 | .924 | .898 | .099 | 8.138 |
| [@_SMOKER3=2] | .075 | 1.185 | .004 | 1 | .949 | 1.078 | .106 | 10.995 |
| [@_SMOKER3=3] | -.547 | 1.102 | .246 | 1 | .620 | .579 | .067 | 5.019 |
| [@_SMOKER3=4] | -.136 | 1.100 | .015 | 1 | .902 | .873 | .101 | 7.542 |
| [@_SMOKER3=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RFSMOK3=1] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RFSMOK3=2] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RFSMOK3=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RACEGR3=1] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RACEGR3=2] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RACEGR3=3] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RACEGR3=4] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RACEGR3=5] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RACEGR3=9] | 0[b] | . | . | 0 | . | . | . | . |
| [DRNKANY5=1] | -11.563 | 453.279 | .001 | 1 | .980 | 9.507E-6 | .000 | .[c] |
| [DRNKANY5=2] | -11.588 | 453.279 | .001 | 1 | .980 | 9.276E-6 | .000 | .[c] |
| [DRNKANY5=7] | .268 | 552.153 | .000 | 1 | 1.000 | 1.307 | .000 | .[c] |
| [DRNKANY5=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_VEGLT1A=1] | .243 | .353 | .472 | 1 | .492 | 1.275 | .638 | 2.546 |
| [@_VEGLT1A=2] | -.049 | .380 | .017 | 1 | .897 | .952 | .452 | 2.004 |
| [@_VEGLT1A=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_FRTLT1A=1] | .651 | .453 | 2.070 | 1 | .150 | 1.918 | .790 | 4.659 |
| [@_FRTLT1A=2] | .608 | .463 | 1.723 | 1 | .189 | 1.836 | .741 | 4.551 |
| [@_FRTLT1A=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RFBMI5=1] | .091 | .327 | .078 | 1 | .779 | 1.096 | .578 | 2.079 |
| [@_RFBMI5=2] | -.224 | .308 | .531 | 1 | .466 | .799 | .437 | 1.461 |
| [@_RFBMI5=9] | 0[b] | . | . | 0 | . | . | . | . |
| [HIVRISK5=1] | -12.466 | .653 | 364.230 | 1 | .000 | 3.855E-6 | 1.072E-6 | 1.387E-5 |
| [HIVRISK5=2] | -11.580 | .000 | . | 1 | . | 9.356E-6 | 9.356E-6 | 9.356E-6 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | [@_RFBMI5=1] | .091 | .327 | .078 | 1 | .779 | 1.096 | .578 | 2.079 |
| | [@_RFBMI5=2] | -.224 | .308 | .531 | 1 | .466 | .799 | .437 | 1.461 |
| | [@_RFBMI5=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [HIVRISK5=1] | -12.466 | .653 | 364.230 | 1 | .000 | 3.855E-6 | 1.072E-6 | 1.387E-5 |
| | [HIVRISK5=2] | -11.580 | .000 | . | 1 | . | 9.356E-6 | 9.356E-6 | 9.356E-6 |
| | [HIVRISK5=7] | -11.681 | 1265.806 | .000 | 1 | .993 | 8.455E-6 | .000 | .[c] |
| | [HIVRISK5=9] | 0[b] | . | . | 0 | . | . | . | . |
| 2 | Intercept | -10.604 | 975.241 | .000 | 1 | .991 | | | |
| | [@_AGEG5YR=1] | 11.183 | 434.360 | .001 | 1 | .979 | 71867.767 | .000 | .[c] |
| | [@_AGEG5YR=2] | 11.119 | 443.606 | .001 | 1 | .980 | 67440.238 | .000 | .[c] |
| | [@_AGEG5YR=3] | 10.762 | 367.444 | .001 | 1 | .977 | 47175.162 | .000 | .[c] |
| | [@_AGEG5YR=4] | 10.090 | 367.443 | .001 | 1 | .978 | 24093.847 | .000 | .[c] |
| | [@_AGEG5YR=5] | -.176 | 408.357 | .000 | 1 | 1.000 | .839 | .000 | .[c] |
| | [@_AGEG5YR=6] | 11.910 | 367.442 | .001 | 1 | .974 | 148753.865 | 2.541E-308 | .[c] |
| | [@_AGEG5YR=7] | 10.792 | 367.442 | .001 | 1 | .977 | 48648.512 | .000 | .[c] |
| | [@_AGEG5YR=8] | 10.920 | 367.442 | .001 | 1 | .976 | 55286.886 | .000 | .[c] |
| | [@_AGEG5YR=9] | 10.845 | 367.442 | .001 | 1 | .976 | 51294.534 | .000 | .[c] |
| | [@_AGEG5YR=10] | 11.019 | 367.442 | .001 | 1 | .976 | 61029.812 | .000 | .[c] |
| | [@_AGEG5YR=11] | 10.963 | 367.442 | .001 | 1 | .976 | 57714.670 | .000 | .[c] |
| | [@_AGEG5YR=12] | 11.031 | 367.442 | .001 | 1 | .976 | 61786.525 | .000 | .[c] |
| | [@_AGEG5YR=13] | 10.403 | 367.442 | .001 | 1 | .977 | 32954.174 | .000 | .[c] |
| | [@_AGEG5YR=14] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACE=1] | -.652 | 1.007 | .419 | 1 | .518 | .521 | .072 | 3.752 |
| | [@_RACE=2] | -1.057 | 1.094 | .932 | 1 | .334 | .348 | .041 | 2.969 |
| | [@_RACE=3] | -1.539 | 1.196 | 1.656 | 1 | .198 | .214 | .021 | 2.237 |
| | [@_RACE=4] | -.215 | 315.594 | .000 | 1 | .999 | .807 | 1.874E-269 | 3.474E+268 |
| | [@_RACE=5] | -.389 | 843.977 | .000 | 1 | 1.000 | .678 | .000 | .[c] |
| | [@_RACE=6] | -.451 | 627.553 | .000 | 1 | .999 | .637 | .000 | .[c] |
| | [@_RACE=7] | -.853 | 1.583 | .290 | 1 | .590 | .426 | .019 | 9.480 |
| | [@_RACE=8] | -.241 | 1.357 | .032 | 1 | .859 | .786 | .055 | 11.223 |
| | [@_RACE=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_SMOKER3=1] | 12.169 | 516.422 | .001 | 1 | .981 | 192739.284 | .000 | .[c] |
| | [@_SMOKER3=2] | 13.192 | 516.422 | .001 | 1 | .980 | 536031.721 | .000 | .[c] |
| | [@_SMOKER3=3] | 12.443 | 516.422 | .001 | 1 | .981 | 253416.789 | .000 | .[c] |
| | [@_SMOKER3=4] | 10.637 | 516.422 | .000 | 1 | .984 | 41649.076 | .000 | .[c] |
| | [@_SMOKER3=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RFSMOK3=1] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RFSMOK3=2] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RFSMOK3=9] | 0[b] | . | . | 0 | . | . | . | . |
| | [@_RACEGR3=1] | 0[b] | . | . | 0 | . | . | . | . |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [DRNKANY5=1] | -13.068 | 453.280 | .001 | 1 | .977 | 2.113E-6 | .000 | .[c] |
| [DRNKANY5=2] | -12.922 | 453.280 | .001 | 1 | .977 | 2.444E-6 | .000 | .[c] |
| [DRNKANY5=7] | -.208 | 552.154 | .000 | 1 | 1.000 | .813 | .000 | .[c] |
| [DRNKANY5=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_VEGLT1A=1] | -.688 | .450 | 2.341 | 1 | .126 | .502 | .208 | 1.213 |
| [@_VEGLT1A=2] | -.821 | .496 | 2.745 | 1 | .098 | .440 | .167 | 1.162 |
| [@_VEGLT1A=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_FRTLT1A=1] | .379 | .710 | .285 | 1 | .593 | 1.461 | .363 | 5.872 |
| [@_FRTLT1A=2] | .811 | .719 | 1.274 | 1 | .259 | 2.250 | .550 | 9.201 |
| [@_FRTLT1A=9] | 0[b] | . | . | 0 | . | . | . | . |
| [@_RFBMI5=1] | 1.780 | .779 | 5.219 | 1 | .022 | 5.933 | 1.288 | 27.330 |
| [@_RFBMI5=2] | .940 | .769 | 1.494 | 1 | .222 | 2.559 | .567 | 11.546 |
| [@_RFBMI5=9] | 0[b] | . | . | 0 | . | . | . | . |
| [HIVRISK5=1] | .777 | 586.453 | .000 | 1 | .999 | 2.174 | .000 | .[c] |
| [HIVRISK5=2] | -.172 | 586.453 | .000 | 1 | 1.000 | .842 | .000 | .[c] |
| [HIVRISK5=7] | -.007 | 1908.278 | .000 | 1 | 1.000 | .993 | .000 | .[c] |
| [HIVRISK5=9] | 0[b] | . | . | 0 | . | . | . | . |

a. The reference category is: 3.

b. This parameter is set to zero because it is redundant.

c. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

## REFERENCES


Aarons, G. A. (2006). Transformational and transactional leadership: Association with attitudes toward evidence-based practice. *Psychiatric services, 57*(8), 1162-1169.

American Cancer Society. (2017a). Breast Cancer Hormone Receptor Status. *Understanding a Breast Cancer Diagnosis* Retrieved from https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html?_ga=2.24704954.1473650592.1547140049-1638978545.1536782261

American Cancer Society. (2017b, 9/6/2017). Lifestyle-Related Breast Cancer Risk Factors, Breast Cancer Risk Factors you Cannot Change *Breast Cancer Risk and Prevention* Retrieved from https://www.cancer.org/cancer/breast-cancer/risk-and-prevention.html

American Heart Association. (2014, 8/15/14). Alcohol and Heart Health. *Healthy Living* Retrieved from http://www.heart.org/en/healthy-living/healthy-eating/eat-smart/nutrition-basics/alcohol-and-heart-health

Bae, J.-M. (2014). The clinical decision analysis using decision tree. *Epidemiology and Health, 36*, e2014025. doi:10.4178/epih/e2014025

Barrett, M. A., Humblet, O., Hiatt, R. A., & Adler, N. E. (2013). Big data and disease prevention: from quantified self to quantified communities. *Big Data, 1*(3), 168-175.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123-1131.

Belle, A., Thiagarajan, R., Soroushmehr, S., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. *BioMed research international, 2015*.

Centers for Disease Control and Prevention. (2016). *2016 Behavioral Risk Factor Surveillance System Survey*

Centers for Disease Control and Prevention. (2018a). *Calculated Variables in the 2017 Behavioral Risk Factor Surveillance System Data File* Retrieved from https://www.cdc.gov/brfss/annual_data/2017/pdf/2017-calculated-variables-version4-508.pdf

Centers for Disease Control and Prevention. (2018b). *Overview: BRFSS 2017*. Retrieved from https://www.cdc.gov/brfss/annual_data/2017/pdf/overview-2017-508.pdf

Chan, Y. H. (2005). Biostatistics 305. Multinomial logistic regression. *Singapore medical journal, 46*(6), 259.

Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine, 28*(3), 660-665.

Davidoff, F., Haynes, B., Sackett, D., & Smith, R. (1995). Evidence based medicine. *BMJ: British Medical Journal, 310*(6987), 1085.

Fox, B. (2011). Using big data for big impact. Leveraging data and analytics provides the foundation for rethinking how to impact patient behavior. *Health Management Technology, 32*(11), 16-16.

Garson, G. D. (2015). Missing values analysis and data imputation. *Asheboro: Statistical Associates Publishing Asheboro, NC*.

Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: a movement in crisis? *Bmj, 348*, g3725.

IBM. (2012). Decision Tree Models. Retrieved from https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_treebuilding.htm

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). *Big data: Issues and challenges moving forward.* Paper presented at the System sciences (HICSS), 2013 46th Hawaii international conference on.

Kent, J. (2018). Big Data to See Explosive Growth, Challenging Healthcare Organizations. *Health IT Analytics*.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs, 33*(7), 1163-1170.

Lash, T. A., & Escobedo, M. R. (2018). Foreword. In: Elsevier.

Li, C., Balluz, L. S., Ford, E. S., Okoro, C. A., Zhao, G., & Pierannunzi, C. (2012). A comparison of prevalence estimates for selected health indicators and chronic diseases or conditions from the Behavioral Risk Factor Surveillance System, the National Health Interview Survey, and the National Health and Nutrition Examination Survey, 2007–2008. *Preventive medicine, 54*(6), 381-387.

Li, X., & Liu, B. (2014). Rule-Based Classification. In: Citeseer.

Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., & Yang, H.-J. (2017). Healthcare predictive analytics for risk profiling in chronic care: a bayesian multitask learning approach. *MIS Quarterly, 41*(2).

McLeod, A., & Dolezel, D. J. D. S. S. (2018). Cyber-analytics: Modeling factors associated with healthcare data breaches. *108*, 57-68.

Mikalef, P., Krogstie, J., van de Wetering, R., Pappas, I., & Giannakos, M. (2018). *Information Governance in the Big Data Era: Aligning Organizational Capabilities.* Paper presented at the Proceedings of the 51st Hawaii International Conference on System Sciences.

Mixymol, V. (2017). Disease Prediction and Risk Analysis using Classification Algorithms. *International Journal, 8*(5).

Monica, K. (2017). Why Are So Few Healthcare Providers Using EHR Data Analytics? *EHR Intelligence*

Moskowitz, A., McSparron, J., Stone, D. J., & Celi, L. A. (2015). Preparing a new generation of clinicians for the era of big data. *Harvard medical student review, 2*(1), 24.

National Cancer Institute. (2015). Risk Factors for Cancer. *Causes and Prevention* Retrieved from https://www.cancer.gov/about-cancer/causes-prevention/risk

National Institute of Diabetes and Digestive and Kidney Diseases. (2016, 11/2016). Risk Factors for Type 2 Diabetes. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes

Neff, G. (2013). Why big data won't cure us. *Big Data, 1*(3), 117-123.

Nichols, H. (2018). The top 10 leading causes of death in the United States. Retrieved from https://www.medicalnewstoday.com/articles/282929.php

Palaniappan, S., & Awang, R. (2008). *Intelligent heart disease prediction system using data mining techniques.* Paper presented at the Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on.

Partnership to Fight Chronic Disease. (2016). *What is the impact of chronic disease in America?* . Retrieved from FightChronicDisease.org http://www.fightchronicdisease.org/sites/default/files/pfcd_blocks/PFCD_US.FactSheet_FINAL1%20%282%29.pdf

Pierannunzi, C., Hu, S. S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC medical research methodology, 13*(1), 49.

Pyo, J. H., Hong, S. N., Min, B.-H., Lee, J. H., Chang, D. K., Rhee, P.-L., . . . Son, H. J. (2016). Evaluation of the risk factors associated with rectal neuroendocrine tumors: a big data analytic study from a health screening center. *Journal of gastroenterology, 51*(12), 1112-1121.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems, 2*(1), 3.

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data, 3*(4), 277-287.

Sarkar, S. K., & Nag, A. (2017). Identifying Patients at Risk of Breast Cancer through Decision Trees. *International Journal of Advanced Research in Computer Science, 8*(8).

Singh, N. K. (2015). Prediction of Breast Cancer using Rule Based Classification. *Applied Medical Informatics, 37*(4), 11-22.

Society of Actuaries. (2016). The State of Predictive Analytics in US Healthcare *Modern Healthcare*(November 2016).

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica, 24*(1), 12-18. doi:10.11613/BM.2014.003

Steinberg, G. B., Church, B. W., McCall, C. J., Scott, A. B., & Kalis, B. P. (2014). Novel predictive models for metabolic syndrome risk: a" big data" analytic approach. *The American journal of managed care, 20*(6), e221-228.

Turnea, M., & Ilea, M. (2018). *Predictive Simulation for Type II Diabetes Using Data Mining Strategies Applied to Big Data.* Paper presented at the The International Scientific Conference eLearning and Software for Education.

Viceconti, M., Hunter, P. J., & Hose, R. D. (2015). Big data, big knowledge: big data for personalized healthcare. *IEEE J. Biomedical and Health Informatics, 19*(4), 1209-1215.

Viceconti, M., Taddei, F., Cristofolini, L., Martelli, S., Falcinelli, C., & Schileo, E. (2012). Are spontaneous fractures possible? An example of clinical application for personalised, multiscale neuro-musculo-skeletal modelling. *Journal of biomechanics, 45*(3), 421-426.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation, 97*(18), 1837-1847.