NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

UTILIZATION ANALYSIS OF DOD HIV PATIENT CARE

by

Xiao C. Ren

June 2018

Thesis Advisor: Andrew T. Anglemyer
Second Reader: Lyn R. Whitaker

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE UTILIZATION ANALYSIS OF DOD HIV PATIENT CARE	**5. FUNDING NUMBERS** ICDRP
6. AUTHOR(S) Xiao C. Ren	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

Human immunodeficiency virus (HIV) remains a major health threat for the United States and the world, and HIV has affected the lives of millions of people including members in the U.S. military. The objective of this thesis is to analyze how HIV-positive active duty members use military healthcare. Using the dataset of active duty members who are currently engaging in the U.S. military HIV Natural History Study (NHS) cohort, the analysis examines the relationship among the annual number of appointments, the top three medical reasons for the appointments, the clinical laboratory data, and the patient's healthcare utilization during a period of five years after seroconversion. The four final models incorporated multiple linear regression, logistic regression, random forests, and cross-validation. These models can provide useful insights and predictions on healthcare utilization so that DoD medical planners can forecast and manage an appropriate amount of HIV care resources in the future.

14. SUBJECT TERMS HIV, patient care utilization in DoD	**15. NUMBER OF PAGES** 75
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

UTILIZATION ANALYSIS OF DOD HIV PATIENT CARE

Xiao C. Ren
Major, United States Air Force
BS, SUNY at Stony Brook, 2002
MBA, Webster University, 2007

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2018

Approved by: Andrew T. Anglemyer
Advisor

Lyn R. Whitaker
Second Reader

Patricia A. Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Human immunodeficiency virus (HIV) remains a major health threat for the United States and the world, and HIV has affected the lives of millions of people including members in the U.S. military. The objective of this thesis is to analyze how HIV-positive active duty members use military healthcare. Using the dataset of active duty members who are currently engaging in the U.S. military HIV Natural History Study (NHS) cohort, the analysis examines the relationship among the annual number of appointments, the top three medical reasons for the appointments, the clinical laboratory data, and the patient's healthcare utilization during a period of five years after seroconversion. The four final models incorporated multiple linear regression, logistic regression, random forests, and cross-validation. These models can provide useful insights and predictions on healthcare utilization so that DoD medical planners can forecast and manage an appropriate amount of HIV care resources in the future.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AIC	Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ART	antiretroviral therapy
Avg	average
CD4	cluster of differentiation 4
CDC	Centers for Disease Control
CSV	comma-separated values
DoD	Department of Defense
DoHHS	Department of Health and Human Services
ICD-9	The International Classification of Diseases, Ninth Revision
HIV	Human Immunodeficiency Virus
IDCRP	Infectious Disease Clinical Research Program
IQR	interquartile range
mL	milliliter
MTF	military treatment facility
MSE	mean square error
NA	not available
NHS	National History Study
MLR	multiple linear regression
PCS	permanent change of station
PG	Personnel-General
PIN	personal identification number
RF	random forests
RSS	residual sum of squares
SD	standard deviation
SECAF	Secretary of the Air Force
SECNAV	Secretary of the Navy
STI	sexually transmitted infection
UNAIDS	Joint United Nations Programme HIV/AIDS

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Human immunodeficiency virus (HIV) remains a major health threat for the United States and the world, and HIV has affected the lives of millions of people, including members in the U.S. military. The key strategy to end this epidemic is to meet the three components in the HIV care continuum for HIV-positive patients around the world (Centers for Disease Control 2017b).

The objective of this thesis is to analyze how HIV-positive active duty members use military healthcare. Using the dataset of active duty members who are currently engaging in the National History Study (NHS) cohort, the analysis examines the relationships among the annual number of appointments, the top three medical reasons for the appointments, the clinical laboratory data, and the patient's healthcare utilization and demographic during a period of five years after seroconversion (Infectious Disease Clinical Research Program 2015a).

The Department of Defense (DoD) healthcare has been meeting the three components of the HIV continuum care. First, all active duty members are mandated to get routine HIV tests. Second, HIV-positive active duty members have open access to healthcare while in active duty status. In addition, patients start antiretroviral therapy (ART) during initial visits in Year 1, and this treatment is aligned with the requirement in the care continuum. Third, the ultimate goal of the HIV continuum care is to achieve HIV viral suppression.

The cohort study results show that patients' viral loads decreased dramatically from Year 1 and during the following four years. The median HIV viral load of patients in Year 1 was 17,008 copies/mL with an interquartile range of 3,647 copies/mL and 41,400 copies/mL, and the median of viral load of patients in Year 2 was 52 copies/mL with an interquartile range of 34 copies/mL and 7,477 copies/mL. In Year 5, the median viral load was 35 copies/mL with an interquartile range of 20 copies/mL to 48 copies/mL. Clearly, the DoD healthcare system is one of the leaders in meeting and achieving the HIV care continuum requirements today.

This analysis further examines the inference statistics among 18 independent variables and the four dependent variables. The 18 independent variables include patient's age, marital status, military service and rank, health risk factors, and laboratory results. The four dependent variables include the average number of visits each year per patient and three binary variables of whether a patient had at least one visit for each of the top three medical reasons.

The preferred model in the primary analysis is the random forests model, and the model provides the most accurate predictions in average number of visits each year per patient. The random forests model has the lowest average mean squared error (MSE) compared to the two other average MSEs from the multiple linear regression model that used the purposeful selection of covariates methodology and the multiple linear regression model that applied the Akaike Information Criterion (AIC). The three average MSEs are derived from the 30 MSEs calculated from cross-validating each model 10 times.

The secondary analysis takes a closer look at the healthcare utilization of three specific medical reasons. The medical reasons are counseling, asymptomatic HIV diagnosis, and psychological stress. The logistic regression models created can predict whether a patient will have at least one visit for each of the three medical reasons or not.

This study has two main findings. First, compared to patients in the other services, patients in the Army have the smallest average number of medical visits annually and a smaller likelihood of having at least one visit for asymptomatic HIV and psychological stress. Second, patients in the Air Force had greater average number of counseling visits compared to patients from the other three services. Second, the likelihood of having at least one counseling visit is significantly higher for patients in the Air Force than those in the other three services.

This research considers various methods and approaches to regression analysis and model building that can be applied to studies of other types of healthcare utilization in the military or private-sector healthcare systems.

References

Centers for Disease Control (2017) Understanding the HIV Care Continuum. Accessed February 16, 2018, https://www.cdc.gov/hiv/pdf/library/factsheets/cdc-hiv-care-continuum.pdf.

Infectious Disease Clinical Research Program (2015) The U.S. Military HIV Natural History Study, 2015 Annual Report, National Institute of Health, ICDRP HIV Research Area, Bethesda, MD.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank Dr. Andrew Anglemyer and Dr. Lyn Whitaker for their assistance, guidance, and patience throughout my thesis process. Also, I am grateful to my family for their efforts and dedication to taking care business at home, so I could complete my thesis on time.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Human immunodeficiency virus (HIV) remains a major health threat for the United States and the world, and HIV has affected the lives of millions of people including members in the U.S. military. The total number of HIV diagnoses among active duty personnel has surpassed 10,000 since the epidemic started (Infectious Disease Clinical Research Program [IDCRP] 2016). Annually, that number rises by about 350 (IDCRP 2016). The Department of Defense (DoD) mandates all military members to receive routine HIV testing (Department of Defense [DoD] 2013), and the military healthcare system provides routine healthcare and conducts clinical researches on the infected population. One of the HIV research programs is the Infectious Disease Clinical Research's (IDCRP) U.S. military HIV National History Study (NHS), started in 1986, which has built a study cohort of more than 6,000 DoD members of whom 1,500 are actively engaging in the cohort (IDCRP 2015a). As the world battles to end the HIV epidemic, in 2015, the Presidential Executive Order on the U.S. HIV/Acquired Immunodeficiency Syndrome (AIDS) strategy emphasized the importance of "improving outcomes at every step of the HIV care continuum" (The White House 2015). Numerous studies and advancements have been completed on the HIV patients in DoD; however, DoD still needs a detailed understanding of the healthcare utilization and the effectiveness of care from care continuum to its active duty HIV-infected population.

The objective of this thesis is to analyze how HIV-positive active duty members utilize military healthcare as to achieve the second component of the HIV care continuum, linking and retaining care. Using the dataset of the active duty members that are currently engaging in the NHS cohort, the analysis examines the relationships between the annual number of appointments, the top three medical reasons for the appointments, the clinical laboratory data, and the patient's healthcare utilization and demographic during a period of five years after seroconversion, "the period of time during which HIV antibodies develop and become detectable" (Aidsmap 2012). The analysis of utilization rates of patients with different military services, demographics, and clinical risk factors will provide practical insights for the DoD to more effectively manage resources of care.

A. HIV CARE CONTINUUM

Despite no cure for HIV, the HIV care continuum has been one of the most effective approaches to slow the spread of virus and, simultaneously, improve the quality of life of the HIV-infected population. The HIV care continuum consists of the following focus areas: testing and diagnosis, linking and retaining in care, and ultimately achieving viral suppression (CDC 2017b).

1. Focus Area One: Testing and Diagnosis

The importance of testing and diagnosis of HIV cannot be overlooked since it allows the infected population to be cognizant of their status, and, thereafter, to seek necessary care that can increase their quality of life and decease the chance of passing on the virus to others (Department of Health and Human Services [DoHHS] 2016). In the United States, approximately 86% of the HIV-infected population received testing and diagnosis in 2011 (DoHHS 2016), and that percentage is down to 60% at the worldwide level (HIV.gov 2017). Despite the increase of the HIV care continuum in recent years, reaching Joint United Nations Programme HIV/AIDS's (UNAIDS) goal of 90% of HIV-positive population that have been diagnosed, received and retained in care, and achieved viral suppression by 2020 remains very challenging (Joint United Nations Programme on HIV/AIDS 2014).

The U.S. military healthcare aims to meet the HIV care continuum requirements. Compared to the general population, active duty military living with HIV receive a higher level of care continuum from their military healthcare system.

The U.S. military has been achieving the first focus area: testing and diagnosis. The Navy and the Marine Corps require their active duty personnel to be tested every 25 months (Secretary of the Navy [SECNAV] 2012), and the Air Force and the Army require their active duty personnel to be tested every 24 months (Secretary of the Air Force [SECAF] 2014; Personnel-General [PG] 2014). In addition to the routine tests, members are required to have additional HIV testing when other conditions are present, such as sexually transmitted infections (STIs), pregnancy, or before starting treatment programs for alcohol

and drugs (SECAF 2014). Clearly, the requirements of testing and diagnosis of HIV in the U.S. military are much stricter and more enforced than for the general population.

2. Focus Area Two: Linking and Retaining Care

The second focus area of linking and retaining care has been more difficult to accomplish than the first focus area. Of the 86% of the HIV-infected population in the United States that received testing and diagnosis in 2011, only 40% of the same population engaged in care (DoHHS 2016).

The military healthcare system provides the initiation and retained care for HIV-positive active duty members. The Air Force instruction (SECAF 2014) states that while on active duty, service members with HIV-positive diagnoses must receive an initial evaluation, then a follow up after six months and then annually. The Navy has similar instructions; however, the Army instruction further emphasizes that the initial evaluation must be followed by a psychosocial assessment and counseling (PG 2014). Again, the mandates by the military to provide care to the active duty HIV-positive members are aligned with the goals of the care continuum.

3. Focus Area Three: Viral Suppression

The third focus area of the care continuum is to achieve viral suppression, an undetectable viral load. Medical care and treatment directly correlate to reduce the HIV viral load, and the most effective treatment of HIV is through antiretroviral therapy (ART) (DoHHS 2016). Studies conclude that, after receiving six months of ART, most people attain viral suppression (CDC 2017a). Despite of the importance of this focus area, only 30% of the HIV-infected population in the United States achieved viral suppression in 2011 (DoHHS 2016) (see Figure 1).

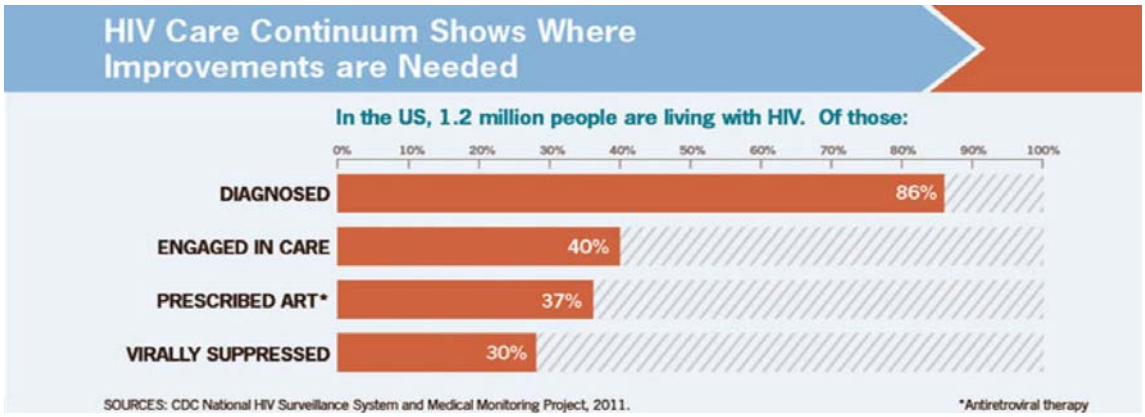

Figure 1. HIV Care Continuum 2011. Source: HIV.gov (2016).

The U.S. military has been administering ART therapy to service members in the early stage of HIV diagnosis. This treatment guideline has helped to improve quality of life of the infected populations, increasing the CD4 cell counts and decreasing the HIV viral load. The ultimate goal of HIV continuum of care is to achieve viral suppression, and the U.S. military healthcare is aiming to achieve it.

Undoubtedly, the HIV care continuum of the HIV-infected population provides effective outcomes in the fight against HIV, and it needs to be more pervasive and accessible not only in the United States but especially in the impoverished regions in other countries where the rates of new HIV infections are even higher.

B. THESIS ORGANIZATION

Chapter II provides background information on the study cohort, the data variables, and the types of statistical approaches used in the thesis. Chapter III provides descriptive statistics of variables in the data set. Chapter IV provides the main results from the primary and secondary analyses. The aim of the primary analysis is to identify predictors of healthcare utilization by using the visit rate, the average visits each year per patient, as a proxy. In the secondary analysis, we identified predictors of the three most frequent reasons of visits since the comorbidities of HIV/AIDS have a direct impact in healthcare utilization and the continuum of care. The final chapter provides the conclusion and recommended objectives for future analysis.

II. DATA INTRODUCTION

Chapter I introduced the importance of and background information on the HIV care continuum in the military and in the general population. This chapter provides background of the HIV study cohort, explanation of the data set, and types of statistical methods used to explore the relationships between the healthcare utilization and factors of clinical results and demographics of the patients.

A. COHORT DESCRIPTION

The IDCRP's U.S. Military NHS is one of the research programs established by a joint service effort between the Air Force, the Army, and the Navy. The aims of this program include studying the drawbacks of ART and HIV, enhancing supervision of care and outcomes to reduce illnesses, and limiting new HIV infections from the correlation between HIV and other STIs (IDCRP 2015a). These aims not only execute the principles of HIV care continuum but also emphasize HIV prevention and aim to achieve a functional cure of HIV. Since the 1986, the NHS's study cohort has accumulated over 6,000 participants (see Figure 2). Currently, about 1,500 participants from active duty members and other DoD beneficiaries (IDCRP 2015a) are engaging in the study cohort, and this thesis focuses solely on the active duty members.

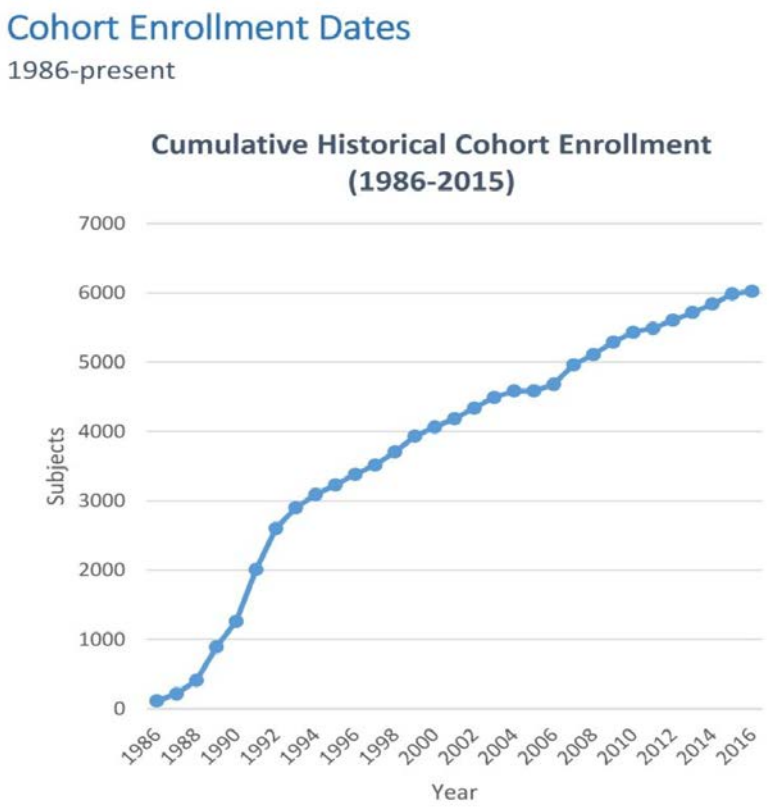

Figure 2. Cumulative Historical Cohort Enrollment. Source: Infectious Disease Clinical Research Program (2015b).

B. DATA OVERVIEW

The data collected in this cohort includes the following categories: demographics, therapeutic, laboratory, repository, clinical, adherence, risk behavior, and neurocognitive screening (Infectious Disease Clinical Research Program [IDCRP] 2015b). The data comes from 11 comma-separated values (CSV) files. Providing privacy to the HIV-positive patients, each patient's identifier is a unique personal identification number (PIN) called "MUCKED_PIN," and the PIN is used in all the CSV files.

Three of eleven CSV files contain demographics data of the patients including patients' gender, race and ethnicity, rank at initial diagnosis, marital status at initial diagnosis, service affiliation, risk of alcohol consumption, and risk of smoking use. One CSV file tracks patients' cluster of differentiation 4 (CD4) cell counts, and another CSV file tracks patients' HIV viral load. One CSV file contains all the International

Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes and medical service lines used by the patients. The last three CSV files contain ART medications included dispense and refill frequencies and patients' adherence rates of these medications. Two CSV files contain members' total number of permanent change of station (PCS) and whether the appointment took place in the same military treatment facility (MTF) or in a new MTF.

C. DATA ELEMENT DESCRIPTION

1. Demographics

Our analysis is based on 796 HIV positive active duty members in the cohort. The gender variable includes male or female. Race and ethnicity have six categories including White, African American, Native Hawaiian or other Pacific Islander, Asian, American Indian or Alaskan Native, and Other. The rank at initial diagnosis has four categories: officer, warrant officer, enlisted, and not applicable. Patients' marital status at initial diagnosis is categorized into cohabitation, married, not married and divorced. The members' service affiliations are the Army, the Navy, the Air Force, and the Marine Corps.

2. Laboratory, Risk, and Clinical Variables

The two main laboratory variables are CD4 cell count and HIV viral load. Hughson stated in 2017 that "CD4 cell count is the number of blood cells in a cubic millimeter of blood, and a higher number of CD4 indicates as stronger immune system." Viral load is the term used to describe the amount of HIV in your blood, and the more HIV there is in your blood, then the faster your CD4 cell count will fall" (Aidsmap 2017). The two risk variables categorize the levels of patients' alcohol and smoke use.

The three main clinical variables are constructed using the three most frequent medical reasons for the visits based on their ICD-9 codes in the dataset (see Table 1). "ICD-9 is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States" (CDC 2015).

Table 1. Three Most Frequent Medical Reasons

ICD-9 Code	Explanation
V6544	Counseling
V08	Asymptomatic HIV
V6289	Psychological Stress

If ICD-9 codes are missing for all visits, then the missing value code not available (NA) is assigned to all three clinical variables. Otherwise, a number was computed to represent number of visits of the three medical reasons on a yearly basis.

D. STATISTICAL METHODS

1. Constructed Variables

In order to analyze the care utilization of the HIV-positive active duty population, based on the original data, a new dataset was created with several new variables. The discrete variable of time representing the number of time periods was constructed. Each time period is 90 days, and there are 20 time periods for each patient. Time period 1 is when a new HIV patient entered the cohort, and we follow the patient's status until time period 20 (i.e., five years) unless the patient left the cohort before that time. The total length of time possible for follow-up is five years. The increment of 90 days allows us to take a more granular look at patients' encounters and their health status in terms of their CD4 cell counts, HIV viral load, different types of appointments, and demographics.

We constructed three additional variables to allow us to take a closer look at the number of visits used by the patients. First variable constructed was the total number of visits per year. Each time period is three months. This variable indicates how many appointments a patient had in a year (time period 1 to time period 4; time period 5 to time period 8, et cetera) based on the total number of appointments from four consecutive time periods. If the information regarding number of visits per year was missing, then the value was NA. Otherwise, we assigned a number to the total number of visits per year.

We constructed the second variable of the total number of visits that a patient had in five years from time period 1 to time period 20; therefore, this variable was the five-year total of the number of visits per year. NA assigned to a year indicated zero visits for the year.

We constructed the third variable of the visit rate per year by dividing the total number of visits per patient by the number of years that the patient stayed in the cohort. The variable provides a very reasonable average of visits each year per patient.

We constructed two sets of categorical variables on CD4 cell counts and HIV viral load count. The CD4 cell count variables have four categories, and they are less than 200, 200 to 349, 350 to 499, and over 500 in the unit of cell per cubic millimeter. Each of the five CD4 cell count variables represents each year. The HIV viral load count variables have two categories, and they are less than 1,000 and greater than or equal to 1,000 in unit of copies per milliliter. Each of the five HIV viral load count variables represents each year.

These constructed variables provide an overview of appointment utilization and allow further exploration of the relationships with other clinical indicators and demographics factors.

2. Univariate Analysis

Our first step in model building uses univariate analysis. Bursac et al. stated in 2008 "the purposeful selection process begins by a univariate analysis of each variable." Using the linear model (lm) function from the R statistical software (R Core Team 2018), this analysis allows us to explore the linear relationships of each dependent variable with each independent variable. For each independent variable, we use a 0.1 level of significance alpha to test the null model where the expected value of the independent variable is constant against the alternative model that includes the independent variable. Bursac et al. stated in 2008 "more traditional levels such as 0.05 can fail in identifying variables known to be important." This step serves as an initial screening step to give a short list of candidate independent variables to be used for further modeling. Because this initial screening is based solely on univariate analyses, it is possible that important independent variables did

not make the short list. Thus, after fitting a multiple regression, we check to see if any discarded independent variables can be added back into the model (Bursac et al. 2008).

3. Multiple Linear Regression Analysis

"In multiple regression, the objective is to build a probabilistic model that relates a dependent variable *y* to more than one independent or predictor variable" (Devore 2014). Our fitted multiple linear regression (MLR) models start with an additive model using the covariates shortlisted in the univariate analysis step.

a. Purposeful Selection of Covariates

The purposeful selection of covariates next uses the backward elimination to remove the variables from the MLR model that have either p-values greater than 0.05 of the hypothesis test that the independent variable is not needed in the presence of the rest of the independent variables or that are not confounding variables. If a removed independent variable caused the coefficients of the other independent variables to change more than 20% from their original coefficients, then that removed independent variable is considered as a confounding variable. Hosmer et al. stated in 2008 "in general, we use a value of about 20 percent as an indicator of an important change in a coefficient." Thus, we include all confounding variables in our model despite their p-values of greater than 0.05. We repeat the backward elimination process until the fitted MLR model contains all variables that have p-values less than 0.05 and all confounding variables (Bursac et al. 2008).

b. Backwards Elimination Using Akaike Information Criterion (AIC)

Additionally, we use a criterion-based backwards elimination procedure where at each step the model with the lowest AIC is selected. We accomplish this with the (step) function from the R statistical software (R Core Team 2018). "The first term of the AIC function is based on RSS [residual sum of squares] which is made smaller by improving the fit" (Faraway 2014). Overfitting of a model increases the penalty term in the function; thus, we choose the model that has the lowest AIC score.

4. Random Forests

Random Forests (RF) are a type of machine learning algorithm. The forest is formed by a number of trees. Each tree generates new predictions, and the random forest averages the predictions (Breiman 2001). Random forest models are a richer class of models than additive multiple regression models. They automatically accommodate interactions among independent variables and non-linear effects of numeric independent variables. Random forests provide a measure of importance for the independent variables; however, unlike MLR, random forests lack the ability to explain the relationships between the dependent and independent variables.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DESCRIPTIVE STATISTICS

Chapter I and II provided the background of the HIV Continuum of Care initiative, military HIV care policies, IDCRP cohort population and history, and statistical methods used for the thesis. This chapter provides useful insights and statistical relationships from the frequency of care, the diagnoses of care, and the laboratory results of the patients in a period of the five years after their initial HIV-positive diagnoses.

A. DEMOGRAPHICS

1. Age

The dataset for the thesis contains 796 active duty military HIV-positive members from the IDCRP cohort. The gender breakdown of the dataset is 772 males and 24 females. Males account for almost 97% of the dataset patient population, and that percentage is higher than the overall active duty male population of 84.5% (Department of Defense [DoD] 2015). The minimum age of the patients is 18, and the maximum age of the patients is 56. The average age is 28.04, and the median age is 26 (interquartile range age of 23–32). The distribution of all the ages is right-skewed (see Figure 3), which is expected considering approximately 40% of active duty military members are less than 26 years old (DoD 2015).

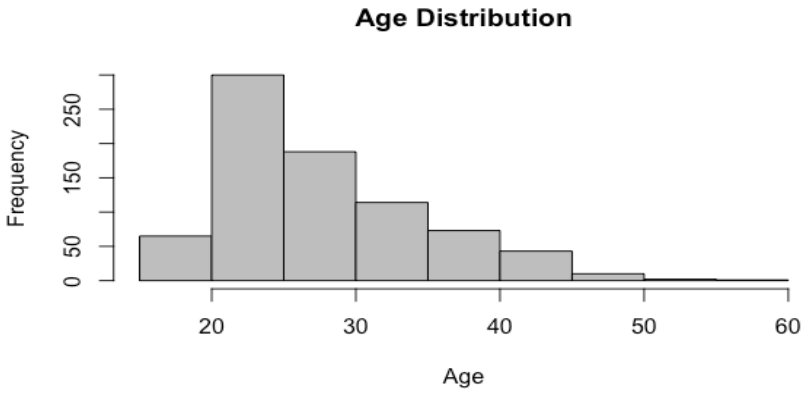

Figure 3. Age Distribution of Patient Cohort. Adapted from Infectious Disease Clinical Research Program (2017).

2. Ranks

The active duty ranks are either enlisted or officers in the dataset. Approximately, 90% of all patients are enlisted, and the other 10% are officers. Those percentages are fairly close to the 82.3% of enlisted and 17.7% of officers in the overall active duty population serving in the military (DoD 2015).

3. Race and Ethnicity

The race and ethnicity variable includes six categories: White, African American, Native Hawaiian or other Pacific Islander, Asian, American Indian or Alaskan Native, and Other. This dataset contains zero American Indian or Alaskan Native; therefore, this category has been excluded in the analysis. African American has the largest proportion of patients, followed by White. African American consists about 17.3% of the total active duty force (DoD 2015) and accounts for 43.21% of the patient population. White consists of about 68.7% of the total active duty force (DoD 2015) and accounts for 32.78% of the patient population. Native Hawaiian or Other Pacific Islander accounts for 1.1% of the total active duty force (DoD 2015) but totals 14.19% of the patient population. The active duty members who are Asian is 4.2% (DoD 2015), and the percentage of the Asians in the patient population is 4.27% (see Figure 4).

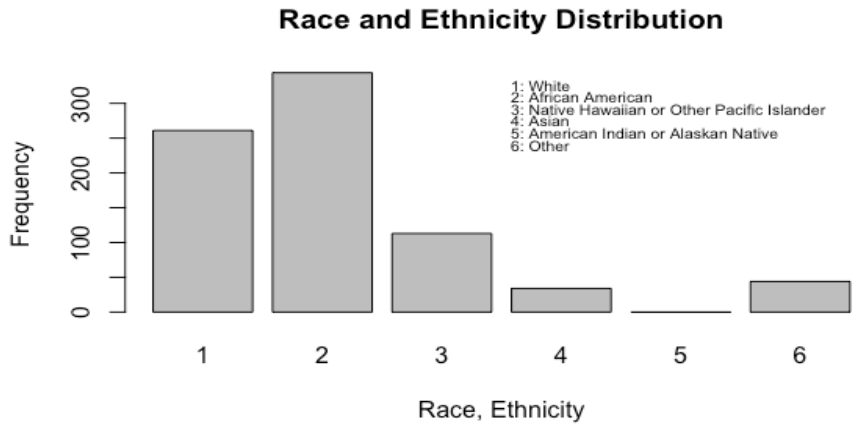

Figure 4. Race and Ethnicity Distribution of Patient Cohort. Adapted from Infectious Disease Clinical Research Program (2017).

4. Marital Status

The marital status of the patients at the time of their initial diagnosis of HIV contains the following categories: Cohabitation, Married, Not Married, and Separated. The percentage of members that were married is 14.7% (see Figure 5), and that is much lower than the 54.3% of all active duty military members that are married (DoD 2015).

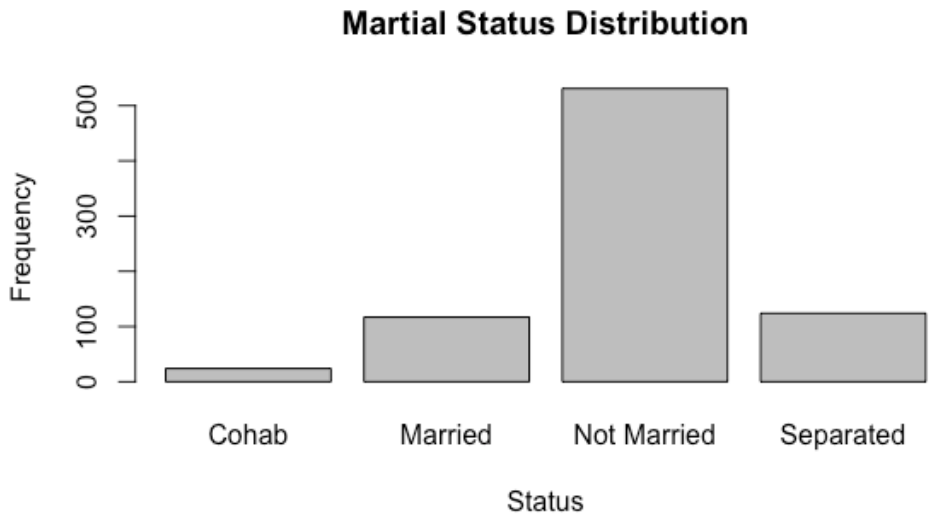

Figure 5. Marital Status Distribution of Patient Cohort. Adapted from Infectious Disease Clinical Research Program (2017).

5. Military Services

The dataset consists of active duty members from the Army, the Navy, the Air Force, and the Marine Corps. Nearly half of the patient population in the dataset serve in the Navy; despite that the Navy accounts for approximately 25% of the total active duty forces (DoD 2015). The second highest proportion of patients serves in the Air Force and is 24.24%, and the Air Force has 23.6% of the total active duty forces (DoD 2015). The Army has the highest proportion of active duty members, nearly 38% (DoD 2015), but accounts for only 18.7% of all patients. Marine Corps also has a lower patient proportion than its proportion of the total active duty forces (see Figure 6).

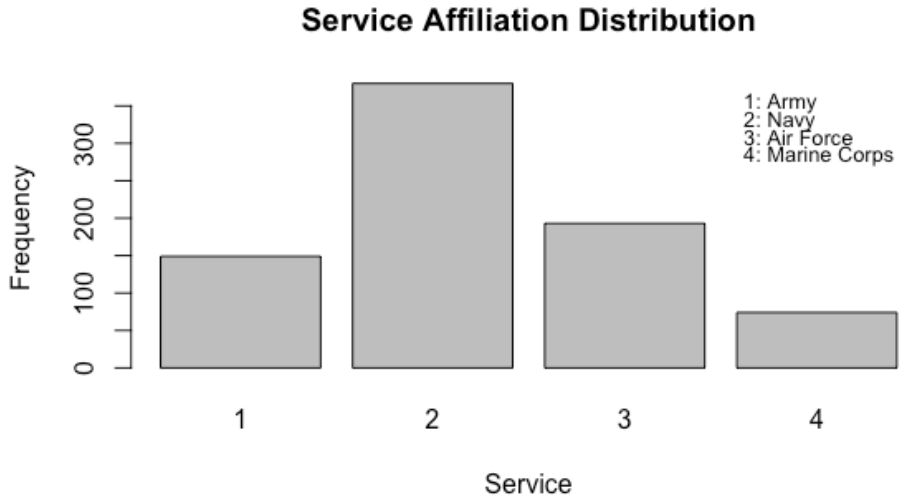

Figure 6. Service Affiliations of Patient Cohort. Adapted from Infectious Disease Clinical Research Program (2017).

B. CLINICAL RISK FACTORS

There are 537 (69.21%) patients not at risk of alcohol consumption, and 234 (30.79%) patients are at risk of alcohol consumption. Out of the 796 patients, 571 (71.73%) are not smokers, and 225 (28.27%) are smokers. The Air Force has the highest percentage (36.56%) of patients who are at risk of alcohol consumption, and the Army has the lowest percentage (25.9%) of patients who are at risk of alcohol consumption. The Marine Corps has the highest percentage (31.08%) of patients who are smokers; the Air Force has the lowest percentage (21.76%) of patients who are smokers (see Table 2).

Table 2. Demographic and Risk Factor Distributions of Patient Cohort by Service

	Army	Navy	Air Force	Marine Corps	Total
Gender					
M	144 (96.64%)	370 (97.37%)	185 (95.85%)	73 (98.65%)	772 (96.98%)
F	5 (3.36%)	10 (2.63%)	8 (4.15%)	1 (1.35%)	24 (3.02%)
Rank					
Officers	21 (14.09%)	36 (9.47%)	28 (14.51%)	3 (4.05%)	88 (11.06%)
Enlisted	128 (85.91%)	344 (90.53%)	165 (85.49)	71 (95.95%)	708 (89.94%)
Avg. Age (SD)	29.72 (7.62)	27.37 (6.43)	28.67 (7.31)	26.43 (5.67)	28.04 (6.89)
Median Age (IQR)	28 (24-34)	26 (22-31.25)	26 (23-34)	25 (22.25-30)	26 (23-32)
Race					
White	39 (26.17%)	120 (31.58%)	77 (39.9%)	25 (33.78%)	261 (32.79%)
African American	75 (50.34%)	163 (42.89%)	75 (38.86%)	31 (41.89%)	344 (43.21%)
Native Hawaiian or Pacific Islander	23 (15.44%)	59 (15.53%)	18 (9.32%)	13 (17.57%)	113 (14.2%)
Asian	4 (2.68%)	17 (4.47%)	11(5.7%)	2 (2.7%)	34 (4.27%)
Other	8 (5.37%)	21 (5.53%)	12 (6.22%)	3 (4.06%)	44 (5.53)
Marital Status					
Cohabitation	3 (2.01%)	10 (2.63%)	8 (4.15%)	3 (4.05%)	24 (3.02%)
Married	35 (23.49%)	40 (10.53%)	28 (14.51%)	14 (18.92%)	117 (14.7%)
Not Married	89 (59.73%)	261 (68.68%)	138 (71.5%)	43 (58.11%)	531 (66.70%)
Separated	22 (14.77%)	69 (18.16%)	19 (9.84%)	14 (18.92%)	124 (15.58%)
Alcohol					
Not at risk = 1	103 (74.1%)	258 (70.68%)	118 (63.44%)	47 (67.14%)	526 (69.21%)
At risk = 2	36 (25.9%)	107 (29.32%)	68 (36.56%)	23 (32.86%)	234 (30.79%)
Smoking					
No = 0	105 (70.47%)	264 (69.47%)	151 (78.24%)	51 (68.92%)	571 (71.73%)
Yes = 1	44 (29.53%)	116 (30.53%)	42 (21.76%)	23 (31.08%)	225 (28.27%)

*Column totals may not add to sample size due to missing data

C. HEALTHCARE UTILIZATION AND CLINICAL RESULTS

The dataset contains 796 patients, and this analysis is based on a period of five years after patients' initial diagnosis of HIV. Not every patient had at least one outpatient visit per year, and the number of patients with at least one outpatient visit per year decreases each year (see Table 3).

Table 3. Patients' Appointment Utilization per Year

sample size n = 796	number of patients with at least 1 outpatient visit per year
Year 1	793
Year 2	726
Year 3	604
Year 4	468
Year 5	389

This trend may result from patients leaving the cohort, getting out the military, being deployed, or having fewer medical appointments because patients become healthier after receiving ART.

Comprehensive statistics of patients' demographics and risk factors with the services breakdown contain a few trends. First, patients in the Army had fewer average number of visits per year than the patients from the other three services in Year 1 to Year 4, a finding that highlights the importance that patients in the Army also had the lowest average CD4 cell counts per year amongst all patients. Second, the average number of visits per year and the median viral load were the highest in Year 1. The average number of visit per year dropped about 30%, and the median viral load dropped more than 99% in Year 2. Clearly, HIV treatment provided in Year 1 made the patients healthier, and that was also supported by their increased average and median CD4 cell counts from Year 1 to Year 2. The median viral load hit the low level of 35 copies/mL with an interquartile range of 20 copies/mL and 48 copies/mL; thus, these results reflected viral suppression (see Table 4).

Table 4. Summary Statistics for Numbers of Appointments and Laboratory Results by Year and Service including Averages (Avg.) and Standard Deviation (SD) or Medians and Interquartile Range (IQR)

	Army	**Navy**	**Air Force**	**Marine Corps**	**Total**
Year 1					
Avg Number Visits (SD)	21.3 (11)	25.5 (12.07)	22.15 (11.07)	23.54 (12.71)	23.72 (11.81)
Avg CD4 (SD)	504.08 (172.29)	535.53 (181.69)	609.3 (223.03)	516.95 (178.26)	545.93 (194.06)
Median CD4 (IQR)	461 (403.5-592)	503 (403-649.5)	586.5 (456-734.25)	502 (376-631)	518 (408.5-656.5)
Avg Viral Load (SD)	34594 (58719)	37427 (116626)	50950 (96555)	26289 (42960)	39154 (97861)
Median Viral Load (IQR)	14754 (3444-40399)	16142 (3455-41200)	20086 (6193-47800)	14156 (3460-28923)	17008 (3647-41400)
Year 2					
Avg Number Visits (SD)	14.62 (9.03)	17.18 (11.59)	17.17 (10.23)	17.56 (13.61)	16.72 (11.05)
Avg CD4 (SD)	563.42 (208.79)	587.13 (205.43)	651.55 (241.98)	564.94 (208.59)	596.57 (217.85)
Median CD4 (IQR)	529 (408-680)	554 (446-693.75)	592.5 (459.25-823.25)	513 (406-707.25)	529 (408-680)
Avg Viral Load (SD)	8949 (25564)	8471 (25495)	13154 (32286)	11419 (27094)	9964 (27469)
Median Viral Load (IQR)	109 (36-6347)	50 (29-5178)	50 (38-11590)	82 (32-9563)	52 (34-7477)
Year 3					
Avg Number Visits (SD)	14.34 (9.19)	16.07 (9.8)	16.43 (10.46)	17.08 (11.51)	15.93 (10.03)
Avg CD4 (SD)	610.35 (244.37)	629.93 (209.11)	673.93 (228.01)	599.9 (221.13)	634.42 (222.46)
Median CD4 (IQR)	587 (455-727.5)	604 (483-757.5)	633 (506-818)	558 (447-767)	602 (477-768)
Avg Viral Load (SD)	9411 (43757)	5911 (35321)	7271 (19944)	6538 (19714)	6934 (32916)
Median Viral Load (IQR)	48 (20-2388)	48 (20-232)	48 (20-1974)	50 (20-1871)	48 (20-654)
Year 4					
Avg Number Visits (SD)	14.72 (8.72)	15.61 (9.66)	17.55 (10.87)	15.78 (9.17)	15.91 (9.77)
Avg CD4 (SD)	624.11 (219.6)	651.78 (225.3)	704.66 (232.39)	603.9 (216.41)	655.77 (226.97)
Median CD4 (IQR)	574.5 (483-769.5)	621.5 (500-756.5)	677 (530-840)	576 (464.5-745.75)	628 (501-785)
Avg Viral Load (SD)	4911 (16020)	3018 (11580)	10773 (73084)	5178 (25306)	5442 (38509)
Median Viral Load (IQR)	48 (20-72)	48 (20-84)	48 (20-53)	48 (20-50)	48 (20-72)
Year 5					
Avg Number Visits (SD)	15.17 (9.78)	15.01 (10.75)	17.93 (11.15)	15.56 (8)	15.81 (10.52)
Avg CD4 (SD)	643.94 (201.53)	712.3 (247.9)	732.93 (246.21)	618.5 (210.65)	698.42 (239.43)
Median CD4 (IQR)	666 (472.5-782.5)	693 (529.25-837)	688.5 (550.25-873.25)	614 (482.75-770.5)	681 (518-831.5)
Avg Viral Load (SD)	2685 (11870)	5522 (31722)	6933 (46910)	282 (988)	4999 (33176)
Median Viral Load (IQR)	48 (20-48)	29 (20-48)	34 (20-48)	48 (20-49)	35 (20-48)

In addition, the average CD4 cell counts have increased each year from Year 1 to Year 5. Overall, the upward trend of average CD4 cell counts and downward trend of average viral load from the patient population indicate that the continuum of care improves patients' quality of life.

IV. RESULTS AND ANALYSIS

Chapter III provided insights and statistical relationships based on the patients' frequency of care, their diagnoses, and the laboratory results. This chapter produces four models to estimate four unique visit utilization measures. The statistical methods for the models are multiple linear regression, multiple linear regression using AIC for variable selection, logistic regression, and random forests. The models not only identify the statistically significant variable coefficients that we can use to find healthcare utilization trends, but also predict future potential healthcare visits for the active duty HIV patients so the DoD can plan and allocate appropriate healthcare resources.

Before conducting analyses, we first randomly separate the overall dataset of 796 patients to a training set of 600 patients and a test set of 196 patients. Thus, we use the training set data to conduct the univariate, multiple linear regression and logistic regression analyses, and use the test set data to assess the final model fits.

A. PRIMARY STUDY

The primary study focuses on a better understanding of the annual healthcare utilization of HIV-positive active duty members. The dependent variable is the patient's visit rate, which is the average number of visits each year per patient. The regression analyses conducted allows us to explore the effects and relationships of the independent variables on the dependent variable.

1. Univariate Analysis

The visit rate is the average number of visits per year for each of the 796 patients over a period of five years. Because some of the patients dropped out of the cohort before the end of the fifth year, we used the average number of annual visits rather than the total number of visits. To calculate the average, we divide the total number of visits for each patient by the number of years that the patient stayed in the cohort. In this univariate analysis, the visit rate is the dependent variable, and the independent variables are the demographics, secondary risk factors, and clinical variables in the dataset. This univariate

analysis is used for initial variable screening. We note that for each coefficient, the standard errors and associated p-values are used as descriptive statistics only.

This univariate analysis includes seven independent categorical variables with p-values less than 0.1. Compared to patients in the Army, patients in the Navy had an estimated 2.2 more visits each year. Additionally, African American patients had an estimated two fewer visits each year when compared to White patients. Patients whose marital status was separated from their spouses had an estimated 4.6 more visits each year than patients who were cohabitating. Patients who were at risk of alcohol consumption had an estimated 2.4 more visits each year than patients who were not at risk. Similarly, patients who smoked had an estimated 1.4 more visits each year than patients who did not smoke. Compared to patients with average HIV viral load less than 1,000 copies/mL, patients with an average viral load higher than that level had 3.5 and 1.9 fewer visits in Year 2 and Year 3, respectively (see Table 5).

Table 5. Visit Rate

Visit Rate (dependent variable)	Coefficient	Standard Error	P-value
Service Affiliation			
Army	Ref	Ref	Ref
Navy	2.23	0.98	0.024
Air Force	1.1	1.11	0.321
Marine Corps	2.1	1.47	0.157
Race			
White	Ref	Ref	Ref
African American	-2.11	0.84	0.012
Native Hawaiian or Other Pacific Islander	-0.88	1.17	0.45
Asian	-1.52	1.83	0.406
Other	-2.09	1.61	0.196
Marital Status			
Cohabitation	Ref	Ref	Ref
Married	1.8	2.22	0.406
Not Married	2.5	2.04	0.207
Separated	4.6	2.2	0.037

(continued)

Table 5 (continued from previous page)

Visit Rate (dependent variable)	Coefficient	Standard Error	P-value
Alcohol Not at risk = 1 At risk = 2	 Ref 2.35	 Ref 0.81	 Ref 0.004
Smoking No = 0 Yes = 1	 Ref 1.39	 Ref 0.82	 Ref 0.089
Avg. VL Year 2 < 1000 => 1000	 Ref -3.5	 Ref 0.75	 Ref <0.001
Avg. VL Year 3 < 1000 => 1000	 Ref -1.91	 Ref 0.84	 Ref 0.024

2. Multiple Linear Regression Model

The backward elimination process starts with removing the average HIV viral load in Year 3 variable since it had the highest p-value. The removed variable causes more than 20% change in coefficient estimates to variables of service affiliation, race and ethnicity, marital status, and risk of smoking; thus, we add the removed variable in the model. Next, we remove the second highest p-value variable, risk of smoking, from the model. We notice both the coefficient estimates of service affiliation and marital status variables changed more than 20%, so the risk of smoking variable is added in the model. Then, we remove the risk of alcohol consumption variable with the third highest p-value. The removed variable caused more than 20% change in coefficient estimates to variables of service affiliation, race and ethnicity, marital status, and average HIV viral load in Year 3; therefore, we include the risk of alcohol consumption variable in the model. Furthermore, we remove the service affiliation variable, and that causes more than 20% change in coefficient estimates to variables of race and ethnicity and average HIV viral load in Year 3; thus, we add the service affiliation variable in the model.

We reinsert two variables that were excluded from the initial univariate screening, because those variables are statistically significant in the presence of other variables in model. The two variables are gender and rank. Thus, the fitted MLR model includes eight independent categorical variables. Five independent variables that have p-values less than 0.05, and the variables are gender, rank, service affiliation, race and ethnicity, and average HIV viral load in Year 2. The model includes three categorical variables that are confounding variables despite their p-values greater than 0.05. Those variables are service affiliation, risk of alcohol consumption, risk of smoking, and average HIV viral load in Year 3 (see Table 6). Again, the coefficient estimates are unbiased, but the standard errors and p-values need to be cautiously regarded.

We explore the possibility of the two-way interactions among independent variables that could improve the model. The results from the analysis of variance F-test show that the p-value for the model that included the two-way interaction terms is 0.56, which is not close to the significant level. Therefore, we choose to drop the interaction terms and proceed with the original model.

We determine the following results while holding all other variables constant. Female patients had an estimated 4.2 more visits each year than the male patients. Patients who were enlisted had an estimated 2.8 more visits each year than patients who were officers. Compared to patients in the Army, patients in the Navy had an estimated 1.9 more visits each year. When compared to White patients, African American patients had an estimated 2.7 fewer visits each year. Additionally, patients in the Other race categorical had an estimated 4 fewer visits each year than the White patients. Patients with the average HIV viral load greater than 1,000 copies/mL used estimated 3.4 fewer visits than patients with the viral load below that level (see Table 6).

Table 6. MLR – Visit Rate

Visit Rate (dependent variable)	Coefficient	Standard Error	P-value	P-value < 0.05
Gender				
M	Ref	Ref	Ref	
F	4.24	2.17	0.05	*
Rank				
Officers	Ref	Ref	Ref	
Enlisted	2.77	1.14	0.016	*
Service Affiliation				
Army	Ref	Ref	Ref	
Navy	1.9	0.98	0.05	*
Air Force	0.42	1.12	0.7	
Marine Corps	1.71	1.46	0.239	
Race				
White	Ref	Ref	Ref	
African American	-2.75	0.86	0.001	*
Native Hawaiian or Other Pacific Islander	-0.83	1.13	0.463	
Asian	0.45	1.71	0.792	
Other	-4	1.65	0.015	*
Alcohol				
Not at risk = 1	Ref	Ref	Ref	
At risk = 2	1.16	0.82	0.158	
Smoking				
No = 0	Ref	Ref	Ref	
Yes = 1	-1.09	0.82	0.185	
Avg. VL Year 2				
< 1000	Ref	Ref	Ref	
=> 1000	-3.44	0.95	0.0003	*
Avg. VL Year 3				
< 1000	Ref	Ref	Ref	
=> 1000	-0.28	1.07	0.791	

3. AIC

The model selected using AIC include three categorical variables of race and ethnicity, marital status, and average HIV viral load in Year 2. Holding other variables constant, the model has three significant trends. First, compared to White patients, African American patients had an estimated average of 2.3 fewer visits, and patients in the Other

race category had an estimated average of 3.9 fewer visits. Second, patients whose marital status was separated had an estimated average of 3.7 more visits each year than patients were in the cohabitation status. Third, comparing to patients with average HIV viral load less than 1,000 copies/mL, patients with the average viral load higher than that mark had an estimated average of 3.6 fewer visits in Year 2.

4. Multiple Linear Regression Diagnostics

We apply backward elimination process with purposeful selection of covariates (Bursac et al. 2008) on the full model that included eight independent variables. We first conduct model diagnostics to check the assumptions of the model.

We examine the assumption of constant variance of the residuals in the model. Figure 7, the plot of residuals versus the fitted values, suggests that the variance of visit rate is increasing some with its expected value. This is confirmed by Figure 8, the normal quantile-quantile (Q-Q) plot of the residuals, which shows that the distribution of residuals is right-skewed.

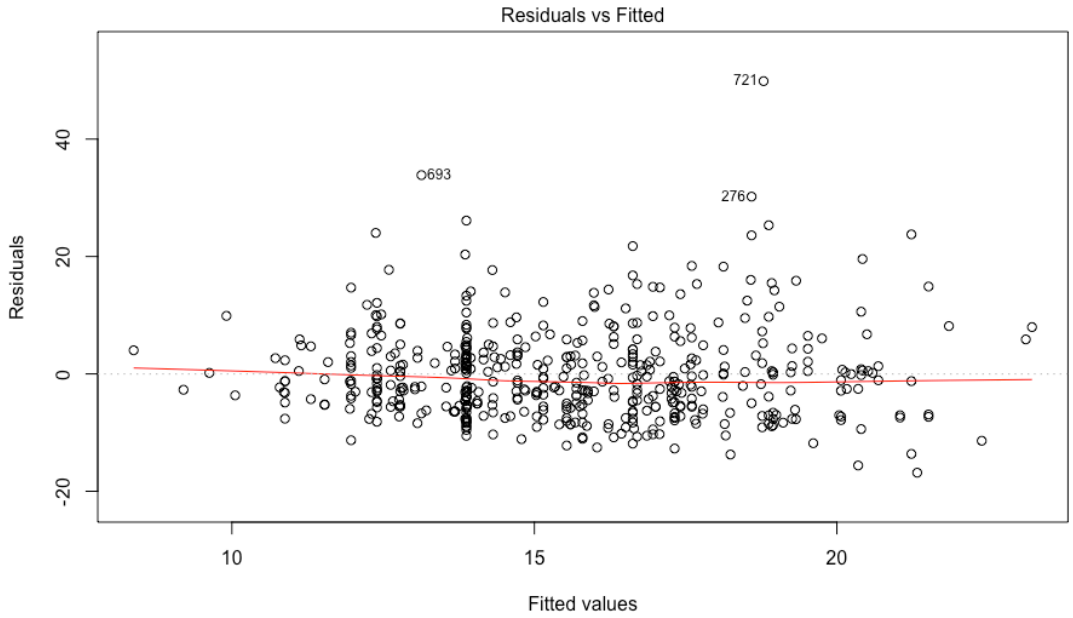

Figure 7. Residuals vs. Fitted Values Plot – MLR Visit Rate

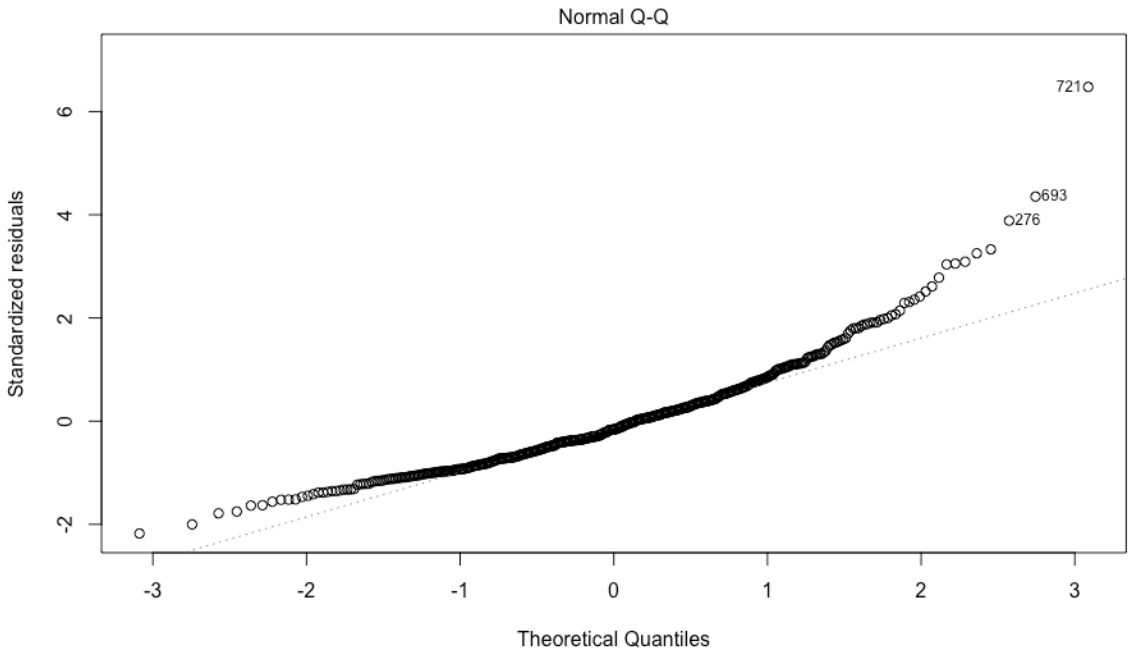

Figure 8. Q-Q Plot for Normality Check – MLR Visit Rate

With heteroscedasticity, the multiple linear regression model can be used to get unbiased estimates of the expected visit rates as a function of the covariates. However, standard error computations and inference results will be affected, but as pointed out by Miller (1986), the effects of heteroscedasticity are not usually dramatic unless heteroscedasticity is unusually severe. In what follows, we proceed with variable selection using the purposeful selection of covariates (Bursac et al. 2008) based on the multiple linear regression model. However, as a check, we also use the same variable selection procedure based on fitting an over-dispersed Poisson regression model to the total number of visits for each patient with log of the number of years as an offset. Using this offset is equivalent to fitting a model with visit rate as the dependent variable (Faraway 2016). The variable selection results are the similar to the results for the multiple linear regression model; thus, we are confident to proceed with our analysis with the multiple linear regression model.

We also check for influential observations in the model. "An influential point is one whose removal from the dataset would cause a large change in the fit" (Faraway 2014). Cook's distance is the tool to check for influential observations, and we consider an influential observation when its Cook's distance is greater than one. Observation 721 has

the highest Cook's distance of 0.15, which indicates that there are no influential observations in the model (see Figure 9).

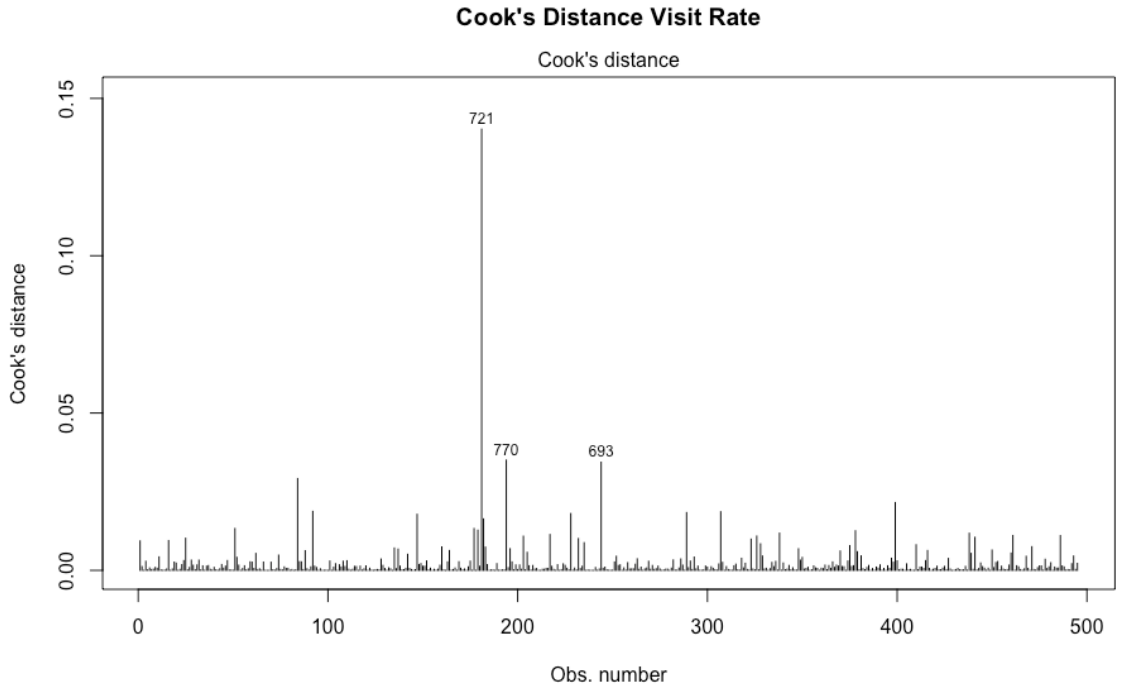

Figure 9. Cook's Distance – MLR Visit Rate

5. Model Selection

We use MLR, AIC, and RF to build the model for predicting the values for the dependent variable of visit rate. The lowest cross-validated mean square error (MSE) is the metric to determine which statistical method is preferred for the model. We use cross-validation because the three models have different complexity, and a likelihood based criteria such as AIC is not appropriate for RF.

To calculate the cross-validated MSE, we further randomly divide the training set of 600 patients into 10 folds (subsets), and each fold contains 60 patients. Then, we fit a model with 9 folds that contain 540 patients and apply the model to make predictions on the remaining fold that contains 60 patients. For example, we use data from fold 1 to fold 9 to fit the first model and use the model to make predictions using the data in fold 10.

Then, we use data of fold 2 to fold 10 to fit the second model, and use the model to make predictions using the data in fold 1. We continue the same process 10 times. The cross-validated MSE is the sum of the MSE computed for each fold. In the end, we select the model with the lowest cross-validated MSE as the final model since our goal is to make the predictions on healthcare utilizations as accurate as possible.

Random forests are a machine learning algorithm that is different from the MLR and AIC; therefore, they produce results that can provide additional insights and comparisons to the results from MLR and AIC. To implement the random forests, we use the *randomForest* package (Liaw and Wiener 2002) in *R*. A random forests model is fitted on the dependent variable and the candidate independent variables. We use the default of 500 trees to form each forest. In our RF model, after the initial increase in MSE, the rapid decrease in the MSE appears within the first few trees. The improvement levels off around 80 trees; thus, the 500 trees are sufficient for this model (see Figure 10).

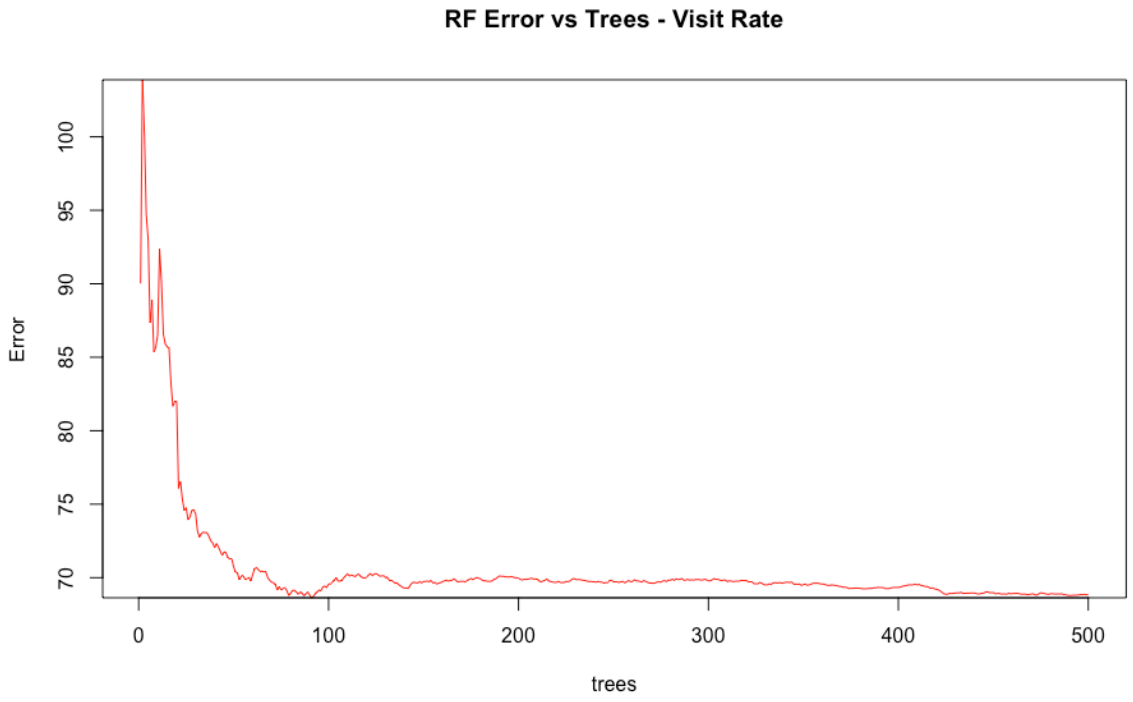

Figure 10. RF Error vs. Trees Graph

"The recommended number of variables randomly sampled as candidates each spilt (mtry) for regression is total number of independent variables divided by three" (Breiman and Cutler 2018). Thus, mtry is set to 6 for the model as there are 18 independent variables accounting for categorical variables with multiple levels.

Using the same ten folds to cross-validate for all three models, we can see that by fold the MSEs of MLR and AIC are quite similar (see Figure 11); however, MSEs from the RF are different. The variations of the MSEs from different folds illustrate the importance of the cross-validation. RF model's cross-validated MSE is 67.25, which is the lowest cross-validated MSE value amongst the three methods; thus, RF is the preferred method for the final model (see Table 7).

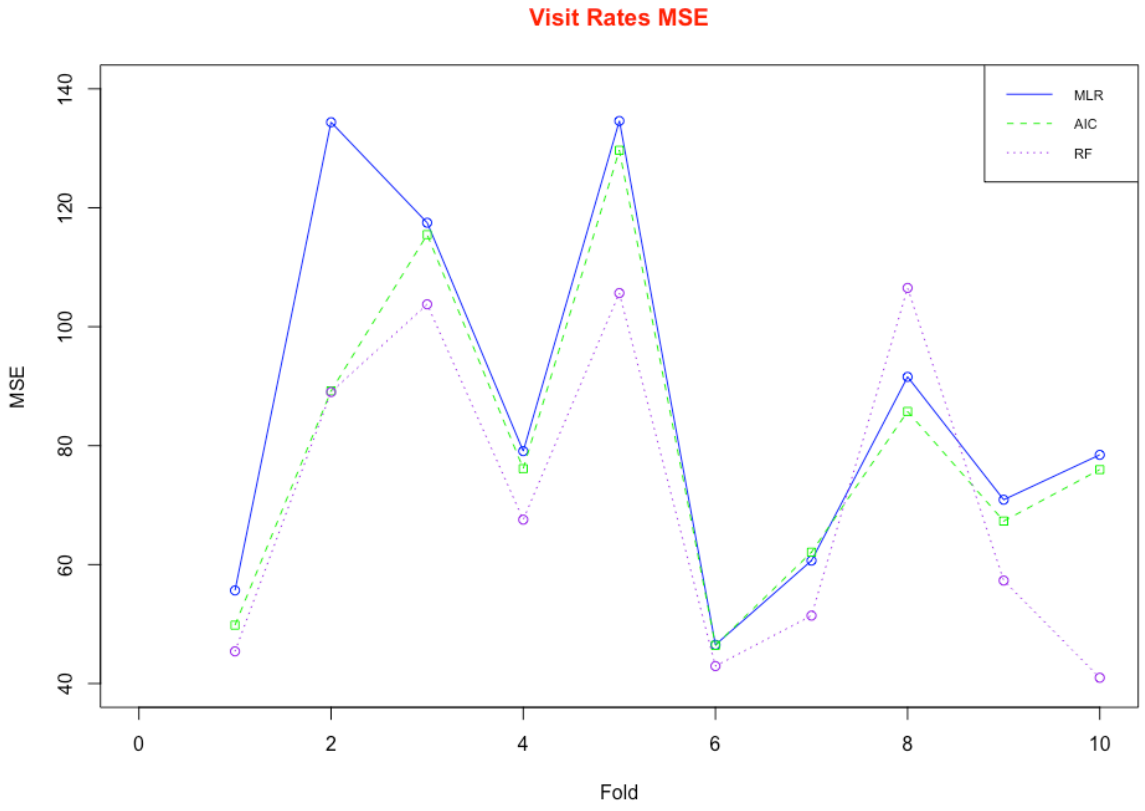

Figure 11. MSEs Breakdown – Visit Rate

Table 7. MSEs – Visit Rate

Visit Rate	
Method	MSE
MLR	81.86
AIC	75.69
Random Forests	67.25

6. Final Result

We include all 600 patients in the training set to build the final model, and we select the method for the model based on the lowest cross-validated MSE. Then, we use the test set of 196 patients to make predictions of healthcare utilization based on the model. Despite RF's inability to explain the regression coefficients, RF provides a measure of variables importance for each independent variable. The top three important variables are military service affiliation, average viral load in Year 1, and average viral load in Year 4 (see Figure 12).

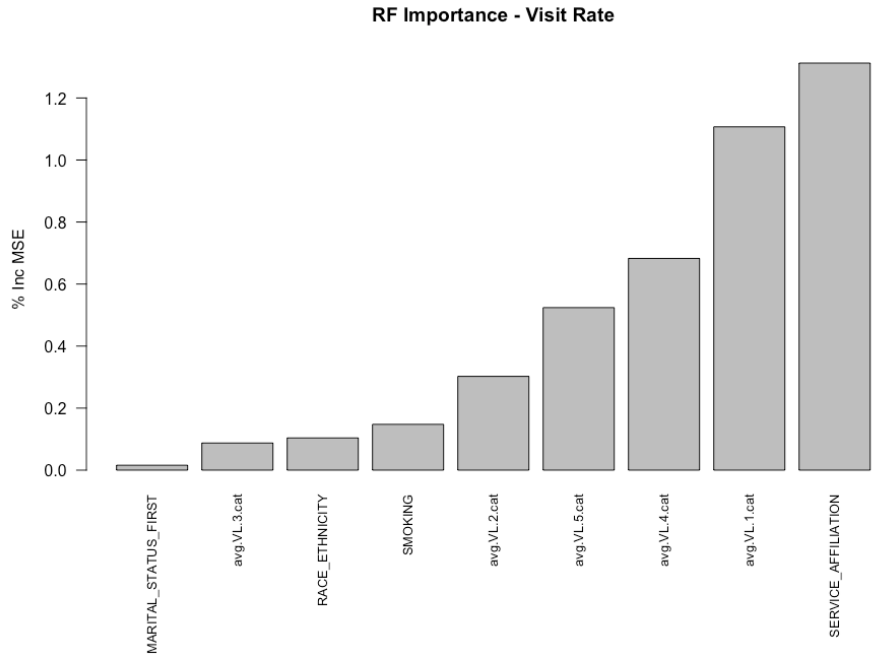

Figure 12. RF Importance Chart

The predictions of the final model derived from the test set data and the actual visits from the test set data show variations between the actual number of visits and the predicted number of visits. Majority of the predicted number of visits are higher than the actual number of visits (see Figure 13).

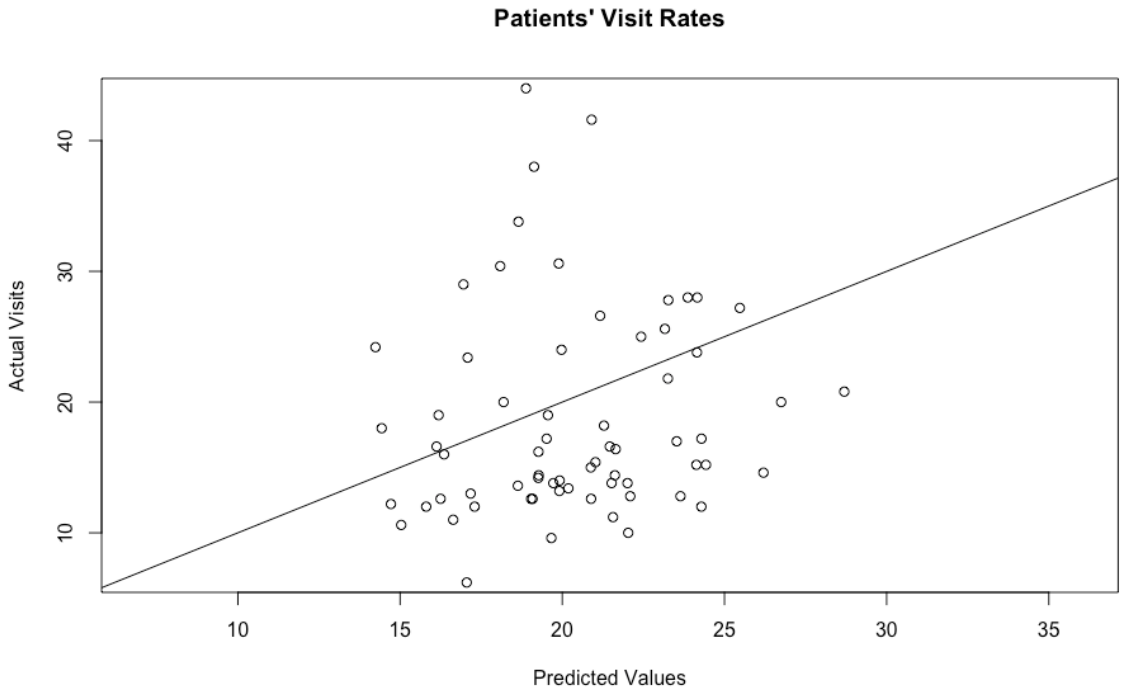

Figure 13. Actual Visits vs. Predicted Values – Visit Rate

B. SECONDARY ANALYSIS

In the secondary analysis, we look at the three most frequent medical reasons for visits using logistic regression. The top three reasons are counseling, asymptomatic HIV diagnosis, and psychological stress. We construct a dependent variable as a binary variable of 0s and 1s for each medical reason. If a patient did not have any of visits for the medical reason, we assign 0 to the dependent variable. On the other hand, if a patient had at least one visit for that medical reason, we assign 1 to the dependent variable. The generalized linear models explore the relationships between the candidate independent variables and the dependent variable. This analysis also follows the purposeful selection of covariates

methodology (Bursac et al. 2008). However, we present the statistical results in percent likelihood of a patient having at least one visit for a particular medical reason.

1. Visits for Counseling

a. Univariate Analysis

This model contains three categorical independent variables with p-values less than 0.1, and the variables are service affiliation, race and ethnicity, and average CD4 cell count in Year 4. Compared to patients in the Army, patients in the Air Force had a 95% (se = 68%) higher chance of having at least one counseling visit, and patients in the Marine Corp had a 31% (se = 59%) higher chance of having at least one counseling visit. Patients in race of Native Hawaiian or Other Pacific Islander had a 35% higher chance to have at least one visit than White patients (se = 57%). Compared to patients with average CD4 cell count less than 200 cells/mm3, patients with average CD4 cell count over 500 cells/mm3 had a 68% higher chance of having at least one visit (se = 61%) (see Table 8).

Table 8. Visits for Counseling

Visits for Counseling (dependent variable)	Coefficient Estimates	Standard Error	P-value
Service Affiliation			
Army	Ref	Ref	Ref
Navy	0.51	0.56	0.871
Air Force	0.95	0.68	<0.001
Marine Corps	0.31	0.59	0.022
Race			
White	Ref	Ref	Ref
African American	0.55	0.56	0.403
Native Hawaiian or Other Pacific Islander	0.35	0.57	0.043
Asian	0.51	0.63	0.924
Other	0.54	0.62	0.712
Avg. CD4 Year 4			
< 200	Ref	Ref	Ref
200-349	-	-	-
350-499	0.69	0.63	0.132
500+	0.68	0.61	0.099

b. Logistic Regression Model

The backward elimination process starts with removing the variable of average CD4 cell count in Year 4 due to its p-value of greater than 0.05; however, after removing the variable, the coefficient estimates of service affiliation variable changed more than 20%, so we add the confounding variable in the model. Also, we reinsert the variable of age, which was excluded from the initial univariate screening, in the model, since this variable is statistically significant in the presence of other variables.

We explore the possibility of the two-way interactions among independent variables that could improve the model. The results from the likelihood ratio test show that the p-value for the model that included the two-way interaction terms is 0.52, which is not close to the significant level. Therefore, we choose to drop the interaction terms and proceed with the original model.

The fitted logistic model includes four independent variables, and they are service affiliation, average CD4 cell count in Year 4, age, and race and ethnicity. Holding all other variables constant, compared to patients in the Army, patients in the Marine Corps had a 19% higher chance of having a least one visit for counseling (se = 62%). The chance of having at least one visit increased by 48% with each additional year in patient's age (se = 50%). Compared to White patients, Native Hawaiian or Other Pacific Islander patients had a 29% higher chance of having a least one visit (se = 60%) (see Table 9).

Table 9. MLR – Visits for Counseling

Visits for Counseling (dependent variable)	Coefficient Estimates	Standard Error	P-value	P-value < 0.05
Service Affiliation				
Army	Ref	Ref	Ref	
Navy	0.33	0.59	0.077	
Air Force	1	0	0.986	
Marine Corps	0.19	0.62	0.004	*
Avg. CD4 Year 4				
< 200	Ref	Ref	Ref	
200-349	-	-	-	
350-499	0.65	0.64	0.284	
500+	0.67	0.62	0.152	
Age	0.48	0.5	0.003	*
Race				
White	Ref	Ref	Ref	
African American	0.56	0.58	0.443	
Native Hawaiian or Other Pacific Islander	0.29	0.6	0.029	*
Asian	0.61	0.7	0.602	
Other	0.54	0.67	0.802	

c. Logistic Regression Diagnostics

For model diagnostics, we first examine the binned version of the deviance residuals and the predicted values of scale of log-odds for the full model (Faraway 2016). Figure 14 does not show any obvious trend in the average residuals. Thus, there is no major concern about the statistical inadequacy of the model. We note that, these residuals are not expected to have mean zero, even with an adequate model.

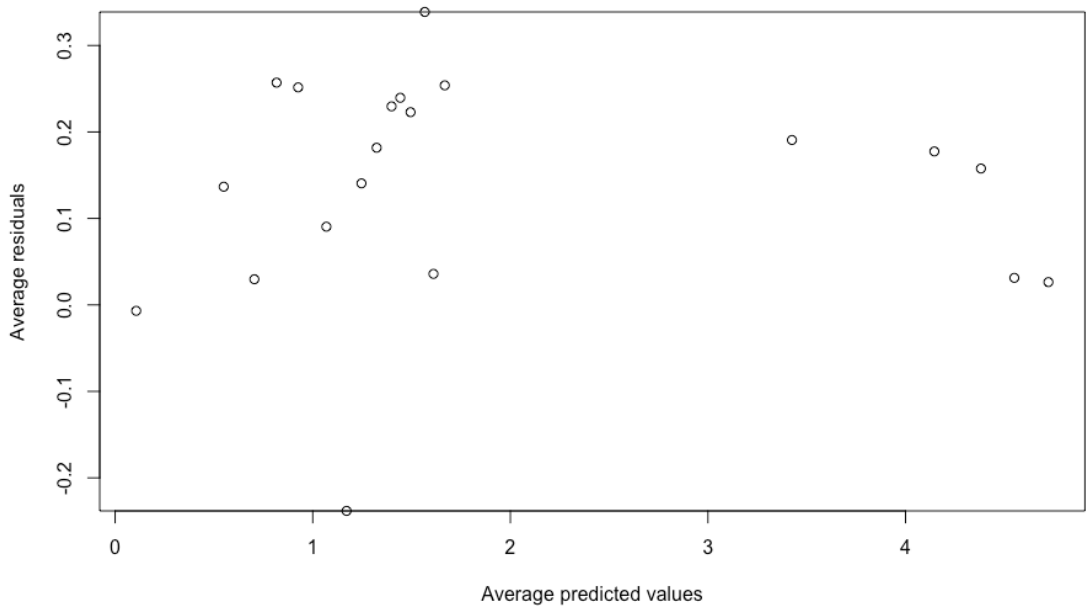

Figure 14. Average Residuals vs. Average Predicted Values Plot – Visits for Counseling

We also check for influential observations in the model. Observation 9 has the highest Cook's distance of less than 0.07, which is below the level to be considered influential; therefore, there are no influential observations in the model (see Figure 15).

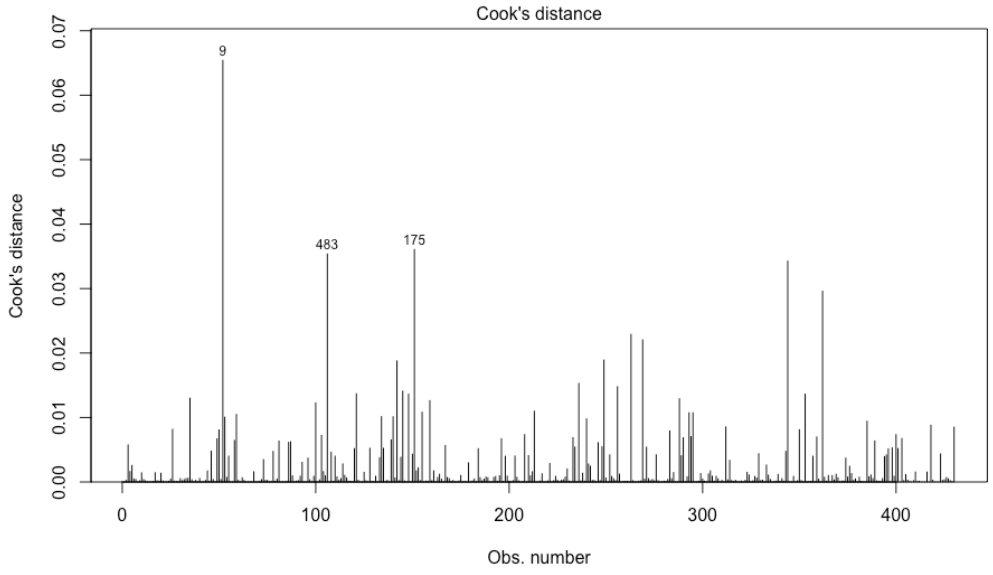

Figure 15. Cook's Distance – Logistic Regression Visits for Counseling

d. Results

We use the test dataset to check the effectiveness of the fitted logistic model at predicting the responses of the dependent variable. At a threshold of 70 percent, the overall classification accuracy rate of the model is at 65.8 percent. Of the 42 patients that did not have any visit for counseling, the model predicts 12 of them correctly (28.6%). The percentage increases to 75.9 percent as the model predicts 117 of 154 patients correctly that they had at least one visit (see Table 10).

Table 10. Confusion Matrix – Visits for Counseling

	Predicted (no visits)	Predicted (at least one visit)
Actual (no visits)	12	30
Actual (at least one visit)	37	117

2. Visits for Asymptomatic HIV

a. Univariate Analysis

This model contains four categorical independent variables with p-values less than 0.1, and the variables are service affiliation, age, race and ethnicity, and average HIV viral load in Year 2. Compared to patients in the Army, patients in the Navy and patients in the Air Force had a 39% (se = 56%) and a 79% (se = 59%) higher chance of having a least one visit for asymptomatic HIV, respectively. The chance of having at least one visit increased by 51% with each additional year in patient's age (se = 50%). African American patients had a 39% higher chance to have at least one visit than White patients (se = 55%). Compared to patients with average HIV viral load greater than 1,000 copies/mL in Year 2, patients with the viral load below that level had a 40% higher chance of having at least one visit (se = 55%) (see Table 11).

Table 11. Visits for Asymptomatic

Visits for Asymptomatic (dependent variable)	Coefficient Estimates	Standard Error	P-value
Service Affiliation			
Army	Ref	Ref	Ref
Navy	0.39	0.56	0.078
Air Force	0.79	0.59	<0.001
Marine Corps	0.53	0.6	0.743
Age	0.51	0.5	0.093
Race			
White	Ref	Ref	Ref
African American	0.39	0.55	0.051
Native Hawaiian or Other Pacific Islander	0.57	0.58	0.389
Asian	0.67	0.65	0.264
Other	0.47	0.61	0.794
Avg. VL Year 2			
< 1000	Ref	Ref	Ref
=> 1000	0.4	0.55	0.063

b. Logistic Regression Model

The backward elimination process starts with removing the variable of age. Then, we remove the race and ethnicity variable because its high p-value, and notice that the coefficient estimates of service affiliation changed more than 20%, so we add the removed variable in the model. Next, we remove the average HIV viral load in Year 2 variable and that change more than 20% to the coefficient estimates in service affiliation, and race and ethnicity. Thus, we add the average HIV viral load in Year 2 variable in the model.

We reinsert the variable of risk of alcohol consumption, which was excluded from the initial univariate screening, in the model because this variable is statistically significant in the presence of other variables.

We explore the possibility of the two-way interactions among independent variables that could improve the model. The results from likelihood ratio test show that the p-value for the model that included the two-way interaction between service affiliation and

risk of alcohol consumption is 0.017, which is significant. Therefore, we include the interaction term in the model.

The fitted logistic model includes five independent variables, and they are service affiliation, race and ethnicity, average viral load in Year 2, risk of alcohol consumption, and the two-way interaction term of service affiliation and risk of alcohol consumption.

We determine the following results while holding all other variables constant. Compared to patients in the Army, patients in the Navy and in the Air Force had a 33% (se = 58%) and 91% (se = 68%) higher chance of having a least one visit for asymptomatic HIV, respectively. African American patients had a 38% (se = 56%) higher chance of having a least one visit than White patients. Patients with risk of alcohol consumption serving in the Air Force had a 13% (se = 72%) higher chance of having a least one visit for asymptomatic HIV than patients with the same risk serving in the Army (see Table 12).

Table 12. MLR – Visits for Asymptomatic HIV

Visits for Asymptomatic (dependent variable)	Coefficient Estimates	Standard Error	P-value	P-value < 0.05
Service Affiliation				
Army	Ref	Ref	Ref	
Navy	0.33	0.58	0.04	*
Air Force	0.91	0.68	0.002	*
Marine Corps	0.61	0.65	0.47	
Race				
White	Ref	Ref	Ref	
African American	0.38	0.56	0.04	*
Native Hawaiian or Other Pacific Islander	0.58	0.59	0.371	
Asian	0.75	0.68	0.15	
Other	0.49	0.62	0.939	
Avg. VL Year 2				
< 1000	Ref	Ref	Ref	
=> 1000	0.4	0.55	0.087	
Alcohol				
Not at risk = 1	Ref	Ref	Ref	
At risk = 2	0.38	0.63	0.367	

(continued)

Table 12 (continued from previous page)

Visits for Asymptomatic (dependent variable)	Coefficient Estimates	Standard Error	P-value	P-value < 0.05
Service Affiliation and Alcohol Interaction				
Army and Alcohol (at risk)	Ref	Ref	Ref	
Navy and Alcohol (at risk)	0.61	0.65	0.442	
Air Force and Alcohol (at risk)	0.13	0.72	0.06	
Marine Corps and Alcohol (at risk)	0.29	0.71	0.344	

c. Logistic Regression Diagnostics

For model diagnostics, we first examine the binned version of the deviance residuals and the predicted values of the scale of log-odds for the full model (Faraway 2016). Figure 16 does not show any obvious trend in the average residuals. Thus, there is no major concern about the statistical inadequacy of the model.

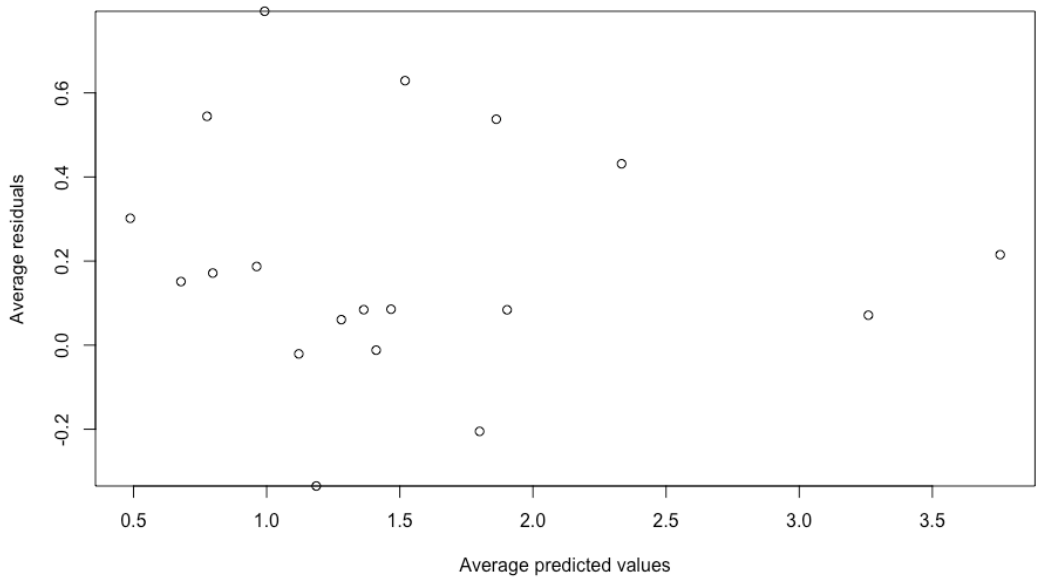

Figure 16. Average Residuals vs. Average Predicted Values Plot – Visits for Asymptomatic HIV

We also check for influential observations in the model. Observation 272 has the highest Cook's distance of less than 0.05, which is below the level to be considered influential; thus, there is no influential observations in the model (see Figure 17).

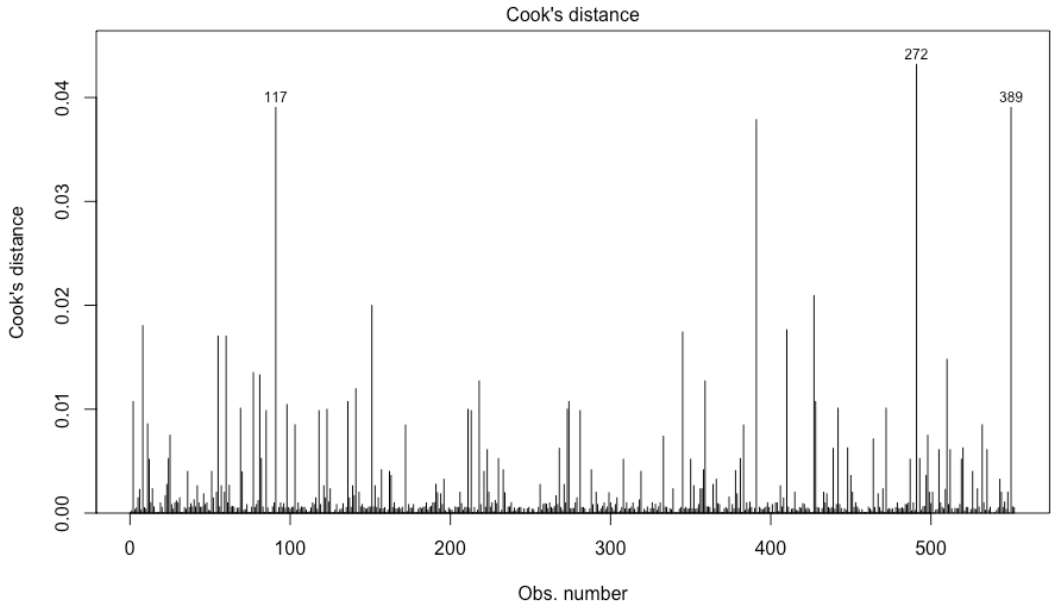

Figure 17. Cook's Distance – Logistic Regression Visits for Asymptomatic HIV

d. Results

We use the test dataset to check the effectiveness of the fitted logistic model at predicting the responses of the dependent variable. At a threshold of 70 percent, the overall classification accuracy rate of the model is at 65.2 percent. Of the 40 patients that did not have any visit for asymptomatic HIV, the model predicts 17 of them correctly (42.5%). The percentage increases to 71.6 percent as the model predicts 101 of 141 patients correctly that they had at least one visit (see Table 13).

Table 13. Confusion Matrix – Visits for Asymptomatic HIV

	Predicted (no visits)	Predicted (at least one visit)
Actual (no visits)	17	23
Actual (at least one visit)	40	101

3. Visit for Psychological Stress

a. Univariate Analysis

This model contains four categorical independent variables with p-values less than 0.1, and the variables are service affiliation, race and ethnicity, marital status, and average HIV viral load in Year 2. When compared to patients in the Army, patients in the Air Force and patients in the Marine Corps had a 12% (se = 58%) and a 67% (se = 58%) higher chance of having a least one visit for psychological stress, respectively. Patients in the Native Hawaiian or Other Pacific Islander had a 70% higher chance to have at least one visit than White patients (se = 56%). Patients who were not married and whose marital status was separated had a 74% (se = 63%) and a 78% (se = 64%) higher chance of having at least one visit when compared to patients who were in cohabitation status. Compared to patients with average HIV viral load in Year 2 greater than 1,000 copies/mL, patients with the viral load below that level had a 42% higher chance of at least one visit (se = 54%) (see Table 14).

Table 14. Visits for Psychological Stress

Visit for Psychological Stress (dependent variable)	Coefficient Estimates	Standard Error	P-value
Service Affiliation			
Army	Ref	Ref	Ref
Navy	0.66	0.55	0.119
Air Force	0.12	0.58	<0.001
Marine Corps	0.67	0.58	<0.001
Race			
White	Ref	Ref	Ref
African American	0.47	0.54	0.639
Native Hawaiian or Other Pacific Islander	0.7	0.56	0.001
Asian	0.52	0.6	0.828
Other	0.56	0.59	0.508
Marital Status			
Cohabitation	Ref	Ref	Ref
Married	0.7	0.64	0.145
Not Married	0.74	0.63	0.067
Separated	0.78	0.64	0.032
Avg. VL Year 2			
< 1000	Ref	Ref	Ref
=> 1000	0.42	0.54	0.091

b. Logistic Regression Model

The variables in the fitted MLR model included service affiliation, race and ethnicity, and average HIV viral load in Year 2. However, unlike the two previous models, the model has no confounding variable, and has only independent variables with p-values less than 0.05. Also, we check for the excluded variables from the univariate screening, and find no excluded variable is statistically significant in the presence of other variables in the model.

We explore the possibility of the two-way interactions among independent variables that could improve the model. The results from the likelihood ratio test show that the p-value for the model that included the two-way interactions is 0.17, which is not at the significant level. Therefore, we choose to drop the interaction terms and proceed with the original model.

We determine the following results while holding all other variables constant. Compared to patients in the Army, patients in the Air Force and in the Marine Corps had a 11% (se = 58%) and 69% (se = 59%) higher chance of having a least one visit for psychological stress, respectively. Patients in the Native Hawaiian or Other Pacific Islander had a 68% higher chance to have at least one visit than White patients (se = 57%). Patients with greater than 1,000 copies/mL in viral load had a 39% higher change of having a visit when compared to patients with the viral load under that level (se = 55%) (see Table 15).

Table 15. MLR – Visits for Psychological Stress

Visit for Psychological Stress (dependent variable)	Coefficient Estimates	Standard Error	P-value
Service Affiliation			
Army	Ref	Ref	Ref
Navy	0.58	0.55	0.128
Air Force	0.11	0.58	<0.001
Marine Corps	0.69	0.59	0.026
Race			
White	Ref	Ref	Ref
African American	0.45	0.55	0.342
Native Hawaiian or Other Pacific Islander	0.68	0.57	0.009
Asian	0.57	0.61	0.551
Other	0.61	0.6	0.268
Avg. VL Year 2			
< 1000	Ref	Ref	Ref
=> 1000	0.39	0.55	0.02

c. Logistic Regression Diagnostics

For model diagnostics, we first examine the binned version of the deviance residuals and the predicted values of the scale of log-odds for the full model (Faraway 2016). Figure 18 does not show any obvious trend in the average residuals. Thus, there is no major concern about the statistical inadequacy of the model.

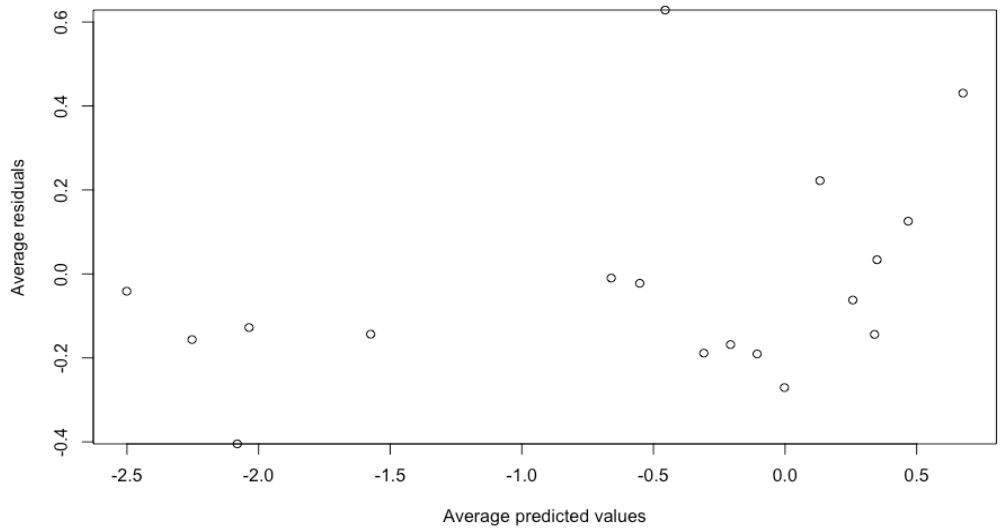

Figure 18. Average Residuals vs. Average Predicted Values Plot – Visits for Psychological Stress

We also check for influential observations in the model. Observation 215 has the highest Cook's distance of less than 0.025, which is below the level to be considered influential; thus, there is no influential observations in the model (see Figure 19).

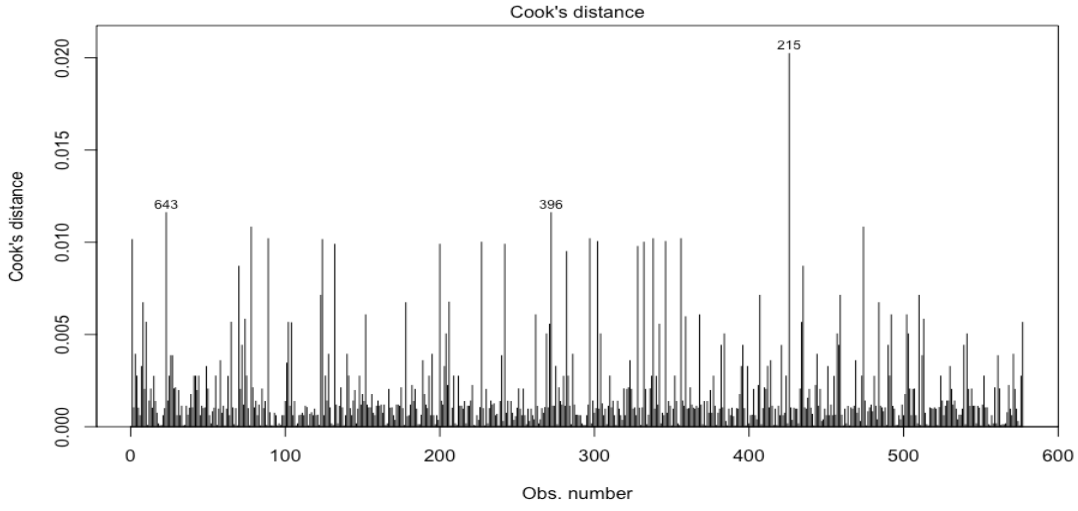

Figure 19. Cook's Distance – Logistic Regression Visits for Psychological Stress

d. Results

We use the test dataset to check the effectiveness of the fitted logistic model at predicting the responses of the dependent variable. At a threshold of 70 percent, the overall classification accuracy rate of the model is at 58.5 percent. Of the 112 patients that did not have any visit for psychological stress, the model predicts 109 of them correctly (97.3%). The percentage decreases to 5 percent as the model predicts 4 of 80 patients correctly that they had at least one visit (see Table 16).

Table 16. Confusion Matrix – Visits for Psychological Stress

	Predicted (no visits)	Predicted (at least one visit)
Actual (no visits)	109	3
Actual (at least one visit)	76	4

V. CONCLUSION

Our goals for this thesis were to recognize the DoD healthcare efforts to achieve the HIV continuum of care focus areas, to analyze how HIV-positive active duty members utilize military healthcare, and to build models that could predict the appointment utilization of the patient population so that DoD medical planners can forecast and manage appropriate amount of HIV care resources in the future.

A. HIV CARE CONTINUUM

DoD healthcare has been meeting the three components of the HIV continuum care. First, all active duty members are mandated to have routine HIV tests. Second, the HIV-positive active duty members have open-access to healthcare while remaining in active duty status. In addition, patients start ART during initial visits in Year 1, and this treatment is aligned with the requirement in the care continuum. Third, the ultimate goal of the HIV continuum care is to achieve HIV viral suppression.

The cohort study results show that the patients' viral loads decreased dramatically from Year 1 and during the following four years. The median HIV viral load of patients in Year 1 was 17,008 copies/mL with an interquartile range of 3,647 copies/mL and 41,400 copies/mL, and the median of viral load of patients in Year 2 was 52 copies/mL with an interquartile range of 34 copies/mL and 7,477 copies/mL. In Year 5, the median viral load was 35 copies/mL with an interquartile range of 20 copies/mL to 48 copies/mL. Clearly, the DoD healthcare system is one of the leaders in meeting and achieving the HIV care continuum requirements today.

B. FINDINGS

This study has two main findings. First, compared to patients in the other services, patients in the Army have the smallest average number of medical visits annually and a smaller likelihood of having at least one visit for asymptomatic HIV and psychological stress. Second, patients in the Air Force had greater average number of counseling visits compared to patients from the other three services. Second, the likelihood of having at least

one counseling visit is significantly higher for patients in the Air Force than those in the other three services.

C. RECOMMENDATIONS

Using various statistical methods and cross-validation, we created four predictive models. Our original dataset contains 796 patients; however, many of the patients were excluded during the model building due to the missing data in one or more independent variables. Consequently, the sample size became even smaller, and that caused less accuracy in the final models. Also, the 18 independent variables were not enough to build models that can accurately predict number of visits even though we constructed ten categorical variables of CD4 cell count levels and HIV viral load levels to form the 18 independent variables. The prevalence of comorbidities of HIV-positive patients is closely related to the overall care utilization; therefore, more data of the comorbidities can improve our models.

Future studies should include more years of patient care data from the sponsor; however, we still need to consider the fact that the overall patient care dataset should only contain the time period when the main HIV clinical care procedures and guidelines are the same through the time period.

If we could predict the care utilization with high levels of accuracy, the DoD would be able to better forecast and manage healthcare resources. Military healthcare intends to increase resources for the future; however, resources for treating the HIV-positive active duty personnel could be reduced, since the rate of new HIV infections should be lowered as the result of continuously implementing the HIV care continuum around the world.

LIST OF REFERENCES

Aidsmap (2012) Seroconversion. Accessed Feburary 16, 2018, http://www.aidsmap.com /Seroconversion/page/1322973/.

Aidsmap (2017) Viral load. Accessed Feburary 16, 2018, http://www.aidsmap.com /CD4-cell-counts/page/1044596/.

Breiman L (2001) Random forests. *Machine Learning.* 45:5–32.

Breiman L, Culter A (2018) Breiman and Culter's random forests for classification and regression. 4.6-14 version. https://cran.r-project.org/web/packages/randomForest/ randomForest.pdf

Bursac Z, Gauss CH, Williams DK, Hosmer DW (2008) Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine.* 3:1–8.

Centers for Disease Control (2015) International classification of diseases, ninth revision, clinical modification. Acessed Feburary 16, 2018, https://www.cdc.gov /nchs/icd/icd9cm.htm.

Centers for Disease Control (2017a) Evidence of HIV treatment and viral suppression in preventing the sexual transmission of HIV. Accessed Feburary 16, 2018, https://www.cdc.gov/hiv/pdf/risk/art/cdc-hiv-art-viral-suppression.pdf

Centers for Disease Control (2017b) Understanding the HIV care continuum. Accessed Feburary 16, 2018, https://www.cdc.gov/hiv/pdf/library/factsheets/cdc-hiv-care-continuum.pdf.

Department of Defense (2013) Human immunodeficiency virus (HIV) in military service members, DoD Instruction 6485.01, Department of Defense, Washington, DC.

Department of Defense (2015) 2015 demographics profile of the military community, Accessed Feburary 16, 2018, http://download.militaryonesource.mil/12038/MOS/ Reports/2015-Demographics-Report.pdf

Department of Health and Human Services (2016) HIV care continuum. Accessed Feburary 16, 2018, https://www.hiv.gov/federal-response/policies-issues/hiv-aids-care-continuum.

Devore J (2014) *Probability and Statistics for Engineering and the Sciences,* 9th ed. (Cengage Learning, Boston, MA).

Faraway J (2014) *Linear Models With R,* 2nd ed. (CRC Press, Boca Raton, FL).

Faraway J (2016) *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models,* 2nd ed. (CRC Press, Boca Raton, FL).

HIV.gov (2017) Global statistics. Accessed Feburary 16, 2018, https://www.hiv.gov/hiv-basics/overview/data-and-trends/global-statistics.

Hosmer D, Lemeshow S, May S (2008) *Applied Survial Analysis: Regression Modeling of Time-to-Event Data,* 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Hughson G, CD4 cell counts. Accessed Feburary 16, 2018, http://www.aidsmap.com/CD4-cell-counts/page/1044596/

Infectious Disease Clinical Research Program (2015a) The U.S. military HIV natural history study. 2015 Annual Report, National Institute of Health, ICDRP HIV Research Area, Bethesda, MD.

Infectious Disease Clinical Research Program (2015b) The U.S. military HIV natural history study. Cohort Profile, Report, National Institute of Health, ICDRP HIV Research Area, Bethesda, MD.

Infectious Disease Clinical Research Program (2016) NHS – human immunodeficiency virus. Accessed Feburary 16, 2018, http://www.idcrp.org/research-area/human-immunodeficiency-virus.

Infectious Disease Clinical Research Program (2017) Data provided to the author via email, March 9.

Joint United Nations Programme on HIV/AIDS (2014) 90–90-90 An ambitious treatment target to help end the AIDS epidemic. Report, UNAIDS, Geneva, Switzerland.

Liaw, A and Wiener, M (2002) Classification and Regression by randomForest. *R News,* 2(3), 18–22.

Miller R (1986) *Beyond ANOVA: Basics of Applied Statistics* (John Wiley & Sons, New York, NY).

Personnel-General (2014) Identification, surveillance, and administration of personnel infected with human immunodeficiency virus. Regulation 600–110, Headquarters: Department of the Army, Washington, DC.

R Core Team (2018) The R project for statistical computing. Accessed: June 3, 2018, https://www.r-project.org.

Secretary of the Air Force (2014) Human immunodeficiency virus program. Instruction 44–178, Headquarters, U.S. Air Force, Washington, DC.

Secretary of the Navy (2012) Management of human immunodeficiency virus, hepatitis B virus and hepatitis C virus infection in the Navy and Marine Corps. Instruction 5300.30E, Department of the Navy, Washington, DC.

The White House (2015) *National hiv/aids strategy for the United States*. Policy, The White House, Washington, DC.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California