NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PREDICTING ARMY POST-IET ATTRITION USING LOGISTIC REGRESSION AND TIME-VARYING COVARIATES

by

Josephine H. Cammack

June 2020

Thesis Advisor: Hong Zhou
Co-Advisor: Samuel E. Buttrey

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2020	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE PREDICTING ARMY POST-IET ATTRITION USING LOGISTIC REGRESSION AND TIME-VARYING COVARIATES	**5. FUNDING NUMBERS**
6. AUTHOR(S) Josephine H. Cammack	
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 Army is trying to reach a force of 500,000 by 2030. Within the next 10 years, the Army needs to play a balancing act of figuring out how many soldiers will retire, attrit, or not reenlist, and how many will leave for medical or other various reasons. Then the Army needs to figure out how many soldiers need to be recruited every year to reach the 500,000 goal. Because of factors such as lower recruiting goals, tightening labor markets, reduced incentives due to a tighter defense budget, and increasing obesity levels, it is getting harder to recruit prospective soldiers. In such an environment, military leaders need to know why soldiers attrit before their first term is complete, and the factors that contribute to this decision. This thesis uses multiple logistic regressions to determine if a soldier will attrit using personnel data from the Person-Event Data Environment database. We discovered that soldiers who attrit have more variables in common by year in contract than by their contract duration. Thus the models are by year in contract due to the changing nature of time-varying covariates. As the year in contract increases, the effects of demographic indicators generally decrease and the effects of medical-related indicators largely increase. This model can help Army G1 predict how many people will be in the military at a given time—knowledge that will also help leaders determine how to prevent attrition and increase the likelihood of success for soldiers.

14. SUBJECT TERMS Army, attrition, attrit, post-IET, logistic regression, time-varying covariates, PED	**15. NUMBER OF PAGES** 93
	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 ARMY POST-IET ATTRITION USING LOGISTIC REGRESSION AND TIME-VARYING COVARIATES

Josephine H. Cammack
Captain, United States Army
BS, U.S. Military Academy, 2011
MA, University of Louisville, 2016

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN APPLIED MATHEMATICS

from the

NAVAL POSTGRADUATE SCHOOL
June 2020

Approved by: Hong Zhou
Advisor

Samuel E. Buttrey
Co-Advisor

Wei Kang
Chair, Department of Applied Mathematics

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Army is trying to reach a force of 500,000 by 2030. Within the next 10 years, the Army needs to play a balancing act of figuring out how many soldiers will retire, attrit, or not reenlist, and how many will leave for medical or other various reasons. Then the Army needs to figure out how many soldiers need to be recruited every year to reach the 500,000 goal. Because of factors such as lower recruiting goals, tightening labor markets, reduced incentives due to a tighter defense budget, and increasing obesity levels, it is getting harder to recruit prospective soldiers. In such an environment, military leaders need to know why soldiers attrit before their first term is complete, and the factors that contribute to this decision. This thesis uses multiple logistic regressions to determine if a soldier will attrit using personnel data from the Person-Event Data Environment database. We discovered that soldiers who attrit have more variables in common by year in contract than by their contract duration. Thus the models are by year in contract due to the changing nature of time-varying covariates. As the year in contract increases, the effects of demographic indicators generally decrease and the effects of medical-related indicators largely increase. This model can help Army G1 predict how many people will be in the military at a given time—knowledge that will also help leaders determine how to prevent attrition and increase the likelihood of success for soldiers.

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

AFMS	Active Federal Military Service
AFQT	Armed Forces Qualification Test
AIT	Advanced Individual Training
ASVAB	Armed Services Vocational Aptitude Battery
CMF	Career Management Field
DA	Department of the Army
DACES	Department of the Army Career Engagement Survey
FS	Force Sustainment
FY	Fiscal Year
GAO	Government Accountability Office
GT	General Technical
HPDT	High Physical Demands Test
IET	Initial Entry Training
LD	Low Density
MACP	Married Army Couples Program
MEDPROS	Medical Protection System
MFE	Maneuver, Fires, and Effects
MOS	Military Occupational Specialties
NCO	Non-commissioned Officer
OPAT	Occupation Physical Assessment Test
OR	Odds Ratio
OS	Operations Support
PDE	Person-Event Data Environment
PHA	Periodic Health Assessment
PID	Person Identifier
PULHES	Physical/Upper/Lower/Hearing/Eyes/Stability-Psychiatric
RAND	Research and Development
TAPAS	Tailored Adaptive Personality Assessment System
TMS	Trunk Muscle Strength
USAPHC	U.S. Army Public Health Command

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank Dr. Hong Zhou and Dr. Sam Buttrey for assisting me in this thesis. This is an interesting topic, and I am grateful for the opportunity to dive deeper into this subject matter. Your professional experience, knowledge, and contributions have tremendously helped me through this. I would also like to thank my family and friends for their encouragement, prayers, and words of wisdom.

Thank you, Andrew, for being my constant companion and biggest cheerleader throughout this time. The last three months have been even more challenging through COVID-19, but we were somehow able to get everything done. Your steadfast support and confidence guided me through the past two years at NPS. Finally, Lillian, thank you for providing me joy, smiles, and laughs throughout the days. Your dad and I will always remember our time at NPS, since I gave birth to you here, and we watched you grow into the beautiful toddler you are today.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This research examines the use of time-varying covariates while using logistic regression to see if it can predict whether Army soldiers will undergo attrition before their initial contract term has ended. A previous thesis experimented using the first-known data in a soldier's record as the input to the logistic regression model (Gobea 2019). This thesis uses multiple logistic regressions that are fitted to each year in contract. Using this approach tackles the problem of how to use time-varying covariates that change year to year, and not just what those inputs are at the beginning of a soldier's career post-IET (initial entry training).

A. BACKGROUND

One of the three lines of effort in the 2018 National Defense Strategy is to build a more lethal force by restoring warfighting readiness and fielding a more lethal force. Former Secretary of Defense James Mattis has stated that the size of the U.S. military force matters, and that the size and skill can be determined by recruiting, developing, and retaining high-quality military and civilian efforts (Mattis 2018, p. 7).

On the recruiting front, 2020 recruiting goals are higher than in the previous two years, and the population of viable recruits is getting smaller (Arkin 2019). The Army did not meet its 2018 initial goal of 76,500 recruits; the year ended with roughly 70,000 recruits (Rempfer 2019). Recruiting goals were lowered after 2018 to be more realistic. The Army met its 2019 fiscal year goal of signing up more than 68,000 active duty soldiers (Rempfer 2019). The goal for 2020 is "north of 68,000," said Major General Frank Muth, head of Army Recruiting Command; this increase in recruiting is intended to achieve the Army's goal of a 500,000-strong active duty force by 2030 (Rempfer 2019). That goal may be hard to achieve during times of improving economic conditions, a tightening labor market, reduced incentives due to a tighter defense budget, and increasing obesity levels among young people since it is harder to recruit prospective soldiers (Vanden Brook 2015; Tice 2016).

Since recruiting numbers have dropped over the past few years, retaining trained soldiers is key not only to reaching a force of 500,000 by 2030 but also to reducing basic combat training (BCT) and advanced individual training (AIT) costs. Army Secretary Mark Esper has stated that he will not use increased enlistment waivers for factors like misconduct or aptitude to reach the 500,000 (Myers 2019). There is also a cost aspect to retention. Back in 2007, Maj. Gen. Thomas Bostick, who was head of the Army's Recruiting Command, said it costs "about $18,000 to bring a soldier to basic training" (Burgess 2007). If one adds in the amount it costs for uniforms, gear, food, classroom learning, pay, moving expenses, support staff pay, combat and technical training, one could be seeing a value of over $44,000 (Olick 2002).

In addition to the monetary amount it takes to train a soldier, time is also an asset that is exhausted as soldiers undergo attrition before their first term of enlistment is complete. Developing soldiers with more specialized skills, such as linguists or those with one of the many medical specialties, can require specific equipment or more classroom instruction time. In order to reduce the amount of time the military spends on recruitment and to support the line of effort to build a more competent lethal force, the Army needs to lower attrition and retain its skilled and trained troops.

Retaining existing members of the military ends up saving money. In 1997, GAO estimated that if services reduced attrition in the first six months of service by 4%, they would see an immediate savings of $4.8 million (GAO 1997). In 2020 dollars, adjusted for inflation, that is approximately $7.7 million. A 10% reduction of attrition would save close to $19.3 million in today's dollars. This includes the cost of transporting, feeding, clothing, housing, medical screening, and training for the first six months. The current post-initial entry training (IET) attrition rate is approximately 24.5% (Devig 2019). Since the Army is trying to increase the size of its force, and recruiting is becoming more difficult, one key to saving time and money in the defense budget is to retain current soldiers.

B. THESIS ORGANIZATION

Chapter I outlines the purpose of this thesis, describes previous works relating to attrition and logistic regression, and introduces that research questions we focus on answering. Chapter II explains the datasets used and the methodology of logistic regression. Afterwards, Chapter III describes the demographic and medical variables in depth. Chapter IV includes the modeling approach, how the model was fit, and an analysis of our findings. Chapter V concludes with a summary of findings and recommendations. In the thesis we use the verb "attrit" as shorthand for "to undergo attrition; to leave the service."

C. PURPOSE OF RESEARCH AND OBJECTIVES

The objective of this thesis is to determine whether a soldier will complete his or her first term, based on multiple characteristics of the soldier, using logistic regression. Knowing this can help senior leaders and retention non-commissioned officers (NCO) identify those soldiers who are most at risk for attriting. Then they will be able to answer questions such as:

- What are the contributing factors for someone deciding to attrit?
- How can a leader or soldier set the conditions to help/persuade someone who is at risk for attrition to stay?
- What policies can be implemented or changed?

On the recruiting front, it is beneficial to know which candidates are best to recruit and will pass basic combat training (BCT) and advanced individual training (AIT). Combined, BCT and AIT are called IET. In addition to knowing which potential recruits can pass IET, recruiters can take it one step further and focus on recruiting those who will pass IET and stay for their whole first term of enlistment. In addition, the Army needs to be able to predict how many soldiers will complete their enlistment term, so it can determine projected force composition, which entails how many soldiers will attrit, retire, get promoted, or leave the service for other reasons. The Army needs to know the numbers of soldiers in specific pay grades and military occupational specialties (MOS) in order to

know how much money it needs from Congress for Soldier pay, benefits, housing, and recruitment.

Another objective of this thesis is to find a more refined way of performing logistic regression utilizing time-varying covariates. The applications of this method of logistic regression are of interest to those wishing to predict binary outcomes whose data has time-dependent variables or whose variables change over multiple time periods.

D. RESEARCH QUESTIONS

Years of research: contracts that started in FY2009–FY2011

- Does the approach of producing multiple one-year logistic regression models with time-varying covariates produce a better output for predicting attrition than just a single logistic regression model?
- Which variables are important in predicting whether a soldier will attrit or not before the end of their enlistment term? Are these variables different for cohorts with different contract lengths?
- Is the one-year attrition behavior of soldiers more similar across different contract lengths or their year in contract?

E. RELATED WORKS

This section reviews how previous research has used logistic regression to analyze attrition. It also describes what variables have been deemed important to previous attrition studies. The second part of this section describes three previous theses that study post-IET attrition and how this thesis expands and differs from previous research.

1. Uses of Logistic Regression

Logistic regression has been used to compare the strength of one predictor to that of another. One such case was done by researchers in 2015 to predict injury attrition in "Trunk muscle strength test to predict injuries, attrition and military ability in soldiers" (Wunderlin et al. 2015). All members of military organizations that are part of the North

Atlantic Treat Organizations are tested on physical ability. One test is the sit-up, and an alternative is the global trunk muscle strength test (TMS). The researchers tested 230 male recruits from the Swiss Army for 13 weeks to investigate whether the TMS is a reasonable alternative to the sit-up test, and to see how those two tests compare in predicting injuries, attrition, and military ability (Wunderlin et al. 2015). The researchers used backward selection and receiver operations characteristic (ROC) curve analysis to compare the power of the sit-up test and TMS to predict attrition, injuries, and military ability. Their variables included 35 categories of anatomical site such as head, upper extremities, trunk, and knee; 18 categories of injury such as fracture, inflammation, and pain; and lastly three injury severity categories of low, moderate, or severe. Attrition was categorized as medical, psychological, and administrative.

A relative odds ratio (OR) was then calculated for the remaining variables. This number allowed the researchers to calculate the impact of each variable in injury risk regardless of its units. They also analyzed the effect size of each variable where 0.02 was a small effect, 0.15 a moderate effect, and 0.35 a high effect. They found out that when the trunk performance test served as the independent variable and injury data as the dependent variable, the analysis revealed a significant discriminative power to predict acute and total injuries for TMS, but not for the sit-up test. Backwards elimination left the following variables for injury risk factors: age, low BMI, low TMS performance, and cigarette smoking (Wunderlin et al. 2015). For military attrition, 42.3% cases were related to medical reasons, 42.3% for psychological reasons, and 15.4% for other reasons. They concluded TMS was the stronger predictor for injuries than the sit-up test, since sit-up performance was excluded as a variable when they did backwards elimination for variable selection for the binary logistic regression model.

Another use of logistic regression is to predict whether someone will attrit or not from a program. Eskreis-Winkler et al. (2014) studied the extent to which what they called "grit," "the tendency to sustain passion and perseverance for long-term goals" (Eskreis-Winkler et al. 2014, p. 1), predicted whether a recruit was going to complete the 24-day Army Special Operations Forces (ARSOF) selection course. About half of the participants who complete the preparatory course do not complete ARSOF. The 824 participants of the

study were from four consecutive cohorts from 2008 and 2009. The researchers excluded approximately 100 participants for medical reasons such as an injury, or for incomplete data. The final population under study was 677 male candidates. They were evaluated on their general intelligence from the Armed Services Vocational Aptitude Battery- General Technical (ASVAB-GT) score, physical fitness from the Army Physical Fitness Test, years of schooling, and grit from an eight-question grit scale questionnaire.

The researchers examined the bivariate relationship between their predictor variables and retention and did a full logistic regression model predicting retention from grit. Just as with the truck muscle strength test, these researchers also standardized their continuous predictors to compare the ORs. Their ORs represented "a change in the odds of retention for one standard deviation change in the predictor variable" (Eskreis-Winkler et al. 2014, p. 3). Their variance inflation factors were all below 2.0; multicollinearity was not an issue. They also performed a hierarchical logistic regression to determine the predictive strength of grit over other predictors variables in their model. Their findings showed years of schooling had a significant relationship with retention, while age did not. In addition, logistic regression showed them that candidates who were grittier were less likely to drop out of ARSOF selection course.

2. Military Attrition Pre-2000

A major part of researching attrition is to find out what variables are important in determining whether a soldier will attrit or not. The RAND corporation and U.S. General Accounting Office (GAO) both focus on factors of early attrition, which is the first six months of a soldier's enlistment. In 1985, Buddin of the RAND corporation published "Analysis of Early Military Attrition Behavior" from a 1979 survey of personnel entering military service (Buddin 1984). GAO's data was from fiscal year (FY) 1994, from the Defense Manpower Data Center.

Recruits who had a history of high unemployment or who changed jobs frequently the year before they enlisted were increase their probability of attrition by 2.2 percentage points(Buddin 1984). Individuals with no prior work experience had a 3.4% higher of attriting than those who did have some work experience. RAND found out that those who

did not graduate high school or have a general educational development (GED) have about an 8% higher rate of attrition than those who did have a high school diploma. Within the three-year enlistment cohort, early attrition increased by one percent for every year by which the recruit was over the age of 17 at enlistment. The best single predictor of attrition was whether an individual has graduated high school or not (RAND 1985).

Older recruits were more prone to attrit early than younger recruits (RAND 1985). This is significant because the cost associated with the loss of a 17-year-old at month 30 in their enlistment vs the cost associated with the loss of a 23-year-old at month 10 of their enlistment is not the same. The cost of recruiting and training the 17-year-old is recovered more than the 23-year-old (RAND 1985). Taken together, the variables of age and previous employment also have predictive value about equal to that of high school graduation status. High school graduation, age, and previous employment all taken all can be used to detect military applicants who are susceptible to early attrition.

Eighty-three percent of separation codes recorded within a soldier's first six months reflected soldiers being medically unqualified for military service, having character or behavior disorders, having fraudulently entered the military, or failing to meet minimum performance criteria (GAO 1997, p. 4). GAO also stated of those who fail to meet minimum performance criteria, they are mostly due to them not being physically prepared or because they lack motivation.

RAND also compared military attrition with civilian separation, since the market for civilian employment directly competes with the market for military jobs. The factors of work history, minority status, and general aptitude had similar effects on civilian job separation and military attrition. In contrast, age, education, and job satisfaction had different effects. This means "older enlistees may be labor market 'misfits'" who then join the military and have a tendency of attriting early (RAND 1985, p. 2). RAND concluded those who dropped out of high school survived longer in the civilian work force than in the military, suggesting that their attitudes and behaviors are less compatible with the military's disciplined lifestyle than civilian employment.

3. Military Attrition Post-2000

Recently, authors have used updated databases to investigate attrition beyond the initial six months of service. Speten (2018) used the Person-Event Data Environment (PDE) to analyze the cohorts of soldiers who joined in FY2005 to FY2010 using demographic data. In 2019, Gobea built on Speten's dataset with the addition of added medical data from FY2008 to FY2010. Speten performed logistic regression with the constant variables taken from the earliest snapshot date, which was assumed to be right after enlistment, and the last snapshot date for variables that changed over time. Speten's conclusions are suspect because one cannot use information at the end of a soldier's career to predict whether he or she will attrit, since we would not have that information at the beginning of the soldier's career. Therefore, Gobea performed logistic regression based on the snapshot of the soldier's data right after graduating from IET.

Speten (2018) found that operations support career fields had the lowest attrition rate, while operational career fields had the highest attrition rate. Enlistees who required an administrative waiver had a lower attrition rate than those who did not require a waiver at all (Speten 2018). He also found that the number of days deployed was associated with a slight decrease in the chance of attrition, and that higher Armed Forces Qualification Test (AFQT) scores were associated with a lower rate of attrition. Naturalized citizenship enlistees and enlistees born outside the U.S. had lower rates of attrition than enlistees born in the U.S. Enlistees that were assigned to multi-component units had a much lower rate of attrition than those assigned to non-deployable units and males in general attrited lower than females (Speten 2018). Speten and Gobea found that, unsurprisingly, soldiers with higher contract durations had higher attrition rates. Gobea also determined that enlistees who were PULHES (Physical capacity, Upper body, Lower body, Hearing, Eyes, Stability/ psychiatric) deployable had a lower probability of attrition than enlistees who were PULHES non-deployable. Higher dental classes, which means poorer dental health, were associated with higher rates of attrition (Gobea 2019).

Devig (2019) performed survival analysis instead of logistic regression; like Gobea, he also used medical variables in addition to the dataset Speten used Devig's (2019) thesis analyzed the time it took for a soldier to attrit. Of importance, he found "time-varying

covariates do affect overall attrition" (p. 58). The variables that were of importance in the split of survival analysis trees were gender, prior service, contract duration, dental readiness class, vision readiness class, and hearing class.

The previous theses above used logistic regression without time-varying covariates. Another thesis analyzed the time it took to attrit using time-varying covariates. This thesis looks at predicting whether a soldier will attrit or not, following the methodology from "The Grit Effect" to determine the outcome. It uses the same starting variables as Gobea and Devig but uses a series of logistic regressions, rather than Devig's survival analysis, to take into account time-varying covariates. The result of the analysis is the binary outcome of whether a soldier will attrit or not within a year, given that he or she survived the previous year(s), and a list of variables and their coefficients that are associated with this outcome.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA AND METHODOLOGY

This section describes where the information for this thesis was gathered from and which datasets were used. It also goes into the categories of variables and how the response variable of attrit was formed. Lastly it goes through the breakdown of how we compile each logistic regression and how the data is split into train and test sets.

A. PERSON-EVENT DATA ENVIRONMENT

The data in this thesis is from the Person-Event Data Environment (PDE). The PDE is a data repository for manpower, service, personnel, financial, health, and medical data for Active Duty, Reserve and National Guard Army personnel (Vie et al. 2013). This platform, compiled by the Army Analytics Group under the Office of the Deputy Under Secretary of the Army, allows commands and external groups to conduct independent research. The remote technology "provides strong protections of human subjects via encoded and deidentified data" (Army Analytics Group 2016, p. 2).

Inside the PDE, personal information about soldiers, such as names, street addresses, phone numbers, and medical record numbers, are removed. Social security numbers are converted to a randomly generated 12-character alphanumeric (Army Analytics Group 2016, p. 30). This 12-character alphanumeric person identifier is known as PID_PDE. Each soldier has a unique Person Identifier (PID_PDE) that is held constant across different datasets within a project. This allows researchers to examine variables and trends across different databases. Researchers can compile their information of interest into one dataset.

Users perform their analysis in a remote desktop; information cannot be downloaded into the local computer, thus keeping all sensitive information in the PDE. Researchers can use statistical software such as R and Toad for Oracle that is made available in the PDE via the remote desktop (Vie et al. 2015). These two programs were used for this thesis. Graphs or charts that do not have personal sensitive information made in the PDE may be removed after examination by security officials.

B. DATASETS USED

The eight datasets from the PDE used in this thesis are the same as the ones Devig (2019) and Gobea (2019) used in their theses. The first six datasets were used by Speten (2018). The main dataset, Active Duty Military Personnel Master, contains demographic factors such as rank, paygrade, education level code, active duty service projected end date, active federal military service base date, primary service occupation code, and additional skill identifier code. This information comes from the Army Human Resources Command. The next dataset used is Active Duty Military Personnel Transaction. This dataset contains the records of Active Duty Soldier's entrance, separation, or reenlistment. It contains specific information such as permanent duty station arrival and departure dates, character of service code, interservice separation code, and separation program designator code. The Military Entrance Processing Command is the third dataset used. This dataset contains information from when a soldier was recruited. It contains variables such as ethnicity, marital status, number of dependents, height, weight, prior service reenlistment codes, and a soldier's AFQT information. The Army Waiver Database contains information such as administrative, medical, and drug/alcohol, and conduct waiver events for entry into military service. The Contingency Tracking System|Overseas Contingency Operations database holds an inventory of all service members deployed in support of overseas contingency operations. It contains the total amount of days and where a soldier has been deployed. The Defense Casualty Information Processing System is information from the Defense Casualty Information Processing System, which holds the record for casualty and mortuary affair cases. It contains data for both wounded and killed-in-action soldiers.

In addition to the previous six datasets, Devig and Gobea added two databases to provide medical indicators of attrition. The first of these is the Periodic Health Assessment (PHA) database. The PHA is a two-part health assessment. The first part is done by the soldier and the second is done by a provider. This dataset has information such as a soldier's height, weight, PULHES, and potential for deployability with six months. Since the PHA form changed in 2016, there were two datasets: one for the old PHA forms and one for the new PHA forms. The two datasets were merged into once since the two forms

consisted of mostly the same questions. The second medical dataset is the Medical Protection System (MEDPROS) dataset, and it consists of medical variables such as dental class, vision ready class, hearing readiness class, immunizations, pregnancy status, and profile codes.

C. VARIABLES USED

This section explains in depth the variables used in this thesis, broken up by numeric, binary, and categorial variables. Most of the variables listed are constant, meaning that they do not change over time. Most demographic information is constant throughout a person's life. Some variables like marriage and soldier's health information change throughout the duration of a soldier's enlistment. The last section of this portion examines time-varying covariates, variables that are not at a constant value.

1. Numeric Variables

There are nine numeric variables considered for this study. Table 1 shows the variable name, the description, and the necessary associated units with the variable. Speten used hostile injury count, nonhostile injury count, deployment count, and days deployed as predictors. We do not use them here since we only have the total value at the end of a soldier's first term. They cannot be used as possible predictors for the logistic regression models since they are not a single fixed value throughout a soldier's first term of enlistment. The number of each can change from one year to the next, and only the total is known for the entire first term of enlistment.

Table 1. Numeric Variables

Variable	Description
Height	Height at enlistment in inches
Weight	Weight at enlistment in pounds
Age	Age at enlistment in years
ASVAB GT Score	The sum of three ASVAB test areas: word knowledge, paragraph comprehension, and arithmetic reasoning
Dependents at Enlistment	Number of dependents at enlistment

2. Binary Variables

Binary variables are variables that only consist of two values. The research examines 17 medically related binary variables, and eight demographic binary variables. Of special note, PULHES nondeployable indicates if someone is nondeployable based on their PULHES status. If a soldier has a 3 or 4 in any of the PULHES categories, then he or she is PULHES nondeployable. Table 2 lists the levels of each PULHES category. The variable Hispanic was created for this thesis. The reason behind creating this variable is in Chapter III, Section B. The last variable in Table 2 is the response variable of attrit. Section D in this chapter discusses this variable in more detail.

Table 2. Binary Variables

Variable	Description and Value
Gender	M: Male F: Female
Hispanic	0: No, 1: Yes
US Citizenship Status Code	C: U.S. Citizen N: Non-U.S. Citizen
Prior Service	0: No, 1: Yes
Medical waiver	0: No, 1: Yes
Drug waiver	0: No, 1: Yes
Conduct Waiver	0: No, 1: Yes
Administrative waiver	0: No, 1: Yes
PULHES Nondeployable	0: No, 1: Yes
Anemia	0: No, 1: Yes
Asthma	0: No, 1: Yes
Back Pain	0: No, 1: Yes
Cancer	0: No, 1: Yes
Chronic Pain	0: No, 1: Yes
Diabetes	0: No, 1: Yes
Epilepsy	0: No, 1: Yes
Headaches	0: No, 1: Yes
Heart Murmur	0: No, 1: Yes
Heart Trouble	0: No, 1: Yes
Hypertension	0: No, 1: Yes
Joint Pain	0: No, 1: Yes
Kidney Disease	0: No, 1: Yes
Liver Disease	0: No, 1: Yes
Mental Health Concerns	0: No, 1: Yes
Pregnancy Status	0: No, 1: Yes
Attrit (Response Variable)	0: No, 1: Yes

3. Categorial Variables

Categorical variables have three or more non-numeric levels in the dataset. There are 33 categorical variables with a total of 130 factor levels. Data in ten variables were grouped together by similarity to reduce the number of levels and to increase the number of observations in each level: the AFQT percentile, age group at enlistment, home of record regions, career management field, career management field group, faith group, hearing readiness, vision readiness, marital status, and race code. Table 3 shows each variable, its levels, and a description of each level.

Table 3. Categorical Variables

Variable	Levels	Level Description
Fiscal Year Group	2009	Soldier joined FY 2009
	2010	Soldier joined FY 2010
	2011	Soldier joined FY 2011
Contract Duration	3	3 Year Contract
	4	4 Year Contract
	5	5 Year Contract
	6	6 Year Contract
AFQT Category Code Percentile	1	93-99%
	2	65-92%
	3A	50-64%
	3B	31-49%
	4A	21-30%
	4B/4C/5	0-20%
Age Group at Enlistment	17-19	Soldiers 17–19 years old at enlistment
	20-22	Soldiers 20–22 years old at enlistment
	23-25	Soldiers 23–25 years old at enlistment
	26-30	Soldiers 26–30 years old at enlistment
	31-35	Soldiers 31–35 years old at enlistment
	36+	Soldiers 36 years and older at enlistment
Education Tier Code at Enlistment	1	High school diploma or have at least 15 college credits
	2	GED or equivalent
	3	No high school diploma, GED, or equivalent

Variable	Levels	Level Description
Ethnicity Affinity Code	22 levels	Level Description in Appendix A
Home of Record Region (Abbreviations are listed in Appendix B)	Midwest	IA, IL, IN, KS, MI, MN, MO, ND, NE, OH, SD, WI
	Northeast	CT, MA, ME, NH, NJ, NY, PA, RI, VT
	South	AL, AR, DC, DE, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV
	Territory	AS, GU, PR, VI
	West	AK, AZ, CA, CO, HI, ID, MT, NM, NV, OR, UT, WA, WY
Career Management Field	9	Interpreter/Translator
	11	Infantry
	12	Engineer
	13	Field Artillery
	14	Air Defense Artillery
	15	Aviation
	18	Special Forces
	19	Armor
	25	Signal
	31	Military Police
	35	Military Intelligence
	37	Psychological Operations
	38	Civil Affairs
	42	Human Resources
	63	Vehicle Mechanic
	68	Health Services
	74	Chemical
	88	Transportation
	89	Ammunition and Ordnance Disposal
	91	Ordnance
	92	Quartermaster
	94	Electronic/Missile Maintenance
	LD	Low Density
Career Management Field Group	Force Sustain-ment (FS)	42,63,68,88,89,91,92,94,LD
	Maneuver, Fires, and Effects (MFE)	11,12,13,14,15,18,19,31,37,38,74

Variable	Levels	Level Description
	Operations Support (OS)	9,25,35
	Other	Entries of NA in the PDE
US Citizenship Origin Code	A	Born in the U.S.
	C	Born outside the U.S.
	N	Naturalized citizen
Race Code		American Indian or Alaskan Native
	1&2	Asian or Pacific Islander
	3	Black or African American
	4	White
Marital Status	D	Divorced
	M	Married
	N	Never Married
	Other	Legally separated, annulled, widow(er)
PULHES_Physical Capacity (Department of the Army [DA] 1994)	1	Good muscular development with ability to perform maximum effort for indefinite periods
	2	Able to perform maximum effort over long periods.
	3	Unable to perform full effort except for brief or moderate periods
	4	Functional level below P3.
PULHES_Upper Body (DA 1994)	1	No loss of digits or limitation of motion; able to do hand to hand fighting.
	2	Slightly limited mobility of joints, or other Musculo-skeletal defects that do not prevent hand–to–hand fighting.
	3	Defects or impairments that require significant restriction of use.
	4	Functional level below U3.
PULHES_Lower Body (DA 1994)	1	No loss of digits or limitation of motion; able to perform long marches, stand over long periods, run.
	2	Slightly limited mobility of joints, or other Musculo-skeletal defects that do not prevent moderate marching, climbing, timed walking, or prolonged effort.
	3	Defects or impairments that require significant restriction of use

Variable	Levels	Level Description
	4	Functional level below L3.
PULHES_Hearing (DA 1994)	1	Audiometer average level for each ear not more than 25 dB at 500, 1000, 2000 Hz with no individual level greater than 30 dB.
	2	Audiometer average level for each ear at 500, 1000, 2000 Hz, or not more than 30 dB, with no individual level greater than 35 dB at these frequencies
	3	Speech reception threshold in best ear not greater than 30 dB HL, measured with or without hearing aid; or acute or chronic ear disease
	4	Functional level below H3.
PULHES_Eyes (DA 1994)	1	Uncorrected visual acuity 20/200 correctable to 20/ 20, in each eye
	2	Distant visual acuity correctable to not worse than 20/40 and 20/70, or 20/30 and 20/100, or 20/20 and 20/ 400.
	3	Uncorrected distant visual acuity of any degree that is correctable not less than 20/40 in the better eye.
	4	Visual acuity below E3.
PULHES_Stability/ Psychiatric (DA 1994)	1	No psychiatric pathology. May have history of a transient personality disorder
	2	May have history of recovery from an acute psychotic reaction due to external or toxic causes unrelated to alcohol or drug addiction.
	3	Satisfactory remission from an acute psychotic or neurotic episode that permits utilization under specific conditions
	4	Does not meet S3 above.
Blood Type	O-	Soldier has O– blood
	O+	Soldier has O+ blood
	A-	Soldier has A– blood
	A+	Soldier has A+ blood
	B-	Soldier has B–blood
	B+	Soldier has B+ blood
	AB-	Soldier has AB– blood
	AB+	Soldier has AB+ blood

Variable	Levels	Level Description
Dental Readiness Class (U.S. Army Public Health Command [USAPHC], 2018)	1	Soldier has had a complete dental exam within the past year; does not require any dental care
	2	Soldier requires some type of dental care or re-check; filling; cleaning; simple extraction
	3	Soldier requires dental care as soon as possible; dental emergency is likely to occur if the condition is not corrected
	4	Soldier requires a complete dental exam; no dental exam within the last 12 months
Hearing Readiness Class (USAPHC, 2017)	1	Soldier's unaided hearing is within H-1 standards for both years
	2	Soldier's unaided hearing is within H-2 or H-3 standards; has hearing aid if required
	3	Soldier's unaided hearing is within H-2 or H-3 standards; does not meet standards with hearing aid; complete audiological evaluation has not been complete
	4	Hearing readiness classification unknown. No hearing test for last 12 months
Vision Readiness Class (USAPHC, 2012)	1	Soldiers whose best-corrected binocular visual acuity is 20/20 or better for all required visual acuity screenings.
	2	Soldiers whose best-corrected binocular visual acuity is worse than 20/20 but at least 20/40 in the poorest of their required visual acuity screenings.
	3	Soldiers who are not optically ready and/or not visually ready
	4	Soldiers who have not been screened within 1 year

4. Time-Varying Covariates

A major difference between Speten and Devig's logistic regression analysis and this thesis is that this thesis uses time-varying covariates. Time-varying covariate values change over time. For example, a soldier who undergoes eye surgery might have his or her vision readiness class change from 2 to 1. Our models consider the value for each time-varying coefficient recorded at the soldier's anniversary. Table 4 shows all time-varying covariates considered in this thesis.

Table 4. Time-Varying Covariates

Anemia
Asthma
Back Pain
Cancer
Chronic Pain
Diabetes
Epilepsy
Headaches
Heart Murmur
Heart Trouble
Hypertension
Joint Pain
Kidney Disease
Liver Disease
Mental Health Concerns
Pregnancy Status
Dental Readiness Class
Hearing Readiness Class
Vision Readiness Class
PULHES_Physical Capacity
PULHES_Upper Body
PULHES_Lower Body
PULHES_Hearing
PULHES_Eyes
PULHES_Stability/Psychiatric
Career Management Field Group
Marital Status

D. RESPONSE VARIABLE

To perform an attrition analysis, the response variable of attrit was constructed using the same techniques Speten and Devig used from the data in the PDE. Soldiers with an Initial Service Separation Code of "1087" were removed since they were discharged from the Army before the completion of their MOS' IET (Speten 2018). Additionally, Soldiers with a "1016" separation code were removed since they served less than 4.5 months, the average duration of IET, and were unqualified for active duty (Devig 2019). This study uses the most common enlistment year terms of three, four, five, and six years. Service term lengths that did not make sense or had very few entries were removed: those were zero, one, two, seven, and eight years.

After removing soldiers who did not complete IET or make it to active duty, soldiers were coded "0" for not-attrit or "1" for attrit for the created response variable based on several factors. Soldiers with an Enlisted Career Status Code of "3" were categorized as not-attrit since they reenlisted. If soldiers did not have a separation code and their initial obligation date was not recorded, they were removed. Soldiers who had separation codes of release from active service, death, officer program, retirement other than medical, other separations or discharges, or transactions (reenlistment) were coded as not-attrit (Devig 2019). If soldiers had missing data in their separation and discharge codes, but did have an initial obligation date, a Calculated Obligation Date was created. The obligation date is computed as the initial contract number of years plus their Active Federal Military Service (AFMS) Base Date. The AFMS Base Date is the date for which DOD Military Service member's creditable Active Military Service begins (PDE 2019). Since an additional snapshot is added to the PDE quarterly, if soldiers did not have a record in the master data three months beyond their Calculated Obligation Date, then those soldiers were assigned the value of attrit. If soldiers did have an updated snapshot date three months after their Calculated Obligation Date, they were coded as not-attrit. The end result for the fiscal year (FY) 2009, FY2010 and FY2011 cohorts combined was 183,932 Soldiers with 44,713 attriting and 139,219 not-attriting.

E. COHORTS AND GROUP NAMING CONVENTION

The cohort for this study consists of all the soldiers who joined Army Active Duty in FY 2009, 2010, or 2011 and who had three, four, five, or six-year contract durations. These years were chosen since they were the three years with the least amount of missing PHA data (Devig 2019). Since previous research has shown contract duration as an important variable to determine attrition, we examine models for the three, four, five, and six-year contract durations separately. We broke the data into 18 groups. The name of each group consists of the contract duration and which year in contract is being analyzed. For example, data set CD3_Y0 is used to examine first-year attrition among soldiers with three-year contracts. The CD3_Y1 is used to examine attrition among soldiers with three-year terms who survived their first year; CD3_Y3 is used to examine attrition among soldiers with three-year terms who survived two years, and so on. Table 5 labels each cohort in the dataset.

Table 5. Cohort Naming Convention

	Year in Contract					
Contract Duration	0 to 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6
3	CD3_Y0	CD3_Y1	CD3_Y2			
4	CD4_Y0	CD4_Y1	CD4_Y2	CD4_Y3		
5	CD5_Y0	CD5_Y1	CD5_Y2	CD5_Y3	CD5_Y4	
6	CD6_Y0	CD6_Y1	CD6_Y2	CD6_Y3	CD6_Y4	CD6_Y5

F. LIMITATIONS AND ASSUMPTIONS

The data in the PDE is only as good as the people who entered it. For this thesis, we assume that the data in the PDE was entered correctly. Some career management field (CMF) data needed to be changed to account for military occupational specialty (MOS) changes, and we assume those changes were correctly made. For example, combat engineers and horizontal engineers changed from 21B to 12B, and 21N to 12N respectively. Therefore, we changed all CMF 21 to CMF 12. Next, to determine which year of service a soldier attritted, every soldier needed an end date. For soldiers who did not have an end date or had the value "NA," their last snapshot date became the soldier's assumed end date. Lastly, since we are using the method Devig created to make the response variable, we assume that his strategy is correct.

Since this study also involved medical and PHA data, we used Devig's discovery of how many entries per fiscal year of accession had missing PHA data. Fiscal years 2005–2008 had approximately 20–80% missing PHA data (Devig 2019), limiting the scope of what years could be used that include medical data. In this thesis, we only analyze FY2009–2011 since these years have the smallest amount of missing PHA data: 2009 has 11.69%, 2010 has 6.57%, and 2011 has 5.61%. Further assumptions that were identified to be made to the data are summarized in Chapter IV, section B.

G. TRAINING AND TEST SETS

The dataset was split into training and test sets. We develop our algorithm using the training set, and assess how it performs with the test set. The combined cohort group of FY2009, FY2010, and FY2011 serve as the training and test sets. A randomly selected 80% of each contract duration by year is used as a training set. The other 20% of each contract duration by year is used as a test set.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DESCRIPTIVE STATISTICS

This chapter contains an overview of the dataset used to produce the logistic regression models. It then shares a summary of what previous authors have found by examining attrition rate with information right after IET with the same variables. The information that follows goes into attrition rates broken down by contract duration. Lastly, the chapter reports new discoveries regarding time-varying covariates and attrition.

A. DESCRIPTIVE STATISTICS RIGHT AFTER IET

Previous authors have examined attrition rates by using non-time-varying covariates and data right after IET. Speten and Devig's analysis of home of record showed West Virginia had the highest attrition rate at approximately 31%, while territories had the lowest attrition rate at 15.04%. The Midwest had an attrition rate of 23.84%, northeast 24.20%, south 26.41%, and west 22.18% (Devig 2019). Out of the four categories of race inside the PDE, Asian or Pacific Islander had the lowest attrition rate at 17.30% (Devig 2019). Soldiers who were married right after IET had a lower attrition rate, at 23%, than those who were not married, divorced, or legally separated, annulled, or widowed (Devig 2019). Across FY2005 to FY2011 females had an average attrition rate of 37.24% while men had an average attrition rate of 22.13% (Devig 2019). While there was no trend for males attriting by FY, females did have a decreasing trend from FY05 to FY10 (Gobea 2019).

Soldiers who were coded as having dental readiness class 3 right after completion of IET only accounted for 3.65% of Devig's FY2005–2011 cohort, but had a disproportionate attrition rate of 54.25%, much higher than those with dental readiness class 1 or 2. For the FY2009–2011 cohort, the attrition rates for soldiers with dental readiness classes 3 and 4, range from 41–57% (Devig 2019). Soldiers with vision readiness classes 3 and 4 in the FY2009-2011 cohort also had a higher attrition rate, averaging at 28.67%, compared to the average among those in classes 1 and 2 of 19%. Unlike vision and dental, hearing readiness class did not show a trend with an increase in class number.

The average attrition rate for FY2009–2011 for class 1 was 19.67%, for class 2 was 17%, for class 3 was 25%, and for class 4 was 15% (Devig 2019).

Devig (2019) discovered that women who become pregnant shortly after the completion of IET have a lower attrition rate at 24.48%, than those who do not become pregnant; the overall attrition rate for females is 37.24%. The lower attrition rate may be attributed to the access to the low-cost healthcare and dentalcare for the service member and her children. Males and females show an upward trend in attrition from FY05 to FY10 when they have a medically non-deployable profile shortly after IET (Gobea 2019). Devig also found this to be the case with his cohort, and he suggests it could be attributed to retention policy changes, but that is hard to quantify (Devig 2019). Every soldier has a PULHES rating, the definition of which can be found in Table 3. There was an increase in attrition associated with physical capacity, upper body, lower body, and psychiatric as the rating increased from 1 to 4 based on the soldiers PULHES rating right after IET. (Devig 2019). Hearing and eyesight did not have a major increase in attrition like the other four categories (Devig 2019).

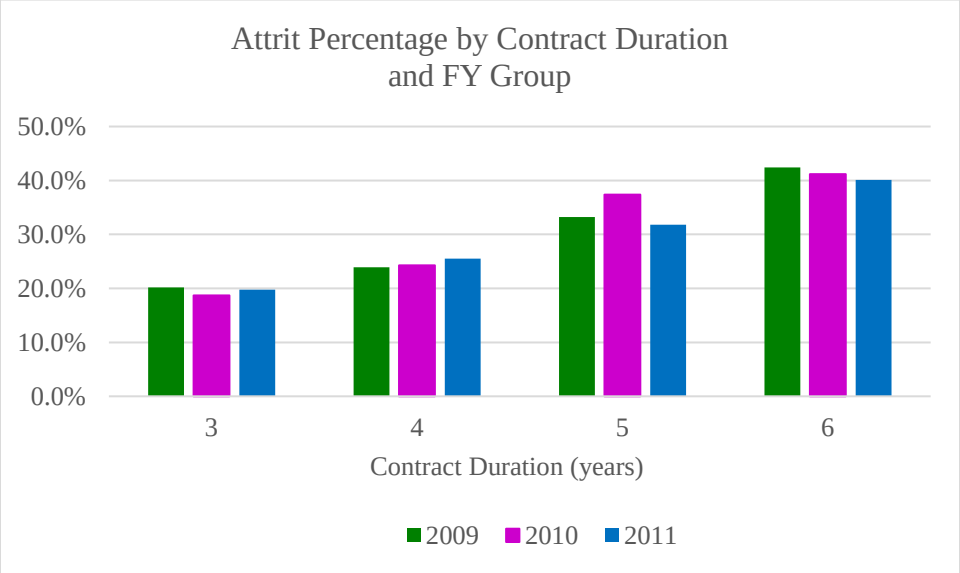
B. ATTRITION RATE BY ACCESSION FISCAL YEAR AND CONTRACT DURATION

Table 6 and Figure 1 have the number and percentage of soldiers who did and did not attrit before the end of their first-term enlistment. The table is broken down by AFMS Cohort FY Group and by contract duration; there is also a total for each cohort FY group and a grand total.

Table 6. Attrition Rate by Accession, Fiscal Year, and Contract Duration

AFMS Cohort FY Group		Contract Duration (years)				
		3	**4**	**5**	**6**	**Total CD**
2009	attrition rate	20.2%	23.9%	33.2%	42.4%	24.9%
	total accessions	30628	19039	5426	5787	60880
2010	attrition rate	18.7%	24.3%	37.4%	41.1%	23.7%
	total accessions	35466	19355	4325	5967	65113
2011	attrition rate	19.8%	25.5%	31.8%	40.1%	24.4%
	total accessions	31870	14876	5448	5745	57939
Total FY09-11	attrition rate	19.5%	24.9%	33.9%	41.2%	24.3%
	total accessions	97964	53270	15199	17499	183932

Original dataset retrieved from Person-Event Data Environment, February 2020.



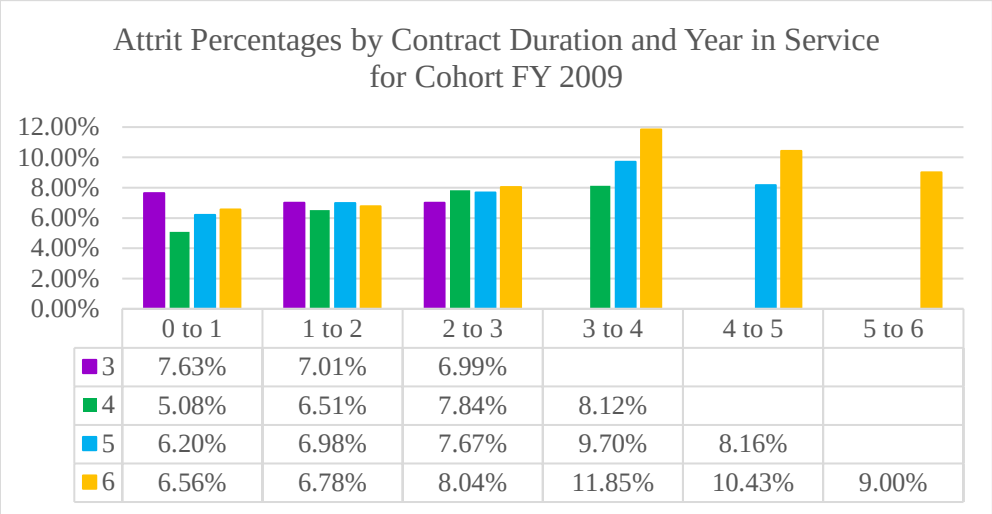
Original dataset retrieved from Person-Event Data Environment, February 2020.

Figure 1. Attrition Rate by Accession, Fiscal Year, and Contract Duration

The no attrit and attrit percentages across the FY2009, FY2010, and FY2011 cohort groups stay roughly constant, except for the 4–6% decrease in the five-year contract duration for FY2010 from FY2009 and FY2011. The attrition rate increases as contract duration increases from 19.50% for three-year contracts, 24.87% for four-year contracts, 33.91% for five-year contracts, to 41.21% for six-year contracts. The total contract duration attrition rate for all FY groups average is 24.31%.

C. ATTRITION RATE BY CONTRACT DURATION AND YEAR IN CONTRACT

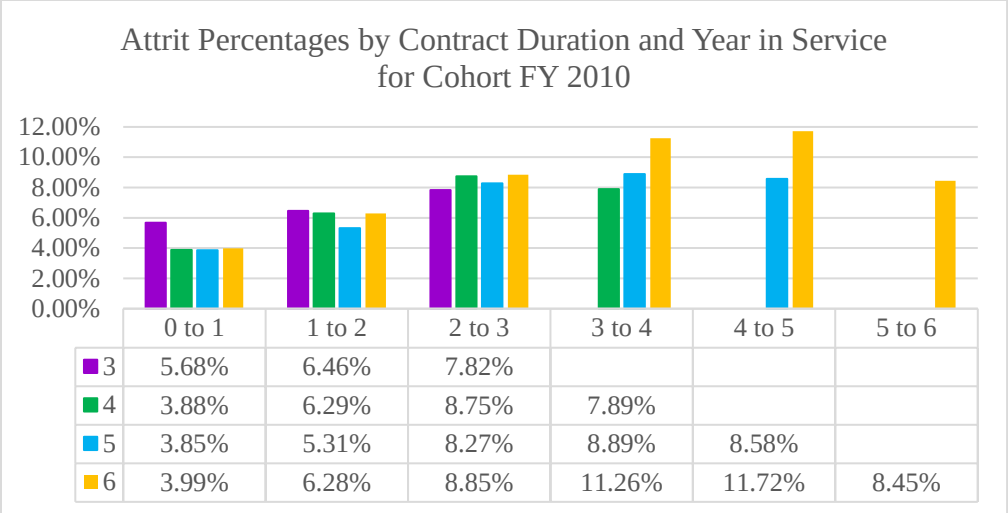
Figures 2, 3, and 4 show the conditional attrition rates by contract duration per year for each AFMS Cohort. That is, each percentage shows the one-year attrition rate just among soldiers who survived to the given starting point. For cohort FY2009, attrition rates for contract duration three (purple) decrease, while for the same contract duration, attrition rates for FY2010 and FY2011 increase. Attrition rates for soldiers with a contract duration of four or five had a steady increase from year zero to the year their contract ended. For contract duration six, there was a general peak of attrition rates during years three to four and four to five, and finally a decrease for the last year in the contract.



	0 to 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6
3	7.63%	7.01%	6.99%			
4	5.08%	6.51%	7.84%	8.12%		
5	6.20%	6.98%	7.67%	9.70%	8.16%	
6	6.56%	6.78%	8.04%	11.85%	10.43%	9.00%

Original dataset retrieved from Person-Event Data Environment, February 2020.

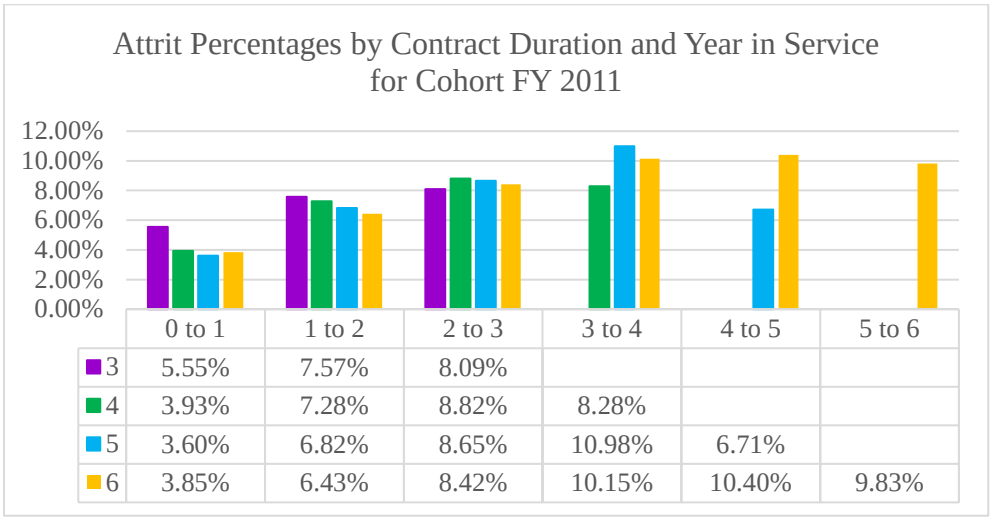
Figure 2. Cohort FY2009 Attrition by Contract Duration and Year in Contract



	0 to 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6
3	5.68%	6.46%	7.82%			
4	3.88%	6.29%	8.75%	7.89%		
5	3.85%	5.31%	8.27%	8.89%	8.58%	
6	3.99%	6.28%	8.85%	11.26%	11.72%	8.45%

Original dataset retrieved from Person-Event Data Environment, February 2020.

Figure 3. Cohort FY2010 Attrition by Contract Duration and Year in Contract



	0 to 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6
3	5.55%	7.57%	8.09%			
4	3.93%	7.28%	8.82%	8.28%		
5	3.60%	6.82%	8.65%	10.98%	6.71%	
6	3.85%	6.43%	8.42%	10.15%	10.40%	9.83%

Original dataset retrieved from Person-Event Data Environment, February 2020.

Figure 4. Cohort FY2011 Attrition by Contract Duration and Year in Contract

D. ATTRITION RATE BY CMF

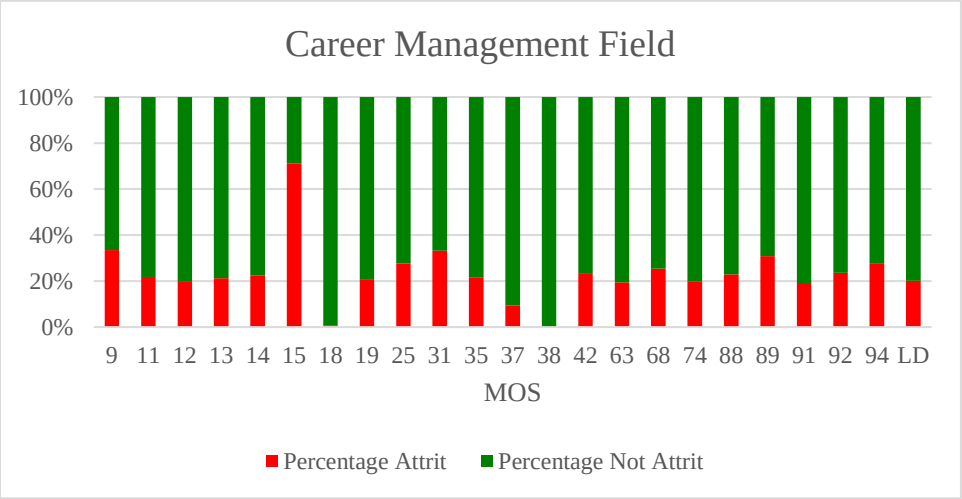
In order to analyze attrition rates by CMF as a time-varying covariate, we looked at the number of soldiers who were ever in a certain CMF. Table 7 accounts for those who started in a CMF or changed into the CMF. It does not include those who switched out of a CMF. The percentage of attrit accounts for anyone who was part of that CMF based on the total at enlistment and changed into CMF.

Table 7. Attrition Rate by CMF

MOS	Not Attrit	Attrit	Total	Percentage Not Attrit	Percentage Attrit
09	7932	3998	11930	66.49%	33.51%
11	29156	8158	37314	78.14%	21.86%
12	13064	3273	16337	79.97%	20.03%
13	8928	2399	11327	78.82%	21.18%
14	2786	806	3592	77.56%	22.44%
15	1104	2733	3837	28.77%	71.23%
18	1015	8	1023	99.22%	0.78%
19	7370	1912	9282	79.40%	20.60%
25	11035	4206	15241	72.40%	27.60%
31	4090	2033	6123	66.80%	33.20%
35	8839	2435	11274	78.40%	21.60%
37	229	24	253	90.51%	9.49%
38	33	0	33	100.00%	0.00%
42	2248	691	2939	76.49%	23.51%
63	2700	651	3351	80.57%	19.43%
68	10802	3676	14478	74.61%	25.39%
74	1913	468	2381	80.34%	19.66%
88	6132	1817	7949	77.14%	22.86%
89	3057	1366	4423	69.12%	30.88%
91	14117	3255	17372	81.26%	18.74%
92	14498	4491	18989	76.35%	23.65%
94	1854	714	2568	72.20%	27.80%
LD	1577	396	1973	79.93%	20.07%

Original dataset retrieved from Person-Event Data Environment, February 2020.

Interpreters, military police, and ammunition/ explosive ordnance (9, 31, and 89) have the highest three attrition rates, all above 30%. The CMFs with the lowest attrition rates are psychological operations (37) at 9.49%, special forces (18) at 0.78%, and civil affairs (38) at 0%. Figure 5 shows the attrition rates per CMF in graphical format.



Original dataset retrieved from Person-Event Data Environment, February 2020.

Figure 5. CMF Attrition Rates

It is very likely that psychological operations, civil affairs, and special forces all have extremely low attrition rates because they recruit from a more highly qualified set of civilians and soldiers. In addition, these CMFs require a certain amount of education before they can qualify for the CMF. The duration of the education at different schools can last up to 18–24 months. Going to these different schools take up a big portion of their initial contract, and some soldiers must reenlist to finish their education. The soldiers' determination and caliber for these three CMFs likely contribute to the low attrition rates.

From this point on, CMFs are grouped into four categories for analysis: Operations Support; Force Sustainment; Maneuver, Fires, and Effects, and None for soldiers who do not have data on their CMF.

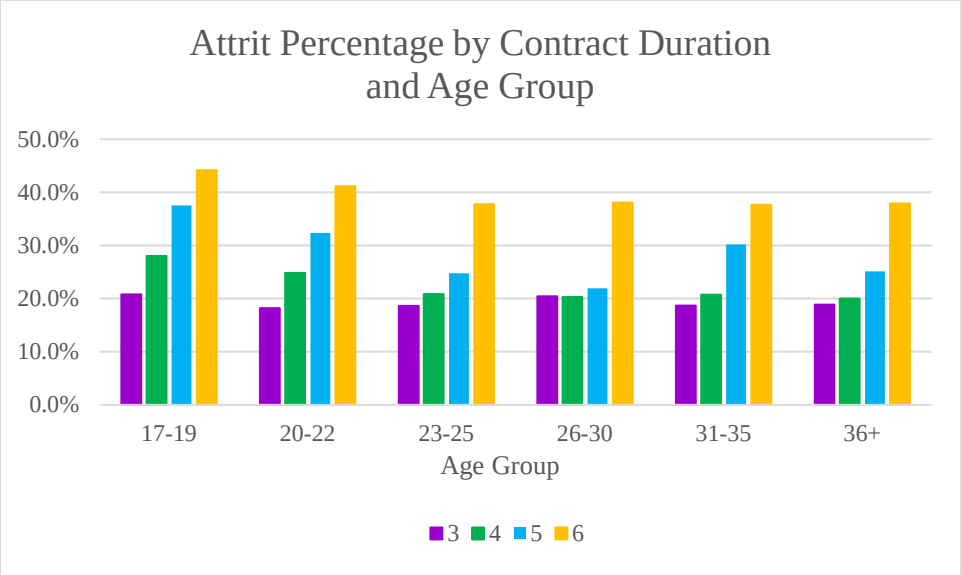
E. ATTRITION RATE BY AGE GROUP AT ENLISTMENT AND CONTRACT DURATION

We already discovered that attrition rates increase as the contract duration increases. Table 8 and Figure 6 show the attrition rates for each age group at enlistment and contract duration. There is no significant increase or decrease for attrition rate as age increases for three-year contract. Attrition rate for four-year contract duration decreases as age increases. For the five-year contract duration, attrition rate decreases as age increases from age 17 to 30, and then attrition increases for the 31–35 age group. The six-year contract attrition rate decreases as age increases.

Table 8. Attrition Rate by Age Group at Enlistment and Contract Duration

Age Group		Contract Duration (years)				
		3	**4**	**5**	**6**	**Total Age Group**
17-19	attrition rate	20.8%	28.0%	37.5%	44.1%	26.6%
	total accessions	36743	21671	6022	6492	70928
20-22	attrition rate	18.2%	24.8%	32.4%	41.2%	23.5%
	total accessions	30982	15784	5226	5368	57360
23-25	attrition rate	18.6%	20.9%	24.8%	37.8%	21.7%
	total accessions	15533	7858	2892	2730	29013
26-30	attrition rate	20.4%	20.3%	21.9%	38.1%	22.3%
	total accessions	9648	4942	1529	1825	17944
31-35	attrition rate	18.7%	20.7%	30.2%	37.6%	22.0%
	total accessions	3280	1875	351	680	6186
36+	attrition rate	18.9%	20.0%	25.1%	37.9%	21.7%
	total accessions	1777	1136	179	404	3496
Total CD	attrition rate	19.5%	24.9%	31.8%	41.2%	24.2%
	total accessions	97963	53266	16199	17499	184927

Original dataset retrieved from Person-Event Data Environment, February 2020.



Original dataset retrieved from Person-Event Data Environment, February 2020.

Figure 6. Attrition Rate by Age Group at Enlistment and Contract Duration

THIS PAGE INTENTIONALLY LEFT BLANK

IV. MODELING AND ANALYSIS

In this chapter we first explain the fundamentals behind logistic regression. Afterwards we examine how the data was prepared and how missing values were handled for use in logistic regression models. Next, we look at how we conducted variable selection for all 18 years using random forests. Then we go over how six final models were chosen to go by year in contract versus contract duration. Finally, we go over diagnostics of the models and our analysis from the findings of each model.

A. MODELING APPROACHES

We use logistic regression to predict whether the probability that a soldier will attrit. Multiple logistic regression allows us to generate a model that predicts the probability of an event happening, which we label P(Y=1), using a function of predictor variables X_1, X_2, …, X_k (Hosmer and Lemeshow 2000). Y is a Bernoulli random variable, where

$$Y = 1 \text{ with probability } p$$
$$Y = 0 \text{ with probability } 1 - p.$$

The expected value of Y, which is the probability that $Y = 1$ – call that p_i – falls between 0 and 1. The logit of p_i – that is, the log of the odds ratio $p_i / (1 - p_i)$ – is then modeled by a linear combination of the predictors. In symbols, the model says

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} \ldots + \beta_k x_{ik.} \quad (1.1)$$

The odds of an event, in our case attrition, are defined as the probability an event occurs divided by the probability it does not. Odds can be interpreted as how much more likely an event will occur vs not occurring. When odds are greater than 1, the event is more likely to occur than not. When odds are less than 1, the event is less likely to occur than not. When the odds are equal to 1, the event is just as likely to occur as to not occur. The log of the odds produces a transformed value that can range from negative infinity to positive infinity; this transformation allows the use of regression analogous to the ordinary linear regression for a continuous response.

The coefficients in the model are estimated using maximum likelihood estimation (MLE). MLE finds coefficients for the predictor variables (β_k) to yield predicted values of p that are, to the extent possible, close to 0 for individuals with do not attrit, and close to 1 for individuals who do (Faraway 2016). Once coefficients are found using logistic regression, predicted probabilities ($0 < p < 1$) can be calculated by inverting the equation above (1.1). The next step is the classify the individual based on a cutoff level. The most popular choice is establishing the cutoff at 0.5. In our models, the cutoff was set as the proportion of soldiers who attrit in the training datasets. More discussion of this is in Section D. Soldiers whose predicted probabilities fall below the cutoff are classified as not attrit, and soldiers with predicted probabilities above the cutoff are defined as attrit.

B. DATA PREPARATION

Yearly snapshots of data were taken from each soldier based on his or her AFMS base date. From there, 18 datasets were organized by contract duration and year within the contract, as seen in Table 5. All 18 datasets have the same covariates listed in Chapter II. We analyzed the percentage of attrit per contract duration per year in contract, and those results are in Figures 1–3.

1. Missing Data Entries

The dataset had numerous blank entries, or entries listed as "NA." The following list explains what changes were made to retain entries that had a NA or missing information.

- Medical variables that are included in the PHA, such as heart murmur, liver disease, hypertension, and pregnancy status were listed as NA unless the soldier had that condition or status. Once the soldier received a diagnosis or a positive lab result, he or she was then listed as Y for yes. Entries that had a "NA" were changed to "N" for no.

- Any missing entries for the dependent quantity at MEPS were changed to the value of zero (for no dependents while enlisting into the Army). Zero was also the mode for that variable.

- Soldiers with missing age information were put into the age group mode, which is “17-19.”
- The few NA’s that were in U.S. citizenship status code variable were placed into the citizen level, which is the mode for this variable.
- Missing entries for education tier code were placed in the mode, which is category 1; soldiers with this category have a high school diploma or at least 15 college credits.
- The mode for home of record region is the south; all NA’s were put into the mode.
- Any missing entries for marital status were placed in the mode, which is “N,” never married.
- Missing information for CMF Group were placed in another created level, which we labeled “None.”
- For numeric variables, the average of each of the 18 datasets was calculated for the variable and used as the value for the NA in each specific dataset.

2. Dental, Hearing, and Vision

From previous theses, dental, hearing, and vision readiness classes are very strong predictors. Dental readiness class was the strongest predictor for many of the 18 datasets when classes 1, 2, 3, and 4 were all used. The class 4’s of dental, hearing, and vision, act like future predictors of attrition, which is not realistic when one is conducting present day analysis. Since soldier who do not get a medical screening or exam in a calendar year are automatically categorized as class 4, an enormous number of soldiers who attrit are categorized as class 4 for dental, hearing, and/or vision. Possible reasons for why soldiers did not receive a medical exam within the previous year include deployment (where they did not have access to a dentist, optometrist, or an audiology exam), pre-deployment

training rotations, lack of providers in a timely manner, miscommunication between medical tracking systems, or that the soldier elected not to receive medical treatment as they were out-processing. In this thesis, we chose to exclude dental, hearing, and vision readiness class 4, and instead focus on classes 1, 2, and 3, since those classes better indicate the soldier's medical health. For soldiers who had dental, hearing, or vision class 4, we chose their previous most immediate medical readiness class before their code turned into a class 4. This indicates their actual medical class based on their most known medical conditions.

3. Purposeful Exclusion

Where there were too few entries, the entries were either omitted or grouped with another level. The "Other" level in variable Marriage had too few entries. Since only a few people were had "Other" as their level by contract duration per year, those entries were removed. In the Year4 data, only two entries in the training set had P_PULHES category 4 had only two entries in the training set. Those two entries were grouped with P_PULHES category 3.

Blood type had too many NA's to be moved into the mode. The total number of NAs exceeded the number of soldiers in categories O–, A–, B–, AB+, and AB–. The amount of NA's was about the same as the number of soldiers in B+, half the number of soldiers in A+ and a third the amount in O+; thus blood type was removed as a predictor.

Further into our research, our models' Variance Inflation Factors (VIF), which detect multicollinearity, identified that ASVAB GT scores and AFQT category code percentiles were correlated. We chose to use ASVAB GT Score and omit AFQT category code percentiles for two reasons. First, ASVAB GT Score was usually the more important predictor of attrition in our 18 models. Secondly, using AFQT category code percentiles variable added more degrees of freedom into the model while ASVAB GT score is a numerical predictor and only has one degree of freedom. Figure 7 shows the relationship between ASVAB GT score and AFQT category code.

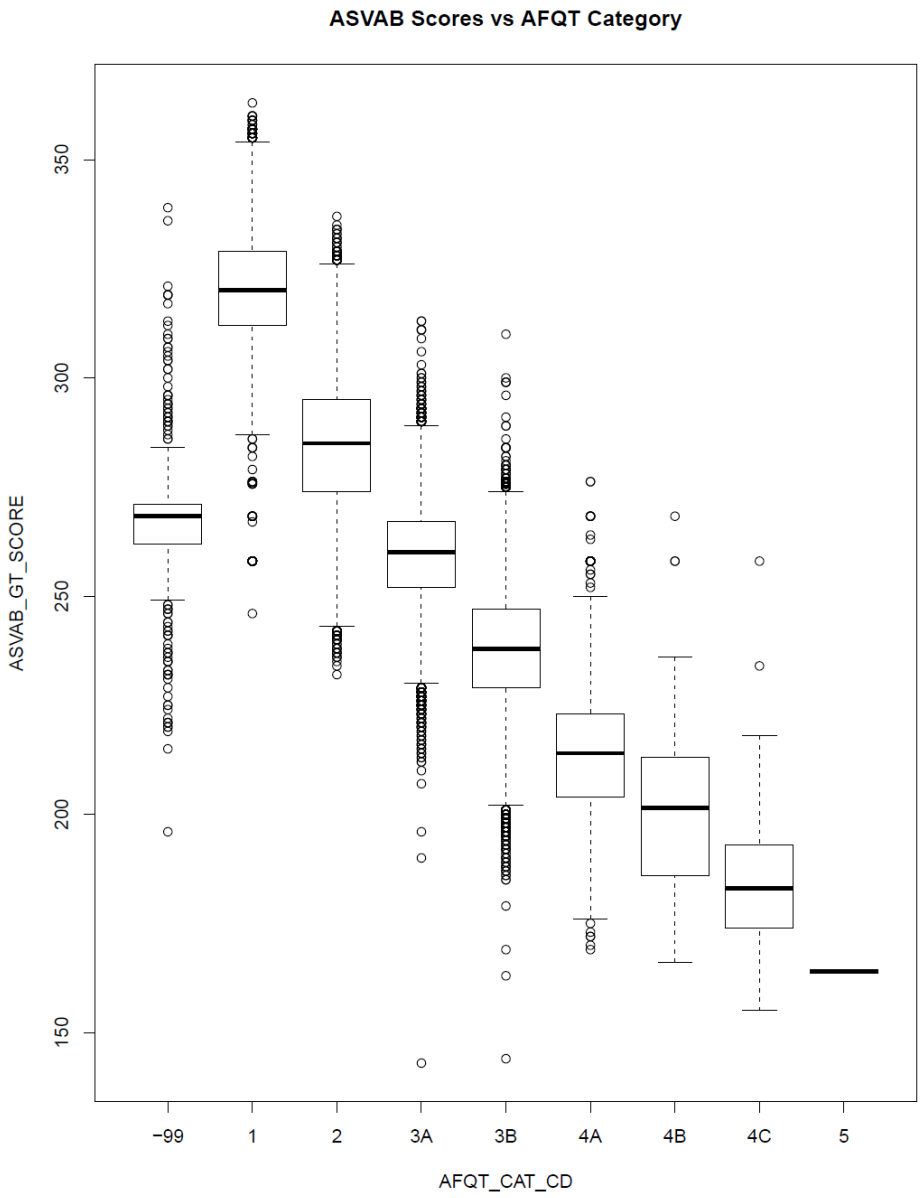


Figure 7. Boxplot of ASVAB GT Score vs. AFQT Category Code

There is also a high VIF and correlation between ethnicity affinity code and race code. Since race does not indicate if one is Hispanic or not, the variable Hispanic was created as a binary variable. We kept race code and Hispanic as our included variables and excluded ethnicity affinity code.

We also excluded fiscal year group since we wanted this model to be used for all year groups in the future. However, one must take note that fiscal year group was a strong predictor in attrition. There is something impacting the strength of year group as a predictor, possibly either economic conditions or policy decisions that may have influenced whether or not soldiers attritted.

C. VARIABLE IMPORTANCE AND MODEL SELECTION

In order to determine which variables are the most significant per year per contract duration, we used variable importance measures from decision trees. Decision trees serve as binary if-then trees. If something is true, then it splits down one side, if it is false, it splits down the other side. Decision trees are prone to overfitting noise in the training set; they can suffer from high variance, so we use bootstrap aggregation, also called bagging. Bagging is one version of random forest, in which a number of trees are constructed from bootstrapped data sets (Kirasich et al. 2018). Bootstrapping allows the computer to create many "new" datasets that are the same size as the original dataset by sampling from the original dataset with replacement. Averaging the bootstrapped datasets reduces variance. While bagging improves the accuracy of a prediction, it also reduces the ability to interpret the model. We are still able to interpret and measure the importance of each variable using relative influence plots.

We computed relative importance to decide which predictors are more effective in predicting our response variable attrit. The relative important plot gives a score to each predictor variable. The larger the score, the more influential it is, and the more important the predictor variable is. When the predictor variable's score is close to zero, then that variable can be dropped from the model.

We performed random forest model selection on the 18 datasets. The top 20 predictors for each dataset can be seen in Appendix C. In order to see whether soldiers' attrition behavior was more similar across years-in-contract or across contract duration, we took the union of the top 15 predictors by year in service and next did the same by contract duration. Based on these results, soldiers who attritted had more variables in common by year in contract than by contract duration. Thus, we chose to perform logistic regression on six models, which are based on the soldier's year in contract. Since we modeled based on year in contract, we added back in contract duration as a variable.

The next step was to merge datasets that had common years together and add in numeric values for contract duration. This gave us our six year-in-contract datasets: Year0, Year1, Year2, Year3, Year4, and Year5. We ran variable importance again through random forest. Not surprisingly, the union of the top 15 predictors previously mentioned gave results similar to those from the top 20 predictors of the merged year-in-contract dataset. Figure 8 shows the random forest variable importance graph for Year1. The graph shows the five most important predictors for Year1 are gender, weight at enlistment, height at enlistment, age group at enlistment, and dental readiness class.

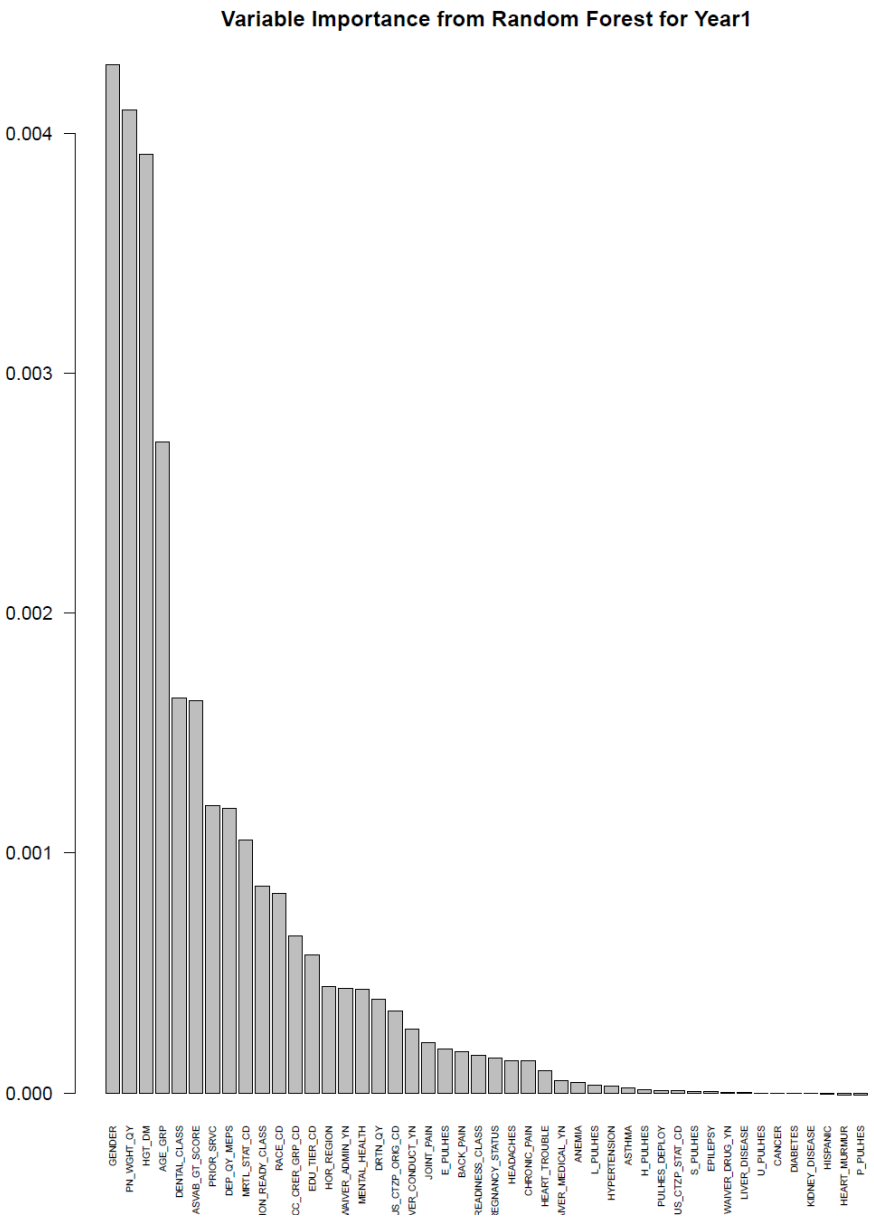


Figure 8. Variable Importance Plot for Year1

Each set of 20 predictors was put into a full logistic regression model for the respective year-in-contract dataset. The full model was then compared to simpler models using Akaike Information Criterion (AIC), which measures penalized prediction error, so it compares the quality of simpler models against the full model. Using forward selection allowed us to start with the intercept-only model and add in the best-performing predictors one at a time. We used the predictors that gave the lowest AIC, and they became our final predictors to go into each logistic regression model.

Tables 9–14 show the logistic regression summary output for each year in contract. Coefficients with greater magnitude have more influence on predicting the logit. Conversely, coefficients that are closer to zero, have less influence. Negative coefficients indicate that a soldier who falls into those categories has a decreased probability of attriting, while positive coefficients indicate a greater chance of attriting. The coefficient for each predictor variable gives us the predicted change in log odds for every one-unit increase, holding all other predictor variables constant. In this model, dental readiness class 3 and vision readiness class 3 have the largest coefficients in magnitude. This indicates soldiers who fall into either or both of those categories have a greater probability of attriting than those who do not. The intercept for the logistic regression model contains the "baseline" levels that do not show up as individual coefficients in the table. For example for Year0 the intercept contains: age group 17–19, gender female, three-year contract duration, dental readiness class 1, CMF group FS, vision readiness class 1, no administrative waiver, race category 12, HOR region Midwest, U.S. citizenship origin code A, not prior service, marital status divorced, and no conduct waiver.

The odds ratio tells us how the odds of attrition change for one unit of change in the predictor variable, holding all other variables constant. For example, the odds of a male soldier undergoing attrition is 0.439 times the odds for a female soldier in their first year of service, when all other variables are held constant. For numeric predictors, the interpretation is based on a unit of increase or decrease. For example, for every unit of increase in ASVAB GT score, the log odds of attrition increases by 0.006. This means there is a 0.6% increase in the odds of attrition for soldiers in their first year of enlistment for each additional ASVAB GT point. Conversely, for the second and third years of enlistment,

an increase in ASVAB GT score has a negative effect on attrition (see Tables 15 and 16). The standard error of the coefficient measures the variability of the estimate in the table. The *z*-value is the regression predictor variable coefficient divided by the coefficient's standard error. A large *z*-value in magnitude indicates high confidence that the "true" underlying predictor variable coefficient is not zero. This relates to Pr(>|z|), the p-value of the two-sided test that the coefficient is zero, where lower values indicate stronger confidence that the true value is non-zero.

Table 9. Logistic Regression Summary Output for Year0

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	–4.195	–	0.342	–12.262	0.000
AGE_GRP20-22	–0.224	0.800	0.032	–7.044	0.000
AGE_GRP23-25	–0.231	0.794	0.039	–5.900	0.000
AGE_GRP26-30	–0.086	0.918	0.046	–1.848	0.065
AGE_GRP31–35	–0.156	0.856	0.074	–2.105	0.035
AGE_GRP36+	–0.262	0.770	0.096	–2.720	0.007
EDU_TIER_CD	0.271	1.311	0.033	8.152	0.000
GENDERM	–0.824	0.439	0.040	–20.447	0.000
DRTN_QY4	–0.403	0.668	0.030	–13.215	0.000
DRTN_QY5	–0.736	0.479	0.048	–15.443	0.000
DRTN_QY6	–0.364	0.695	0.046	–7.985	0.000
PN_WGHT_QY	0.001	1.001	0.000	2.572	0.010
HGT_DM	0.001	1.001	0.005	0.108	0.914
DENTAL_CLASSD2	–0.994	0.370	0.063	–15.801	0.000
DENTAL_CLASSD3	3.507	33.356	0.065	54.231	0.000
CMF_GRP_CDMFE	–0.010	0.990	0.033	–0.318	0.751
CMF_GRP_CDNone	1.808	6.097	0.039	46.176	0.000
CMF_GRP_CDOS	0.299	1.349	0.040	7.512	0.000
ASVAB_GT_SCORE	0.006	1.006	0.000	13.366	0.000
VISION_READY_CLASSV2	1.645	5.180	0.238	6.905	0.000
VISION_READY_CLASSV3	2.497	12.152	0.111	22.455	0.000
WAIVER_ADMIN_YN	0.430	1.538	0.054	8.015	0.000
RACE_CDR3	–0.289	0.749	0.067	–4.281	0.000
RACE_CDR4	–0.008	0.992	0.061	–0.129	0.897
HOR_REGIONNortheast	0.000	1.000	0.044	–0.008	0.994

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
HOR_REGIONSouth	0.018	1.018	0.033	0.543	0.587
HOR_REGIONTerritory	–0.490	0.612	0.136	–3.615	0.000
HOR_REGIONWest	-0.212	0.809	0.038	-5.531	0.000
US_CTZP_ORIG_CDC	-0.254	0.775	0.057	-4.464	0.000
US_CTZP_ORIG_CDN	-0.063	0.939	0.083	-0.759	0.448
PRIOR_SRVC	0.221	1.247	0.076	2.886	0.004
MRTL_STAT_CDM	0.012	1.012	0.095	0.122	0.903
MRTL_STAT_CDN	-0.172	0.842	0.093	-1.850	0.064
DEP_QY_MEPS	-0.088	0.915	0.021	-4.291	0.000
WAIVER_CONDUCT_YN	-0.107	0.899	0.058	-1.835	0.067

Table 10. Logistic Regression Summary Output for Year1

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	–1.593	–	0.309	–5.149	0.000
AGE_GRP20-22	–0.209	0.812	0.026	–8.116	0.000
AGE_GRP23-25	–0.516	0.597	0.038	–13.671	0.000
AGE_GRP26-30	–0.527	0.590	0.048	–10.880	0.000
AGE_GRP31-35	–0.618	0.539	0.081	–7.611	0.000
AGE_GRP36+	–0.598	0.550	0.102	–5.867	0.000
EDU_TIER_CD	0.399	1.490	0.029	13.812	0.000
GENDERM	–0.690	0.501	0.036	–19.283	0.000
DRTN_QY4	–0.047	0.954	0.026	–1.805	0.071
DRTN_QY5	–0.050	0.951	0.042	–1.200	0.230
DRTN_QY6	–0.012	0.988	0.040	–0.300	0.764
PN_WGHT_QY	0.001	1.001	0.000	3.151	0.002
HGT_DM	0.001	1.001	0.005	0.297	0.767
PRIOR_SRVC	1.466	4.330	0.048	30.258	0.000
DENTAL_CLASSD2	–0.263	0.769	0.027	–9.756	0.000
DENTAL_CLASSD3	1.607	4.987	0.067	24.052	0.000
MENTAL_HEALTHY	1.580	4.857	0.089	17.683	0.000
VISION_READY_CLASSV2	0.510	1.665	0.053	9.581	0.000
VISION_READY_CLASSV3	0.770	2.161	0.078	9.897	0.000

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
JOINT_PAINY	0.446	1.562	0.051	8.752	0.000
US_CTZP_ORIG_CDC	–0.354	0.702	0.053	–6.707	0.000
US_CTZP_ORIG_CDN	–0.245	0.783	0.084	–2.907	0.004
HOR_REGIONNortheast	0.075	1.077	0.040	1.862	0.063
HOR_REGIONSouth	0.028	1.028	0.030	0.926	0.354
HOR_REGIONTerritory	–0.580	0.560	0.136	–4.246	0.000
HOR_REGIONWest	–0.103	0.902	0.035	–2.996	0.003
ASVAB_GT_SCORE	–0.004	0.996	0.000	–7.550	0.000
WAIVER_CONDUCT_YN	0.270	1.310	0.046	5.862	0.000
CMF_GRP_CDMFE	0.009	1.009	0.026	0.348	0.728
CMF_GRP_CDNone	–0.135	0.873	0.058	–2.318	0.020
CMF_GRP_CDOS	–0.061	0.940	0.036	–1.694	0.090
MRTL_STAT_CDM	–0.234	0.791	0.096	–2.440	0.015
MRTL_STAT_CDN	–0.229	0.796	0.096	–2.391	0.017
DEP_QY_MEPS	–0.033	0.967	0.019	–1.746	0.081
RACE_CDR3	0.090	1.094	0.062	1.456	0.145
RACE_CDR4	0.110	1.117	0.058	1.905	0.057

Table 11. Logistic Regression Summary Output for Year2

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	–0.444	–	0.292	–1.519	0.129
AGE_GRP20-22	–0.234	0.791	0.025	–9.365	0.000
AGE_GRP23-25	–0.512	0.599	0.035	–14.599	0.000
AGE_GRP26-30	–0.525	0.592	0.043	–12.069	0.000
AGE_GRP31-35	–0.564	0.569	0.069	–8.137	0.000
AGE_GRP36+	–0.776	0.460	0.094	–8.250	0.000
GENDERM	–0.320	0.726	0.037	–8.705	0.000
DRTN_QY4	0.138	1.148	0.025	5.528	0.000
DRTN_QY5	0.188	1.206	0.039	4.812	0.000
DRTN_QY6	0.199	1.220	0.037	5.346	0.000
PN_WGHT_QY	0.006	1.006	0.000	14.121	0.000
HGT_DM	–0.027	0.974	0.004	–5.977	0.000
CHRONIC_PAINY	0.890	2.436	0.054	16.525	0.000
PULHES_DEPLOY	1.463	4.321	0.065	22.572	0.000
BACK_PAINY	0.389	1.475	0.037	10.617	0.000
MENTAL_HEALTHY	0.911	2.487	0.065	14.008	0.000

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
PRIOR_SRVC	1.321	3.747	0.056	23.521	0.000
PREGNANCY_STATUSY	0.947	2.578	0.076	12.465	0.000
ASVAB_GT_SCORE	–0.003	0.997	0.000	–6.842	0.000
HOR_REGIONNortheast	0.027	1.027	0.039	0.686	0.493
HOR_REGIONSouth	0.076	1.079	0.029	2.587	0.010
HOR_REGIONTerritory	–0.451	0.637	0.121	–3.709	0.000
HOR_REGIONWest	–0.079	0.924	0.033	–2.364	0.018
US_CTZP_ORIG_CDC	–0.225	0.798	0.047	–4.796	0.000
US_CTZP_ORIG_CDN	–0.440	0.644	0.086	–5.122	0.000
JOINT_PAINY	0.278	1.320	0.039	7.120	0.000
CMF_GRP_CDMFE	0.069	1.071	0.025	2.716	0.007
CMF_GRP_CDNone	–0.107	0.898	0.056	–1.908	0.056
CMF_GRP_CDOS	–0.167	0.846	0.036	–4.699	0.000
HEADACHESY	0.259	1.295	0.048	5.352	0.000
RACE_CDR3	0.175	1.191	0.058	3.008	0.003
RACE_CDR4	0.060	1.062	0.055	1.097	0.272
MRTL_STAT_CDM	–0.203	0.816	0.078	–2.611	0.009
MRTL_STAT_CDN	–0.194	0.824	0.078	–2.485	0.013

Table 12. Logistic Regression Summary Output for Year3

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	–0.927	–	0.412	–2.248	0.025
AGE_GRP20-22	–0.120	0.887	0.037	–3.243	0.001
AGE_GRP23-25	–0.274	0.760	0.048	–5.704	0.000
AGE_GRP26-30	–0.391	0.676	0.059	–6.620	0.000
AGE_GRP31-35	–0.547	0.579	0.094	–5.845	0.000
AGE_GRP36+	–0.691	0.501	0.120	–5.736	0.000
GENDERM	–0.439	0.645	0.051	–8.533	0.000
DRTN_QY5	0.311	1.365	0.041	7.528	0.000
DRTN_QY6	0.399	1.490	0.039	10.248	0.000
PN_WGHT_QY	0.008	1.008	0.001	13.510	0.000
HGT_DM	–0.030	0.970	0.007	–4.629	0.000
CHRONIC_PAINY	0.998	2.713	0.061	16.492	0.000
PULHES_DEPLOY	1.524	4.590	0.099	15.406	0.000
MENTAL_HEALTHY	0.824	2.279	0.072	11.450	0.000
PRIOR_SRVC	1.219	3.383	0.098	12.399	0.000

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
EDU_TIER_CD	0.422	1.525	0.044	9.574	0.000
BACK_PAINY	0.362	1.437	0.047	7.688	0.000
CMF_GRP_CDMFE	0.169	1.184	0.038	4.424	0.000
CMF_GRP_CDNone	–0.154	0.857	0.074	–2.072	0.038
CMF_GRP_CDOS	–0.060	0.942	0.046	–1.307	0.191
S_PULHESS2	1.090	2.974	0.406	2.685	0.007
S_PULHESS3	1.536	4.648	0.269	5.707	0.000
S_PULHESS4	1.856	6.399	1.178	1.576	0.115
ASVAB_GT_SCORE	–0.003	0.997	0.001	–5.633	0.000
JOINT_PAINY	0.240	1.271	0.050	4.844	0.000
U_PULHESU2	0.676	1.967	0.137	4.933	0.000
U_PULHESU3	–0.066	0.936	0.148	–0.448	0.654
MRTL_STAT_CDM	–0.146	0.864	0.086	–1.703	0.089
MRTL_STAT_CDN	–0.162	0.851	0.087	–1.869	0.062

Table 13. Logistic Regression Summary Output for Year4

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	0.343	–	0.709	0.484	0.628
AGE_GRP20-22	–0.074	0.929	0.063	–1.165	0.243
AGE_GRP23-25	–0.214	0.807	0.081	–2.644	0.008
AGE_GRP26-30	–0.377	0.686	0.104	–3.601	0.000
AGE_GRP31–35	–0.066	0.936	0.154	–0.429	0.668
AGE_GRP36+	–0.439	0.645	0.223	–1.962	0.049
DRTN_QY6	0.389	1.476	0.052	7.380	0.000
PN_WGHT_QY	0.012	1.012	0.001	11.748	0.000
HGT_DM	–0.057	0.945	0.011	–5.126	0.000
GENDERM	–0.375	0.687	0.093	–4.021	0.000
CHRONIC_PAINY	0.837	2.309	0.085	9.752	0.000
PULHES_NONDEPLOY	1.337	3.808	0.218	6.112	0.000
MENTAL_HEALTHY	0.793	2.210	0.102	7.699	0.000
BACK_PAINY	0.396	1.486	0.069	5.705	0.000
RACE_CDR3	0.166	1.181	0.166	0.997	0.318
RACE_CDR4	0.437	1.548	0.151	2.890	0.003
ASVAB_GT_SCORE	–0.005	0.995	0.001	–4.253	0.000
HEADACHESY	0.237	1.267	0.091	2.581	0.009
S_PULHESS2	0.661	1.937	0.714	0.925	0.354

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
S_PULHESS3	1.710	5.529	0.467	3.656	0.000
S_PULHESS4	–0.122	0.885	1.663	–0.073	0.941
JOINT_PAINY	0.177	1.194	0.072	2.444	0.014
U_PULHESU2	0.528	1.696	0.192	2.740	0.006
U_PULHESU3	0.378	1.459	0.244	1.548	0.121
P_PULHESP2	0.242	1.274	0.157	1.539	0.123
P_PULHESP3	0.496	1.642	0.245	2.020	0.043
HOR_REGIONNortheast	0.116	1.123	0.103	1.121	0.262
HOR_REGIONSouth	0.154	1.166	0.075	2.053	0.040
HOR_REGIONTerritory	–0.504	0.604	0.298	–1.691	0.090
HOR_REGIONWest	0.117	1.124	0.084	1.379	0.167
DEP_QY_MEPS	–0.061	0.941	0.036	–1.667	0.095
US_CTZP_ORIG_CDC	–0.236	0.790	0.125	–1.883	0.059
US_CTZP_ORIG_CDN	–0.186	0.830	0.209	–0.888	0.374

Table 14. Logistic Regression Summary Output for Year5

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	–0.988	–	1.010	–0.978	0.328
AGE_GRP20-22	–0.084	0.920	0.094	–0.888	0.375
AGE_GRP23-25	–0.205	0.815	0.116	–1.769	0.077
AGE_GRP26-30	–0.466	0.627	0.144	–3.243	0.001
AGE_GRP31-35	–0.246	0.782	0.201	–1.225	0.220
AGE_GRP36+	–0.727	0.484	0.283	–2.570	0.010
PN_WGHT_QY	0.010	1.010	0.001	6.423	0.000
HGT_DM	–0.051	0.950	0.017	–3.072	0.002
GENDERM	–0.163	0.850	0.155	–1.054	0.292
CHRONIC_PAINY	0.970	2.638	0.102	9.536	0.000
PULHES_DEPLOY	1.967	7.151	0.177	11.096	0.000
MENTAL_HEALTHY	0.803	2.232	0.131	6.107	0.000
PRIOR_SRVC	1.920	6.821	0.280	6.848	0.000
HEADACHESY	0.440	1.553	0.118	3.721	0.000
L_PULHESL2	0.497	1.644	0.251	1.983	0.047
L_PULHESL3	–1.375	0.253	0.564	–2.439	0.015
RACE_CDR3	0.336	1.400	0.244	1.380	0.168
RACE_CDR4	0.509	1.664	0.220	2.311	0.021
HOR_REGIONNortheast	–0.223	0.800	0.172	–1.298	0.194

Predictor Variable	Coefficient	Odds Ratio	Std. Error	z value	Pr(>|z|)
HOR_REGIONSouth	0.049	1.050	0.109	0.447	0.655
HOR_REGIONTerritory	–0.930	0.394	0.533	–1.744	0.081
HOR_REGIONWest	0.097	1.102	0.126	0.775	0.439

D. MODEL DIAGNOSTICS AND PERFORMANCE

We performed model diagnostics to determine if our models were a good fit and tested the models' performance. The first diagnostic test was to test for multicollinearity using computed VIFs. As a rule of thumb, a VIF value that exceeds five or ten can indicate a problematic level of collinearity (James et al. 2017 p. 101). Our final models for each year-in-contract dataset had coefficients with VIF's all lower than five. Table 15 shows the VIFs and degrees of freedom for Year2.

Table 15. Variance Inflation Factors for Year2

Coefficient	VIF	Df
AGE_GRP	1.32	5
GENDER	1.90	1
DRTN_QY	1.19	3
PN_WGHT_QY	1.64	1
HGT_DM	2.11	1
CHRONIC_PAIN	1.25	1
PULHES_NONDEPLOY	1.01	1
BACK_PAIN	1.28	1
MENTAL_HEALTH	1.10	1
PRIOR_SRVC	1.00	1
PREGNANCY_STATUS	1.11	1
ASVAB_GT_SCORE	1.31	1
HOR_REGION	1.16	4
US_CTZP_ORIG_CD	1.15	2
JOINT_PAIN	1.26	1
OCC_CRER_GRP_CD	1.34	3
HEADACHES	1.21	1
RACE_CD	1.39	2
MRTL_STAT_CD	1.17	2

Cook's distance identifies potential observations that serve as influential outliers. The greater the value of Cook's distance, the more those observations could be outliers. We examined if observations were true outliers, such as the P_PULHES category 4 in Year4, then we the data was reorganized to fit into the nearest level. Figure 9 shows the graph of Cook's distance for all the observations in dataset Year2 based on the model. The three greatest values are identified by observation number. In the final model of Year2, there were no influential points, as the three observations were not outliers.

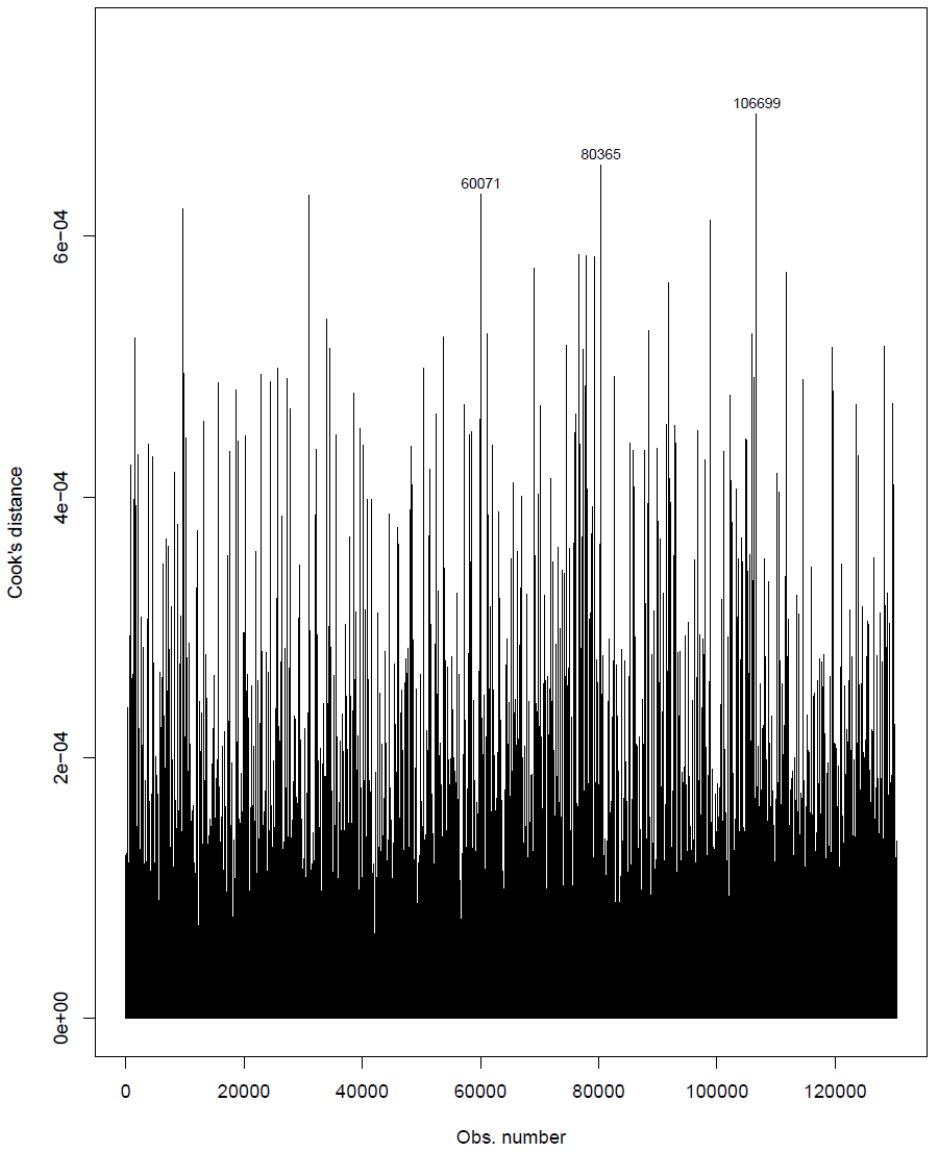


Figure 9. Cook's Distance for Year2

One measure of performance of each model can be seen in a confusion matrix. It shows how the actual not attrit and attrit numbers in the test set vs the number of not attrit and attrit predicted by the model. Ideally, our models would accurately predict probabilities of attrition, producing high predicted probabilities of attrition for soldier who undergo attrition, the TP block. However, since attrition is a rare event, it is not very surprising that almost all the predicted probabilities are quite small, smaller than the usual 0.5 cutoff. In this case, we chose to model using the proportion of soldiers who attrit in the training set. However, the models can still be useful if they help rank-order soldiers by probability of attrition, because identifying the soldiers most at risk provides useful information to commanders, manpower analysts and recruiters.

Tables 16–21 show the confusion matrices for Year0 to Year5.. In Year0, the top left block of 33,694 is the true negative (TN) number of soldiers that did not attrit and were predicted not to attrit. The upper right number of 1,315 is the number of false positive (FP) outcomes: for these soldiers, the model predicted attrit while the soldier did not in fact attrit. The lower left block of 1,258 is the number of false negative (FN) outcomes. These arise when the model predicted not attrit for a soldier who did attrit. The true positive (TP) 692, on the lower left block, represents instances where the model predicted attrit and the soldier did attrit. The accuracy of the model can be measured by the ratio of TP + TN to the total number of soldiers in the table. A higher accuracy indicates a better model. The accuracies of models Year0, Year1, Year2, Year3, Year4, and Year5 are 93.0%, 88.9%, 87.3%, 86.9%, 86.8%, and 86.4% respectively.

Table 16. Confusion Matrix for Year0

Year0	**Predicted Class**		
Actual Class		Not Attrit	Attrit
	Not Attrit	33694	1315
	Attrit	1258	692

Table 17. Confusion Matrix for Year1

Year1	**Predicted Class**		
Actual Class		Not Attrit	Attrit
	Not Attrit	30653	1939
	Attrit	1940	435

Table 18. Confusion Matrix for Year2

Year2	**Predicted Class**		
Actual Class		Not Attrit	Attrit
	Not Attrit	27886	2037
	Attrit	2115	558

Table 19. Confusion Matrix for Year3

Year3	**Predicted Class**		
Actual Class		Not Attrit	Attrit
	Not Attrit	11996	928
	Attrit	942	352

Table 20. Confusion Matrix for Year4

Year4	**Predicted Class**		
Actual Class		Not Attrit	Attrit
	Not Attrit	4142	320
	Attrit	330	141

Table 21. Confusion Matrix for Year5

Year5	**Predicted Class**		
Actual Class		Not Attrit	Attrit
	Not Attrit	1901	151
	Attrit	156	54

The boxplots shown in Figures 10–15 are drawn from the test set. They show soldiers grouped in order of highest to lowest predicted probably of attrition. For example, in Year0, the first (leftmost) group shows the 150 soldiers with the highest predicted probability of attrition, based on the model. Each boxplot is overlaid by a red dot that shows the actual proportion of soldiers in that group who underwent attrition. When the dot is below the center of the boxplot, the average of the predicted probabilities in that group was larger than it should have been, and when the dot is above, the average was smaller. The lift number in red at the bottom of the graph is the ratio of the number of soldiers who attrit in each specified group to the expected number of soldiers who attrit in a group of the same size selected at random.

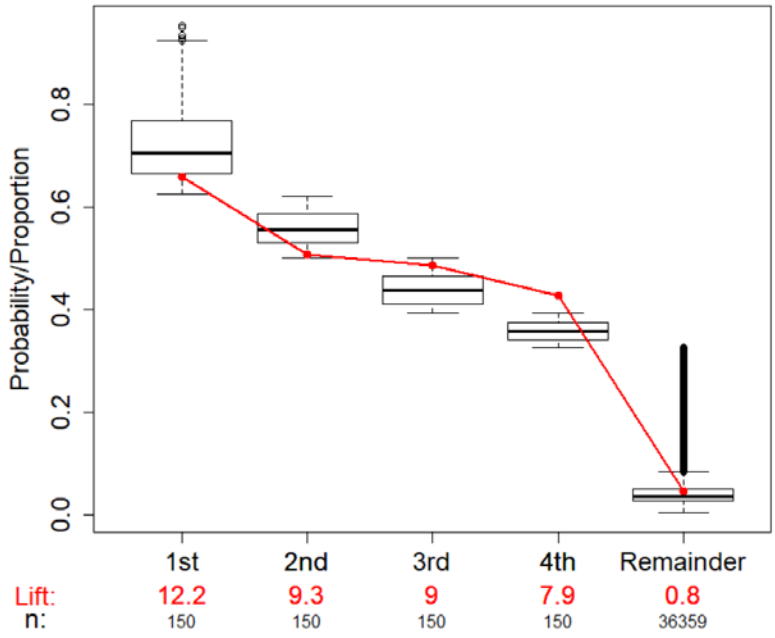


Figure 10. Observed and Predicted Attrition, Grouped by Predicted Probability for Year0

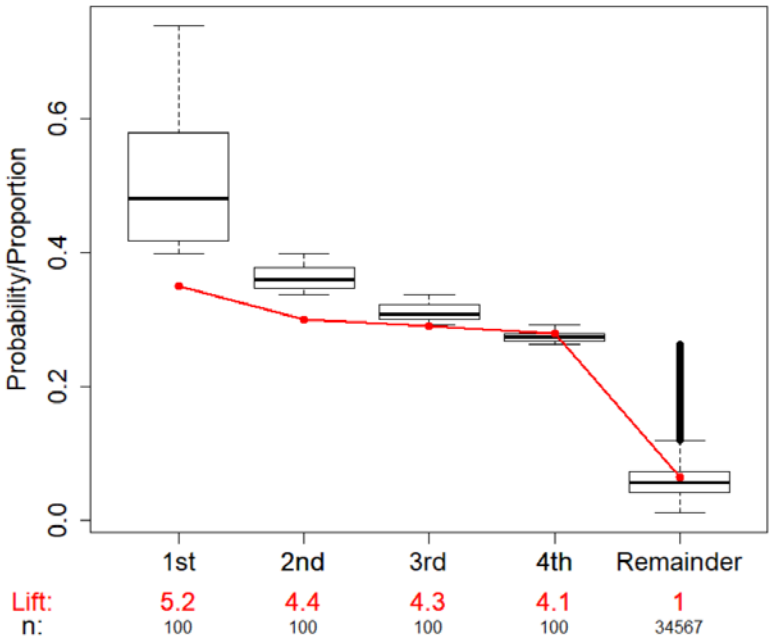


Figure 11. Observed and Predicted Attrition, Grouped by Predicted Probability for Year1

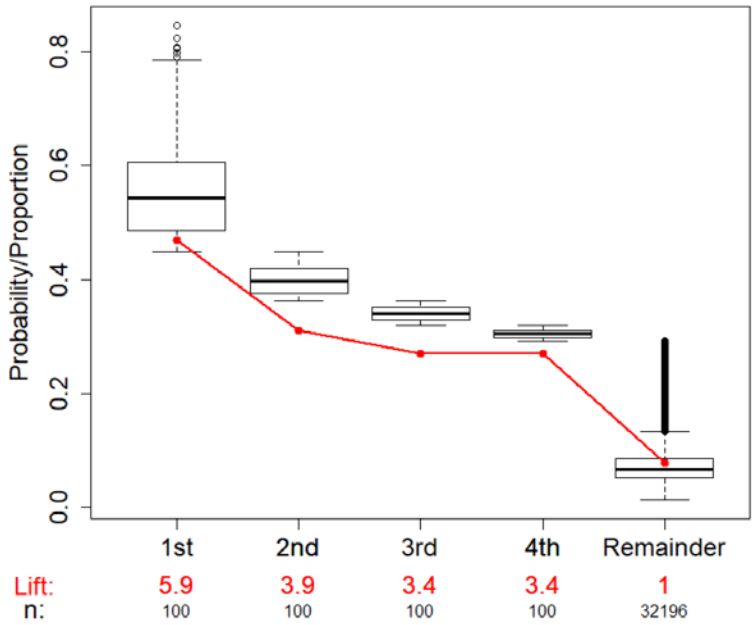


Figure 12. Observed and Predicted Attrition, Grouped by Predicted Probability for Year2

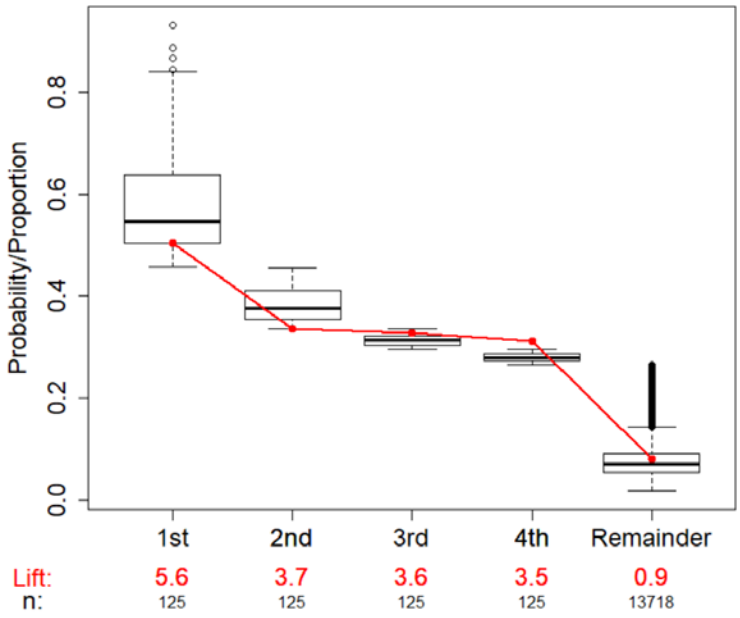


Figure 13. Observed and Predicted Attrition, Grouped by Predicted Probability for Year3

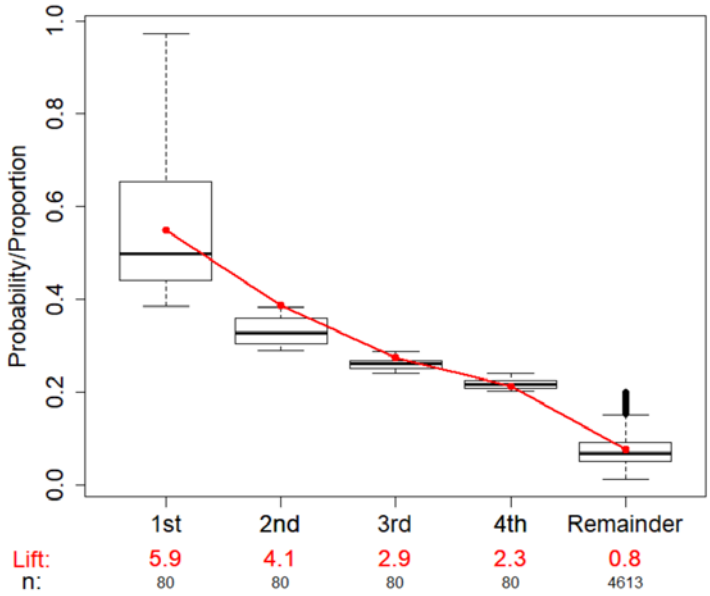


Figure 14. Observed and Predicted Attrition, Grouped by Predicted Probability for Year4

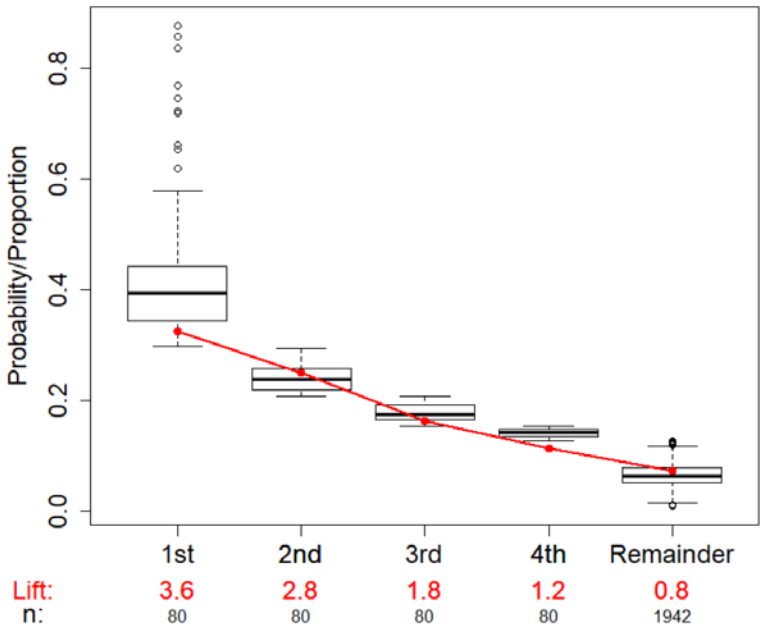


Figure 15. Observed and Predicted Attrition, Grouped by Predicted Probability for Year5

E. ANALYSIS OF FINDINGS

We discovered earlier soldiers who attrit have more variables in common by year in contract than by their contract duration; thus we chose six models by year in contract. The changes of predictor variables tell us modeling by year is important to get a true picture of what variables are important for soldier attrition.

There are similarities and differences of the predictor variables across all years of service. Across all years except Year5, contract duration continues to be an important predictor. Within the first two years of a soldier's enlistment, soldiers with a four-, five-, or six-year enlistment have a smaller chance of attriting than soldier with a three year contract. Conversely, during the third, fourth, and fifth year in service, the higher the contract duration, the greater chance of attriting. Age groups 26–30, 31–35, and 35+ have a lower probability of attriting. The 17–19 age group, which is in the intercept, has the highest probability of attriting. Prior service, as found in previous research, continues to be a strong predictor. Those who have prior service have a higher rate of attrition than those who do not. Weight is a predictor variable in all six models. It suggests that the heavier a

soldier is, the more chance there is for them to attrit. For the later years, a soldier's height suggests that the taller a soldier is, the less chance there is for them to attrit. Males have a lower probability of attriting than females in all six years.

Education tier code and ASVAB GT score are both indicators in the initial years of a soldier's contract, but by Year5, they have both fallen off as predictor variables. This may be due to the fact that jobs require six-year contracts are more specialized and those soldiers undergo more education in the Army, where they learn different problem solving techniques and life skills.

One note of interest is the dental readiness class and vision readiness class. Since we redefined the use medical readiness class 4, it only shows up as an important predictor variable for the first two years of service, whereas before removing level 4, it showed up in most of the 18 per term per year models. Dental readiness class 3 is the strongest predictors of attrition for the first two years. This may be because dental hygiene is an everyday procedure. Soldiers who cannot take care of their teeth possibly lack discipline, which is very much needed in the Army.

As the year in contract increases, effects of demographic indicators generally decrease and the effects of medically-related indicators generally increase. When the binary predictor variables of mental health, pregnancy status, joint pain, chronic pain, back pain, headaches, and PULHES nondeployable are present, soldiers have a higher rate of attrition. Chronic pain and PULHES nondeployable were two of the highest magnitude coefficients across all years in which they were present. Higher levels of S_PULHES indicated a greater chance of attrition. Conversely, higher levels of U_PULHES and L_PULHES did not indicate higher levels of attrition.

V. CONCLUSION

One of the key areas of research for this thesis was to see if constructing multiple logistic regressions would generate the ability to model attrition using time-varying covariates. After cleaning the data set, we organized the master dataset by breaking up the dataset into six groups, one for every year in contract. Variable importance for each of the datasets was determined using the random forest, and AIC was used as a further indicator for variable selection. We discovered that soldiers who attrit have more variables in common by year in contract versus their contract duration, thus we chose to produce models based on year in contract. We were able to use time-varying covariates to fit a logistic regression model to each of the six datasets to predict attrition one year at a time, given their survival for the previous year. Different time-varying covariates, such as a soldier's medical condition and marriage status, were identified as important predictor variables of attrition depending on the year in contract.

A. RECOMMENDATION FOR COMMANDS

These models can help command identify soldiers as highest risk of attrition. The calculations and output of this model will have to be computed by someone who has access to medical and demographic data. The PDE has such information, but since the PDE goes by PID, the analyst cannot put a name to the PID. Once commands use the models, names of identified soldiers can be given to retention NCOs and leaders within the company to help discover reasons why a soldier wants to leave the military and craft a plan to help the soldier stay or reenlist. Manpower analysts can use this to determine what recruiting efforts are needed in order to keep the force at a certain level. Recruiters can also get a sense of who will more likely last for their entire first term of enlistment.

Just like Duckworth's work of using grit as an indicator of attrition in the Army Special Operations Selection Course, a statistic of a soldier's grit can be determined and collected. The Army has soldiers take the Global Assessment Tool (GAT), which assess their physical and psychological health based on social, emotional, spiritual, family, and

physical dimensions. Questions from the GAT may be able to determine grit, or questions can be added to sections to make a comprehensive grit assessment.

B. RECOMMENDATIONS FOR FURTHER RESEARCH

Conducting this research has led us to more questions and identified information gaps that could be useful as attrition predictors. Previous studies and interviews have shown that a potential predictor of military service is knowing or being related to someone who has served in the military (Philipps et al. 2020). The percentage of recruits who have a prior relationship with the military is increasing. While the Army would like to recruit from all of America's population, those who have family who have been in the military may be more apt to stay in the military and may have a more successful military career as a result of receiving guidance from those family members. One such example includes recruits from Fayetteville, NC, which is home to Fort Bragg. Even though Manhattan has eight times as many people than Fayetteville, Fayetteville brought in twice as many military enlistments (Philipps et al. 2020). (However, this data for Fayetteville includes soldiers who have decided to stay for second- and third-term enlistments). Information on family history of military service might well be a useful predictor of attrition.

Further research should also attempt to determine why females attrit more than males, with emphasis on the dual military spouse. Whether a soldier is part of the Married Army Couples Program (MACP) may possibly be an indicator of attrition. The military tries its best to keep couples together, but sometimes that does not happen. Combined with the number of dependents, a difficult childcare situation, and housing situation, for some couples enrolled in MACP, it has proven to be difficult for both service members to stay in the military (Grisales 2019). The one who departs the military is usually the female service member.

Dental, vision, hearing class 4 needs to be examined every time one conducts attrition analysis. Especially now, during the 2019 Novel Coronavirus (COVID-19) pandemic, soldiers are not going into dental offices and medical clinics for routine checkups. Appointments are made for emergency medical issues only. More soldiers will be classified as class 4 for any three of these medical readiness classes if their last checkup

was more than a year ago. Thus, class 4 should not be taken at face value as a strong predictor.

Fiscal year of accession was an important predictor variable before we removed it so that the model could be used for future predictions. This may indicate that Army policy or extermal economic conditions for each year may play an important role in whether a soldier decides to attrit or not. We may see an increase in retention and possibly recruitment due to the impact of COVID-19 economic downturn. On the other hand, other soldiers may not reenlist if they feel a strong sense of duty to be closer to home to help their family. External factors, such as the job market and the nation's will to be engaged during specific fiscal years should be researched to see why attrition varies for that year.

In a few years, analysts will be able to see if the incorporation of new tests and assessment contribute to the decrease of attrition. The Occupation Physical Assessment Test (OPAT) and High Physical Demands Test (HPDT), decreases the amount of attrition due to physical reasons. The OPAT, which started in 2017, is given to recruits to assess their physical fitness (Lopez 2016). The HPDT is conducted during AIT and the tasks are MOS-specific and gender-neutral (McIntyre 2016). Any soldier who wishes to change MOS will need to pass the MOS-specific HPDT before doing so. Lastly, the three-year pilot study of the Tailored Adaptive Personality Assessment System (TAPAS), which started in FY2020, aims to identify potential soldiers who will outperform what their AFQT score suggests, may also predict potential performance, behaviors, and attrition of recruits (Brading 2020). TAPAS scores may, along with AFQT, be another good indicator of attrition.

New data will also be collected with the Department of the Army Career Engagement Survey (DACES). Soldiers will be asked to take this survey every year during their birth month and when they out-process. The survey aims to let Soldiers communicate their concerns anonymously and direct to Army leadership (Army Talent Management Task Force 2020). Soldiers will also be able to request more information about career opportunities and be contacted further if they choose to be. The information gathered from DACES will help inform retention efforts, which directly effects Army first-term attrition.

Along with the addition of OPAT, HPDT, and TAPAS testing during recruitment and AIT, the DACES may provide necessary indicators of attrition and retention.

APPENDIX A. ETHNICITY AFFINITY CODE

Levels	Description
AA	Asian Indian
AB	Chinese
AC	Filipino
AD	Guamanian (Chamorros)
AF	Japanese
AG	Korean
AI	Vietnamese
AJ	Other Asian descent
AK	Mexican
AL	Puerto Rican
AM	Cuban
AN	Latin American
AO	Other Hispanic descent
AP	Aleut
AQ	Eskimo
AR	U.S./Canadian Indian tribes
AS	Melanesian
AT	Micronesian
AU	Polynesian
AV	Other Pacific island descent
BG	Other
BH	None [Not associated with any particular ethnic affinity]

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. HOME OF RECORD STATE

Levels	Description
AL	Alabama
AK	Alaska
AZ	Arizona
AR	Arkansas
CA	California
CO	Colorado
CT	Connecticut
DE	Delaware
DC	District of Columbia
FL	Florida
GA	Georgia
HI	Hawaii
ID	Idaho
IL	Illinois
IN	Indiana
IA	Iowa
KS	Kansas
KY	Kentucky
LA	Louisiana
ME	Maine
MD	Maryland
MA	Massachusetts
MI	Michigan
MN	Minnesota
MS	Mississippi
MO	Missouri
MT	Montana
NE	Nebraska
NV	Nevada
NH	New Hampshire
NJ	New Jersey
NM	New Mexico
NY	New York
NC	North Carolina
ND	North Dakota

Levels	Description
OH	Ohio
OK	Oklahoma
OR	Oregon
PA	Pennsylvania
RI	Rhode Island
SC	South Carolina
SD	South Dakota
TN	Tennessee
TX	Texas
UT	Utah
VT	Vermont
VA	Virginia
WA	Washington
WV	West Virginia
WI	Wisconsin
WY	Wyoming
AS	American Samoa
GU	Guam
PR	Puerto Rico
VI	Virgin Islands

APPENDIX C. TOP 20 PREDICTORS FOR 18 DATASETS

CD3_Y0	CD3_Y1	CD3_Y2
CMF Group	Height at Enlistment	Weight at Enlistment
Dental Class	Gender	Height at Enlistment
ASVAB GT Score	Weight at Enlistment	Gender
Age Group at Enlistment	Age Group at Enlistment	Age Group at Enlistment
Gender	ASVAB GT Score	ASVAB GT Score
Height at Enlistment	Dependents at Enlistment	Dependents at Enlistment
Weight at Enlistment	Dental Class	Race Code
Marital Status	Prior Service	CMF Group
Education Tier	Marital Status	PULHES Deployable
Vision Class	Vision Class	Marital Status
Race Code	Race Code	Prior Service
Dependents at Enlistment	CMF Group	Back Pain
Admin Waiver	Education Tier	Chronic Pain
Conduct Waiver	Mental Health	US Citizenship Origin Code
Home of Record Region	US Citizenship Origin Code	Mental Health
US Citizenship Origin Code	Home of Record Region	Joint Pain
Prior Service	Admin Waiver	Dental Class
US Citizenship Status Code	Joint Pain	Pregnancy Status
Hispanic	Back Pain	Home of Record Region
Drug Waiver	Hearing Class	Hearing Class

CD4_Y0	CD4_Y1	CD4_Y2	CD4_Y3
Dental Class	Gender	Weight at Enlistment	Weight at Enlistment
Gender	Height at Enlistment	Height at Enlistment	Height at Enlistment
Height at Enlistment	Weight at Enlistment	Gender	Gender
Weight at Enlistment	Age Group at Enlistment	Age Group at Enlistment	PULHES Nondeployable
Age Group at Enlistment	Dental Class	ASVAB GT Score	ASVAB GT Score
Vision Class	Dependents at Enlistment	Race Code	Age Group at Enlistment
ASVAB GT Score	Prior Service	Chronic Pain	Chronic Pain
Race Code	ASVAB GT Score	Dependents at Enlistment	Mental Health
Dependents at Enlistment	Marital Status	Marital Status	Race Code
Marital Status	Race Code	Pregnancy Status	Dependents at Enlistment
Admin Waiver	Admin Waiver	Mental Health	Marital Status
Education Tier	Vision Class	Home of Record Region	U_PULHES
Home of Record Region	Home of Record Region	PULHES Nondeployable	P_PULHES
CMF Group	Education Tier	Joint Pain	S_PULHES
Prior Service	E_PULHES	CMF Group	US Citizenship Origin Code
US Citizenship Origin Code	Mental Health	Back Pain	Education Tier
Conduct Waiver	US Citizenship Origin Code	Prior Service	Prior Service
US Citizenship Status Code	CMF Group	Education Tier	Back Pain
Medical Waiver	Conduct Waiver	US Citizenship Origin Code	Headaches
Hispanic	Headaches	Conduct Waiver	Home of Record Region

CD5_Y0	CD5_Y1	CD5_Y2
Dental Class	Weight at Enlistment	Height at Enlistment
Gender	Gender	Weight at Enlistment
Height at Enlistment	Height at Enlistment	Gender
Weight at Enlistment	Age Group at Enlistment	Age Group at Enlistment
Age Group at Enlistment	ASVAB GT Score	CMF Group
ASVAB GT Score	Dental Class	ASVAB GT Score
CMF Group	CMF Group	Dependents at Enlistment
Vision Class	Prior Service	Chronic Pain
Dependents at Enlistment	Dependents at Enlistment	Joint Pain
Admin Waiver	Marital Status	Headaches
Marital Status	Vision Class	Back Pain
US Citizenship Origin Code	Race Code	Race Code
Education Tier	Admin Waiver	PULHES Nondeployable
Race Code	Education Tier	Marital Status
Prior Service	Pregnancy Status	Mental Health
Conduct Waiver	Home of Record Region	Prior Service
Home of Record Region	Conduct Waiver	Home of Record Region
Medical Waiver	Headaches	Admin Waiver
Hearing Class	Joint Pain	Pregnancy Status
US Citizenship Status Code	US Citizenship Origin Code	Heart Trouble

CD5_Y3	CD5_Y4
Weight at Enlistment	Weight at Enlistment
Gender	Height at Enlistment
Height at Enlistment	Mental Health
Age Group at Enlistment	PULHES Nondeployable
Chronic Pain	Age Group at Enlistment
PULHES Nondeployable	Chronic Pain
ASVAB GT Score	P_PULHES
CMF Group	Gender
Dependents at Enlistment	ASVAB GT Score
Mental Health	U_PULHES
Marital Status	Joint Pain
P_PULHES	Back Pain
Joint Pain	CMF Group
U_PULHES	S_PULHES
Headaches	Race Code
Back Pain	Home of Record Region
Heart Trouble	Admin Waiver
Race Code	Headaches
Conduct Waiver	L_PULHES
Medical Waiver	Prior Service

CD6_Y0	CD6_Y1	CD6_Y2
Dental Class	Weight at Enlistment	Height at Enlistment
Weight at Enlistment	Height at Enlistment	Weight at Enlistment
Gender	Gender	Gender
Height at Enlistment	Age Group at Enlistment	Age Group at Enlistment
Age Group at Enlistment	Prior Service	Chronic Pain
Dependents at Enlistment	Dental Class	Dependents at Enlistment
CMF Group	Vision Class	Joint Pain
Marital Status	ASVAB GT Score	Marital Status
ASVAB GT Score	Dependents at Enlistment	Back Pain
Prior Service	Marital Status	Race Code
Education Tier	Admin Waiver	ASVAB GT Score
Admin Waiver	CMF Group	Mental Health
US Citizenship Origin Code	Race Code	Prior Service
Race Code	Education Tier	Headaches
Vision Class	Mental Health	CMF Group
Home of Record Region	Home of Record Region	Admin Waiver
Conduct Waiver	E_PULHES	PULHES Nondeployable
US Citizenship Status Code	US Citizenship Origin Code	Education Tier
Drug Waiver	Anemia	E_PULHES
E_PULHES	Medical Waiver	Home of Record Region
CD6_Y3	**CD6_Y4**	**CD6_Y5**
Weight at Enlistment	Weight at Enlistment	Weight at Enlistment
Height at Enlistment	Height at Enlistment	Height at Enlistment
Gender	Chronic Pain	Chronic Pain
Chronic Pain	Joint Pain	PULHES Nondeployable
Age Group at Enlistment	PULHES Nondeployable	Dependents at Enlistment
Dependents at Enlistment	Age Group at Enlistment	Age Group at Enlistment
PULHES Nondeployable	Dependents at Enlistment	Marital Status
Prior Service	S_PULHES	Mental Health
U_PULHES	Gender	Prior Service
Marital Status	Mental Health	Joint Pain
Mental Health	U_PULHES	Headaches
Pregnancy Status	Back Pain	U_PULHES
S_PULHES	P_PULHES	CMF Group
Back Pain	Race Code	Home of Record Region
P_PULHES	Headaches	S_PULHES
Race Code	CMF Group	P_PULHES
ASVAB GT Score	Home of Record Region	Gender
Anemia	Pregnancy Status	Back Pain
L_PULHES	Heart Trouble	US Citizenship Origin Code
Heart Trouble	L_PULHES	L_PULHES

LIST OF REFERENCES

Arkin W (2019) Fewer Americans want to serve in the military. Cue Pentagon panic. *The Guardian* (April 10). https://www.theguardian.com/commentisfree/2019/apr/10/fewer-americans-serve-military-pentagon-panic.

Army Analytics Group (2016) Supplemental information: Person-event data environment. Version 2, Army Analytics Group, Fairfield, CA.

Army Talent Management Task Force (2020) Soldiers sound off to Army leadership in new survey. Army.mil (May 7). https://www.army.mil/article/235366/soldiers_sound_off_to_army_leadership_in_new_survey.

Brading T (2020) New entrance test to increase Soldier quality, reduce attrition. Army.mil (January 6). https://www.army.mil/article/231249/new_entrance_test_to_increase_soldier_quality_reduce_attrition.

Buddin R (1984) Analysis of early military attrition behavior. Report R-3069-MIL, RAND Corporation, Santa Monica, CA. https://doi.org/10.7249/R3069.

Burgess L (2007) Army reaches recruiting goals at increasing costs to taxpayers. *Stars and Stripes* (September 6). https://www.stripes.com/news/army-reaches-recruiting-goals-at-increasing-costs-to-taxpayers-1.68551.

Devig, A. L. (2019) Predicting U.S. Army enlisted attrition after initial entry training using survival analysis. Master's thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA. http://hdl.handle.net/10945/62725.

Eskreis-Winkler L, Shulman E, Beal S, Duckworth A (2014) The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology* 5(36), https://doi.org/10.3389/fpsyg.2014.00036.

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

GAO (1997) Military attrition: DOD could save millions by better screening enlisted personnel. Report GAO/NSIAD-97-39, United States General Accounting Office, Washington, DC, https://www.gao.gov/assets/160/155698.pdf.

Gobea, G.A. (2019) Predicting U.S. Army first-term attrition after initial entry training, part II. Master's thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA. http://hdl.handle.net/10945/64167.

Grisales C (2019) Military recruitment, retention challenges remain, service chiefs say. *Stars and Stripes* (May 16). https://www.stripes.com/military-recruitment-retention-challenges-remain-service-chiefs-say-1.581364.

Hosmer D, Lemeshow S (2000) *Applied Logistic Regression*, 2nd ed. (John Wiley and Sons, New York, NY).

James G, Witten D, Hastie T, Tibshirani R (2017) *An Introduction to Statistical Learning* (Springer, New York).

Lopez C (2016) Army to administer four-part OPAT to recruits. Army.mil (June 3). https://www.army.mil/article/168882/army_to_administer_four_part_opat_to_recruits.

Mattis, J. (2018) Summary of the 2018 national defense strategy of the United States of America. Department of Defense, https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf.

McIntyre C (2016) Army tests new fitness standards at Fort Sill. Army.mil (November 3). https://www.army.mil/article/177774/army_tests_new_fitness_standards_at_fort_sill.

Myers M (2019). After 2018's recruiting shortfall, it will take a lot longer to build the Army to 500K. *Army Times* (March 14). https://www.armytimes.com/news/your-army/2019/03/14/after-2018s-recruiting-shortfall-it-will-take-a-lot-longer-to-build-the-army-to-500k/.

Olick D (2002). An Army of one carries a high price. CNBC (October 21). http://www.nbcnews.com/id/3072945/t/army-one-carries-high-price/#.XmH1rahKguU.

Person-Event Data Environment (2019) Accessed February 24, 2020, https://pde.army.mil/Main.

Philipps D, Arango T (2020) Who signs up to fight? Make up of U.S. recruits shows flaring disparity. *The New York Times* (January 14). https://www.nytimes.com/2020/01/10/us/military-enlistment.html.

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Accessed January 21, 2020, https://www.R-project.org/.

RAND (1985) RAND research brief: Analysis of early military attrition behavior. RB-2001-2, RAND Corporation, Santa Monica, CA. https://doi.org/10.7249/RB2001-2.

Rempfer K (2019) New in 2020: What you should know about the Army's recruiting push in the new year. *Army Times* (December 30). https://www.armytimes.com/news/your-army/2019/12/30/new-in-2020-what-you-should-know-about-the-armys-recruiting-push-in-the-new-year/

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

Tice, J (2016) Army recruiting market tightens but service expect to make 2016 goal. *Army Times* (February 23). https://www.armytimes.com/news/your-army/2016/02/23/army-recruiting-market-tightens-but-service-expects-to-make-2016-goal/.

Vanden Brook, T (2015) Army hits its target for recruits in 2015. *USA Today* (October 2015). https://www.usatoday.com/story/news/nation/2015/10/01/army-recruiting/73166250/.

Wunderlin S, Roos L, Roth R, Faude O, Frey F, Wyss T (2015) Trunk muscle strength tests to predict injuries, attrition and military ability in soldiers. *The Journal of Sports Medicine and Physical Fitness* 55(5):535–543.

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