2019-12

PREDICTING U.S. ARMY FIRST-TERM ATTRITION AFTER INITIAL ENTRY TRAINING, PART II

Gobea, Gabriel A.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PREDICTING U.S. ARMY FIRST-TERM ATTRITION AFTER INITIAL ENTRY TRAINING, PART II

by

Gabriel A. Gobea

December 2019

Thesis Advisor: Lyn R. Whitaker
Second Reader: Jon Alt (TRAC-Monterey)

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 2019	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE PREDICTING U.S. ARMY FIRST-TERM ATTRITION AFTER INITIAL ENTRY TRAINING, PART II	**5. FUNDING NUMBERS**
6. AUTHOR(S) Gabriel A. Gobea	
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)

The U.S. Army currently stands at an active duty strength of 476,000. MG Joseph Callaway, commander of Personnel Management at Army headquarters, stated recently that the Army missed its recruiting mission and is in danger of not reaching its end strength of 483,500. The Army's shortfall comes from a strong economy and increased competition from the private sector, which can pay more. The Army is growing its force to meet the high demand for deployments to continue the fight against the war on terrorism. In order to increase its force, the Army must not only recruit new personnel but also ensure that the civilians it recruits complete their first-term obligation contract. This thesis continues the work of Speten in 2018 and uses the Army's Person-Event Data Environment (PDE) to build a logistic regression model to predict attrition among active duty enlisted soldiers. This research uses demographic and medical factors from the PDE to identify soldiers with the highest probability of failure. We use random forests to identify important predictors of attrition and use those predictors to fit a simple additive logistic regression model. The result shows that PULHES Non-deployable, Dental Class, Contract Duration, Unit Type, Medical Non-deployable, Hearing Class, Gender, Smoker, Education Tier, and Marital Status are the most influential factors that contribute first-term attrition.

14. SUBJECT TERMS Army, attrition, survival analysis, logistic regression, random forest, enlisted, retention, first-term, classification, predict, tree	**15. NUMBER OF PAGES** 73
	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.

PREDICTING U.S. ARMY FIRST-TERM ATTRITION AFTER INITIAL ENTRY TRAINING, PART II

Gabriel A. Gobea
Major, United States Army Reserve
BS, University of Texas at Tyler, 2002
MEd, Lamar University, 2011

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 2019

Approved by: Lyn R. Whitaker
Advisor

Jon Alt
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The U.S. Army currently stands at an active duty strength of 476,000. MG Joseph Callaway, commander of Personnel Management at Army headquarters, stated recently that the Army missed its recruiting mission and is in danger of not reaching its end strength of 483,500. The Army's shortfall comes from a strong economy and increased competition from the private sector, which can pay more. The Army is growing its force to meet the high demand for deployments to continue the fight against the war on terrorism. In order to increase its force, the Army must not only recruit new personnel but also ensure that the civilians it recruits complete their first-term obligation contract. This thesis continues the work of Speten in 2018 and uses the Army's Person-Event Data Environment (PDE) to build a logistic regression model to predict attrition among active duty enlisted soldiers. This research uses demographic and medical factors from the PDE to identify soldiers with the highest probability of failure. We use random forests to identify important predictors of attrition and use those predictors to fit a simple additive logistic regression model. The result shows that PULHES Non-deployable, Dental Class, Contract Duration, Unit Type, Medical Non-deployable, Hearing Class, Gender, Smoker, Education Tier, and Marital Status are the most influential factors that contribute first-term attrition.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AC	U.S. Army Active Component
AUC	Area Under the Curve
ARD	Army Resiliency Directorate
CAR	Center of Accessions Research
CPL	Corporal
DMDC	Defense Manpower Data Center
DoD	Department of Defense
GAO	Government Accountability Office
HIV	Human Immunodeficiency Virus
IET	Initial Entry Training
LIK	Likelihood
MEDPROS	Medical Protection System
MEPS	Military Entrance Processing Stations
NPS	Naval Postgraduate School
OASD	Office of the Assistant Secretary of Defense
PDE	Person-even Data Environment
PHA	Periodic Health Assessment
PID	Person Identifier
PII	Personally Identifiable Information
PV1	Private
PV2	Private Second Class
RAND	Research and Development Corporation
ROC	Receiver Operating Characteristics
SGT	Sergeant
SSG	Staff Sergeant
UIC	Unit Identification Codes
USACHPPM	U.S. Army Center for Health Promotion and Preventive Medicine

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The U.S. Army currently stands at an active duty strength of 476,000. MG Joseph Callaway, commander of Personnel Management at Army headquarters, stated recently that the Army missed its recruiting mission and is in danger of not reaching its end strength of 483,500 (Myers 2018). Recruiting shortfall comes from a robust civilian economy and increased competition from the private sector, which can pay more. The Army wants to grow its force to meet the high demand for deployments to continue to fight the war on terrorism. In order to increase its force, the Army must not only recruit new personnel but also ensure that the civilians it recruits complete their first-term obligation contract.

This thesis uses the Person-Event Data Environment (PDE) to build a logistic regression model that can predict attrition among active duty enlisted soldiers. It continues the work of Speten (2018) by using both demographic and medical factors from the PDE to identify soldiers with the highest probability of failure. To limit the number of predictors, we take two approaches. We fit a lasso-regularized logistic regression model. We also fit a random forest to identify important predictors and then use those predictors to fit an additive logistic regression model. The performance of the two model fits is comparable on an independent hold-out validation set, but the logistic regression model using the variables identified using the random forest is more easily interpreted and has greater potential for improvement.

The data used during this research comprises of the cohort dataset created during the first part of the first-term attrition study by Speten (2018). Using medical tables from the PDE, medical data was merged to the original cohort dataset, which added the medical data to soldiers entering basic training during fiscal years (FY) 2005 to FY 2010. FY 2005 through 2007 have missing medical data that present some challenges to building a logistic regression model. Thus we construct a new cohort using records of soldiers entering basic training only during FY 2008 and FY 2009. The FY 2010 data is used as a test set to assess how well the final model forecasts. The new FY 2008–FY 2009 cohort is split into a training dataset and a validation dataset using 80% for the training dataset and 20% for the validation dataset.

The training dataset produces a logistic regression model with 19 (random forest selected) variables. Using the validation dataset, the logistic regression model has an accuracy of 86% and a misclassification rate of 14%. Model accuracy drops to 83.7% when used to predict FY 2010. Further, because the logistic regression provides good attrition probability estimates, these results suggest that the model is best used for identifying groups of soldiers with high (or low) attrition rates rather than predicting attrition for individual soldiers. These results are slightly better than those of Speten (2018), but our model uses both medical and demographic variables and only uses variables whose values predict attrition rather than those whose values may be a consequence of attrition.

The results from the logistic regression model show that the addition of medical variables is important for predicting attrition. Results from the model show that the medical and demographic factors that influence U.S. Army first-term attrition are PULHES Non-deployable, Dental Class, Contract Duration, Unit Type, Medical Non-deployable, Hearing Class, Gender, Smoker, Education Tier, and Marital Status. The work and analysis performed during this research makes it possible to build, test, and validate a working logistic regression model that can be used to predict U.S. Army first-term attrition with 86% accuracy.

References

Myers M (2018) The Army is supposed to be growing, but this year, it didn't at all. *Army Times.* Accessed February 1, 2019, https://www.armytimes.com/news/your-army/2018/09/21/the-army-is-supposed-to-be-growing-but-this-year-it-didnt-at-all

Speten , K (2018) Predicting U.S. Army first-term attrition after initial entry training. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/59593

ACKNOWLEDGMENTS

I want to thank my wife, Elizabeth, for all her love, support, and encouragement throughout my time here at Naval Postgraduate School (NPS). I could not have accomplished all of this without you by my side. You gave me strength when I was weak, you picked me up when I would fall, and gave me the courage to press on when things got tough. Your love and support helped me to make it through. Thank you for being who you are and for always loving me no matter what.

I would also like to thank Professor Lyn Whitaker for all her help throughout this research. Your encouragement, mentoring, and guidance help me to accomplish this thesis. There are no words that can ever express how thankful I am for all you did for me during my time at NPS.

Thank you to Dr. Alt and MAJ Tony Smith from TRAC Monterey. None of this would be possible without you giving me the opportunity to work on this research project.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The ability to recruit and retain soldiers will continue to be a challenge for the Army. The Army invests resources and thousands of dollars in government funds to train recruits to become U.S. Army soldiers. The Army, as well as the other services, struggle with early attrition. Army recruits who fail to meet their initial contract are costly for the U.S. Army and make it harder for the Army to meet an end strength of 490,000 by 2019 (Bushatz 2018). Recruits who fail to meet their initial contract obligation have a significant impact not only to force numbers but also cost the Army tens of thousands of training dollars and equipment.

This thesis is a continuation of the work Speten (2018) and arguments the work of Devig (2019). This research will focus on using demographic and medical data from active duty soldiers to identify the primary demographic and medical factors that can be used to predict first-term attrition among soldiers who complete Initial Entry Training (IET).

A. BACKGROUND

The U.S. Army Active Component (AC) currently stands at a strength of 476,000. According to MG Joseph Callaway, head of personnel management at Army headquarters, the Army did not meet their accession target and will miss their target end strength of 483,500 (Myers 2018). The Army plans on increasing its fighting force to meet the high demand of deployments to continue fighting the war on terrorism. The Army must not only recruit new soldiers but must also keep those recruits from leaving the Army before completing their first-term obligation.

B. PREVIOUS RESEARCH

Military attrition continues to put a strain on the Department of Defense (DoD). Despite the increase of highly qualified recruits, about a third of the new recruits will leave the military before they are able to complete their first-term obligations (Government Accountability Office [GAO] 1998). Although research on military attrition is extensive, research concerning how medical factors contribute to military attrition after IET is limited.

Further, with the exception of the work of Devig (2019) and our work conducted using the same fiscal year (FY) 2005 through FY 2010 U.S. Army assessions, these studies are based on data that predates our earliest records by at least twenty years. However, earlier studies, such as the study by the U.S. Army Center for Health Promotion and Preventive Medicine (USACHPPM) (Knapik et al. 2004) give insights into how medical factors influence Army attrition. Additional research on military attrition conducted mainly by two organizations: GAO (GAO 1997, 1998) and The Research and Development Corporation (RAND) (Buddin 1981) provide insights on how the military struggles with first-term attrition. Work on military attrition focusing on using demographic and administrative factors (e.g., Speten 2018) show that in general significant factors that contribute to service members leaving the military include gender and general education. The following is an overview of some of that research to get a better understanding of U.S. Army attrition. We do not review Speten (2018) or Devig (2019) explicitly but refer to these studies as needed throughout this thesis.

1. U.S. Army Center for Health Promotion and Preventive Medicine

The Center of Accessions Research (CAR) tasked USACHPPM with conducting a literature review on military attrition. Their research (Knapik et al. 2004) looks at a wide range of factors influencing attrition with the primary focus on looking into health-related factors.

In its technical report, USACHPPM finds that overall three-year military attrition steadily rose from 26% in 1985 to 31% in 1995. Approximately one-third of the attrition occurs within the first six months of service, and approximately 26% of that population attrite due to medical or physical problems. This implies that most of these service members attrite during their basic training or technical school. Attrition is higher for those receiving medical waivers for hearing problems, back disorders (Army only), prior knee injuries (Army only), depression, and a skin/cellular tissue disorder (Knapik et al. 2004). Mental health-related factors also contribute to attrition in basic training and in advance training personnel. USACHPPM's research concludes that one way to reduce attrition is to prescreen individuals before they enter the service.

2. Government Accountability Office

The GAO provides auditing and investigation services for the U.S. Congress, and is often referred to as the "congressional watchdog." GAO helps Congress oversee federal programs to ensure accountability to the citizens of the United States. In 1997, Congress requested the GAO review military attrition rates of first-term, active duty personnel who separate within the first six months of their enlistments (GAO 1997).

Speten (2018) writes that GAO engaged in many studies dealing with military attrition across all military branches. GAO uses data collected from the Defense Manpower Data Center (DMDC) on recruits from 1986 to 1994. The GAO concludes that one-third of enlistees leave the service before completing their first-term contract. GAO analysts observe that 55% of the service member who attrit within in the first six months are either separated for a medical condition or fail to meet performance standards.

The following year, GAO expanded on its attrition study to improve recruiting systems. In that same year, GAO (1998) reports it conducted several studies to investigate why enlisted first-term attrition remains constant when there is an increase in qualified recruits. GAO (1998) recommends improving medical screening at Military Entrance Processing Stations (MEPS). The intent is to have mechanisms in place to identify past medical problems or mental health problems. GAO (1998) also recommends having incentive systems for recruiters. Many of the services measure recruiting success by the number they can enlist each year. GAO made these recommendations to Congress to reduce first-term attrition by recruiting a qualified applicant that is physically and medically able to finish their first tour of duty.

3. RAND Corporation: "The Role of Serving Experience in Post-Training Attrition in the Army and Air Force" by Richard Buddin

Office of the Assitant Secretary of Defense (OASD) sponsored RAND's study (Buddin 1981) to gain insights into first-term enlisted attrition with the purpose of developing strategies and solutions to the manpower problems DoD faces now and in the future. High attrition rates are costly due to the number of resources and equipment that are invested in training and equipping individual service members only for them to leave

before finishing their first term. It costs the services more to lose a technically qualified specialist than to lose a trainee (Buddin 1981). RAND wanted to gain insights on what individual characteristics and military environments affect post-training attrition.

RAND (Buddin 1981) base its analysis on the FY 1975 cohort file constructed by DMDC. This file contains data on nonprior military accessions for FY 1975 (Buddin 1981), and a multivariate attrition model to describe the effects individual characteristics and military environment have on attrition. RAND's model uses demographic and administrative variables to see which characteristics influence post-training attrition rates.

Buddin (1981) shows that some of the influencing factors are region of origin, age at enlistment, education, family status, race, mental aptitude, and family status. The results also show that recruits without high school diplomas are 10% more likely to leave before their enlistment term than recruits with high school diplomas. Married recruits are 3%–8% less likely than single recruits to leave before their first-term, and Army recruits who enter the service before 18 years of age have 5%–7% higher attrition rates than Army recruits who enter at age 18 years or older. RAND presented their results to policymakers in order for them to implement policy changes that would lower first-term attrition.

C. OBJECTIVE AND ORGANIZATION

We use personnel data and software in the Person-Event Data Environment (PDE) to investigate post-IET first-term attrition among AC soldiers enlisting in the U.S. Army from FY 2005 to FY 2010. We do this by using the information available about the soldier completing IET. This is a follow-on research to previous work of Speten (2018). Using the PDE, Speten (2018) focuses on finding relationships between first-term attrition and demographic and administrative variables. We add medical variables constructed from the databases in the PDE, but unavailable to Speten (2018) with a focus on those variables that can be used to identify groups of soldiers with the highest probability of failure in order to implement preventative measures. In particular, using logistic regression to estimate attrition probabilities, we find that using medical variables improves the accuracy of predicting post-IET first-term attrition.

The thesis consists of five chapters. Chapter II provides a description of the data and the methodology for constructing models used to predict attrition. In Chapter III, the data and variables are explored using descriptive statistics. Chapter IV gives a discussion on logistic regression as well as our analysis and findings. Chapter V summarizes and concludes the thesis and provides recommendations.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA AND METHODOLOGY

A. PERSON-EVENT DATA ENVIRONMENT

The PDE, built and adminstered by the Army Analytical Group (AAG), gives the analyst access to databases from many sources, access to metadata describing each database, and a set of analytical tools to conduct their research. The PDE provides a centralized data warehouse for a soldier's service, financial, and medical data that can be accessed by the researcher (Speten 2018).

The PDE is designed to provide a comprehensive and accessible data repository and analytical environment (Jensen 2016). In this environment, data such as personnel data and medical data can be stored, quality-controlled, and secured to protect Personally Identifiable Information (PII). This system generates a person identifier (PID) field unique to each individual and constant across databases used for a particular project allowing the researcher to merge data across databases easily. Querying data now becomes more manageable in the PDE by using the PID as the key to link databases.

In this research, datasets constructed by Speten (2018) provide a starting point. The dataset used by Speten (2018) contains primarily administrative and demographic variables for the cohort all AC soldiers enlisting from FY 2005 through FY 2010 who do not attrite in IET. This cohort dataset contains a few AC medical variables such as health information from a MEPS. Using medical databases from the PDE not available to Speten (2018), we merge soldier medical records to the existing cohort data to bring more predictive power to our data set.

B. DESCRIPTION OF RESEARCH DATA

1. Datasets Used

Our research uses the cohort dataset constructed by Speten (2018). This dataset contains 414,766 observations and 63 variables and is described in detail by Speten (2018). The variables consist of demographic and administrative variables from the soldier's individual service record that is managed by the Army Human Resource Command (HRC)

in the Total Army Personnel Database (Speten 2018). The original dataset also includes soldier data from the Military Entrance Processing Command (MEPCOM) system from MEPCOM databases contained within the PDE. The PDE databases record values of some time-varying variables such as marital status as "snapshots" at fixed times such as on the last day of each quarter in an FY. Others, such as promotions or discharges, are recorded as transactions with a description of the transaction and the transaction date. Since IET length varies among soldiers, immediate post-IET time-varying variable values are taken to be the value recorded on a date closest to the IET completion date. During the building of our cohort, we remove 4,607 soldiers because, as noted by Devig (2019), we notice that their separation code is "1016" and they serve less than 4.5 months after their start date which implies that they are discharged during IET. The new total for the cohort is 410,159.

The first database used from PDE to merge medical data to the cohort dataset is the Physical Health Assessment (PHA) database. This database consists of medical data collected from soldiers during their annual PHA. The PHA is a screening tool used by the Army to evaluate the individual soldier's combat readiness. PHA data includes a review of current medical conditions, vision screening, measurement and documentation of vitals (height, weight, blood pressure), and self-reporting health status. The PHA data also contains behavioral health screenings or medical profiles that can keep a soldier from deploying. One of the most critical pieces of information that we get from the PHA data table is PULHES information for each soldier. PULHES stands for Physical capacity, Upper extremities, Lower extremities, Hearing, Eyes, and Stability/Psychiatric and measures a soldier's medical fitness in each of these cases (Army-Portal 2011).

The second database used is the Medical Protection System (MEDPROS) Individual Medical Readiness database. The MEDPROS data contains the complete overall readiness profile for all soldiers. It consists of overall medical readiness with regards to dental and vision exams, Human Immunodeficiency Virus (HIV) screening, and PHA documentation and examinations. The MEDPROS readiness profiles are constructed using all medical data available on a soldier, and gives an overall determination to see if a soldier is deployable or nondeployable.

2. Medical Variables Used

After joining medical fields with the original cohort dataset, our cohort dataset contains 410,159 observations with 105 variables. There are 63 original variables and 42 new variables that we get from the medical databases. The 42 new variables contain three fields not used during our analysis because two contain dates, and a third gives separation codes. The remaining 39 variables consist of ten categorical variables with more than two levels, and 29 binary variables representing categorical variables with two levels.

a. Binary Variables

Binary variables are used to represent categorical variables with two levels, "yes" and "no," Table 1 shows 29 the binary medical variables that are in our cohort dataset. The majority of these variables come out of the PHA database. Medical Nondeployable Profile (MEDICAL_NONDEPLOYABLE_PROFILE) and Limited Duty Profile (LIMITED_DUTY_PROFILE) are variables that a medical provider will enter into MEDPROS. The rest of the variables are self-report medical conditions that a soldier will answer during their annual PHA.

The PHA database contains many blanks or missing values. We assume if the field is a "yes/no" selection and came from soldier self-reporting medical conditions, the blank or missing value should be "no." An example of this is when a soldier is answering if he or she is allergic to bee stings. If the soldier leaves this blank, assume he or she is not allergic to bee stings.

Table 1. Binary Variable Definitions

Variable Name	Description	Code
MEDICAL_NONDEPLOYABLE_PROFILE	Soldier Has Medical Nondeployable Profile	0: No, 1: Yes
LIMITED_DUTY_PROFILE	Soldier Has Limited Duty Profile (medical)	0: No, 1: Yes
CH_ANEMIA_SOLDIERHAS	Current Health: Soldier Has Anemia	0: No, 1: Yes
CH_ASTHMA_SOLDIERHAS	Current Health: Soldier Has Asthma	0: No, 1: Yes
CH_BACKPAIN_SOLDIERHAS	Current Health: Soldier Has Back Pain	0: No, 1: Yes
CH_CANCER_SOLDIERHAS	Current Health: Soldier Has Cancer	0: No, 1: Yes
CH_CHRONICPAIN_SOLDIERHAS	Current Health: Soldier Has Chronic Pain and Soldier Currently Treated	0: No, 1: Yes
CH_DIABETES_SOLDIERHAS	Current Health: Soldier Has Diabetes	0: No, 1: Yes
CH_EPILEPSY_SOLDIERHAS	Current Health: Soldier Has Epilepsy	0: No, 1: Yes
CH_HEARTMURMUR_SOLDIERHAS	Current Health: Soldier Has Heart Murmur	0: No, 1: Yes
CH_HEARTTROUBLE_SOLDIERHAS	Current Health: Soldier Has Heart Trouble, Symptom	0: No, 1: Yes
CH_HYPERTENSION_SOLDIERHAS	Current Health: Soldier Has High Blood Pressure	0: No, 1: Yes
CH_JOINTPAIN_SOLDIERHAS	Current Health: Soldier Has Joint Pain	0: No, 1: Yes
CH_KIDNEY_SOLDIERHAS	Current Health: Soldier Has Kidney Disease	0: No, 1: Yes
CH_LIVER_SOLDIERHAS	Current Health: Soldier Has Liver Disease	0: No, 1: Yes
CH_THYROID_SOLDIERHAS	Current Health: Soldier Has Thyroid Disease	0: No, 1: Yes
CH_ULCERS_SOLDIERHAS	Current Health: Soldier Has Ulcers	0: No, 1: Yes
EVAL_EKGREF	Soldier Has EKG Referral	0: No, 1: Yes
CH_STROKE_SOLDIERHAS	Current Health: Soldier Has Stroke	0: No, 1: Yes

Variable Name	Description	Code
CH_TUBERCULOSIS_SOLDIERHAS	Current Health: Soldier Has Tuberculosis	0: No, 1: Yes
ALLERGY_BEESTINGS_ALLERGY	Soldier Has Allergy to Bee Stings	0: No, 1: Yes
ALLERGY_CODEINE_ALLERGY	Soldier Has Allergy to Codeine	0: No, 1: Yes
ALLERGY_IODINE_ALLERGY	Soldier Has Allergy to Iodine	0: No, 1: Yes
ALLERGY_LATEX_ALLERGY	Soldier Has Allergy to Latex	0: No, 1: Yes
ALLERGY_PENICILIN_ALLERGY	Soldier Has Allergy to Penicillin	0: No, 1: Yes
PH_DOESCHEW	Soldier Uses Smokeless Tobacco	0: No, 1: Yes
PH_ISSMOKER	Soldier Smokes Cigarettes	0: No, 1: Yes
PULHES_DEPLOYABLE	Soldier is PULHES Deployable	0: No, 1: Yes
OVH_PROFILE_SM_ANS	Question: Are you on a profile or do you have a medical condition that keeps you from taking any part of the APFT,_x000D_ requires you to take alternate APFT event, or keeps you from doing your military job duties?	0: No, 1: Yes

b. Categorical Variables

Categorical variables consist of 3 or more levels. In our cohort dataset, the PULHES, Hearing Class, and Dental Class variables each have four levels. For PULHES variables 1, 2, 3, and 4 represent “no,” “some,” “significant,” and “severe” limitations respectfully. For Hearing and Dental Class 1, 2, 3 represent increasining severity, but 4 represents lack of knowledge or no exam in the previous year. Table 2 shows the categorical variables for our cohort dataset.

Table 2. Categorical Variables Definitions

Variable Name	Description	Level
DENTAL_CLASS	Dental Class	1. The soldier does not require treatment 2. The soldier has some oral conditions that will not result in an emergency in 12 months. 3. The soldier has oral conditions that will result in an emergency withing 12 months. 4. The soldier has not had an exam in the last 13 months.
HEARING_READINESS_CLASS	Indicates hearing readiness classification	1. The soldier has no hearing limitations 2. The soldier has some hearing limitations to activities. 3. The soldier has significant hearing limitations 4. The soldier has not had an exam in the last year.
PULHES_PFIELD	PULHES: P Field	1. The soldier has no limitations 2. The soldier has some limitations to activities. 3. The soldier has significant limitations 4. The soldier is severely limited.
PULHES_UFIELD	PULHES: U Field	1. The soldier has no limitations 2. The soldier has some limitations to activities. 3. The soldier has significant limitations 4. The soldier is severely limited.
PULHES_LFIELD	PULHES: L Field	1. The soldier has no limitations 2. The soldier has some limitations to activities. 3. The soldier has significant limitations 4. The soldier is severely limited.
PULHES_HFIELD	PULHES: H Field	1. The soldier has no limitations 2. The soldier has some limitations to activities. 3. The soldier has significant limitations 4. The soldier is severely limited.
PULHES_EFIELD	PULHES: E Field	1. The soldier has no limitations 2. The soldier has some limitations to activities. 3. The soldier has significant limitations 4. The soldier is severely limited.
PULHES_SFIELD	PULHES: S Field	1. The soldier has no limitations 2. The soldier has some limitations to activities. 3. The soldier has significant limitations 4. The soldier is severely limited.
FH_FATHER_CHEMDEPTYPE	Family History: Father, Chemical Dependency Type	0. None 1. Alcohol 2. Other
FH_MOTHER_CHEMDEPTYPE	Family History: Mother, Chemical Dependency Type	0. None 1. Alcohol 2. Other

The table is adapted from Army-Portal.com (2011)

C. METHODOLOGY

We first provide a detailed description of the method used by Speten (2018) to define the "attrit" and "non-attrit" response variable. Using the original cohort, we then merge medical predictor variables to the existing cohort dataset. Medical variables, such as those indicating current health status, may change over the course of a soldier's first term. Others, such as those documenting allergies, may change as pre-existing medical conditions are discovered. For those medical variables corresponding to changes in the medical condition, we use values from a soldiers earliest post-IET transaction or approved PHA form. This gives us variables that measure (as well as we can) a soldier's medical status immediately after IET as they arrive at their first unit. Finally, we will split our data into a training set to train a logistic regression model, and a test set to validate the model.

1. Building Response Variable

The cohort starts with 429,908 unique soldier records that represent all the enlisted soldiers that arrive at basic training in FY 2005 to FY 2010 (Speten 2018). Of those soldiers, 11,704 are removed for having a "1087" Service Separation Code (ISVC_SEP_CD) which is discharged from the Army before completing IET. Another 4,607 are also removed for having a "1016" service separation code indicating unqualified for active duty and for serving for less than 4.5 months.

Next, the Enlisted Career Status Code (ELN_CRER_STAT_CD) is used to determine if a soldier completes their first term. Soldiers who had a "3" Enlisted Career Status Code are treated as "non-attrit" because the code represents reenlistment. We classify 175,970 (42.5%) soldiers with the code "3" as "non-attrit" which leaves a remaining 237,627 soldiers. We are able to split these soldiers into two categories, those with good Separation and Discharge Codes (SPD_CD) and those without. Out of the 1,869 soldiers without a good SPD_CD, 139 are removed because of missing entries and no value for their initial obligation date. The rest of the soldiers are missing their end date but do have Basic Active Service Date (AFMS_DT). We calculate a Calculated Obligation Date (CALC_OBL_DT) by adding the number of years soldier would contract for by the Army

to their Basic Active Service Date. So if a soldier's Basic Active Service Date is 11/1/2016 and he or she enlists for three years, then their Calculated Obligation Date is 11/1/2019.

We use the Calculated Obligation Date for each soldier and compare it to the last quarterly date that data is updated in a soldier's record, also known as "(the last) snapshot" date. If the Calculated Obligation Date is greater than the snapshot date, then the soldier is classified at "attrit," otherwise we classify them as "non-attrit." Using this logic, we classify 1,528 (0.37%) as "attrit" and 202 (0.049%) as "non-attrit. Of the 235,758 soldiers that have good Separation and Discharge codes, 109,126 (26.4%) have good codes and are classified as "non-attrit," and 3,299 are removed because they do not have a value for initial obligation duration. With the remaining 123,333 soldiers, again using the Calculated Obligation Data, we are able to classify 21,302 (5.2%) as "non-attrit" and 102,031 (24.1%) as "attrit." Figure 1 summarizes the methodology of classification as a flowchart.

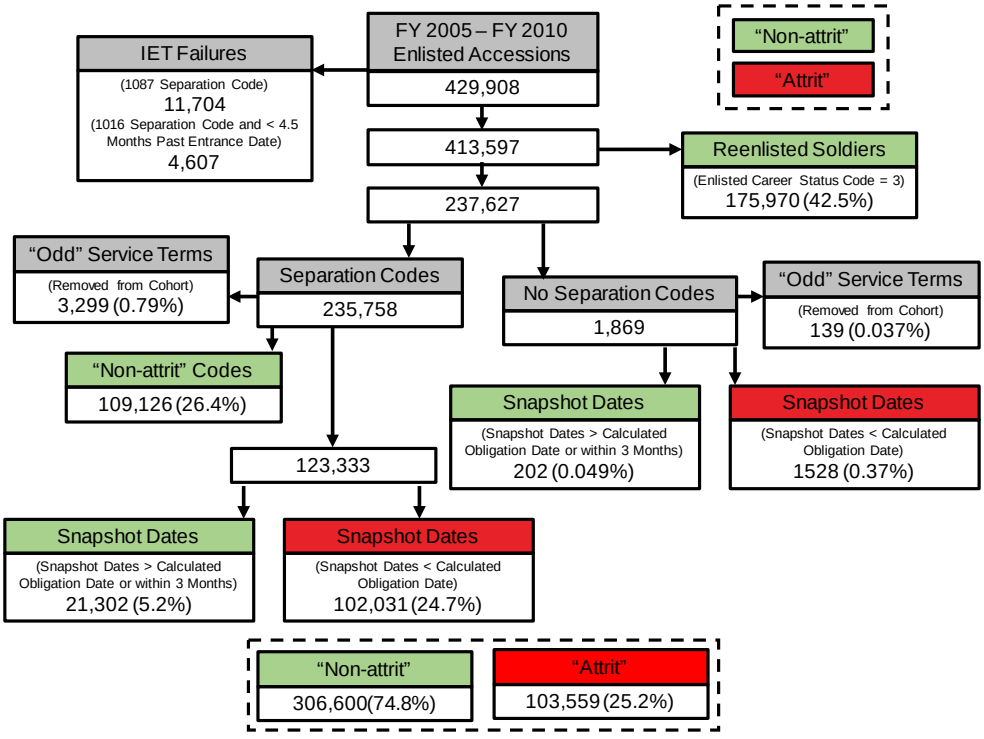

Figure 1. Flowchart Summary of Response Variable

2. Merging Predictor Variables

We use 23 of the predictor variables constructed by Speten (2018). We only use those that can be constructed on or before the end of IET among the variables selected by Speten (2018 p. 47–49)as important for predicting post-IET first-term attrition. This gives us a dataset with only a total of 62 predictor variables.

As stated previously, the PHA database has a lot of missing values for records of the early FY assessions. For the 67,215 soldiers who enlisted in FY 2005, 38.4% have missing PHA data, and for the 75,279 soldiers who enlisted in FY2006, 32.3% have missing PHA data. In order to reduce the number of soldier records with missing data, we choose to only look at FY 2008 and FY 2009 for model fitting, reserving FY 2010 for to assess the final model. This reduces the number of observations for model fitting to 130,772. We use a function (Rstudio Team 2018) called missRanger() from the R package missRanger of Mayer (2019) to impute missing values by a chained random forest. The missRanger() function allows us to fit both random forest models and logistic regression models using all 130,772 records.

3. Training, Validation, and Test Sets

Once imputing is complete, our last step in preparing our data for analysis by splitting our FY 2008–FY 2009 cohort into a training set used to train our model and a validation set which will validate our model. We will do this by using a randomly selected 80% of the data for training and 20% of the data for validation. The training set consists of 104,618 observation, and the valistion set consists of 26,154. The FY 2010 test set is only use to asses the final model fit and consists of 66,806 observations.

D. LIMITATIONS AND ASSUMPTIONS

One limitation of our research is working within the PDE system. In order to keep PII data secured, our analysis had to be done while logged into the PDE system. The PDE system does go offline during business hours for regular updates or repairs which during business hours limits the time to do analysis. In addition, Unit Identification Codes (UIC)

in the PDE are scrambled to protect PII. Thus, we cannot use IET location or first unit as predictor variables.

Another limitation is that our study works with data from soldiers who asses from FY 2005 through FY 2010. This cohort contains 410,159 observations, but missing values in the medical variables, force us to restrict attention to FY 2008 and FY 2009 data. Another limitation is that we are only working with AC Army enlisted soldiers. We do not look at Reserve Component, National Guard, or Officers in our study.

An assumption for this study is all data that comes from the PDE is accurate and gives a complete representation of the soldier's medical and administrative information. Second, we assume that the methodology used to construct the response variable and predictor variables in both Speten (2018) and our work is reasonable.

III. DESCRIPTIVE STATISTICS

A. DATASET OVERVIEW

In this chapter, we describe the 401,159 observations that make up our cohort dataset. We begin by first computing the average attrition rate by accession fiscal year, and then show attrition rates as they relate to non-deployable soldiers. Finally, we will see how attrition relates to the different levels for Dental Class, Hearing Class, and other demographic variables.

Table 3 shows the breakdown of the cohort dataset by accession fiscal year and by the response variable "non-attrit" and "attrit." The cohort has an attrition rate of 25.2% across all fiscal years. The attrition rate does change slightly from fiscal year to fiscal year, and never exceeds 26.4 %.

Table 3. Full Cohort Dataset-Attrition Rates by Accession Fiscal Year

	FY05	FY06	FY07	FY08	FY09	FY10	Total
Non-Attrit	50,199 (74.7%)	56,213 (74.7%)	51,575 (73.6%)	50,870 (73.6%)	46,272 (75.0%)	51,471 (77.0%)	306,600 (74.8%)
Attrit	17,105 (25.3%)	19,066 (25.3%)	18,512 (26.4%)	18,224 (26.4%)	15,335 (25.0%)	15,335 (23.0%)	103,559 (25.2%)

Next, we compare attrition rates by gender for each of the accession fiscal years. Across all fiscal years, males have an attrition rate average of 22.8% while females show a slightly higher attrition average of 38.5%. The results show that female attrition rates are always higher than males. The average male attrition rate ranges from 22.1% to 24.0% and does not seem to show an upward or downward trend. Female attrition rates range from 32.5% to 41.9% and also show a downward trend from FY 2005 to FY 2010. Figure 2 shows a graph of the attrition rates by gender.

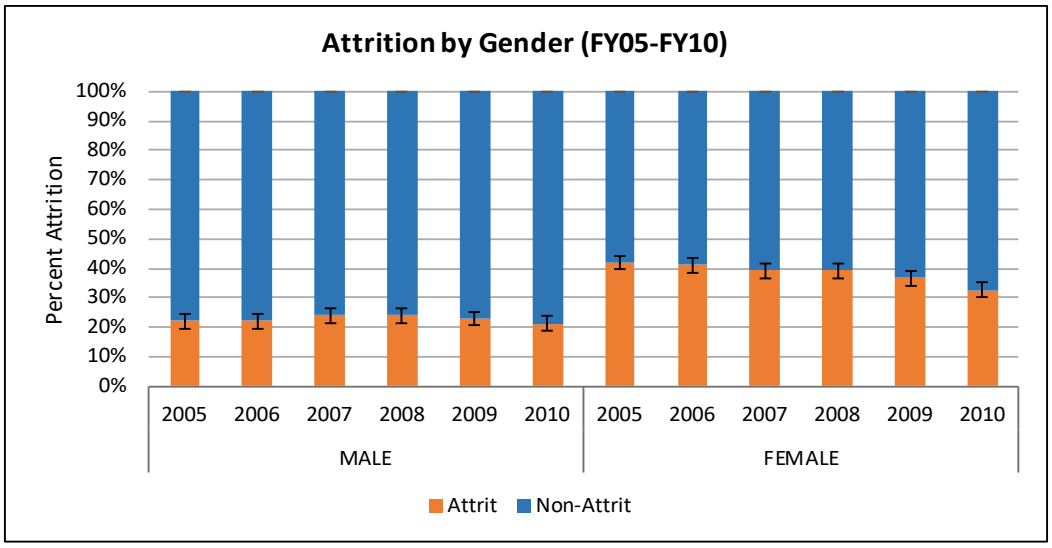

Figure 2. Attrition Rate by Gender and Accession Fiscal Year

B. SUMMARY OF NON-DEPLOYABLE SOLDIERS

For an individual soldier to be deployable to a combat environment, the soldier must be able to carry and shoot their assigned weapon, must be able to wear helmet/body armor, be able to pass a physical fitness test, and be able to operate in austere areas that regularly experience significant environmental conditions (Cox 2018).

A soldier is medically non-deployable when he or she cannot perform those duties due to a medical condition. Figure 3 shows attrition rates by gender for medically non-deployable soldiers as they arrive at their first unit. Male attrition rates range from 21.6% to 47.5% and have an average attrition rate of 32.6 %. There is also an increase or upward trend in attrition rates from FY 2005 to FY 2010. Female attrition rates range from 32.8% to 42.9% and have an average attrition rate of 36.6%. Females soldiers also show an increase or upward trend from FY 2005 to FY 2010. The results show that females who are medically non-deployable have a higher attrition rate than non-deployable male attrition rates and that attrition rates for males and females that are medically non-deployable are increasing yearly.

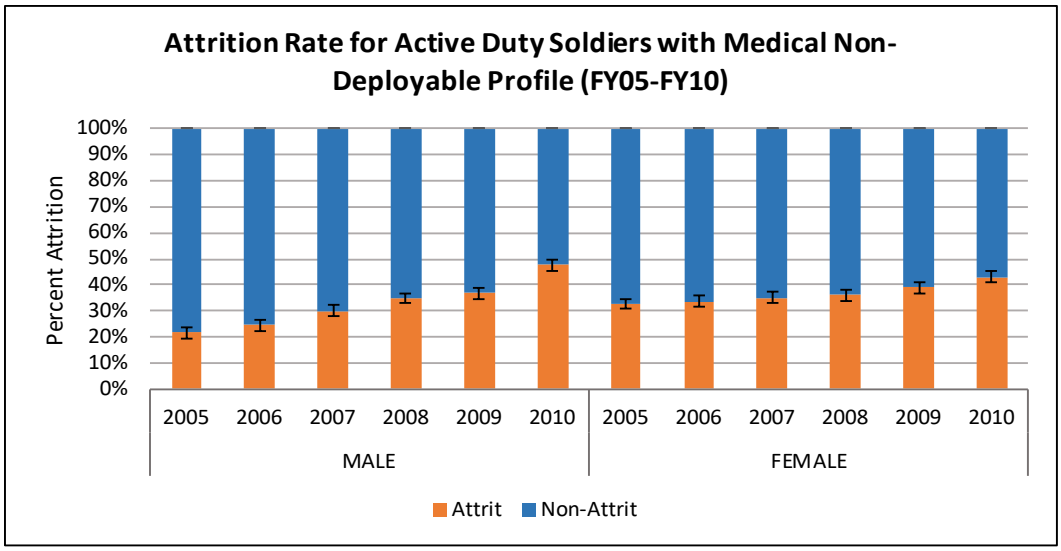

Figure 3. Attrition Rates for Medically Non-deployable Soldiers

A soldier is PULHES non-deployable as they arrive at their first unit when he or she receives a rating of three or four in any of the PULHES categories. In Figure 4, we see the attrition rates for males and females that are PULHES non-deployable. Male PULHES non-deployable attrition rates range from 45.1% to 66.8% and have an average attrition rate of 55.5% across all fiscal years. Female PULHES non-deployable attrition rates range from 64.2% to 70.4% and have an average attrition rate of 67.2%. From the results, we see that female PULHES non-deployable soldiers have a steady attrition rate across all fiscal years and remain higher than males. Finally, the results show that soldiers who are PULHES non-deployable have a higher chance of getting out of the service before their first term than medically non-deployable soldiers.

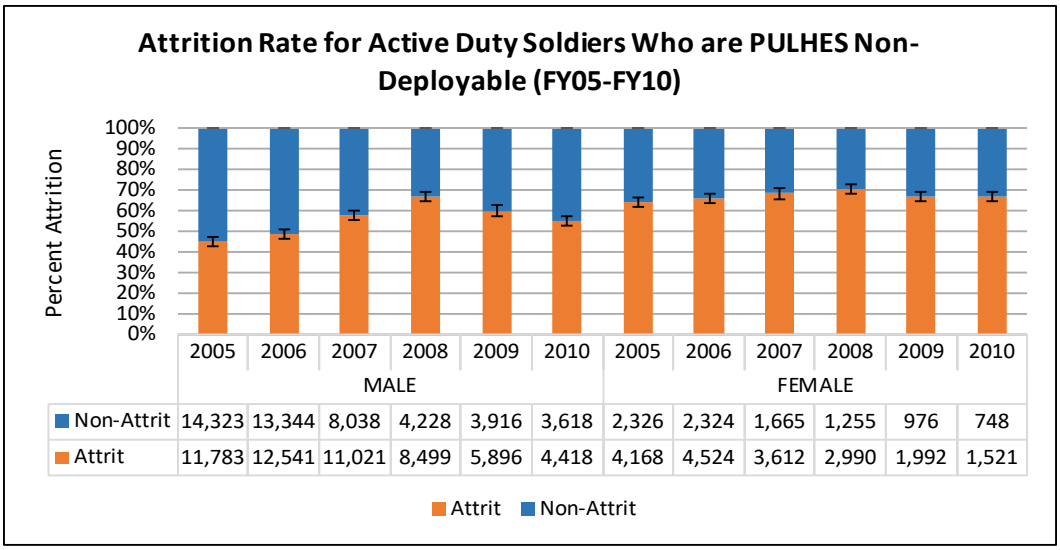

Figure 4. Attrition Rate for PULHES Non-Deployable Soldiers

C. SUMMARY OF DENTAL AND HEARING CLASS

Soldiers are non-deployable as they arrive at their first unit if their Dental or Hearing classification is three or four. In Figure 5, we can observe the attrition rate for all the dental classes by fiscal year. Dental Class One attrition rates range from 5.7% to 16.0% and with an average attrition rate of 10.7% across all fiscal years. Dental Class One has the lowest attrition rate for each fiscal year, which is what we expect to see. Dental Class Two has an average attrition rate of 19.7% and ranges from 8.2% to 34.3%. Dental Class Three attrition rates have a range of 48.6% to 61.1% with an average attrition rate of 55.4%. These results show that Dental Class Three has the largest attrition rate for every fiscal year. Soldiers who fail to schedule an annual check-up or miss their appointment will receive a classification of four until they see the dentist, which could explain the lower attrition rate for Class Four. Dental Class Four attrition rates range from 13.9% to 57.9% and have an average attrition rate of 36.0% across all fiscal years. Dental Class Four has the second-highest attrition rate and shows an increase in attrition rate every fiscal year.

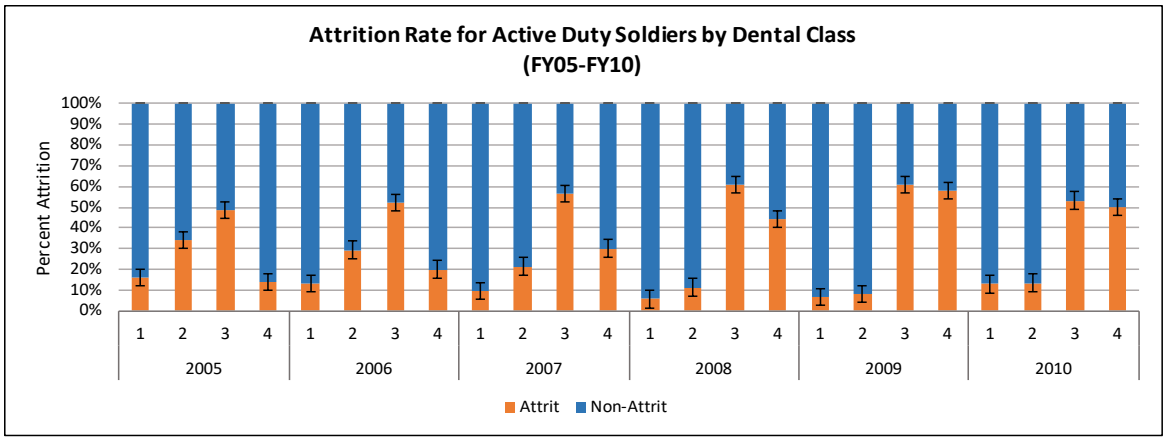

Figure 5. Attrition Rate by Dental Class

Attrition rates for Hearing Class are lower than Dental Class rates. The single highest attrition rate is 36.2%, which is in FY2007 for Hearing Class Three. The results suggest that soldiers are not attiring due to hearing. Hearing Class One attrition rates range from 24.1% to 27.4% with an average attrition rate of 26%. Class Two has an average attrition rate of 11.8 and range from 6.2% to 14.6%. Class Three has an average of 27.1% and ranges from 23.0% to 32.2%. Hearing requires an annual check-up for all soldiers and classifies soldiers with a missing examination with a hearing classification of four. Hearing Class Four has an attrition average of 18.3% across all fiscal years and a range of 15.8% to 20.2%. Hearing Class Four shows little variation over fiscal years. These results show that many soldiers who get out are either Hearing Class One or Class Three (see Figure 6).

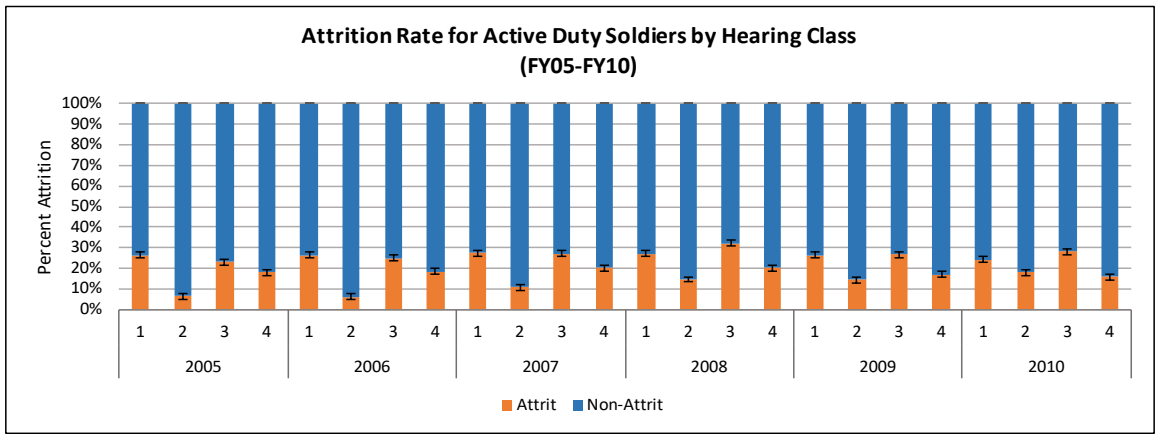

Figure 6. Attrition Rate by Hearing Class

D. OTHER SUMMARY STATISTICS

Note that all variables in this section are recorded around the time of IET completion, or arrival to their first unit. We can see from Figure 7, soldiers with rank Private (PV1) have the highest attrition rate per fiscal year. The average attrition rate for PV1 is 30.6% across all fiscal years. The second-largest is Private Second Class (PV2). PV2 average attrition rate is 25.0%. The result shows that soldiers in the ranks of Corporal (CPL), Sergeant (SGT), and Staff Sergeant (SSG) have very little chance of getting out of the service before their term is up. Very few soldiers have the rank of CPL, SGT, or SSG when they report to their first unit, which explains the minimal attrition rate for these ranks.

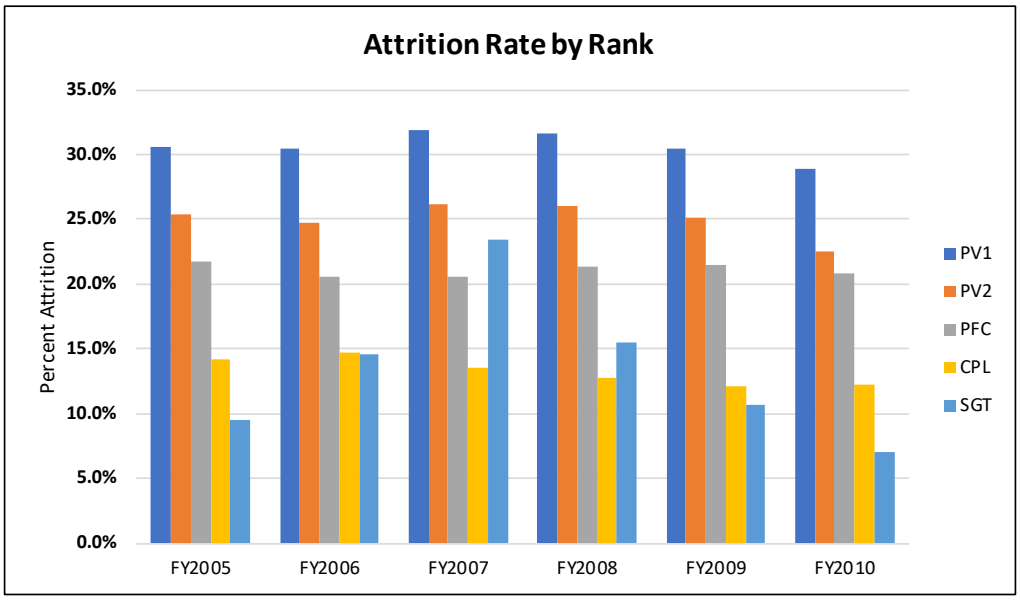

Figure 7. Attrition by Rank

Figure 8 shows that attrition rate for soldiers by their marital status at the time they arrived at their first unit. From the graph, one can see that soldiers that have never married have the highest attrition rate. Soldiers who never marry have an average attrition rate of 31.9%, while those that are married and divorce have an average of 18.2% and 20.8%, respectively. The results show that soldiers that are not married have the highest attrition rate, but do not show how getting married or divorced impacts post-IET first-term attrition.

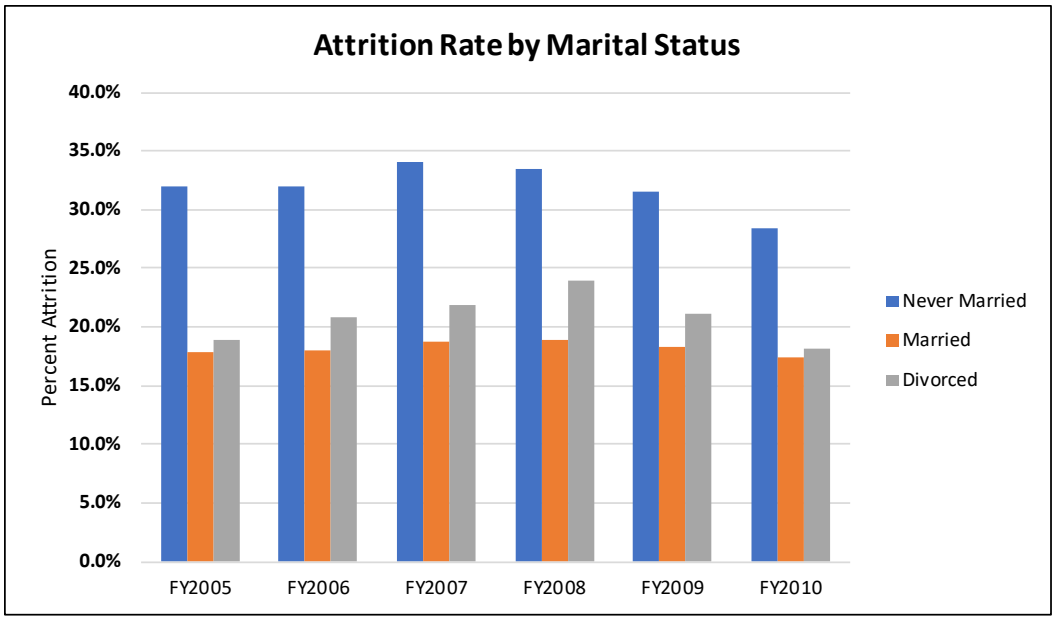

Figure 8. Attrition by Marital Status

THIS PAGE INTENTIONALLY LEFT BLANK

IV. LOGISTIC REGRESSION MODELING AND FINDINGS

In this chapter, we describe our approach to fitting a model to estimate the probability of post-IET first-term attrition and our findings.

A. MODELING APPROACHES

The response variable for this cohort is binary taking values zero or one for non-attrit or attrit respectfully. Let n be the number of observations in the cohort then the response variable is modeled as independent Bernoulli random variables Y_i where $P_i = P(Y_i = 1)$ for i = 1, …, n. Following Speten (2018), we ultimately fit a logistic regression model (Faraway 2016a) where the log-odds or logit of P_i is expressed as a linear predicitor

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_i x_{ik} \ ,$$

where k is the number of predictors, $x_{i1}, x_{i2}, \ldots x_{ik}$ are the values of the predictor variables for the i^{th} observation, and $\beta_0, \beta_1, \ldots \beta_k$ are the coefficients to be estimated using maximum likelihood.

We choose additive logistic regression as our binary regression model because it follows Speten (2018); given estimated values of the coefficients, it is easy to use; and it tends to predict well. By additive, we meant that interaction terms or transformations of numeric predictors are not included in the logistic regression model. In addition, the logistic regression model serves as a starting point for models that accommodate large numbers of observations and predictors.

Our cohort dataset is large both in numbers of observations and in numbers of predictor variables. We note that the actual number of predictors *k* in the logistic regression model with 62 predictor variables is much larger than 62. Categorical variables with l levels, require $l - 1$ binary predictors. For example, the categorical variable Dental Class with levels 1, 2, 3, 4 has three associated binary variables Dental Class 2, Dental Class 3, and Dental Class 4 which take values 1 if an observation has the specified Dental Class and zero otherwise. When all three binary wariables are zero, then the observation has Dental Class 1, often call the reference level. In short, with no many multi-level categorical

variables, the number of predictors in the logistic regression model is too large to take a traditional model-fitting or variable selection approach.

Instead, we use algorithmic models to reduce the number of predictors for the logistic regression fit. This simplifies the model and guards against over-fitting. We take two approaches. The first "regularizes" the coefficients of the logistic regression model by maximizing the likelihood (LIK) with a penalty for coefficient magnitudes. I.e., we solve

$$\max_{\beta_0,\beta_1,\ldots,\beta_k} \log(LIK) + \lambda \sum_{j=1}^{k} |\beta_j|,$$

(where the predictors have been standardized to have variance 1). We choose the lasso or L1-Norm penalty. We use the lasso penalty because it shrinks un-needed coefficients to zero (Friedman et al. 2010). The hyper-parameter λ chosen by cross-validation governs how much the likelihood is penalized.

In the second approach, we fit a random forest (Breiman et al. 2002) to estimate the probability of attrition based on all predictors. We do not use the random forest fit to predict probabilities directly because the random forest cannot be expressed simply in closed-form as can the logistic regression fit. Rather we use a measure of variable importance computed while fitting the random forest to select only the most "important" variables. We do this because the importance of a variable from a random forest fit accounts for potential interactions, and for continuous variables, it accounts for non-linearities as well. Breiman (2002) describes random forests and how variable importances are computed. The variables with the largest variable importance are then used in a logistic regression model.

The lasso-regularized logistic regression fit and the logistic regression based on the variables selected using the random forest are then compared using the test set. Thus selecting the better of the two models is not based on the statistical inference, but rather on how well they predict on an independent hold-out set and their ease of use.

B. MODEL FITTING

In this section, we describe the specifics of our model fitting approach.

1. Model Building

The model used for this thesis is fit using R Studio (Rstudio Team 2018). First, we fit a full model using all 62 variables from our training dataset. We begin by using lasso regularization to select our model. Here, we reduce our model complexity by shrinking model coefficients to zero to reduce variance, but at the same time introducing some bias into our model. Using the R package glmnet (Friedman et al. 2010), we compute cross-validated lasso-logistic regression misclassification rates for 100 values of λ. Figure 9 shows the misclassification rates plotted against log(λ). Across the top of the Figure 9 plot are the numbers of non-zero coefficiants. When the number of non-zero coefficients is low (i.e., lamda is large and the number of predictors used is small), the cross-validated misclassification rate is large indicating a poor fit. As more predictors are used, (i.e., λ decreases) the misclassification decreases. The left-most verticlal dotted grey line in Figure 9 indicates the value of λ with the smallest cross-validated misclassification rate. We use the 1-standard error (1-SE) rule λ indicated by the right-most vertical dotted line. This is the largest λ with cross-validated misclassification rate error within on SE of the "best" λ.

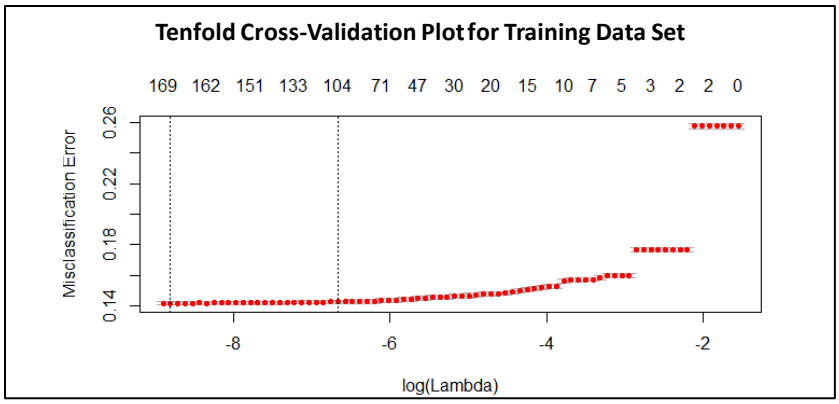

Figure 9. Cross-Validation Lasso Regularization Plot

Our second approach to building a simplified model, is to fit a random forest model using the R package ranger (Write et al. 2017) with the intent to identify variables to be used in a logistic regression model. The random forest fit gives a measure of variable importance for each variable. We then select the 20 most important variables (see Figure

11) and fit a logistic regression model using those 20 variables. In Figure 10, only one variable importance is assigned to each multi-level categorical variable regardless of the number of levels. This makes interpretation of predictors more manageable.

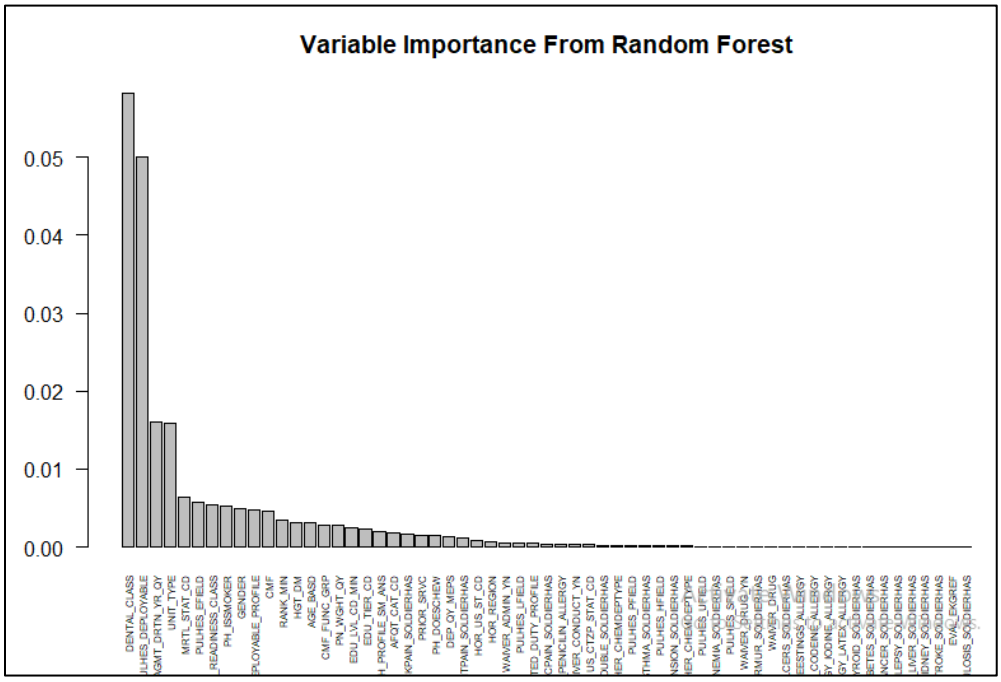

Figure 10. Variable Importance Graph for Random Forest

2. Model Selection

In this section, we compare the lasso-logistic regression with the logistic regression whose predictors are selected using the random forest. We refer to the later fit as the "logistic regression fit." To compare the two model fits, we used the validation dataset to compute validation set misclassification rates for both models. We also plot the validation set Receiver Operating Characteristic (ROC) curves to compare the models. The Area Under the Curve (AUC) from the ROC gives us a statistic compare the two models.

Table 4 gives the validation set confusion matrix for the lasso-logistic regression. An observation in the validation set is classified as attrit if its estimated probability of attrition is greater than 0.5. From Table 4, among the 6,561 attrites in the test set, 2,781 or

42.4% are classified incorrectly. Of the 19,593 non-attrites, only 977 or 5.0% are classified incorrectly. The overall validation set misclassification rate is 14.4%.

Table 4. Validation Dataset Confusion Matrix for Lasso-Logistic Regression

	Observed Non-Attrit	**Observed Attrit**
Predicted Non- Attrit	18,616	2,781
Predicted Attit	977	3,780

Table 5 gives the confusion matrix for the logistic regression fit. The validation set overall misclassification rate is 14%. From Table 5, among the 6,561 attrites in the validation set, 2,446, or 37.3% are classified incorrectly. Of the 19,593 non-attrites only 1,221 or 6.2% are classified incorrectly. Although the misclassification rates for the two models fits are about the same, the second model fit balences the misclassification rates among attrites and among non-attrites more evenly.

Table 5. Validation Dataset Confusion Matrix for Logistic Regression

	Observed Non-Attrit	**Observed Attrit**
Predicted Non- Attrit	18,372	2,446
Predicted Attit	1,221	4,115

Figure 11 shows the two ROC curves for the lasso-logistic regression and the logistic regression. Comparing the ROC curves, we see that the logistic regression is performing as well as the lasso-logistic regression. The AUC of the second logistic regression is slightly below the lasso-logistic regression. We select the logistic regression since it is more easily interpreted and constructing a more complex model (e.g., adding interactions) using its predictor variables has the greater potential for imporving the model fit.

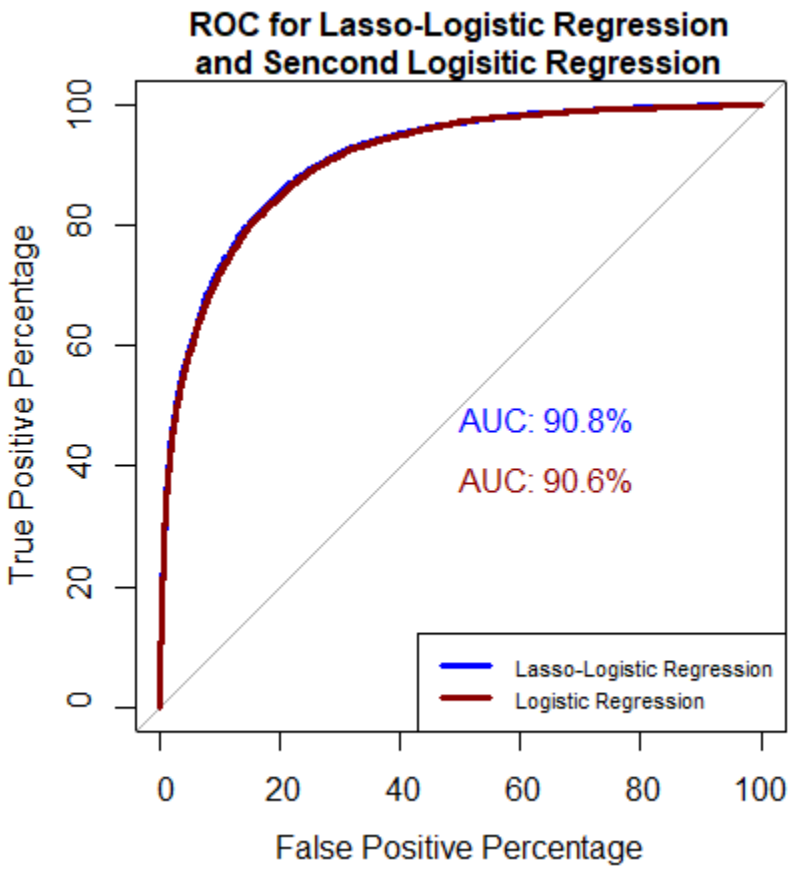

Figure 11. Testset ROC for The Lasso-Logistic Regression and Logistic Regression

3. Model Diagnostics

In choosing and assessing model performance, we rely on how well the fitted models predict on hold-out sets, either using the validion set or cross-validation. Although we do not rely on formal statistical inference nor on estimated standard errors, we do perform some (but not all) traditional diagnostics to check to see if the logistic regression fit in the previous section is reasonable.

We first check for outliers and for influential observations. We also check for multicollinearity among the predictors (Kassambara 2017). Although residuals for logistic regression models are not expected to be normally distributed, to check for outliers, we use the half normal plot of standardized residuals as shown in Faraway (2016b). We can see in Figure 12 that the half normal plot shows no evidence of any outliers. To check for

influential values, we calculate Cook's distance for each observation. All Cook's distance are less than one. Using the rule of thumb (see e.g Faraway 2016a) that infuencial observations have Cook's distance greater than one, we don't identify any observations as being as unduly infuencial.

Figure 12. Half Normal Plot for Standardized Residuals from the Logistic Regression

By using the R package car (Fox and Weisberg 2019), we check for multicollinearity (Kassambara, 2017). Using the vif() function from the car package, we compute the variance inflation factors (VIF) for each predictor variable. For VIF values, the rule of thumb is that a VIF that exceeds 5 or 10 is a sign of multicollinearity. Table 6 shows the results of the vif() function. There are no VIF values that exceed 5 or 10, which implies that there is no strong multicollinearity in our model.

Table 6. VIF Values for Logistic Regression Model

Variable Name	VIF Value
DENTAL_CLASS	1.19
PULHES_DEPLOYABLE	1.25
ASVC_AGMT_DRTN_YR_QY	2.44
UNIT_TYPE	1.35
MRTL_STAT_CD	1.20
PULHES_EFIELD	1.03
DENTAL_CLASS	1.19
PH_ISSMOKER	1.14
HEARING_READINESS_CLASS	1.06
GENDER	1.89
MEDICAL_NONDEPLOYABLE_PROFILE	1.08
CMF	3.44
RANK	2.85
HGT_DM	1.99
AGE_BASD	1.35
PN_WGHT_QY	1.57
EDU_TIER_CD	1.16
EDU_LVL_CD_MIN	2.67
OVH_PROFILE_SM_ANS	1.11

We do not check to see if numeric variables need to be transformed, or if interaction terms are needed. However, we do test the goodness of fit of the model. Using a technique described by Faraway (2016a), we construct a plot that graphs the observed proportion of attrites against the corresponding mean predicted attrition probability. First, we estimate

the linear predictor for each observation from our model, to partition the observations into 300 equally sized bins based on their estimated linear predictors. The number of attrition events is computed, and the mean of the predicted probability is calculated for each bin. To capture variability, an approximate 95% confidence interval for the expected proportion attriting in each bin is also calculated. Figure 13 shows that the second logistic regression fit does a fair job of estimating the proportion of attrits in each of the 300 bins. We can see in Figure 13 that the predicted probabilities tend to under estimate the actual proportions when the predicted probability is less than 0.50 and over estimate when the predicted probability is greater than 0.50. This suggests that a more complex model, one perhaps with interaction terms might predict probabilities with less bias than the additive logistic regression model. Using the Hosmer-Lemeshow Test, as Faraway (2016a) does in his book, we get a p-value of .99 for the null hypothesis that the fit is adequate. This indicates that there is no evidence of lack of fit.

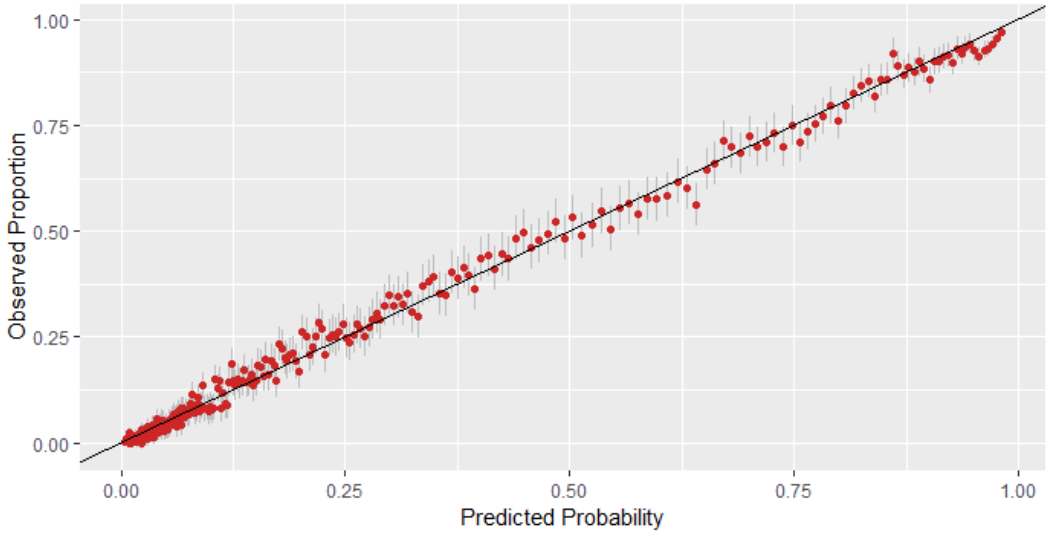

Figure 13. Plot of Binned Predicted Probability and Observed. Proportions for Logistic Regression Model. Adapted from Faraway (2016).

The Figure 13 plot is consistent with an AUC of 90.6%. The AUC can be interpreted as: select, at random, two soldiers from the population, one that attrites and one that does not; 0.906 is the probability that the attriter has the larger estimated probablilty of the two. Figure 13 and the high AUC indicate that the logistic regression model based on FY 2008–FY 2009 data estimates FY 2008–FY 2009 attrition probabilities well. The disappointing 14% misclassification rate is a consequence of fact that many of those estimated probablilities are close to 0.50.

C. FINDINGS

This section covers our analysis that was conducted in order to answer the research questions that were introduced in Chapter I.

1. Demographic and Medical Factors that Influence Active Duty Army Soldier Attrition

To find the variables that have the most significant impact on the logistic regression fit, we use the importance score appropriate for an additive logistic regression fit using R package caret (Kuhn et al. 2019). This importance score uses the absolute value z-statistic for each of the model parameters, which gives us relative importance for each predictor (Kuhn et al. 2019). The results from the variable importance scores confirm that there are demographic and medical factors that contribute to active duty soldier first-term attrition. The variables that have the most impact on first-term attrition are PULHES Non-Deployable, Dental Class Four, and Contract Duration of six years. Our analysis of the logistic regression model shows that the top ten most influential variables are PULHES Non-deployable, Dental Class, Contract Duration, Unit Type, Medical Non-deployable, Hearing Class, Gender, Is smoker, Education Tier, and Marital Status. Table 7 shows the first 25 variable importance scores; the rest are shown in Appendix A.

Table 7. Logistic Regression Variable Importance Table

Variable Name	Importance Score
PULHES_DEPLOYABLE	80.04
DENTAL_CLASS4	75.62
ASVC_AGMT_DRTN_YR_QY6	57.42
DENTAL_CLASS3	46.16
ASVC_AGMT_DRTN_YR_QY5	43.70
UNIT_TYPETDA	35.09
ASVC_AGMT_DRTN_YR_QY4	31.69
MEDICAL_NONDEPLOYABLE_PROFILE	23.30
HEARING_READINESS_CLASS4	20.35
GENDERM	18.48

Continued in Appendix A

Our analysis also includes studying the logistic regression summary output. Table 8 and its continuation in Appendix B show each variable's estimated coefficient, odds ratio, and probability. Also included as descriptive statistics in Tabel 8 are the coefficient standard errors and the p-values (Pr(|z|)) for the two-sided test that the coefficient is zero. Using Table 8 helps simplify our interpretation of the variables. The positive estimates increase the probability of first-term attrition, while the negative estimates decrease the probability of first-term attrition. For this model, the most important variable is PUHLES_DEPLOYABLE. Keeping all of the other variables fixed, a soldier who is PULHES deployable, this will have a smallerdecrease the estimated probability of attrition than one who is non-deplyable. Table 8 confirms that demographic and medical factors that influence first-term attrition are PULHES Non-deployable, Dental Class, and Contract duration.

Table 8. Logistic Regression Variable Summary

Variable Name	Estimated Coefficient	Std. Error	Odds Ratio	Probability	Pr(>|z|)
DENTAL_CLASS2	0.42	0.04	1.52	0.60	< 0.001
DENTAL_CLASS3	2.49	0.05	12.01	0.92	< 0.001
DENTAL_CLASS4	2.79	0.04	16.31	0.94	< 0.001
PULHES_DEPLOYABLE	-1.87	0.02	0.15	0.13	< 0.001
ASVC_AGMT_DRTN_YR_QY4	0.81	0.03	2.25	0.69	< 0.001
ASVC_AGMT_DRTN_YR_QY5	1.70	0.04	5.46	0.85	< 0.001
ASVC_AGMT_DRTN_YR_QY6	2.18	0.04	8.88	0.90	< 0.001
UNIT_TYPEMULTI	0.44	0.13	1.56	0.61	< 0.001
UNIT_TYPETDA	0.97	0.03	2.64	0.73	< 0.001
MRTL_STAT_CDM	0.03	0.06	1.03	0.51	>.05
MRTL_STAT_CDN	0.70	0.06	2.01	0.67	< 0.001
MRTL_STAT_CDOther	0.57	0.31	1.76	0.64	0.069
PULHES_EFIELD3	-1.66	0.54	0.19	0.16	0.002

Continued on Appendix B

2. Comparing Second Logistic Regression Model to Speten (2018) Logistic Regression Model

The goal for this thesis is to develop a logistic regression model that can be used to predict or forecast active duty soldier post-IET first-term attrition using both medical and demographic data. To do this, we compare both our new logistic regression model and Speten (2018) model using FY 2010 dataset to compute FY 2010 dataset misclassification rates for both models. We also plot the FY 2010 dataset ROC curves to compare the models. AUC from the ROC is also used to compare the two models. Speten (2018) model uses variables *Number of Days Deployed*, *Max Time in Grade*, *Unit Type Max*, and *Max Rank* that are not used in the new logistic regression model. They are not used in the new

regression model because these variables are not available when a soldier reports to their first unit after IET. An example of this, *Number of Days Deployed*, is not available to soldiers reporting to units after IET who have not yet deployed. These types of variables are strongly related to attrition, because, for example, soldiers who complete their first term have more opportunity to deploy that those who attrite. However, without accounting for time until attrition or first term completion it is difficult to ascertian how much of the relationship is a consequence of attrition and how much is a predictor or attrition.

Table 9 gives the FY 2010 dataset confusion matrix for Speten (2018). An observation in the FY 2010 dataset is classified as attrit if its estimated probability of attrition is greater than 0.5. From Table 9, among the 15,335 attrites in the FY 2010 dataset, 8070 or 52.6% are classified incorrectly. Of the 51,471 non-attrites, only 4,735 or 9.2% are classified incorrectly. The overall FY 2010 dataset misclassification rate for Speten (2018) model is 19.7%.

Table 9. FY 2010 Dataset Confusion Matrix for Speten (2018) Model

	Observed Non-Attrit	**Observed Attrit**
Predicted Non- Attrit	46,739	8,070
Predicted Attit	4,735	7,265

Table 10 shows the confusion matrix for the new new logistic regression fit. The FY 2010 dataset overall misclassification rate is 16.3% which is lower than the misclassification rate for Speten (2018) model. From Table 10, among the 15,335 attrites in the test set, 9,304, or 60.7% are classified incorrectly. Of the 51,471 non-attrites only 1,597 or 3.1% are classified incorrectly. From the misclassifications rates, we can see that the new logistic regression model is preforming better .

Table 10. FY 2010 Dataset Confusion Matrix for New Logistic Regression Model

	Observed Non-Attrit	**Observed Attrit**
Predicted Non- Attrit	49,874	9,304
Predicted Attit	1,597	6,031

Comparing the ROC curves, we see that the new logistic regression model is performing as well as Speten (2018) moodel. The AUC of the new logistic regression is slightly above Speten (2018) model. The new logistic regression model performs better by 1.3% (see Figure 14). This would imply that with the addition of the medical variables, new logistic regression model is at least as good as the Speten (2018) model even though our new model only includes variables whose values are available immediately post-IET.

Both FY 2010 AUCs are lower than the FY 2008–FY 2009 validation set AUCs and similarly the FY 2010 miscalssifiation rates are higher than those of the validateion set. Further, the decrease in AUC to 82.6% means that the model fit may be used to identify groups of soldiers who have a greater chance of attrition, but should not be used to predict whether an individual soldier will attrit or not.

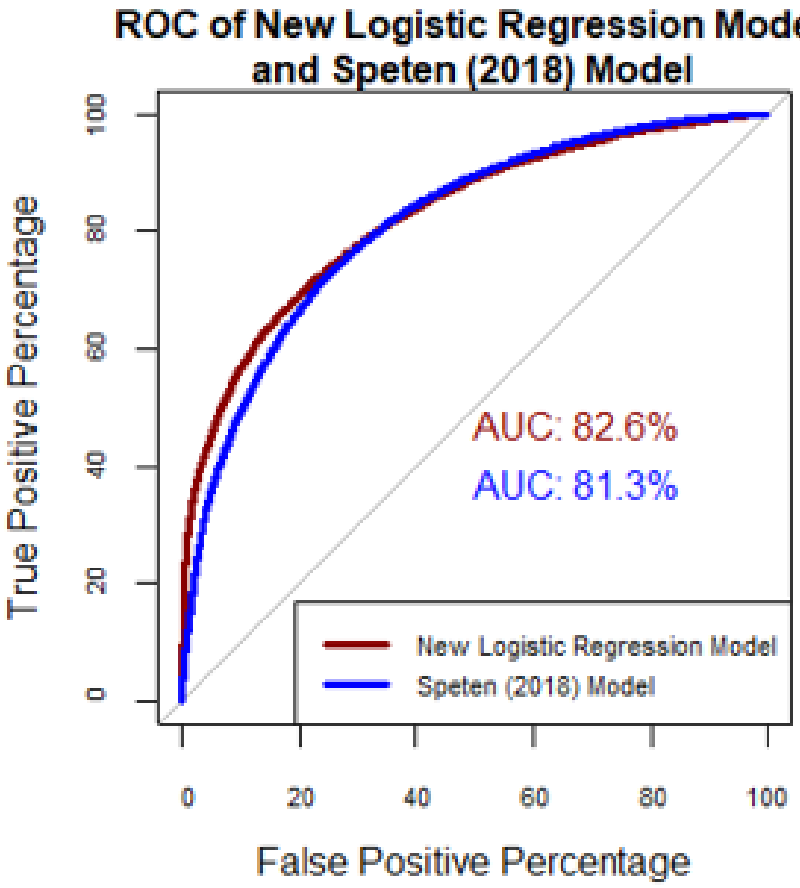

Figure 14. Comparison of New Logistic Regression Model to Speten (2018) Model

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY

A. CONCLUSIONS

Speten (2018) has challenges with using separation codes but is able to establish a methodology for identifying soldiers who completed their first contract term among soldiers who complete IET (Speten, 2018). Our research incorporates Speten's (2018) work and adds medical variables to the demographic data with the intent to find the combination of medical and demographic factors that contribute to post-IET first-term attrition.

1. Data Preparation

Using the cohort data that contains soldiers who entered basic training during FY 2005 to FY 2010, our research leverages the medical data available in the PDE and constructs a new cohort dataset to study post-IET first-term attrition with regards to medical and demographic variables.

The medical data creates some challenges due to missing medical records during the early fiscal years. This research uses data for soldiers assessing in FY 2008 and FY 2009 data to construct a training and validation dataset to train and validate a logistic regression model that can be used to predict post-IET first-term attrition probabilities using medical and demographic data from the PDE. The FY 2010 data is reserved for final model testing.

2. Analysis of Logistic Regression Model

From the analysis of the logistic regression model, we conclude that the medical and demographic factors that most contribute to soldier first-term attrition are PULHES Non-deployable, Dental Class, Contract Duration, Unit Type, Medical Non-deployable, Hearing Class, Gender, Is smoker, Education Tier, and Marital Status.

The analysis also shows that the logistic regression model fit using FY 2008–FY 2009 data research can predict active duty soldier post-IET first-term attrition with an accuracy of 86% on the independent FY 2008–FY 2009 valudation set and that the accuracy decreases to 83.7% for the FY 2010 dataset. The FY 2010 ROC curves give an

AUC of 82.6%. These results are similar to those of Speten (2018), however unlike Speten's (2018) model, our logistic regression model includes medical variables and only uses variables that can be used to predict attrition.

B. RECOMMENDATIONS

The logistic regression model fit during this research has an accuracy of 86%. A recommendation to Army Resiliency Directorate (ARD) is to implement this logistic regression model into analytical tools that can be created within the PDE to help identify groups of active duty soldiers who are at risk of leaving before their contract obligation is complete. From this research, ARD can use the information to build analytical tools within the PDE. The impact of this tool can aid Army leadership with active duty soldier retention, as well as recruiting soldiers that will have a low probability of attrition. ARD can improve soldier resiliency as well as by identifying soldiers who are at risk of failure. Then through interventions, help soldiers finish their contractual obligation or provide a path to reenlistment.

APPENDIX A. VARIABLE IMPORTANCE SCORES

Table 11. Variable Importance Table

Variable Name	Importance Score
PULHES_DEPLOYABLE	80.04
DENTAL_CLASS4	75.62
ASVC_AGMT_DRTN_YR_QY6	57.42
DENTAL_CLASS3	46.16
ASVC_AGMT_DRTN_YR_QY5	43.70
UNIT_TYPETDA	35.09
ASVC_AGMT_DRTN_YR_QY4	31.69
MEDICAL_NONDEPLOYABLE_PROFILE	23.30
HEARING_READINESS_CLASS4	20.35
GENDERM	18.48
PH_ISSMOKER	15.00
EDU_TIER_CD2	14.59
MRTL_STAT_CDN	12.45
DENTAL_CLASS2	11.11
CMF68	10.81
PN_WGHT_QY	10.78
EDU_TIER_CD3	8.91
CMF15	8.74
RANK_MINPV1	8.54

Variable Name	Importance Score
CMF35	8.52
PULHES_EFIELD2	8.18
CMF42	7.54
AFQT_CAT_CD3B	7.13
CMF91	6.93
RANK_MINPV2	6.71
AFQT_CAT_CD3A	6.71
CMF18	5.77
RANK_MINSSG	5.16
AFQT_CAT_CD4C	4.43
EDU_LVL_CD_MINCLG	4.23
RANK_MINPFC	4.07
HEARING_READINESS_CLASS3	4.05
HGT_DM	3.98
AFQT_CAT_CD2	3.97
EDU_LVL_CD_MINHS	3.95
CMF63	3.82
CMF92	3.79
CMF25	3.70
UNIT_TYPEMULTI	3.39
CMF35	8.52
AFQT_CAT_CD4A	3.27

Variable Name	Importance Score
AFQT_CAT_CD4A	3.27
CMF37	3.19
PULHES_EFIELD3	3.10
CMF31	3.02
RANK_MINSGT	2.73
CMF56	2.51
CMF27	2.26
CMF14	2.22
CMF46	2.22
EDU_LVL_CD_MINGRAD	2.01
CMF89	2.00
CMF74	2.00
MRTL_STAT_CDOther	1.82
CMF13	1.81
OVH_PROFILE_SM_ANS	1.71
HEARING_READINESS_CLASS2	1.68
CMF94	1.27
AFQT_CAT_CD5	1.25
CMF12	1.06
PULHES_EFIELD4	1.03
CMF19	1.00
CMF88	0.83

Variable Name	Importance Score
AFQT_CAT_CD4B	0.81
MRTL_STAT_CDM	0.60
AGE_BASD	0.55
CMF36	0.17
CMF79	0.11
CMF38	0.07
CMF29	0.07
CMF51	0.04

APPENDIX B. LOGISTIC REGRESSION VARIABLE SUMMARY

Table 12. Continuation of Logistic Regression Variable Summary

Variable Name	Estimated Coefficient	Std. Error	Odds Ratio	Probability	Pr(>|z|)
HEARING_READINESS_CLASS4	-0.62	0.03	0.54	0.35	< 0.001
GENDERM	-0.65	0.04	0.52	0.34	< 0.001
MEDICAL_NONDEPLOYABLE	0.67	0.03	1.95	0.66	< 0.001
CMF12	-0.05	0.04	0.95	0.49	>.05
CMF13	-0.08	0.05	0.92	0.48	>.05
CMF14	-0.17	0.08	0.84	0.46	0.026
CMF15	-0.50	0.06	0.61	0.38	< 0.001
CMF18	-3.40	0.59	0.03	0.03	< 0.001
CMF19	-0.05	0.05	0.95	0.49	>.05
CMF25	-0.15	0.04	0.86	0.46	< 0.001
CMF27	-0.34	0.15	0.71	0.42	0.024
CMF29	-12.30	174.50	0.00	0.00	>.05
CMF31	-0.16	0.05	0.85	0.46	0.003
CMF35	-0.48	0.06	0.62	0.38	< 0.001
CMF36	-11.48	68.89	0.00	0.00	>.05
CMF37	-0.94	0.29	0.39	0.28	< 0.001
CMF38	-12.62	177.70	0.00	0.00	>.05
CMF42	-0.51	0.07	0.60	0.38	< 0.001
CMF46	-0.55	0.25	0.57	0.37	0.027

Variable Name	Estimated Coefficient	Std. Error	Log-Odds	Probability	Pr(>|z|)
CMF56	-0.46	0.18	0.63	0.39	0.012
CMF63	0.31	0.08	1.37	0.58	< 0.001
CMF68	-0.51	0.05	0.60	0.38	< 0.001
CMF74	-0.18	0.09	0.83	0.45	0.045
CMF79	-12.53	118.50	0.00	0.00	>.05
CMF88	-0.04	0.05	0.96	0.49	>.05
CMF89	-0.20	0.10	0.82	0.45	0.045
CMF91	-0.31	0.04	0.73	0.42	< 0.001
CMF92	-0.15	0.04	0.86	0.46	< 0.001
CMF94	-0.10	0.08	0.91	0.48	>.05
RANK_MINPFC	0.28	0.07	1.33	0.57	< 0.001
RANK_MINPV1	0.58	0.07	1.79	0.64	< 0.001
RANK_MINPV2	0.46	0.07	1.58	0.61	< 0.001
RANK_MINSGT	0.42	0.15	1.52	0.60	0.006
RANK_MINSSG	1.25	0.24	3.49	0.78	< 0.001
HGT_DM	-0.02	0.00	0.98	0.50	< 0.001
AGE_BASD	0.00	0.00	1.00	0.50	>.05
PN_WGHT_QY	0.00	0.00	1.00	0.50	< 0.001
EDU_TIER_CD2	0.37	0.03	1.44	0.59	< 0.001
EDU_TIER_CD3	0.71	0.08	2.03	0.67	< 0.001
EDU_LVL_CD_MINCLG	0.41	0.10	1.50	0.60	< 0.001
EDU_LVL_CD_MINGRAD	0.38	0.19	1.46	0.59	0.044

Variable Name	Estimated Coefficient	Std. Error	Log-Odds	Probability	Pr(>|z|)
OVH_PROFILE_SM_ANS	-0.06	0.04	0.94	0.48	>.05
AFQT_CAT_CD2	0.20	0.05	1.23	0.55	< 0.001
AFQT_CAT_CD3A	0.36	0.05	1.43	0.59	< 0.001
AFQT_CAT_CD3B	0.38	0.05	1.47	0.59	< 0.001
AFQT_CAT_CD4A	0.26	0.08	1.30	0.57	< 0.001
AFQT_CAT_CD4B	0.34	0.42	1.40	0.58	>.05
AFQT_CAT_CD4C	1.80	0.41	6.07	0.86	< 0.001
AFQT_CAT_CD5	1.52	1.21	4.57	0.82	>.05

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Army-Portal.com (2011) Physical Profile Serial System (PULHES). Accessed March 28, 2019, http://www.army-portal.com/jobs/pulhes.html

Breiman L, Cutler, A, Liaw A, Wiener M (2002) *Breiman and Cutler's Random Forests for Classification and Regression*, Version 4.6-14. Accessed May 5, 2019, https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Buddin R (1981) The role of service experience in post-training attrition in the Army and Air Force. Report R-2682-MRAL, RAND Corporation, Santa Monica, CA, https://www.rand.org/pubs/reports/R2682.html

Bushatz A (2018) Army secretary wants to boost active-duty end strength above 500,000. Accessed January 9, 2019, https://www.military.com/dodbuzz/2018/08/09/army-secretary-wants-boost-active-duty-end-strength-above-500000.html

Cox, M (2018) The Army releases deploy-or-out rules for administratively sidelined troops. Accessed March 25, 2019, https://www.military.com/daily-news/2018/11/13/army-releases-deploy-or-out-rules-administratively-sidelined-troops.html

Devig, A (2019) Predicting U.S. Army enlisted attrition after initial entry training using survival analysis. Master's thesis, Naval Postgraduate School, Monterey, CA.

Faraway, J (2016a) *Extending the Linear Model with R,* 2nd ed. (Taylor & Francis Group, Boca Raton, FL).

Faraway, J (2016b) *R Package: Functions and Datasets for Books by Julian Faraway,* Version 1.0.7, Access May 23, 2019, https://CRAN.R-project.org/package=faraway

Friedman J, Hastie T, Tibshirani R, Simon N, Balasubramanian N, Junyang Q (2010) Regularization paths for generalized linear models via coordinate descent., Version 2.0-18. Accessed May 5, 2019, https://www.rdocumentation.org/packages/glmnet/versions/2.0-18

Fox, J, Weisberg S (2019) *An R Companion to Applied Regression*, Third Edition., Version 3.0-5. Sage Publications, Washington, DC. Accessed May 23, 2019, https://www.rdocumentation.org/packages/car/versions/3.0-5

Government Accountability Office (1997) Military attrition: DoD could save millions by better screening enlisted personnel. Report GAO-97-39, Washington, DC, https://www.gao.gov/assets/160/155698.pdf

Government Accountability Office (1998) Military attrition: DoD needs to better analyze reasons for separation and improve recruiting systems. Report GAO/T-NSIAD-98-117, Washington, DC, https://www.gao.gov/assets/110/107274.pdf

Jensen D (2016) Supplemental Information: Person-Event Data Environment. Unpublished technical report.

Kassambara A (2017) Logistic Regression Assumptions and Diagnostics in R , Accessed May 23, 2019, http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

Knapik J, Jones B, Hauret S, Piskator E (2004) Review of the literature on attrition from the military services and strategies to reduce attrition. Technical Report 12-HF-01Q9A-04, U.S. Army Center for Health Promotion and Preventive Medicine, https://apps.dtic.mil/dtic/tr/fulltext/u2/a427744.pdf

Kuhn, M, Wing J, Steve Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, the R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C, Hunt T (2019) *Classification and Regression Training,* Version 6.0-84, Accessed May 23, 2019, https://CRAN.R-project.org/package=caret

Myers M (2018) The Army is supposed to be growing, but this year, it didn't at all. *Army Times.* Accessed February 1, 2019, https://www.armytimes.com/news/your-army/2018/09/21/the-army-is-supposed-to-be-growing-but-this-year-it-didnt-at-all

Mayer, M (2019) missRanger: Fast imputation of missing values, version 1.0.5. Accessed May 5, 2019, https://CRAN.R-project.org/package=missRanger

RStudio Team (2018) RStudio: Integrated development for R. RStudio. Accessed February 5, 2019, http://www.rstudio.com/.

Speten , K (2018) Predicting U.S. Army first-term attrition after initial entry training. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/59593

Wright M, Ziegler A (2017) A fast implementation of random forests. Version 0.11.2, Accessed May 5, 2019, https://cran.r-project.org/web/packages/ranger/ranger.pdf

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California