NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PREDICTING U.S. ARMY ENLISTED ATTRITION AFTER INITIAL ENTRY TRAINING USING RANDOM SURVIVAL FORESTS

by

Nicholas R. Lazzarevich

March 2022

Thesis Advisor: Samuel E. Buttrey
Second Reader: Lyn R. Whitaker

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE March 2022	3. REPORT TYPE AND DATES COVERED Master's thesis

4. TITLE AND SUBTITLE PREDICTING U.S. ARMY ENLISTED ATTRITION AFTER INITIAL ENTRY TRAINING USING RANDOM SURVIVAL FORESTS	5. FUNDING NUMBERS
6. AUTHOR(S) Nicholas R. Lazzarevich	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	10. SPONSORING / MONITORING AGENCY REPORT NUMBER

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

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

13. ABSTRACT (maximum 200 words)

The U.S. Army requires models that predict the proportion of post–Initial Entry Training (IET) soldiers who complete their initial term of service, and which assess the risk of attrition prior to completion at various points during these terms. The Army struggles to access sufficient recruits to maintain approved personnel levels due to economic competition and a shrinking population of candidates who are both willing and eligible for recruitment. Roughly 24% of soldiers who complete IET fail to complete their initial term of service. Modeling post-IET attrition and identifying factors that contribute to attrition will allow the Army to access soldiers with lower risk of attrition and assess policies to address attrition throughout the initial term. Continuing work done by Devig in 2019 with survival analysis, this research utilizes the randomForestSRC R package by Ishwaran and Kogalur in 2020 to build a series of random survival forests, allowing us to approximate effects of time-varying covariates (TVC). This research uses data stored in the Person-Event Data Environment and consists of demographics, deployments, medical readiness, and initial entry data. Using fiscal year (FY) 2010 as a training set and FY 2011 as a test set, we find that two of the top 10 predictors are medical while the rest are demographic, and four are TVC. The final models perform well for predicting cohort attrition at various points during the first term, but not for attrition of individuals.

14. SUBJECT TERMS Army, enlisted, attrition, survival analysis, machine learning, medical data, predict, random forests, random survival forests, Person-Event Data Environment	15. NUMBER OF PAGES 99
	16. PRICE CODE

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

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

PREDICTING U.S. ARMY ENLISTED ATTRITION AFTER INITIAL ENTRY TRAINING USING RANDOM SURVIVAL FORESTS

Nicholas R. Lazzarevich
Major, United States Army
BSCE, Bucknell University, 2008
MS, Missouri University of Science and Technology, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2022

Approved by: Samuel E. Buttrey
Advisor

Lyn R. Whitaker
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The U.S. Army requires models that predict the proportion of post–Initial Entry Training (IET) soldiers who complete their initial term of service, and which assess the risk of attrition prior to completion at various points during these terms. The Army struggles to access sufficient recruits to maintain approved personnel levels due to economic competition and a shrinking population of candidates who are both willing and eligible for recruitment. Roughly 24% of soldiers who complete IET fail to complete their initial term of service. Modeling post-IET attrition and identifying factors that contribute to attrition will allow the Army to access soldiers with lower risk of attrition and assess policies to address attrition throughout the initial term. Continuing work done by Devig in 2019 with survival analysis, this research utilizes the randomForestSRC R package by Ishwaran and Kogalur in 2020 to build a series of random survival forests, allowing us to approximate effects of time-varying covariates (TVC). This research uses data stored in the Person-Event Data Environment and consists of demographics, deployments, medical readiness, and initial entry data. Using fiscal year (FY) 2010 as a training set and FY 2011 as a test set, we find that two of the top 10 predictors are medical while the rest are demographic, and four are TVC. The final models perform well for predicting cohort attrition at various points during the first term, but not for attrition of individuals.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AAG	Army Analytics Group
ACT-MAST	Active Duty Military Personnel Master
ACT-TRAN	Active Duty Personnel Transaction
AFQT	Armed Forces Qualification Test
AIT	Advanced Individual Training
ASVAB	Armed Services Vocational Aptitude Battery
AWD	Army Waiver Database
BCT	Basic Combat Training
BMI	Body Mass Index
CMF	Career Management Field
CTC-OCO	Contingency Tracking System – Overseas Contingency Operations
DCIPS	Defense Casualty Information Processing System
DOD	Department of Defense
FY	Fiscal Year
GT	General Technical
HOR	Home of Record
IET	Initial Entry Training
KM	Kaplan-Meier
MEDPROS	Medical Protection System
MEPCOM	Military Entrance Processing Command
MOS	Military Occupational Specialty
OOB	Out-of-Bag
PDE	Person-Event Data Environment
PHA	Periodic Health Assessment
PULHES	Physical, Upper, Lower, Hearing, Eyesight, Psychiatric
PID	Personal Identification
PII	Personally Identifiable Information
RFL	Research Facilitation Laboratory
TVC	Time-Varying Covariates
USAREC	U.S. Army Recruiting Command

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Every year, the U.S. Army must recruit 60,000 to 80,000 soldiers to its ranks, and every year 24% of soldiers who complete Initial Entry Training (IET) will fail to complete their first term of service (Devig 2019). In 2018, the Army failed to meet its recruitment goal of 76,500, which itself was a reduction from an original goal of 80,000, and it spent $429 million in enlistment bonuses trying to achieve the mission (South 2019). While it surpassed its recruitment goal in 2019, the COVID-19 pandemic forced the Army to reduce its 2020 recruitment goal from 68,000 to 62,000, which it narrowly met (Dickstein 2020). Now facing a fiscally constrained environment, the Army must seek to retain the soldiers currently serving as well as recruiting future soldiers in order to meet current and future end strength goals.

This research used the Person-Event Data Environment (PDE) to access, analyze, and model the data. The PDE is a cloud-based data repository created by the Army Analytics Group (AAG) and its Research Facilitation Laboratory (RFL) in 2006 to store sensitive soldier, family member, veteran, and Department of Defense (DOD) civilian employee data (Vie et al. 2013, p. 1). The data is stored in numerous relational databases using personal identification numbers (PIDs) as keys.

This research used a cohort built by Devig (2019), which included soldiers who started service from fiscal year (FY) 2005 through FY 2011 and were in the ranks of Private (E1) through Staff Sergeant (E6) at the time of enlistment. Devig's (2019) cohort started with 488,971 soldiers, of which 465,822 successfully completed IET. For this research, we used FY 2010 as the training set and FY 2011 as the test set, due to large amounts of missing medical data in previous years. FY 2010 and FY 2011 have a combined 124,002 unique soldiers. Overall, this research used 67 different covariates, including 36 constant covariates and 31 time-varying covariates (TVCs). Both constant and time-varying covariates come in one of two forms: numeric or categorical. This research used six numeric covariates, 32 categorical variables with more than two factor levels, and 29 binary categorical variables. Our research replaced dental, hearing, and vision readiness class 4

data, which was included by Devig (2019), but constitutes future knowledge. The final cohort attrition rate is 24.27%, slightly below Devig's (2019) rate of 24.46%.

Our research fits a random survival forest that grows multiple survival trees and then uses the average of the predictions from all the member trees to make a prediction for a given entry. Specifically, we use the randomForestSRC R (R Core Team 2017) package of Ishwaran and Kogalur (2020). The survival tree analysis used by Devig (2019) allowed for time-varying covariates, but random survival forests did not allow for them at the time this research began. We approximate the effects of TVCs by stratifying our cohort by term length and year of term, growing 18 different random forests and then stitching them together.

Using the threshold measure developed by Ishwaran et al. (2010) for variable selection, we identify the importance variables for inclusion in each model, ranging from 25 to 49 variables. After tuning model parameters for performance, we used sampling with replacement and grew production forests with 850 trees each, a minimum of 10 observations per node, up to three random split points per variable, and considering only 75 out of a possible 365 unique split times per model.

Analysis of the final models' terms show that the 10 most important variables are Career Management Field (CMF), both at enlistment and after IET, dental readiness class, Body Mass Index (BMI) and age at enlistment, AFQT category, Home of Record (HOR) state, PULHES lower extremities after IET, marital status, and gender. Only two of these variables are medical, and four of them are TVCs. Devig (2019) also found that dental readiness was highly important, but he included dental class 4s. Our models do not provide good prediction of attrition for individuals, but perform well for predicting attrition of cohorts.

References

Devig A (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.

Dickstein C (2020) Army hits 2020 recruiting, retention goals amid pandemic, but top officials say more diversity needed. Stars and Stripes. Accessed March 29, 2021, https://www.stripes.com/news/us/army-hits-2020-recruiting-retention-goals-amid-pandemic-but-top-officials-say-more-diversity-needed-1.648068.

Ishwaran H, Kogalur U (2020) Fast unified random forests for survival, regression, and classification (RF-SRC). R package version 2.9.3, https://cran.r-project.org/package=randomForestSRC.

Ishwaran H, Kogalur U, Gorodeski E, Minn A, Lauer M (2010) High-dimensional variable selection for survival data. *Journal of the American Statistical Association* 105(489):205-217, https://doi.org/10.1198/jasa.2009.tm08622.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

South T (2019) Rising costs, dwindling recruit numbers, increasing demands may bring back the military draft. *Military Times*. Accessed March 29, 2021, https://www.militarytimes.com/news/your-military/2019/11/19/rising-costs-dwindling-recruit-numbers-increasing-demands-may-bring-back-the-draft/.

Vie L, Griffith K, Scheier L, Lester P, Seligman M (2013) The Person-Event Data Environment: leveraging big data for studies of psychological strengths in soldiers. *Front. Psychol*. 4(934), https://doi.org/10.3389/fpsyg.2013.00934.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

Thank you to my family for sticking through NPS. Knowing everything that was going to happen, I don't know if we still would have chosen NPS, but I know that without you all, I never would have finished. Thank you to Dr. Sam Buttrey, first for agreeing to be my thesis advisor, and then remaining my advisor through my needed extensions. Your expertise, patience, and humor made this thesis possible. Finally, thank you to the entire OR department for continuing to support my efforts long after my physical departure from NPS.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Every year, the U.S. Army must recruit 60,000 to 80,000 soldiers to its ranks, and every year 24% of soldiers who complete Initial Entry Training will fail to complete their first term of service (Devig 2019). In 2018, the Army failed to meet its recruitment goal of 76,500, which itself was a reduction from an original goal of 80,000, and it spent $429 million in enlistment bonuses trying to achieve the mission (South 2019). While it surpassed its recruitment goal in 2019, the COVID-19 pandemic forced the Army to reduce its 2020 recruitment goal from 68,000 to 62,000, which it narrowly met (Dickstein 2020). Now facing a fiscally constrained environment, the Army must seek to retain the soldiers currently serving as well as recruiting future soldiers in order to meet current and future end strength goals.

A. PURPOSE

New Army recruits conduct Initial Entry Training (IET), which consists of Basic Combat Training (BCT), Advanced Individual Training (AIT), and potentially follow-on training. IET lasts anywhere from a few months to over a year. Soldiers who complete IET will continue in service until they complete their contracted term and separate, reenlist for an additional contracted term, or attrit from service prior to completing their contracted term. In order to maintain Congressionally approved end-strength levels, the Army must balance recruiting, retention, retirement, and attrition.

To account for IET and post-IET first-term attrition, the Army recruits larger cohorts than otherwise required. This requires additional resources, such as money and recruiters. Every soldier who serves as a recruiter is a soldier who cannot serve in the deployable force such as in a brigade combat team or security force assistance brigade.

The Army defines an accession as "a recruit who signs a contract and ships to initial entry training" (U.S. Army Recruiting Command [USAREC] 2021). In fiscal year (FY) 2012, the Army accession mission was 58,000 (USAREC 2021). It successfully accessed 60,490 recruits using 7,896 recruiters, which means each recruiter accounted for roughly

eight accessions (USAREC 2021). Each accession cost the Army roughly $22,300, which is the estimated cost of only accession, not other costs such as training or recruit pay while in training (USAREC 2021). Assuming that 12% of recruits will fail to complete IET (Power 2019), that 24% of recruits who do complete IET will fail to complete their first-term, and that the Army accounted for attrition when establishing the accession mission, this means the Army may have only required 38,790 accessions. The assumed additional 19,210 accessions cost the Army roughly $428M and took 2,401 soldiers out of the deployable Army to serve as recruiters.

This research seeks to identify administrative, demographic, and medical factors that predict first-term attrition. Specifically, this research will use survival analysis and random forests to identify these factors at various points during a soldier's first term.

B. RELATED WORK

This research continues the work of Speten (2018), Gobea (2019), Devig (2019), and Cammack (2020). Currently, there are no available studies of military attrition utilizing random survival forests. Therefore, we will provide a brief overview of survival analysis and random forests following a review of previous research.

1. Previous Research

Speten (2018) used administrative and demographic soldier data from FY 2005 through FY 2010 to predict first-term attrition using a logistic regression model. He identified that roughly 24% of first-term soldiers who complete IET will attrit prior to completing their contract. His model found that deployment history, initial contract length, and marital status were the most important factors in predicting first-term attrition. Speten did not have access to soldier medical data and did not account for time-varying covariates (TVC) in his model. Instead, he captured values for these variables at the start and end of the soldier's first term. Of note, Speten's work included an in-depth review of previous military attrition studies.

Gobea (2019) also used a logistic regression model, but gained access to soldier medical data for inclusion in modeling. He removed deployment history (which Speten

identified as the most important predictor) from his modeling since this information is not available at the beginning of a soldier's first term and cannot be used to predict attrition. He also narrowed his data to FY 2008 through FY 2010 due to a large proportion of missing medical data in previous years. Together with Devig, Gobea improved on the method for determining if a soldier attrits. His model found that PULHES non-deployable, dental class, and initial contract length were the most important factors in predicting first-term attrition. Like Speten, he did not account for TVCs in his model, instead using only start values for these variables.

Devig (2019) worked from the same data as Gobea, but used a survival tree for modeling. Significantly, the survival tree allowed him to account for time-varying covariates. He found that dental class, vision class, and PULHES codes were the most important factors in predicting first-term attrition. Of note, Devig's work included a review of previous military attrition studies that used survival analysis. Devig specifically recommended future work using survival random forest classification for military attrition.

Cammack (2020) worked from the same data as Devig and Gobea. She continued the work of Gobea by utilizing logistic regression and also incorporating TVCs, accomplished by taking snapshots at the start of each year of a contract term. Cammack used data from FY 2009 through 2011. She initially stratified by both contract term length (3–6 years) and year of term (year 0–5), but eventually stratified by only year of term based on variable importance. She found contract duration, age group, prior service, weight at enlistment, height at enlistment, and gender are all strong predictors of attrition across the years. She also found that, in general, demographic predictors became less important and medical predictors became more important as year increased, with chronic pain, PULHES nondeployable, and PULHES psychiatric being the strongest predictors. Of note, Cammack recoded all class 4 ratings for dental, vision, and hearing readiness. A class 4 indicates a soldier has not had a check within the last year, but does not provide information of their actual class. She recoded these to the observed class level prior to becoming a class 4, resulting in dental and vision readiness being important predictors in only the first two years, instead of all six years as they were with class 4 included.

2. Survival Analysis and Random Forests

Survival analysis is used to analyze the amount of time until an event or events occur. Originally developed to study time until death in largely medical related fields, other disciplines also found survival analysis useful to describe the distribution of times until an event. For this research, the event is soldier attrition from the Army during the first term. Sprent and Smeeton (2007) provide background on survival analysis. Survival data often includes right-censored entries. An entry is considered right-censored if at the observed time the event of interest (here, attrition) has not been seen, so that the full lifetime is unknown. In our research, because we fit separate models for different terms lengths, the data is not right-censored.

The Kaplan-Meier (KM) estimator is a nonparametric estimator of the survival function (Kaplan and Meier 1958). The survival function *S(t)* is the probability of "surviving" past time t, that is, the probability that the event of interest has not occurred by time *t*. Figure 1 shows the KM curves of the survival functions of male and female soldiers from FY 2010 who were 18 years old at enlistment and in the 15-series Career Management Field (CMF), which is aviation. The vertical axis is the proportion of the population alive at time *t*, and the horizontal axis is time in years since enlistment. Every time a soldier attrits from service, the plot steps down to reflect the new lower proportion still "alive." The flat portion at the beginning of the plots represents IET; we removed IET attritions from the data. The plots represent 292 females and 2,392 males. As the population size increases, the estimator becomes smoother and approaches the actual survival function. This is observed in the relative smoothness of the male KM curve versus the stepped female KM curve.

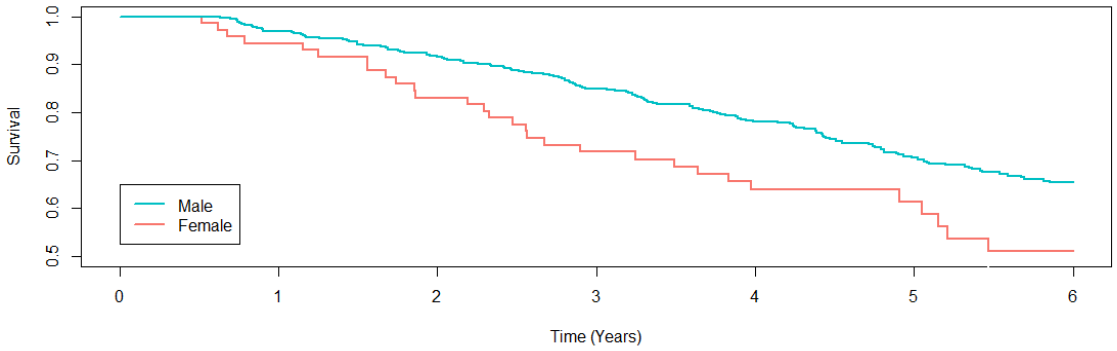

Figure 1. Example Kaplan-Meier Curves

A classification tree is a binary tree grown by recursive splits, with all observations starting in a root node and then progressing down diverging branches to final terminal nodes, or leaves (Ishwaran et al. 2008). At each node, data splits on a yes or no question, such as "is the soldier male," and then follows the corresponding branch. Splits are chosen to increase node purity while maintaining other parameters such as minimum node size and maximum tree depth. If the response variable is whether a soldier attrits, then an ideal split results in all soldiers who attrit in one terminal node, and all soldiers who do not attrit in another terminal node. Trees present an easy to understand graphic of the model, they do not require complex mathematical expressions to transform numeric variables or include variable interactions, and they are robust to both noise and outliers. Trees tend to have low bias, but high variance. This makes them excellent candidates for ensemble models such as random forests, because ensemble models maintain the low bias while reducing the variance and improving prediction (Ishwaran et al. 2008).

Figure 2 is an example classification tree for FY 2010 soldiers where attrition is the response variable and term length, gender, age, weight, and height are the predictor variables. The bubble at each node shows the predicted class ("0" equals not attrit), the predicted probability of attrition, and the percentage of observations in that node. For the bottom left terminal node, which are soldiers with a 3- or 4-year term length, the node predicts not attrit with 21% predicted attrition and 83% of observations in the node. The bottom right node holds soldiers with a 6-year term length who weigh 217 pounds or more

at enlistment. It predicts attrit with 56% predicted attrition and 1% of observations. Of note, although age and height were also included, the tree never splits on them.

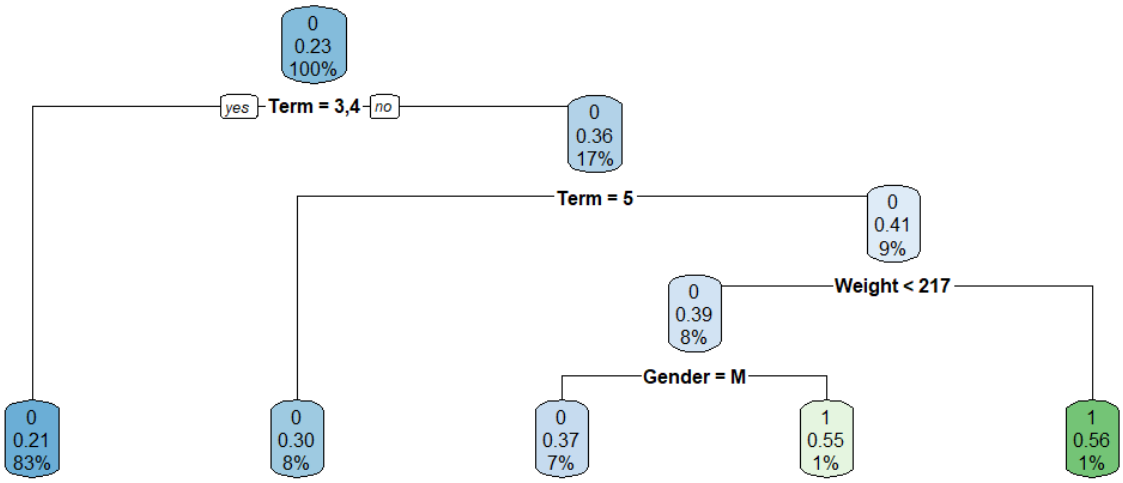

Figure 2. Example Classification Tree

In survival trees, split locations are chosen by log-rank, which is a statistic from a nonparametric hypothesis test that compares the survival distributions of the two sides of the potential shift (Ishwaran and Kogalur 2020). After calculating log-rank for all potential split locations, the algorithm chooses the split with the strongest evidence to support rejecting the null hypothesis. This may not choose a strong split, but it will choose the best available split.

Devig (2019) fit a model using a survival tree. In his model (see Figure 3), all entries are alive at time zero. In terms of a tree, this is the root or parent node. As the tree progresses, each line represents a path to a different terminal node in the tree, where the final estimated *S(t)* are the KM estimates based on the entries falling in each terminal node. The gray lines show the estimated survival curves for each leaf. The FY curves in color represent the aggregated KM curves for all soldiers in the respective FY. Survival trees produce strong, easily understood visuals, but they are sensitive to the dataset, and even small changes can produce significantly different results.

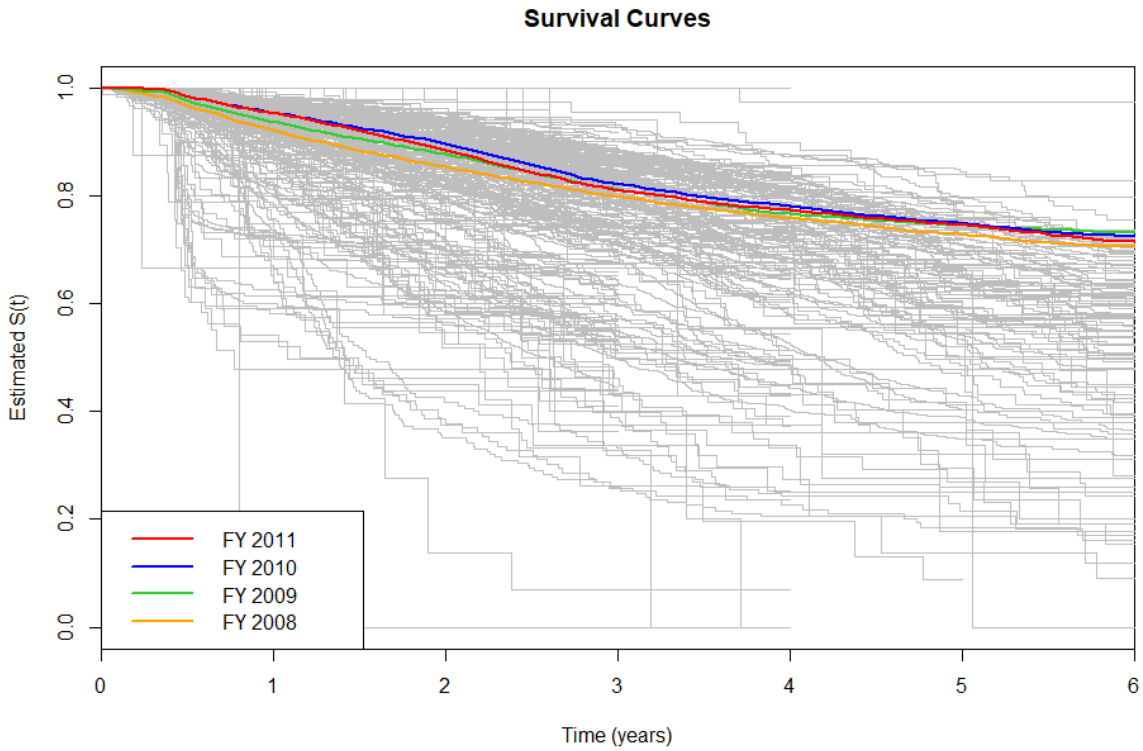

Figure 3. Example Classification Tree Survival Curves. Source: Devig (2019).

According to Breiman (2001), "random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest." As the name suggests, a random forest is a group of multiple trees, each grown using randomly sampled data from the original data. This reduces the sensitivity to the data seen in tree models and allows for training the model without requiring an explicit validation dataset.

The algorithm accomplishes randomized sampling of observations through bootstrap aggregating, also called bagging (Ishwaran et al. 2008). Bagging samples the original dataset with replacement, producing a sample dataset of the same size as the original dataset. Sampling with replacement results in some observations appearing multiple times, while others do not appear at all. Unselected observations are said to be Out-of-Bag (OOB). Each tree in a forest will have its own OOB sample that serves as a validation set for that tree. Random feature selection for splitting further randomizes the

model. Deterministic splitting considers every possible split point of every feature. For high dimensional data, deterministic splitting is computationally intensive and time consuming. Instead, at each node the algorithm randomly selects a set number of features and considers them for splitting. For classification trees, the number of features randomly selected defaults to the square root of the total number of features. If features are numeric or categorical with a large number of factors, users can also randomly choose a given number of split locations within each feature to further reduce time and computation.

Our research fits a random survival forest that grows multiple trees, similar to Devig's in Figure 3, and then uses the average of the predictions from all the member trees to make a prediction for a given entry. Specifically, we use the randomForestSRC R (R Core Team 2017) package of Ishwaran and Kogalur (2020). The survival tree analysis used by Devig (2019) allowed for time-varying covariates, but random survival forests did not allow for them at the time this research began. We will approximate the effects of TVCs by stratifying our cohort by term length and year of term, growing 18 different random forests and then stitching them together. We will discuss our modeling method in Chapter IV.

C. THESIS OUTLINE

This thesis is divided into five chapters. Chapter II provides an overview of the data and variables, to include how the cohorts were formed and how the response was constructed. It also discusses how the data was reshaped from a wide format to a long format and then broken into sub-cohorts. Chapter III provides descriptive statistics of the cohorts. Chapter IV discusses how we formed our model and assesses its performance. Finally, Chapter V summarizes the findings and provides recommendations for future work.

II. DATA INTRODUCTION

A. PERSON-EVENT DATA ENVIRONMENT

This research used the Person-Event Data Environment (PDE) to access, analyze, and model the data. The PDE is a cloud-based data repository created by the Army Analytics Group (AAG) and its Research Facilitation Laboratory (RFL) in 2006 to store sensitive soldier, family member, veteran, and Department of Defense (DOD) civilian employee data (Vie et al. 2013, p. 1). The data is stored in numerous relational databases using personal identification numbers (PIDs) as keys.

The PDE secures data in several ways. All PDE users must first receive authorization first to access the PDE in general, and then additional authorization for each requested database. Users log in to the PDE using a remote desktop, and cannot access the internet or copy and paste anything in or out of the desktop. Next, the PDE either removes or anonymizes personally identifiable information (PII) such as personal addresses, unit zip codes, and social security numbers. This research used the PIDs to combine multiple records from multiple datasets to form a master dataset for analysis.

The PDE provides several statistical tools for data analysis. This research exclusively used R (R Core Team 2017), which is an open-source statistical software package. It includes both a base statistical code and numerous add-in packages for in-depth statistical analysis. Although users cannot export any data from the PDE, they can request export of other products, such as code or plots. When requested, AAG reviews the products for sensitive information and then exports them to the user if none is present.

B. DATASETS USED

This research used the same datasets as Devig (2019) and Gobea (2019); six datasets contain administrative and demographic data, and two datasets contain medical data. Speten (2018) did not have access to the medical datasets.

The six administrative and demographic datasets are the Active Duty Military Personnel Master (ACT-MAST), the Active Duty Military Personnel Transaction (ACT-

TRAN), the Military Entrance Processing Command (MEPCOM), the Army Waiver Database (AWD), the Defense Casualty Information Processing System (DCIPS), and the Contingency Tracking System – Overseas Contingency Operations (CTS-OCO). The ACT-MAST dataset provides quarterly snapshots of soldier administrative data such as marital status, military occupation specialty (MOS), and rank. It also includes the date a soldier began military service and the duration of the term of service, which we use to determine the contractual end date for a soldier's first term. The ACT-TRAN dataset contains changes in a soldier's status, and includes important data for determining attrition such as reenlistment and separation codes. The MEPCOM dataset contains demographic data at the time of enlistment. This is useful for filling in missing soldier data in the early entries of other datasets. It also contains scores from the Armed Services Vocational Aptitude Battery (ASVAB). The AWD dataset contains data on whether the soldier received an enlistment waiver. The DCIPS dataset contains data on injuries incurred during deployments. Finally, the CTS-OCO dataset contains data on overseas deployments.

The two medical datasets are the Physical Health Assessment (PHA) and the Medical Protection System (MEDPROS). The PHA is actual two separate datasets, one for the old format and one for the new. The data from both was merged into a single dataset. The PHA is an annual assessment of a soldier's physical and mental condition. The MEDPROS dataset contains data on medical readiness and deployability status. Both Devig (2019) and Gobea (2019) found that the Physical/Upper/Lower/Hearing/Eyesight/ Psychiatric (PULHES) status in MEDPROS was important for predicting attrition. MEDPROS data updates whenever there is a change, not on an annual basis like the PHA data.

C. COHORT DESCRIPTION

The cohort built by Devig (2019) included soldiers who started service from fiscal year (FY) 2005 through FY 2011 and were in the ranks of Private (E1) through Staff Sergeant (E6) at the time of enlistment. A FY is defined as starting on 1 October of the previous year and ending on 30 September of the actual year. For example, FY 2010 started on 1 October 2009 and ended on 30 September 2010. Devig's (2019) cohort started with

488,971 soldiers. Of those soldiers, 465,822 successfully completed IET. For this research, we used FY 2010 as the training set and FY 2011 as the test set, due to large amounts of missing medical data in previous years. FY 2010 and FY 2011 have a combined 124,002 unique soldiers. Of note, a total of 1,183 soldiers (28 in FY 2010 and 22 in FY 2011), all attrits, were removed from Devig's (2019) cohort. These soldiers had missing end dates and end ages, making it impossible to conduct the data transformation needed for survival analysis.

D. METHODOLOGY

This section describes the covariates used in this research, the method used to construct the response variable, survival analysis specific variables and format, the research's limitations and assumptions, and formation of the training and test sets.

1. Variables Used

This research used two kinds of covariates, constant and time-varying covariates (TVC). Constant covariates remain the same throughout a soldier's term and include data such as gender, age at enlistment, race, and whether he or she has prior service. TVCs may change throughout a soldier's term and include data such as rank, marital status, and PULHES data. While TVCs may change, we may only use the information known at a given time. As an example, one cannot use the maximum rank achieved by a soldier in his or her first term to assess his or her likelihood of attrition at the beginning of the term. Overall, this research used 67 different covariates, including 36 constant covariates and 31 TVCs.

Both constant and time-varying covariates come in one of two forms: numeric or categorical. This research used six numeric covariates, all discrete. Next are categorical variables. If a categorical variable only has two factor levels, it is known as binary. This research used 61 categorical variables with a total of 336 different factor levels. Thirty-two categorical variables are multi-factor and 29 are binary. Variables are described in Table 1. Constructed means that a variable was built from existing data fields. Collapsed means that the number of factor levels of a variable were reduced by lumping small levels

together. Fiscal year accession and service agreement duration, both marked with "*" in Table 1, were used to stratify the data into training and test sets, but were not used as covariates for analysis. Additionally, variables marked with "**" are variables used in this research that were not included in Devig's (2019) research.

Table 1. Summary of Variables. Adapted from Devig (2019).

Variable	Type	Constructed	Collapsed	Factor Levels	Time-Varying
Administrative Waiver	Binary	No	No	2	No
AFQT Category Code	Categorical	No	Yes	6	No
Age at Enlistment	Numeric	Yes	No	N/A	No
Age Group at Enlistment	Categorical	Yes	No	4	No
Anemia	Binary	No	No	2	Yes
Asthma	Binary	No	No	2	Yes
ASVAB GT Score	Numeric	Yes	No	N/A	No
Back Pain	Binary	No	No	2	Yes
Blood Type	Categorical	No	No	8	No
Body Mass Index (BMI)**	Categorical	Yes	No	N/A	No
BMI Category**	Categorical	Yes	Yes	4	No
Cancer	Binary	No	No	2	Yes
Career Management Field (CMF) Code at Enlistment	Categorical	No	Yes	18	No
CMF Code after IET**	Categorical	No	Yes	18	Yes
CMF Functional Group	Categorical	Yes	No	3	No
Chronic Pain	Binary	No	No	2	Yes
Citizenship Origination Code	Categorical	No	Yes	4	No
Citizenship Status Code	Binary	No	Yes	2	No
Conduct Waiver	Binary	No	No	2	No
Dental Class	Categorical	No	No	4	Yes
Deployment**	Binary	Yes	Yes	2	Yes
Diabetes	Binary	No	No	2	Yes
Drug Waiver	Binary	No	No	2	No
Education Tier Code at Enlistment	Categorical	No	No	3	No
Epilepsy	Binary	No	No	2	Yes
Ethnic Affinity Code	Categorical	No	No	22	No
Faith Group Code	Categorical	No	Yes	53	No

Variable	Type	Constructed	Collapsed	Factor Levels	Time-Varying
Fiscal Year Accession*	Categorical	No	No	7	No
Gender	Binary	No	No	2	No
Headaches	Binary	No	No	2	Yes
Hearing Readiness Class	Categorical	No	Yes	4	Yes
Heart Murmur	Binary	No	No	2	Yes
Heart Trouble	Binary	No	No	2	Yes
Height at Enlistment	Numeric	No	No	N/A	No
Hispanic**	Binary	Yes	Yes	2	No
Home of Record Region	Categorical	Yes	No	5	No
Home of Record State/Territory	Categorical	No	No	56	No
Hostile Injury**	Binary	Yes	Yes	2	Yes
Hypertension	Binary	No	No	2	Yes
Joint Pain	Binary	No	No	2	Yes
Kidney Disease	Binary	No	No	2	Yes
Limited Duty Profile	Binary	No	No	2	No
Liver Disease	Binary	No	No	2	Yes
Marital Status Code	Categorical	No	Yes	4	Yes
Medical Waiver	Binary	No	No	2	No
Mental Health Concerns	Binary	No	No	2	Yes
Nondeployable Profile	Binary	No	No	2	No
Non-Hostile Injury**	Binary	Yes	Yes	2	Yes
Number of Days Deployed**	Numeric	Yes	No	N/A	Yes
Number of Dependents at Enlistment	Numeric	No	No	N/A	No
Pregnancy Status	Binary	No	No	2	Yes
Prior Service	Binary	No	No	2	No
PULHES*** after IET	Categorical	No	No	4	Yes
PULHES*** at Enlistment	Categorical	No	No	3	No
Race Code	Categorical	No	No	5	No
Service Agreement Duration*	Categorical	No	No	4	No
Vision Readiness Class	Categorical	No	Yes	4	Yes
Weight at Enlistment	Numeric	No	No	N/A	No

* Variables used only for stratification of training and test sets.

** Variables not included by Devig (2019).

*** Includes six variables: Physical, Upper, Lower, Hearing, Eyesight, and Psychiatric.

a. Numeric Variables

We used seven numeric variables, described in Table 2. Devig (2019) did not use number of days deployed or BMI in his research. Speten (2018) originally constructed and used number of days deployed in his analysis, with it being a significant indicator in his attrition model. However, this information would not be available to the model at the beginning of a soldier's first term, which invalidates its inclusion in the model. This is also the reason Devig chose to exclude number of days deployed from his analysis. We chose to use this variable as a TVC. In order to do so, we went back to the original data and reconstructed it as we did other TVCs. All soldiers, including those with prior service, were given an initial value of 0; this value may increase as time progresses. In this way, we only consider changes in the first term, and not what may have happened in previous enlistments.

Table 2. Numeric Variables. Adapted from Devig (2019).

Variable	Description
Age at Enlistment	Recruit's age at time of enlistment. Constructed using birth date and date of enlistment.
ASVAB GT Score	General Technical score of ASVAB consisting of three knowledge areas: word knowledge, paragraph comprehension, and arithmetic reasoning.
BMI at Enlistment	Recruit's BMI at enlistment. Constructed from weight and height.
Height at Enlistment	Recruit's height at time of enlistment in inches.
Number of Days Deployed**	TVC that captures total number of days in a deployed status up to time *t*.
Number of Dependents at Enlistment	Number of dependents at the time of enlistment.
Weight at Enlistment	Weight at time of enlistment in pounds.

** Variables not included by Devig (2019).

b. Categorical

We used 32 multi-factor categorical variables with 278 total factor levels, described in Table 3. Devig (2019) used the same categorical variables with four exceptions. First,

we included CMF code both at enlistment and after IET; Devig (2019) did not use it after IET. Second, we added the factor level of "None" to citizenship origination code. Previously, all non-citizens were listed as "NA" or missing in the data. Since random forests cannot directly handle missing values, we added the level to ensure these soldiers were not later imputed to a level indicating citizenship. Next, we included BMI, binning into categories defined by the Centers for Disease Control and Prevention (Centers for Disease Control [CDC] 2021). The final exception was to CMF factor levels, which we discuss in the next section.

Table 3. Categorical Variables. Adapted from Devig (2019).

Variable	Description	Levels	Level Description
AFQT Category Code	Categories based on percentiles between 1 and 99.	I	93-99%
		II	65-92%
		IIIA	50-64%
		IIIB	31-49%
		IVA	21-30%
		IVB+	1-20%
Age Group at Enlistment	Age groups based on age at enlistment.	1	17-24 years old
		2	25-34 years old
		3	35-44 years old
		4	45+ years old
Blood Type	Blood type of enlistee.	A	All eight unique blood types
		A+	
		AB	
		AB+	
		B	
		B+	
		O-	
		O+	
BMI Category**	BMI as defined by the CDC (2021).	Underweight	< 18.5
		Normal	18.5-24.9
		Overweight	25-29.9
		Obese	> 30
CMF Code at Enlistment / after IET**	Assigned occupational code	Multiple	See Table 4
CMF Functional Group	Categories of CMF codes that perform similar functions.	FS	Force Sustainment
		MFE	Maneuver, Fires, and Effects

Variable	Description	Levels	Level Description
		OS	Operations Support
Citizenship Origination Code	Code that indicates enlistee's U.S. citizenship origination.	A	Born in the U.S.
		C	Born outside the U.S.
		N	Naturalized citizen
		None	Not a citizen
Dental Class	Dental Readiness determined by kevel of treatment needed.	1	No treatment needed
		2	Require non-urgent dental treatment or reevaluation
		3	Require urgent dental treatment
		4	No dental exam in last 13 months; require immediate dental exam
Education Tier Code at Enlistment	Indicates high school completion of equivalent.	1	High school diploma
		2	GED or equivalent
		3	No high school diploma, GED, or equivalent
Ethnic Affinity Code	Code that represents cultural background.	AA	Asian Indian
		AB	Chinese
		AC	Filipino
		AD	Guamanian
		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)
Faith Group Code	Code that represents the faith	Multiple	See Appendix A

Variable	Description	Levels	Level Description
	the enlistee identifies (collapsed for all faiths with less than 100 observations)		
Fiscal Year Accession*	The fiscal year the soldier enlisted in the Army.	2005	The year the soldier joined.
		2006	
		2007	
		2008	
		2009	
		2010	
		2011	
Hearing Readiness Class	Hearing readiness determined by level of treatment needed.	1	Hearing test current; no hearing issues
		2	Hearing test; minor issues
		3	Minor hearing issues need evaluated by audiologist
		4	No hearing tests within past 13 months; requires immediate hearing test
Home of Record Region	Geographical region where enlistee's home of record is located.	Midwest	States in regions defined by the U.S. Postal Service
		Northeast	
		South	
		Territory	
		West	
Home of Record State/Territory	Home of record state code.	Multiple	See Appendix B
Marital Status Code	Legal marriage status.	N	Never married
		M	Married
		D	Divorced
		Other	Legally separated, annulled, widow(er)
PULHES after IET	A qualifier of an enlistee's physical profile and stamina, upper extremities, lower extremities, hearing, eyesight, and psychiatric and emotional profile following the completion of IET.	1	High level of fitness
		2	Possess a medical condition that limits some activities
		3	Possess a medical condition that requires significant limitations
		4	Performance of military duty must be drastically limited
		1	High level of fitness

Variable	Description	Levels	Level Description
PULHES at Enlistment	A qualifier of an enlistee's physical profile and stamina, upper extremities, lower extremities, hearing, eyesight, and psychiatric and emotional profile upon enlistment in the Army.	2	Possess a medical condition that limits some activities
		3	Possess a medical condition that requires significant limitations
Race Code	Code indicating race of enlistee.	1	Other (American Indian/ Alaska Native/ Native Hawaiian/ Pacific Islander)
		2	Asian
		3	Black/African American
		4	White
Service Agreement Duration*	Contract duration of first term enlistment.	3, 4, 5, 6	First term duration
Vision Readiness Class	Vision readiness determined by level of treatment needed.	1	Vision screening up-to-date; best corrected vision is 20/20 or better
		2	No screening required; best corrected vision is 20/40 or better
		3	Not optically or visually ready; comprehensive vision screening needed.
		4	No vision screening within past 13 months; requires immediate vision screening

* Variables used only for stratification of training and test sets.

** Devig (2019) only used CMF at Enlistment, not after IET.

Table 4 describes the 18 CMF factor levels. Two CMF codes changed during the fiscal years in the cohort. CMF 63 (Vehicle Mechanics) was recoded to CMF 91 (Ordnance) effective 1 October 2009, and CMF 21 (Engineer) was recoded to CMF 12 effective 1 October 2010. As previously discussed, we include both CMF at enlistment as a constant variable, and CMF after IET as a TVC. Because of this, we needed to ensure all

observations throughout the cohort were coded the same. This way the recoding in the Army system does not show as a CMF change for the soldier.

Table 4. Career Management Fields. Adapted from Devig (2019).

CMF Codes	Description
11	Infantry
12	Engineer
13	Field Artillery
14	Air Defense Artillery
15	Aviation
19	Armor
25	Signal
31	Military Police
35	Military Intelligence
42	Human Resources
68	Health Services
74	Chemical
88	Transportation
89	Ammunition/Explosive Ordnance
91	Ordnance/Vehicle Mechanics
92	Quartermaster
94	Electronic/Missile Maintenance
LD	Multiple

c. Binary Variables

We used 29 binary variables, described in Table 5. Hostile injury, non-hostile injury, and deployments are all binary TVCs not included by Devig (2019). We originally constructed all three variables as numeric TVCs. Hostile and non-hostile injury both had very small levels of occurrence, and the majority that did experience injuries only had one, so we transformed them into a binary variable. All soldiers start as a "No" and become a "Yes" if injured while deployed during their first term. For deployments, we maintained number of days deployed as a numeric TVC, and also added deployments as a binary TVC. Again, all soldiers start as a "No" and become a "Yes" if they deploy. Both Devig (2019) and Gobea (2019) included nondeployable and limited duty profiles as constant variables

in their research. Similar to Speten's (2018) use of days deployed, these variables represent future knowledge making it erroneous to include them in modeling. We were unable to convert the nondeployable and limited duty profiles into TVCs due to no corresponding date information in the source dataset, and therefore excluded them from any modeling.

Table 5. Binary Variables. Adapted from Devig (2019).

Variable	Description
Administrative Waiver	N: Did not receive a waiver Y: Received a waiver
Conduct Waiver	
Drug Waiver	
Medical Waiver	
Anemia	N: Has not been diagnosed with the condition Y: Has been diagnosed with the condition
Asthma	
Back Pain	
Cancer	
Chronic Pain	
Diabetes	
Epilepsy	
Headaches	
Heart Murmur	
Heart Trouble	
Hypertension	
Joint Pain	
Kidney Disease	
Liver Disease	
Mental Health Concerns	
Citizenship Status Code	C: U.S. Citizen N: Non-U.S. Citizen
Deployments**	N: Has not deployed in first term. Y: Has deployed in first term.
Gender	M: Male F: Female
Hispanic*	N: Not Hispanic Y: Ethnic code of AK, AL, AM, AN, or AO
Hostile Injury	N: Has not sustained injury while deployed in first term. Y: Has sustained injury while deployed in first term.
Non-Hostile Injury	
Limited Duty Profile	N: Has not had the profile Y: Has had the profile
Nondeployable Profile	

Variable	Description
Pregnancy Status	N: Is not currently pregnant Y: Is currently pregnant
Prior Service	0: Has no prior service 1: Has prior service

** Variables not included by Devig (2019).

d. Time-Varying Covariates

We used 31 time-varying covariates, shown in Table 6. CMF code after IET, hostile injury count, non-hostile injury count, and number of days deployed, all marked with an "**," were not included in Devig's (2019) research. Inclusion of these variables has already been discussed.

Table 6. Time-Varying Covariates. Adapted from Devig (2019).

Variable		
Anemia	Epilepsy	Liver Disease
Asthma	Headaches	Marital Status Code
Back Pain	Hearing Readiness Class	Mental Health Concerns
Cancer	Heart Murmur	Non-Hostile Injury**
CMF Code after IET**	Heart Trouble	Number of Days Deployed**
Chronic Pain	Hostile Injury**	Pregnancy Status
Dental Class	Hypertension	PULHES after IET (all 6)
Deployment**	Joint Pain	Vision Readiness Class
Diabetes	Kidney Disease	

** Variables not included by Devig (2019).

e. Unique Survival Analysis Variables and Format

Survival analysis requires transforming data from a wide format to a long format. In a wide format, each unique soldier has a single row in the data. Whenever a TVC changes, we added two new columns for that soldier, one showing the new value of the TVC and the other the time it changed. In the long format, a soldier may have multiple rows of data. Instead of adding new columns whenever a TVC changes, the long format

adds a new row for that soldier. This format worked for the survival tree modeling conducted by Devig (2019). However, our survival random forests cannot handle TVCs directly, requiring a different approach. We further transform the long data so each row corresponds to a given year for a given soldier, acting as a snapshot for the beginning of that year. For example, if a soldier enlists for three years and does not attrit, he or she will have three rows in the data, with each row being a snapshot of his or her data at the beginning of that year. If the same soldier were to attrit in the second year of his or her term, he or she would only have two rows in the data.

Survival analysis also requires additional variables for each observation. In order to determine when an enlistee attrits, one needs a start age and end age, both calculated in years. All our observations have a start age of 0 that corresponds with their enlistment date. The end age depends on the contract length of the first term as well as if the soldier completes his or her first term. Soldiers who enlist for three years and complete their first term have an end age of three years. Soldiers who instead attrit after two years have an end age of two years. Start and end age are further broken down by year within a first term to tstart and tstop, respectively. There are also two response variables, attrit and status. Attrit indicates if a soldier attrits during the first term. Status indicates if a soldier attrits in a given year of his or her first term. For both response variables, "0" indicates survival and "1" indicates attrition. For example, if an enlistee eventually attrits, but not until the third year of his or her contract, then attrits is "1," while status is "0" for the first and second years, and "1" for the third year. These additional variables allow us to track both if and when a soldier attrits.

Table 7 shows an example of the long format for two soldiers. Both soldier A and B enlist for a 3-year first term. Soldier A serves all three years, indicated by attrit equals "0," end age equals service duration equals 3, and the three rows of data. Soldier B attrits during the second year of his or her first term, indicated by attrit equals "1," end age not equaling service duration, and only two rows of data. Status equals "0" for soldier B's first year, indicating survival, but equals "1" during his or her second year, indicating attrition. Additionally, tstop in the second year equals 2.5, also indicating attrition.

Table 7. Example of Long Format with Survival Variables

Soldier	Service Duration	Attrit	status	tstart	tstop	End Age	Time-Varying Covariates
A	3	0	0	0	1	3	Snapshot at enlistment
A	3	0	0	1	2	3	Snapshot at start of 2nd year
A	3	0	0	2	3	3	Snapshot at start of 3rd year
B	3	1	0	0	1	2.5	Snapshot at enlistment
B	3	1	1	1	2.5	2.5	Snapshot at start of 2nd year

2. Response

We use the response variable built by Devig (2019), as shown in Figure 4. As previously mentioned, we removed 1,183 soldiers due to missing data making it impossible to calculate end dates and ages required for survival analysis. We calculated the additional response variable, status, determining whether a soldier attrits in a given year while transforming our long data from each row indicating a TVC change to each row indicating a year of a soldier's first term. Devig (2019) provides an in-depth discussion of building the response variable in his thesis.

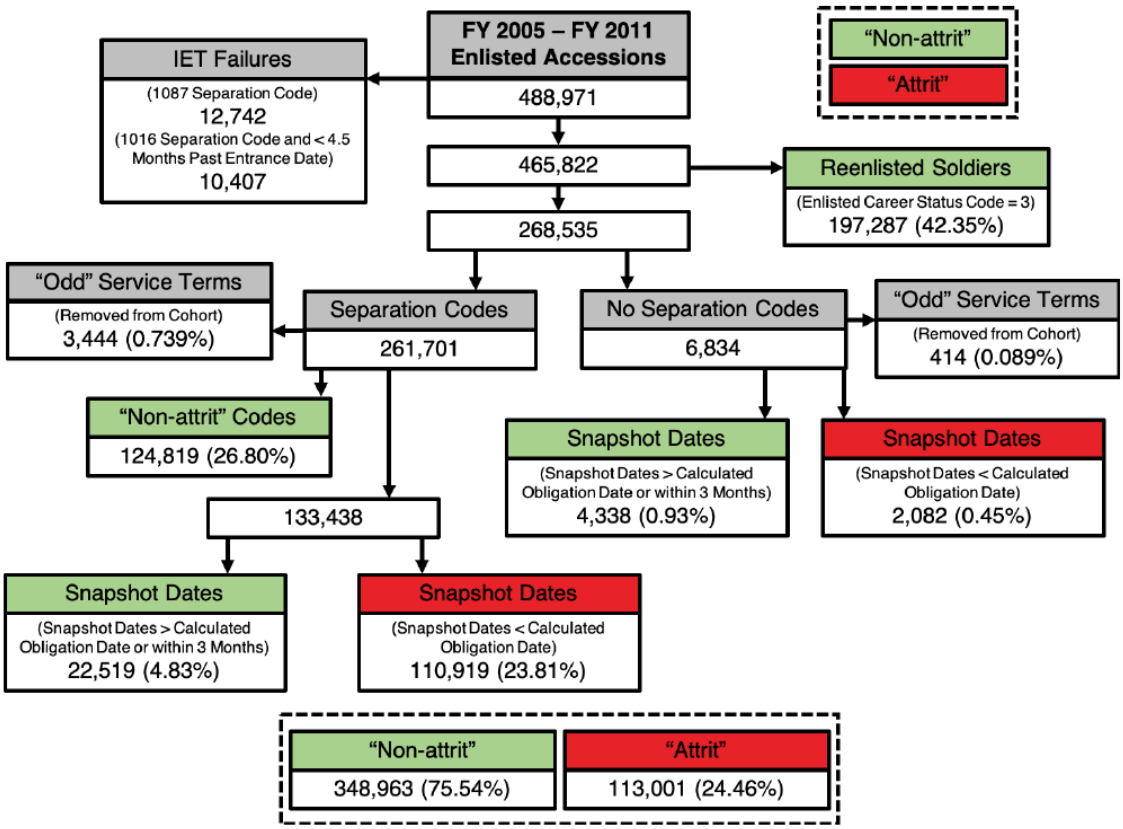

Figure 4. Building the Response Variable Flowchart. Source: Devig (2019).

3. Limitations and Assumptions

First, we assumed the data compiled from the various databases represents a complete and accurate picture of the cohort. Next, we assumed that Devig's (2019) response variable methodology, which itself included multiple assumptions, was accurate.

Our data had large amounts of missing data, especially in the earlier FYs. Survival random forests cannot handle missing data within an observation; therefore if an observation has missing data the forest must either discard the observation or attempt to impute the missing data by considering other observations. Prior to imputing the missing data, we attempted to infer values based on known data in other variables, thereby reducing the amount of imputation required. We further discuss our data inference and imputation of missing data in Chapter IV.

4. Training and Test Sets

We used the FY 2010 cohort as the training set and the FY 2011 cohort as the test set. We chose to discard FY 2008 and FY 2009 due to higher levels of missing values, both before and after inference. FY 2010 and FY 2011 have very similar profiles with respect to missing values. We used imputation, which we discuss further in Chapter IV, to fill in the remaining missing values.

Because survival random forests currently cannot account for TVCs, we stratified the training set by contract term and year of contract term, resulting in 18 separate training sets; three sets for 3-year terms, four sets for 4-year terms, etc. For example, the first set for 3-year contract terms consists of all soldiers alive immediately after IET; the second set consists of all soldiers alive at the start of the second year; and the third set consists of all soldiers alive at the start of the third year. In each set, the TVCs are treated as time constant, and set to their values at the start of the training set year. In this way we approximate changes to TVCs through the first term. Table 8 shows the stratification of training and test sets by term length and year of term.

Table 8. Training and Test Set Stratification

Term Length	Year of Term	Population
3-Year	1st Year	Soldiers with 3-year contract alive after IET
	2nd Year	Soldiers with 3-year contract alive at start of 2nd year
	3rd Year	Soldiers with 3-year contract alive at start of 3rd year
4-Year	1st Year	Soldiers with 4-year contract alive after IET
	2nd Year	Soldiers with 4-year contract alive at start of 2nd year
	3rd Year	Soldiers with 4-year contract alive at start of 3rd year
	4th Year	Soldiers with 4-year contract alive at start of 4th year
5-Year	1st Year	Soldiers with 5-year contract alive after IET
	2nd Year	Soldiers with 5-year contract alive at start of 2nd year
	3rd Year	Soldiers with 5-year contract alive at start of 3rd year
	4th Year	Soldiers with 5-year contract alive at start of 4th year
	5th Year	Soldiers with 5-year contract alive at start of 5th year
6-Year	1st Year	Soldiers with 6-year contract alive after IET
	2nd Year	Soldiers with 6-year contract alive at start of 2nd year
	3rd Year	Soldiers with 6-year contract alive at start of 3rd year
	4th Year	Soldiers with 6-year contract alive at start of 4th year
	5th Year	Soldiers with 6-year contract alive at start of 5th year
	6th Year	Soldiers with 6-year contract alive at start of 6th year

III. DESCRIPTIVE STATISTICS

A. INTRODUCTION

Our cohort attrition closely matched Devig (2019) with the only deviation being the 1,183 attrition records removed due to missing end ages. The cohort attrition rate is 24.27%, slightly below Devig's (2019) rate of 24.46%. Table 9 shows the attrition rates and raw numbers for the cohort.

Table 9. Attrition Rate by Fiscal Year of Enlistment. Adapted from Devig (2019).

	Fiscal Year						
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Non-attrit	77.58%	76.36%	74.58%	74.08%	75.17%	76.74%	75.63%
	50,520	56,293	51,353	50,578	45,705	50,712	43,802
Attrit	22.42%	23.64%	25.42%	25.92%	24.83%	23.26%	24.37%
	14,601	17,430	17,501	17,701	15,097	15,373	14,115

This chapter provides descriptive statistics for some variables and their relationship to the response variable. Rather than rehashing statistics covered by Devig (2019) and Speten (2018), we endeavor to provide descriptive statistics from new perspectives, stratified where possible by service term obligation length and focused primarily on the time aspect of the variables. Figure 5 shows attrition rates by FY and service term obligation length. With the exception of FY 2005, all years showed increased attrition rates as term length increased, with the effect becoming more pronounced in later fiscal years. Finally, we further discuss stratification of data into training and test sets.

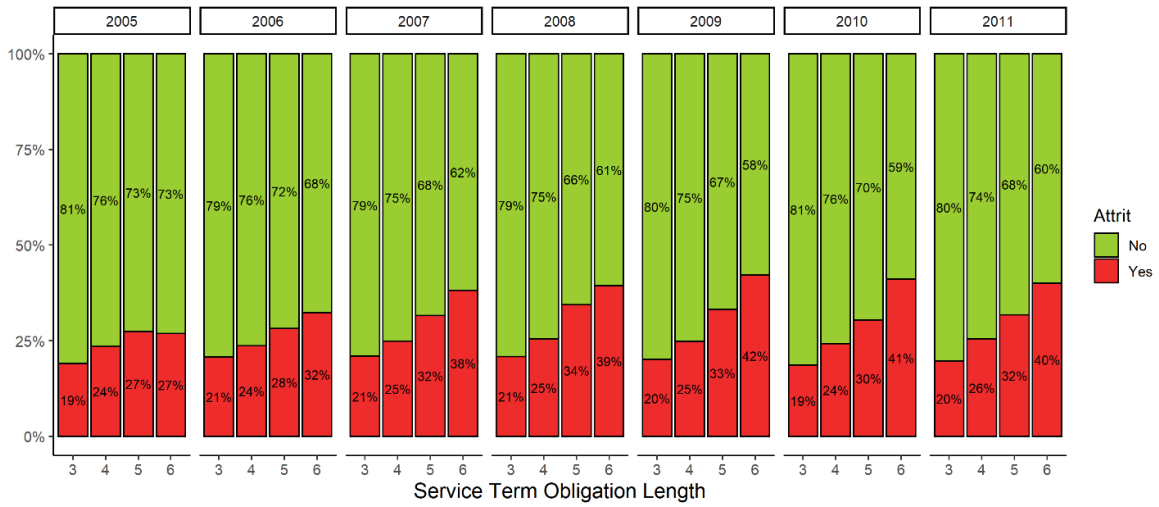

Figure 5. Attrition Rates by Service Term Obligation Length

B. DEMOGRAPHIC VARIABLES

1. Gender

Figure 6 shows attrition rates by gender, and stratified by FY and term length. The green shaded bars represent non-attrits, and the red shaded bars represent attrits, with the sizes of the bars representing the percent comprised by the group. The lighter shaded bars represent females, and the darker shaded bars represent males. Finally, the percentages within the bar represent the attrition or non-attrition rates conditioned by gender. A bar will not display a percentage if the bar represents less than 5% of the population. Using FY 2010 and a 5-year term length as an example: female attrits and non-attrits each make up less than 10% of the total population, while male attrits make up roughly 25% and male non-attrits the remainder. From the percentages, we see females have a 45% attrition rate while males have only a 28% attrition rate. The data is presented in this way to show both attrition rates and population proportion for subgroups, both on a single chart. In this way we minimize the chance of an extreme attrition rate for a small group being perceived as more important than it actually is. As an example, if a notional third gender were present and experienced a near 100% attrition rate, but represented less than 1% of the total population, it would barely show at all. The figure shows that females represent a smaller proportion than males but have higher attrition rates across the cohorts.

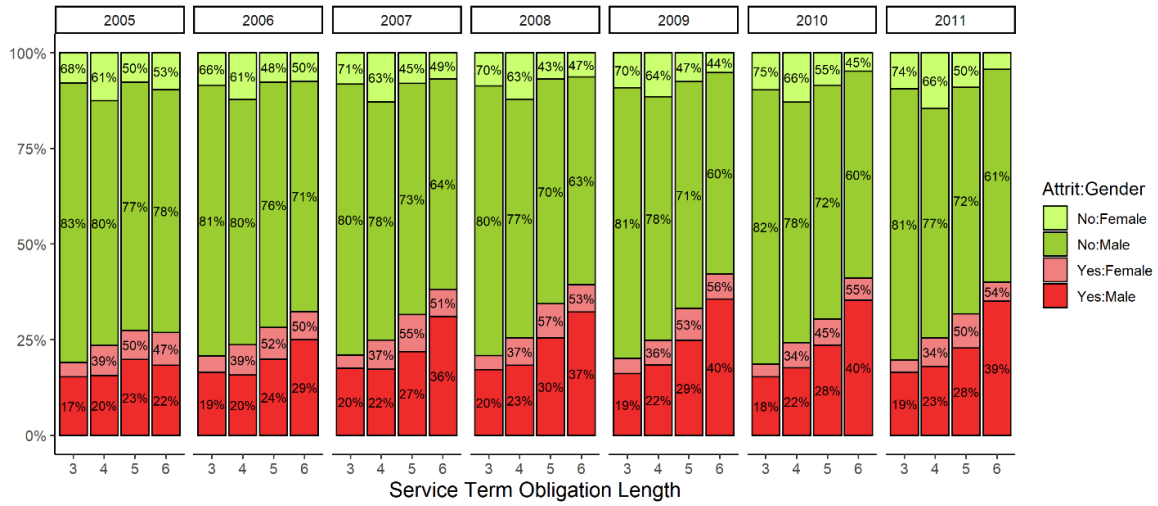

Figure 6. Attrition Rates by Gender

Figure 7 shows when the average attrition occurred by gender. Of note, this average only includes soldiers who attrit; we excluded non-attrits from the average since we already know they serve 100% of their term. With the exception of FY 2005 3-years terms, the average female attrition completed a lower percentage of their term than the average male attrition. The percentage completed increases as FY increases for both genders. Of note, the distance between females and males in 3-year terms stays relatively consistent across the cohort, while longer term lengths have a greater gap variation.

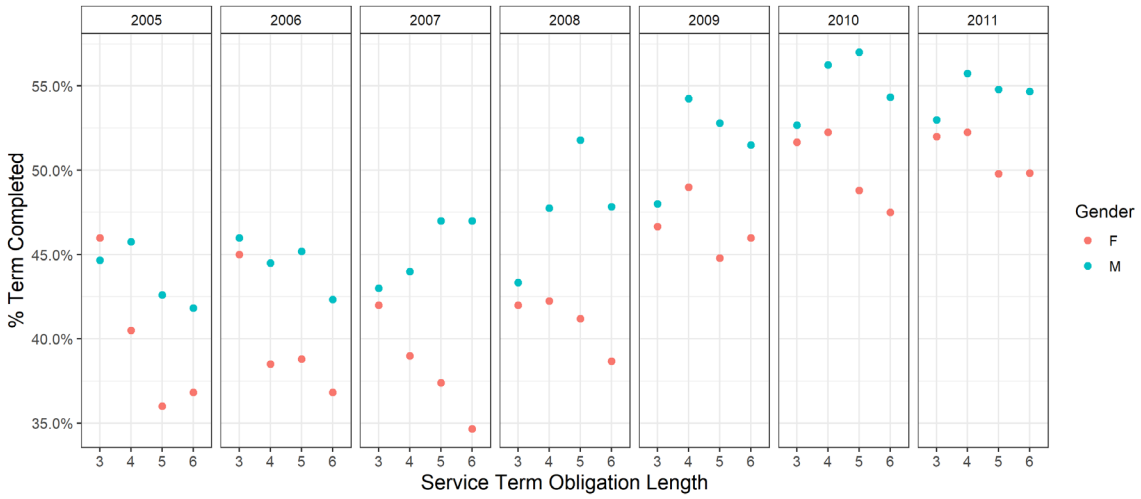

Figure 7. Percent of Term Completed for Attrition by Gender

2. Body Mass Index Category

We constructed Body Mass Index (BMI) and BMI category variables from height and weight at enlistment for our research. Figure 8 shows attrition rates by BMI category at enlistment, stratified by term length. With the exception of underweight, attrition rates for all BMI categories increase with term length. The normal weight category represents 51% of enlistees and has the second lowest attrition rates in all term lengths. The underweight and obese categories represent 3% and 11% of enlistees, respectively. The overweight category represents 35% of enlistees and has the lowest attrition rates in all term lengths except 6 years. BMI does not reflect body composition, which can result in individuals with low body fat but higher weight falling in the overweight category. This likely explains the seeming contradiction of the overweight category's lower attrition rates than the normal category.

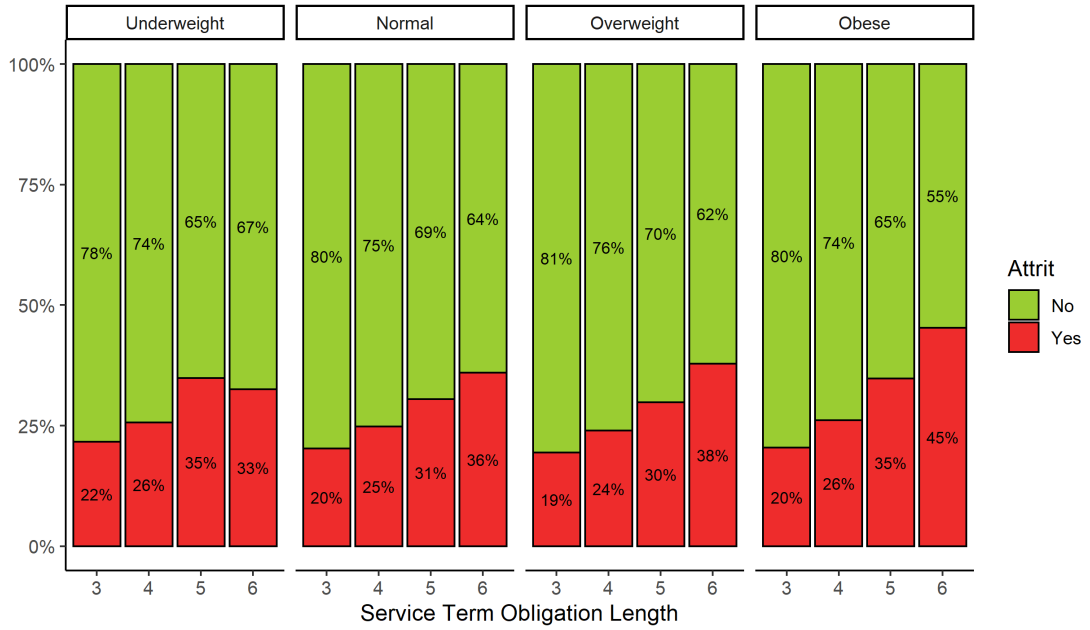

Figure 8. Attrition Rates by Body Mass Index Category

3. Career Management Field

We considered Career Management Field (CMF) as a time-varying covariate (TVC), looking at it both at enlistment and after IET. Devig (2019) observed that Military Police (CMF-31) had the highest attrition rates, and Ordnance (CMF-63) had the lowest. As noted in Chapter II, we recoded CMF-63 to CMF-91, which had a 22% attrition rate in Devig (2019) and a 21% rate in our study.

Figure 9 shows attrition rates by CMF, stratified by whether or not soldiers changed CMF during their first-term. The top chart uses CMF at enlistment as the basis, which we will call transfer out, and the bottom chart uses the final CMF, either at attrition or term completion, which we will call transfer in. To further explain, the bars for Armor (CMF-19) in both charts are nearly identical. This means a very small proportion of soldiers transfer into or out of Armor during their first term. The top chart for Military Intelligence (CMF-35) shows that the majority of soldiers who enlist as Military Intelligence remain so throughout their first term, while the bottom chart shows that roughly a third of soldiers who end their first term as Military Intelligence transferred in during the term.

Low Density (LD), Explosive Ordnance (CMF-89), and Ordnance/Mechanics (CMF-91) experienced the greatest number of transfers out. Military Intelligence, Health Services (CMF-68), and Infantry (CMF-11) experienced the greatest amount of transfers in. Transferring either in or out of LD and Explosive Ordnance decreased attrition rates, which possibly reflects greater job satisfaction in these areas following a change. Transferring out of Military Intelligence or Health Services increased attrition rates, while transferring into them increased attrition rates. These CMFs have longer IETs and require higher ASVAB scores, so transfers out possibly represent soldiers who failed to meet standards and were involuntarily transferred to another CMF, resulting in decreased job satisfaction and higher attrition.

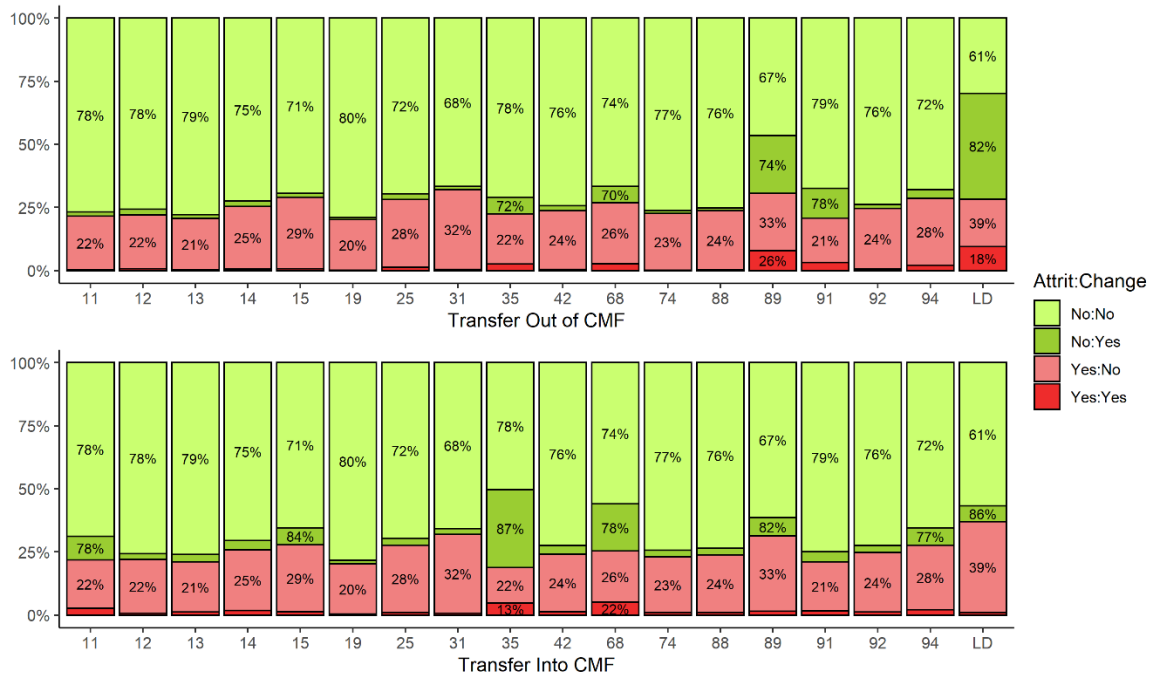

Figure 9. Attrition Rates by CMF

C. DEPLOYMENT-RELATED VARIABLES

Speten (2018) used hostile injury count, non-hostile injury count, number of days deployed, and number of deployments as numeric variables in his analysis. These variables

were based on their count at the end of a soldier's first term, information not available at the beginning of the term when the model was evaluating them. Devig (2019) did not use these variables because they require future, unknowable information. We reconstructed hostile injury count, non-hostile injury count, and number of days deployed for inclusion as TVCs where each soldier starts at "0."

1. Hostile and Non-hostile Injury

Our analysis included hostile and non-hostile injuries as TVCs, which were not included by Devig or Speten. Figures for both variables are essentially indistinguishable from Figure 5, and are therefore not included. While these variables were originally built as numeric, we later recoded them as binary due to the small number of injuries observed. Soldiers injured by hostile action where less likely to attrit than those uninjured across all FYs, while those injured by non-hostile actions were equal or more likely to attrit than those uninjured. While it is unclear from the data, a soldier's perception of his or her injury may explain this seeming contradiction. A soldier injured by hostile action may hold the enemy responsible, while a soldier injured by non-hostile action may hold his or her unit or the Army responsible for his or her injury, resulting in a higher attrition.

Although both hostile and non-hostile injuries appear to impact attrition, they represent a very small percentage of the total population. Hostile injuries account for only 1.88% of the cohort, and non-hostile injuries account for 0.75%. As such, we do not expect them to rate as important variables during modeling.

2. Deployments and Number of Days Deployed

Speten (2018) found that number of days deployed was an important indicator for attrition and used it in his final model. We considered deployments as a TVC, looking both at the raw days deployed as well as a binary variable.

Figure 10 compares the average percent of completed term that soldiers spent deployed for both attrits and non-attrits. As an example, a non-attrit soldier who completes a 3-year term and spends one year deployed spent 33% of his or her first term deployed, on avergae. A soldier who attrits after two years and spends one year deployed spent 50%

of his or her first term deployed. Non-attrits have higher a higher completion percentage across all FYs and service terms than attrits with one exception: FY 2008 6-year term soldiers who attrit completed on average 36% of their first term. This abnormally high percent is likely caused by erroneous data entry or outlier soldiers who spend the majority of their term deployed before they attrit. The average for attrits stays relatively constant near 7.5% throughout the cohorts, while non-attrits start around 20% in FY 2005 and decrease to 10% by FY 2011. The decrease in deployments following the 2008 surge in Iraq likely explains the decrease in average for both groups.

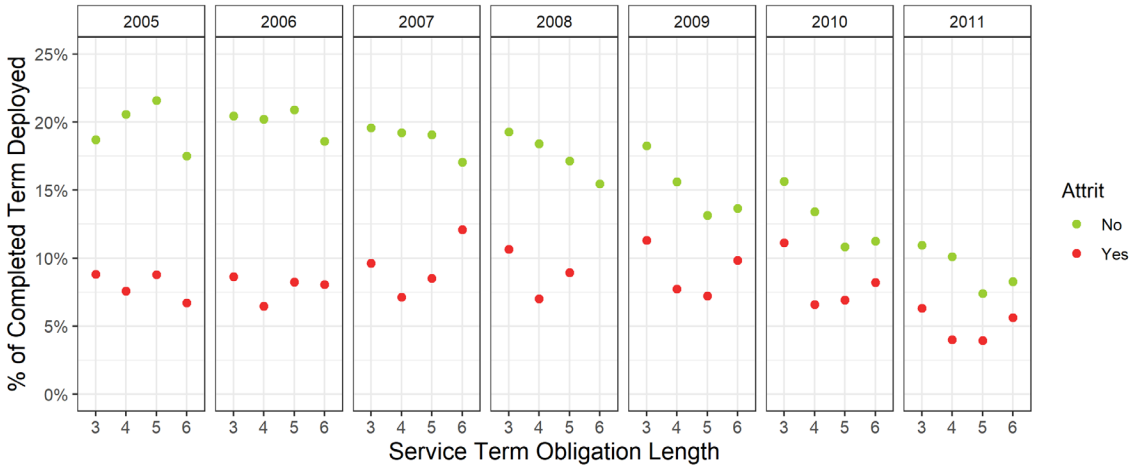

Figure 10. Percent of Completed Term Spent Deployed

Figure 11 looks at deployments as a binary variable. The attrition rates increase as service term increases, with the only exception being a decrease in attrition for deployers from a 5-year term to a 6-year term. The overall rates for both deployers and non-deployers remains relatively constant across the cohort.

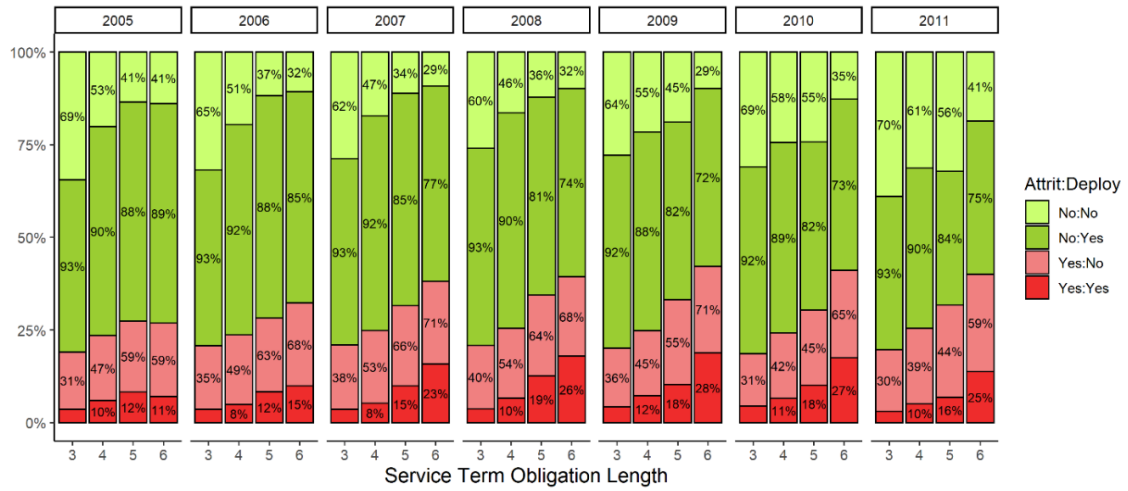

Figure 11. Attrition by Deployment

D. MEDICAL VARIABLES

1. Dental, Hearing, and Vision Readiness

We considered dental, hearing, and vision readiness as TVCs. As a reminder, class 1 generally means no issues, class 2 means minor issues, class 3 means major issues, and class 4 means the soldier requires an exam to determine his or her class. As briefly discussed in Chapter I and further discussed in Chapter IV, dental, hearing, and vision readiness class 4 levels were recoded to their previous level, if known, or "NA" if unknown.

Devig (2019) found dental readiness was an important variable for predicting attrition. Figure 12 shows attrition rates for the final dental class, without class 4. Class 2 represents the largest proportion of the population in all years, and its attrition rate ranges from 2% to 39%. Class is next largest proportion, and has attrition rates ranging from less than 1% to 33%. Class 3 is the smallest proportion, and has attrition rates ranging from 21% to 62%. In FY 2010 and FY 2011, all classes experience increasing attrition rates as term length increases.

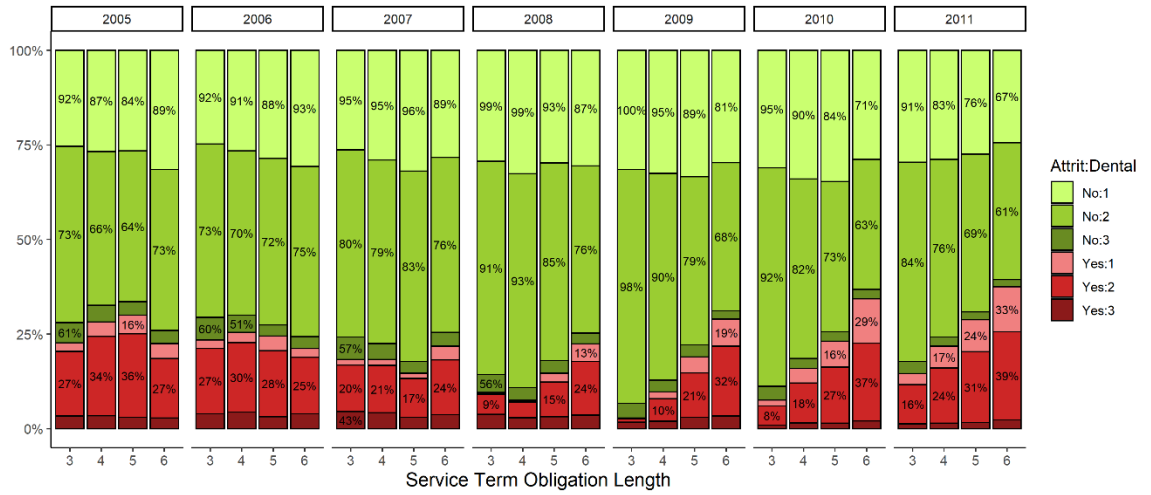

Figure 12. Attrition by Dental Readiness, Without Class 4

The hearing readiness attrition is essentially indistinguishable from Figure 5 after we removed class 4 observations, and is not included. Class 1 represents at least 90% of the population in the all FYs and ranges in attrition rate from 10% to 36%. Classes 2 and 3 each represent 5% or less of the population, with class 2 attrition rates ranging from 1% to 36% and class 3 ranging from 11% to 50%. Similar to dental readiness, attrition rates increase as term length increases.

Figure 13 shows attrition rates for the final vision class, without class 4. Class 1 represents roughly 90% of the population in all FYs and ranges in attrition rate from 10% to 39%. Class 2 represents roughly 7% of the population and ranges in attrition rate from 8% to 40%. Class 3 represents roughly 3% of the population and ranges in attrition rate from 14% to 45%. We again observe increasing attrition rates as term length increases.

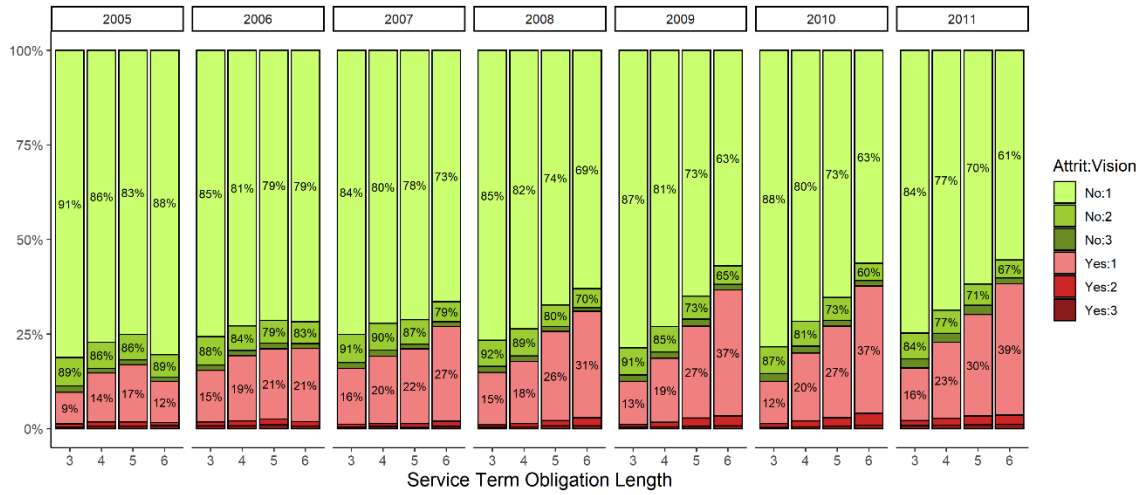

Figure 13. Attrition by Vision Readiness Class, Without Class 4

2. PULHES

We considered the six PULHES variables both at enlistment and as TVCs. Devig (2019) included all PULHES as TVCs in his final model, but did not include PULHES at enlistment. As a reminder, level 1 indicates a high level of fitness, level 2 indicates a medical condition that limits some activities, level 3 indicates a medical condition that requires significant limitations, and level 4 indicates drastically limited performance of military duty. Table 10 shows the minimum proportion of the population represented by class 1 for each PULHES, as well as the class 1 attrition rate range across the cohort. Of the categories, only eyesight and lower extremities warrant discussion. Of note, Devig (2019) only chose to include eyesight in his models based on importance, with the remaining PULHES categories included to keep them together.

Table 10. PULHES Class 1 Statistics

Category	Level 1 Minimum Proportion of Population	Level 1 Cohort Attrition Rate Range
P	95%	19-42%
U	96%	19-42%
L	85%	19-42%
H	97%	19-42%

Category	Level 1 Minimum Proportion of Population	Level 1 Cohort Attrition Rate Range
E	80%	18-40%
S	98%	19-42%

Figure 14 shows attrition rates for lower extremities as a TVC using the last known readiness level. Level 1 represents from 85% to 99% of the population, while level 2 represents up to 10%. Levels 3 and 4 represent the remainder. Level 1 attrition rate increases as term length increases for all years. 3- and 4-year term attrition rates remain relatively constant across the cohort, while 5- and 6-year terms both increase in the middle years and slightly decrease in the end years. Level 2 attrition rates also increase as term length increases, except for FY 2005, and generally increase across the cohort. 6-year terms show the greatest increase in attrition rate, rising from 5% in FY 2005 to 33% in FY 2011.

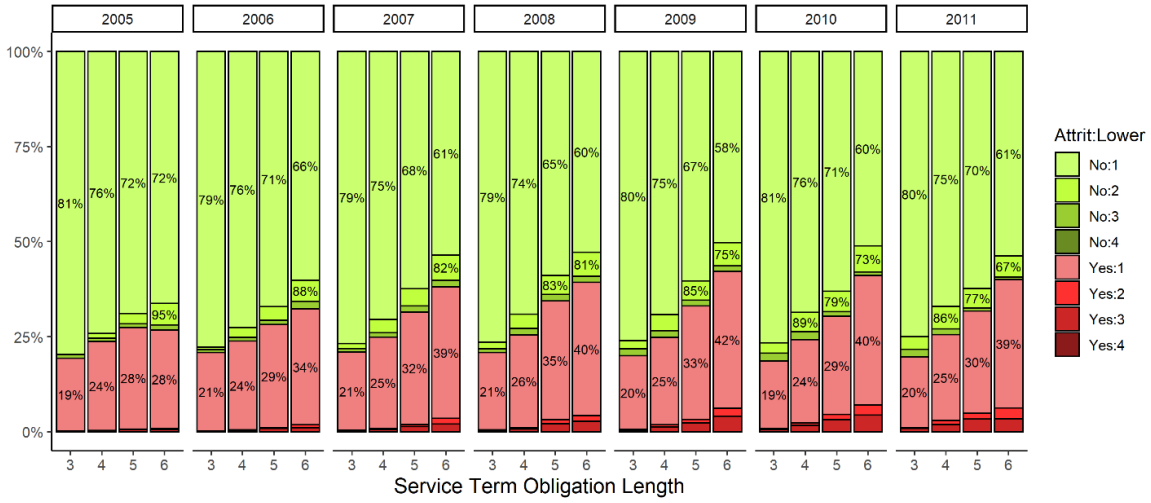

Figure 14. Attrition by PULHES-Lower Extremities

Figure 15 shows attrition rates for eyesight as a TVC using the last known readiness level. Level 1 represents from 80% to 95% of the population, while level 2 represents from 5% to 19% of the population. Levels 3 and 4 represent the remainder. Level 1 attrition rate increases as term length increases for all years except FY 2005. All term length attrition

rates remain relatively constant across the cohort except 6-year terms, which increases from 25% in FY 2005 to 38% in FY 2011. Level 2 attrition rates increase as term length increases, and across the cohort, rising from 19% for a 3-year term in FY 2005 to 72% for a 6-year term in FY 2011.

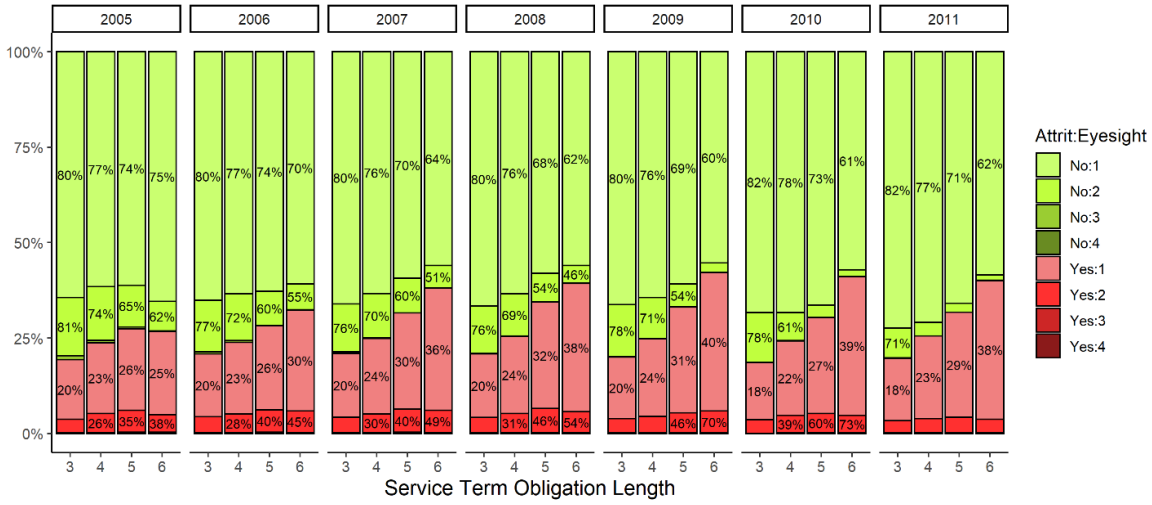

Figure 15. Attrition by PULHES-Eyesight

E. TRAINING AND TEST SET STRATIFICATION

Using FY 2010 as a training set and FY 2011 as a test set, we stratified the dataset by first term length and year of first term, resulting in 18 separate datasets by FY. We chose to discard FY 2008 and FY 2009 due to higher levels of missing values. Devig's (2019) survival tree models both split early on service term obligation length, indicating the importance of this variable. We chose to stratify by term length because of its importance, we know it at the time of enlistment, and it remains constant throughout the first term.

Figure 16 shows the distribution of attrition by FY and first term length, and binned by year of term. As an explanation, looking at FY 2010 and a 3-year term length, we see that 30% of the overall attrition occurs in the first year, 33% in the second, and 37% in the third. For a 3-year first term length, a uniform distribution would mean roughly 33% of total attrition occurs during every year of the term. Similarly, uniform distribution for a 4-

year term corresponds with 25%, 5-year with 20%, and 6-year with 16%. Fiscal years 2005 through 2009 attrition all skewed early, with a greater proportion occurring in the early years of the term. For FY 2010 and FY 2011, the distribution is much closer to uniform. This is a surprising result since we might expect soldiers to become less likely to attrit as they approach the end of their term, resulting in greater attrition in the early years of a term as we observed in the early FYs.

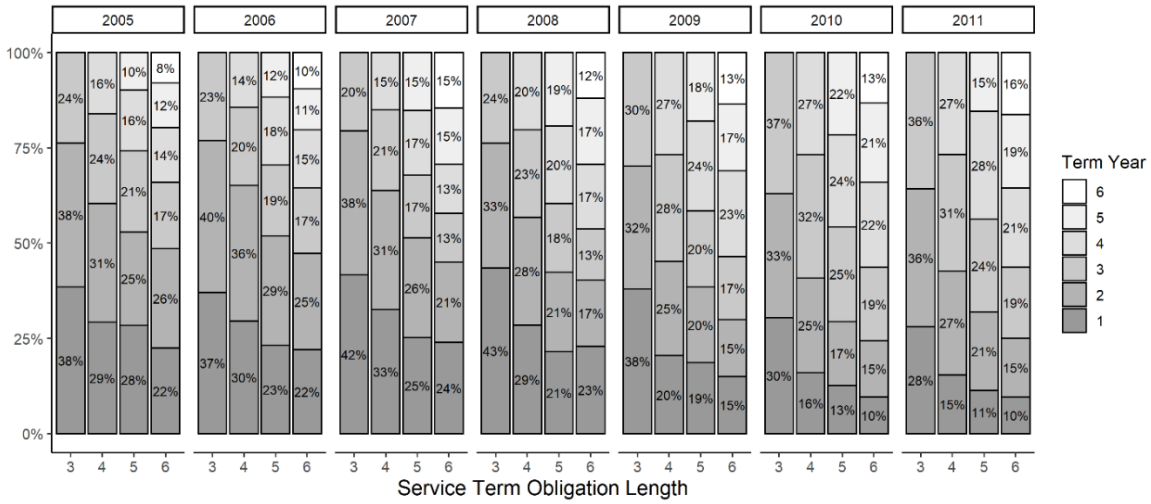

Figure 16. Attrition by Fiscal Year, Term Length, and Binned by Term Year

Figure 16 also illustrates our modeling stratification as well as estimation of TVCs. For FY 2010, each block represents a training set for a model. TVCs for each block remain constant, and only change between blocks as a snapshot at the beginning of that block. Chapter IV discusses modeling, assesses model fit, and evaluates model performance on the test set in depth.

IV. MODELING AND ANALYSIS

This chapter discusses the random survival forest modeling approach, data preparation including handling of missing data, variable selection based on importance, modeling diagnostics and performance, and analysis of findings. In this chapter we will use a short hand nomenclature to reference datasets and models. As an example, we will refer to the 3-Year Term, Year 0 dataset or model as 3T0Y, and the 6-Year Term, Year 5 as 6T5Y.

A. MODELING APPROACH

We use a random survival forest to identify important predictors of attrition and to actually predict attrition. Ishwaran et al. (2008) provide the high-level description of the random survival forest algorithm below, which we discussed in greater detail in Chapter I. Of note, CHF is the Cumulative Hazard Function, which is the negative log of the survival function, 1 – log ($S(t)$).

1. Draw B bootstrap samples from the original data. Note that each bootstrap sample excludes on average 37% of the data, called out-of-bag data (OOB data).
2. Grow a survival tree for each bootstrap sample. At each node of the tree, randomly select p candidate variables. The node is split using the candidate variable that maximizes survival difference between daughter nodes.
3. Grow the tree to full size under the constraint that a terminal node should have no fewer than $d_0 > 0$ unique deaths.
4. Calculate a CHF for each tree. Average to obtain the ensemble CHF.
5. Using OOB data, calculate prediction error for the ensemble CHF. (Ishwaran et al., 2008, p. 844)

After growing the forest using the training data and validating using the OOB data, we pass the test data to the forest for prediction and assessment. Each observation in the test data passes through each tree in the forest, with each tree voting either “attrit” or “not attrit” based on the terminal node of the observation. The majority vote from the trees becomes the overall vote of the forest. Although the default level for a vote is simple majority (i.e., 51%), we can adjust the level up or down based on the characteristics of the

model and desired levels of true positive and false positive rates. This prediction method works best for predicting outcomes for entire cohorts.

The majority vote method provides a simple yes or no, without any measure of the vote strength. For example, if a forest has 100 trees, and an individual receives 51 votes for attrit, then the forest votes attrit. However, the model treats this prediction the same as an individual where all 100 trees vote attrit. Predicting outcomes for individual soldiers requires a different method.

For predicting individuals, instead of recording the vote of each tree's terminal node, the forest records the probability of survival estimate by the node. For example, if a terminal node classifies a soldier as attrit with 51% of training observations, the forest records 0.51 for that tree. If instead all observations in the terminal node attrit, then the forest records 1.0 for that tree. For a forest with 100 trees, each soldier will have 100 different scores from 0 to 1, providing a point estimate with confidence intervals for each soldier. Our research focuses on prediction of cohorts, not individual soldiers, so we will use the majority vote method to evaluate our predictions.

We assess model fit during model parameter tuning using Harrell's Concordance Index, or C-Index. The C-Index categorizes all measurable observation pairs as either concordant or discordant based on their relative risks and attrition time, and returns the ratio of concordant pairs to measurable pairs. This provides a one-dimensional goodness of fit metric for survival analysis. The model produces a risk of attrition for each soldier based on his or her ensemble terminal node statistic (Schmidt et al. 2016). For every pair of soldiers, it compares their risk scores and attrition times. If the soldier with the higher risk score attrits first, then the pair is concordant; if not, the pair is discordant. Pairs where both soldiers do not attrit are excluded. Pairs where only one soldier attrits are concordant if the soldier who attrits has a higher risk, and discordant if not. A C-Index of one indicates perfect association, while a C-Index of 0.5 or less indicates non-informative prediction (Schmidt et al. 2016). The OOB error rate equals one minus the C-Index.

B. DATA PREPARATION

Prior to variable selection, we handle missing data and consider class 4 entries in dental, hearing, and vision readiness to determine the appropriateness of keeping them in the data. We also purposefully exclude several variables based on reasonableness from a modeling and policy standpoint.

1. Missing Data

Our data contains a large proportion of missing data. Random survival forests cannot handle missing data, which leads us to establish a methodology for filling in the missing data. First, we use existing data to make inferences on the missing data. Next, we use random forests to impute the remaining missing data. Table 11 shows the percent missing values by FY both before and after inference.

Table 11. Percent Missing Values by Fiscal Year before and after Inference

Inference	Cohort	FY 2005	FY 2006	FY 2007	FY 2008	FY 2009	FY 2010	FY 2011
Before	13.5%	30.3%	22.8%	15.0%	9.5%	6.3%	4.4%	4.1%
After	1.5%	2.6%	2.3%	1.7%	1.5%	1.0%	0.7%	0.7%

- We used Home of Record (HOR) State/Territory to fill in 400 missing values for HOR Region. Specifically, American Samoa (AS) was incorrectly excluded from the Territory category of HOR Region.
- We used Career Management Field (CMF) at Enlistment and CMF after IET (initial value) to fill in missing values in the other. An analysis showed that 91.9% of entries matched between the two categories, making it a reasonable assumption. This reduced 32,051 missing values in CMF at Enlistment and 1,665 in CMF after IET.
- Comparing U.S. Citizenship Origination and Status showed that 3,048 missing Origination values were all Non-Citizens in Status. We created a

new category of "None" in Origination to account for these observations. We also moved 15 Origination observations of born in the U.S. and Status "NA" to Status of "Citizen."

- For all binary medical covariates (pregnancy, profile, anemia, asthma, etc.), we assumed that a missing value was actually a "No," which removed nearly 150,000 missing values in each of these covariates, and over 400,000 in pregnancy.

- Finally, we used PULHES at Enlistment to fill in roughly 145,000 missing values in each of the PULHES after IET initial values, assuming that any significant change over that period, such as a PULHES code of 3, would result in the soldier attriting during IET.

Overall, inference significantly reduced missing data in all years, reducing the cohort missing data from 13.5% to just 1.5%. Table 12 shows changes in missing data from inference for the covariates with highest levels.

Table 12. Percent (%) Missing Data before and after Inference, all Variables > 1%

Variable	% Missing Before Inference	% Missing After Inference
Pregnancy Status	89%	0%
Hearing Readiness	35%	34%
PULHES after IET	32%	1%
Binary Medical	32%	0%
Vision Readiness	15%	15%
Dental Class	12%	12%
Blood Type	9%	9%
CMF at Enlistment	8%	< 1%
U.S. Citizenship Origination	3%	3%
ASVAB GT Score	3%	3%
U.S. Citizenship Status	3%	3%
Education Tier	3%	3%
PULHES at Enlistment	2%	2%
Height at Enlistment	1%	1%

Variable	% Missing Before Inference	% Missing After Inference
HOR Region	1%	1%
HOR State / Territory	1%	1%

Following inference of missing data, our data from FY 2010 and FY 2011 contained roughly 0.7% missing values in constant covariates or initial entries for time-varying covariates (TVC). As a reminder, we use FY 2010 as our training set and FY 2011 as our test set. We used the impute() function from the randomForestSRC R package of Ishwaran and Kogalur (2020) to fill in these missing values. Imputation is an iterative process that grows a minimum of two random forests. With small amounts of missing data, two iterations are sufficient (Ishwaran et al. 2008, p. 855).

In the first random forest, missing data is filled in prior to each split by randomly selecting a value from the distribution of the in-bag cases without missing values. After splitting into left and right daughter nodes, imputed data returns to missing, and new values are selected from the new in-bag distribution. This continues until no possible splits remain. Final values are selected in each terminal node, with the average selected for continuous variables and the mode selected for integer and categorical variables. These final values are then used to grow a second forest. The forest is grown as if there are no missing values. Once grown, new values are again selected for missing values in the terminal nodes. This process can be repeated multiple times (Ishwaran and Kogalur 2020).

Imputation can be conducted either once for the overall dataset, or every time a new forest is grown. We chose to impute all missing data prior to training any model. This both saved model training time, and provided a consistent dataset for training. We conducted data imputation for FY 2010 and FY 2011 separately so as not to inadvertently use training data to impute test data.

2. Dental, Hearing, and Vision Readiness Class 4

As discussed previously and by Cammack (2020), dental, hearing, and vision readiness class 4s represent a lack of information. A soldier is coded as class 4 if he or she

has not received medical screening within the previous 12 months. Once screened, the soldier is classified as class 1, 2, or 3. Chapter III showed high attrition rates for class 4 soldiers, with especially large proportions of class 4 soldiers in the early FYs. The high attrition rates are possibly caused by soldiers identified for early separation neglecting to complete screenings during out processing. We chose instead to focus on classes 1 through 3, since these likely provide a more accurate assessment of the soldier's medical status.

For all observations of dental, hearing, or vision readiness class 4, we recoded the soldier to the previous known class level, or "NA" if the soldier did not have any recorded class. We completed imputation of missing data previously discussed both with and without class 4 entries in order to observe the impact on variable importance.

3. Purposeful Exclusion

Prior to conducting variable importance and selection, we reviewed the variables to determine if their inclusion was reasonable from a modeling and policy standpoint. Variables considered unreasonable were purposefully excluded from variable importance and selection. Of note, with the exception of blood type, all variables excluded below were included during data imputation.

- Blood type contained a high number of missing values. It was not reasonable to impute missing blood types based on our available data. Also, it is unreasonable to expect the Army to set policy based on blood type, regardless of the strength of evidence to do so.

- Faith group is a large categorical variable. We initially considered collapsing it into broader categories, but ultimately chose to exclude it from the data. Similar to blood type, it is unreasonable to expect the Army to set policy based on a soldier's faith.

- Ethnic affiliation is a categorical variable with a large number of factor levels, and it strongly correlates with race code, which is a categorical with four factor levels. Following Cammack (2020), we include a

Hispanic binary factor as well as maintained race code, but we exclude ethnic affiliation.

- Nondeployable and limited duty profiles are binary, constant variables. As discussed in Chapter II, inclusion of these variables amounts to use of future knowledge. We therefore exclude them from modeling.

- Accession FY and term length are both categorical variables used for stratification of datasets, and are excluded from modeling.

C. VARIABLE IMPORTANCE AND SELECTION

High-dimensional survival analysis requires careful consideration of variables used for modeling (Ishwaran et al. 2010). Because the algorithm randomly selects a subset of variables for consideration at each split point, it may ultimately split on a noise variable. A general technique is to fit initial models using all variables, use a variable importance measure to select the most important variables, then refit the model using only those variables.

1. Importance Measures

Breiman (2001) introduces measures of variable importance based on random forests. For one such measure, he first builds a random forest and calculates its OOB prediction error. Next, he randomly permutes variables in the OOB sample and recalculates prediction error. The difference between the original prediction error and the "permuted" prediction error is the variable importance. The larger the difference, the more important the variable is for prediction. This technique has some potential issues. First, since it is based on prediction error, it will vary depending on what statistic is used for prediction error. Next, the difference for a variable may be incorrectly interpreted as the amount of variability in the model explained by that variable. Finally, since it relies on randomness to increase noise, it is difficult if not impossible to understand why a variable is important (Ishwaran et al. 2010).

Ishwaran et al. (2010) introduced a variable importance measure called threshold. After growing a forest, the minimal depth for each variable is determined for each tree. Minimal depth is defined as the minimum distance from the root node to a node that splits on that variable. If a variable first splits at the root node, then its minimal depth is 0. Minimal depths for a variable are determined for each tree and then averaged across the forest, forming the variable threshold. The average does not include trees where a variable does not split. Finally, all variables are averaged to form the forest threshold. A variable is considered important if the variable threshold is less than the forest threshold. The idea behind threshold for variable selection is that important variables should consistently split early throughout the forest. This technique was chosen because it depends only on tree topology, and is therefore easy to explain and visualize.

We used the rfsrc() function from the randomForestSRC R package of Ishwaran and Kogalur (2020) to grow our forests for 36 total datasets, 18 that included dental, hearing, and vision readiness class 4s, and 18 that did not. But first, we evaluated several methods for handling memory issues in the R environment.

2. Handling Large Survival Problems

Our datasets include 63 variables, range in observations from 35,449 (3T0Y) to 3,835 (6T5Y), and include up to 365 unique attrition times, one for each day of the year. As the algorithm seeks to find the best split locations, it calculates and compares the survival distributions of each daughter node for each potential split location, choosing the split with the strongest difference between survival distributions. For large survival problems, Ishwaran and Kogalur (2020) recommend reducing the number of potential attrition times. In order to prevent memory issues within the R environment, we used three different methods to reduce the size of the forests, running all 18 datasets without readiness class 4 levels for each method. We used the OOB error rates to compare performance across the methods, and then used the overall best method to fit models on the 18 datasets with class 4 levels. This was done to observe the impact on variable selection of removing class 4 levels; final production models will be grown without class 4 levels. Each method used

the rfsrc() function default values except as noted; all methods used 500 trees instead of the default 1000. First we will define some terms, then define the methods.

- Sampling with replacement (SWR) – given a dataset with *N* observations, creates a new dataset of size *N* for each tree by random sampling with replacement from the original dataset. This is called bootstrap aggregating, or bagging.
- Sampling without replacement (SWOR) – given a dataset with *N* observations, randomly sample 63% of observations from the original dataset for each tree (Ishwaran and Kogalur 2020).
- ntime – the number of unique attrition times considered for each tree. If left blank, the algorithm will consider all available times (Ishwaran and Kogalur 2020).
- nsplit – the number of random split points to consider for a variable selected for consideration at a given node (Ishwaran and Kogalur 2020). The algorithm defaults to 10; setting it to zero results in deterministic splitting, which is computationally and time intensive.
- Binning – grouping attrition times into a given number of bins. If binning into quarters, then any attrition that occurs in the first three months of a soldier's first will be set to 0.25, second three months to 0.5, the third to 0.75, and the last to 0.99, resulting in only 4 unique attrition times instead of a possible 365.

The methods used are as follows:

1. SWR, attrition times binned into quarters.
2. SWOR, ntime = 75.
3. SWR, ntime = 75.

After running all models, method three produced the lowest error rate in eleven of eighteen models, method two in six of eighteen, and method one in the remaining one. Method three was also second best in three of the seven models where it was not best. We used method three to fit all future models, including the 18 datasets that include dental, hearing, and vision class 4s, models for tuning parameters, and final production models.

3. Variable Selection

Comparing variable importance for data with and without dental, hearing, and vision readiness class 4 resulted in minimal differences. The biggest difference was hearing readiness, which was in the top 20 important variables for 15 models with class 4, but no models without. Vision readiness decreased from 12 models when readiness class 4s were included, to two models without class 4s, and height at enlistment and headaches both increased by four models. The remaining variables either stayed the same, or differed by three or less. Figure 17 shows variable importance for 6T5Y, both with and without class 4. The red lines represent the forest threshold, roughly steady for both at 12.4. Categories in blue are below the forest threshold, and therefore considered important. Removing class 4s decreased the relative importance of hearing readiness class while all other variables remained roughly unchanged.

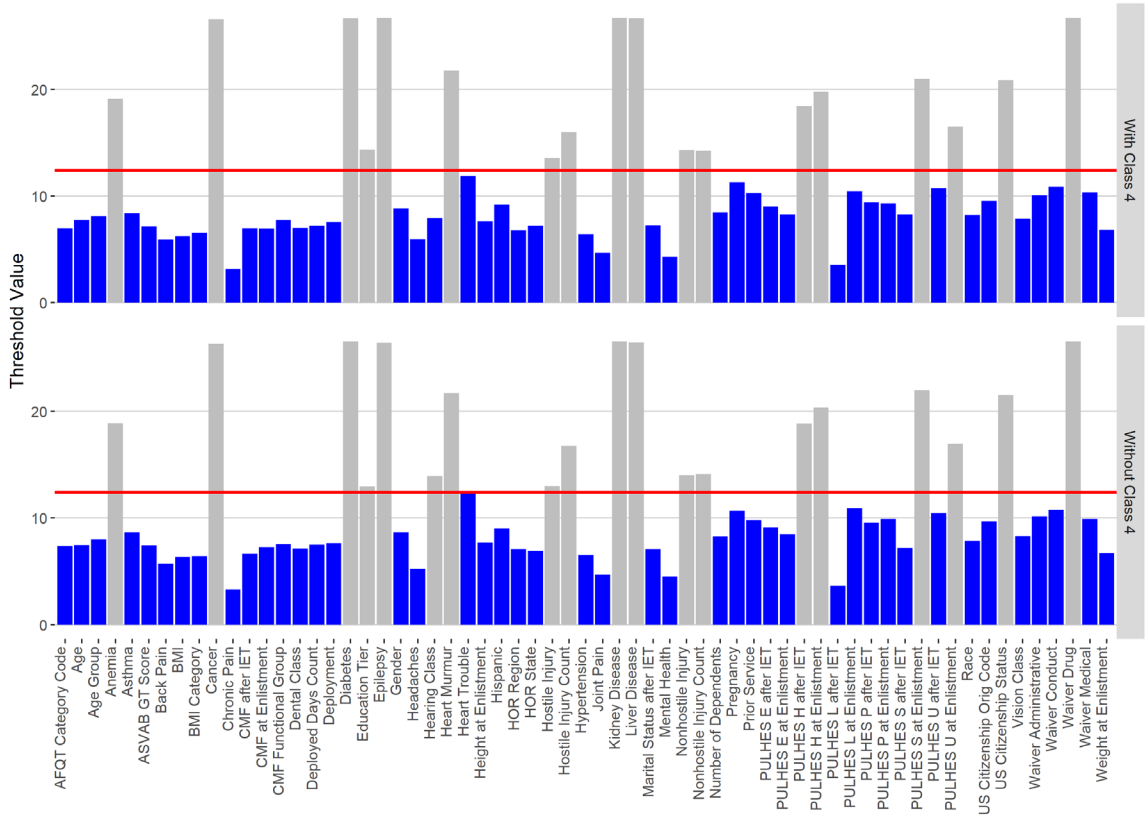

Figure 17. Variable Importance for 6-Year Term, Year 5, With and Without Class 4

Figure 18 shows variable importance from initial modeling. The x-axis shows the number of times the variable appears in the top 20 important variables for a model, with 18 meaning the variable is in the top 20 for all models. The y-axis gives the average rank for the variable. As an example, AFQT is in the top 20 variables in all models, and its rank ranges from four in the 5T1Y model, to 18 in the 6T5Y model, with an average rank of 11. Circle markers represent demographic variables, and triangle markers represent medical covariates. Blue markers represent constant covariates, and red markers represent TVCs. Numeric variables are shown in green, and categorical variables with five or more levels are shown in purple.

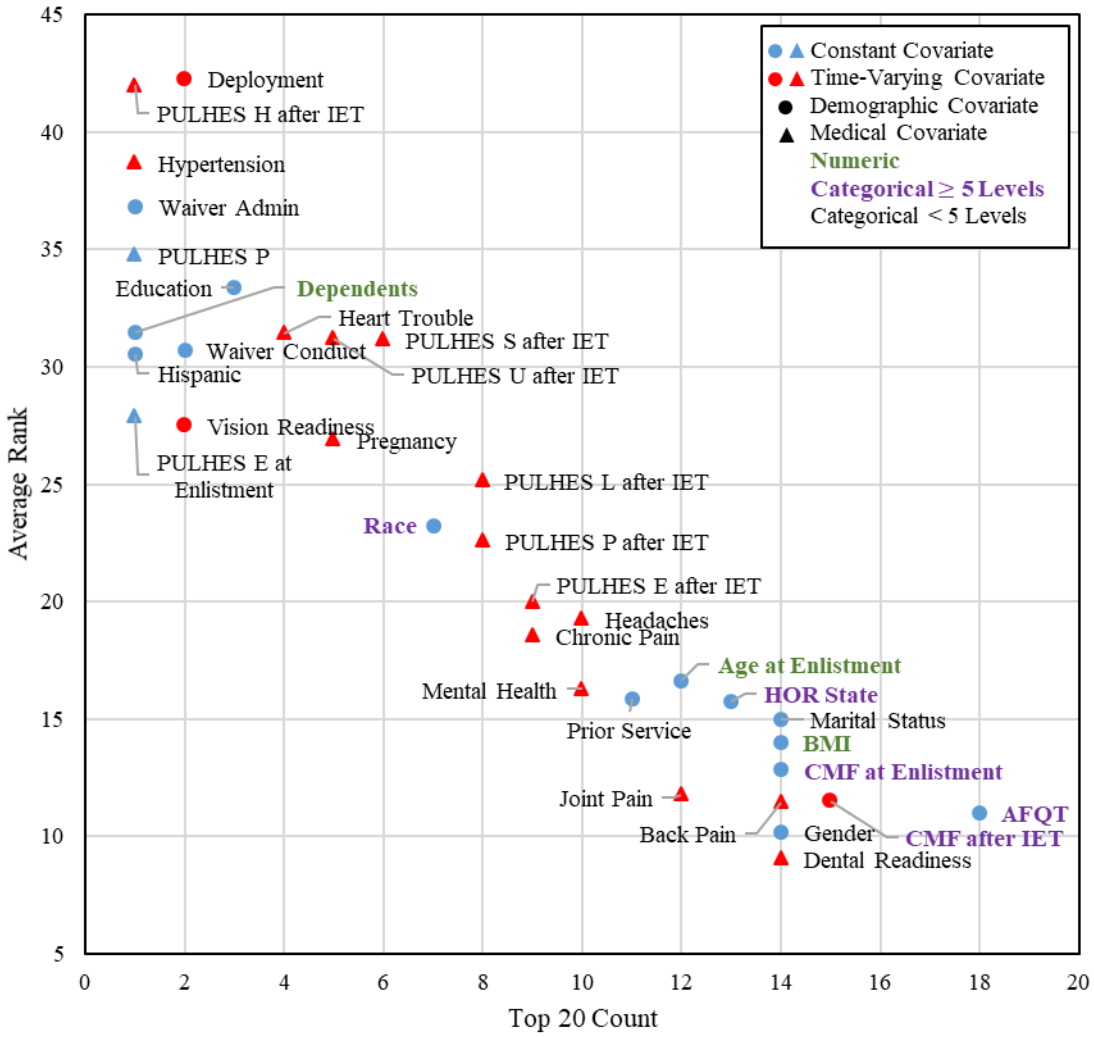

Figure 18. Initial Modeling Variable Importance

Overall, only AFQT was in the top 20 for all models; 21 variables were never in the top 20 and were excluded from the Figure. Variables in the bottom right corner are more important, overall, than variables in the top left corner. The proportions of TVCs and constant covariates in the top 20 match their proportions in the 63 overall variables, and they appear uniformly spread throughout the field. When stratifying by term length, demographic variables are more important than medical variables, and constant variables are more important than TVCs. When stratifying by year of term, demographic and constant variables are more important during the first three years, while TVCs are more

important during the last three years and demographic and medical variables are equal in importance.

We created several variables by aggregating numeric or large categorical variables into smaller categorical variables, making it inappropriate to include both in the models. In the event that both variables were important, we chose the variable that was important in more models and had a lower average threshold value. Overall, this resulted in exclusion of eight additional variables. Table 13 shows comparisons of these variables. We also excluded height and weight at enlistment since BMI is constructed from them and included in the model. We also chose to exclude ASVAB GT score, which Cammack (2020) determined strongly correlated with AFQT.

Table 13. Related Variables in Initial Modeling

Variable	Number of Models in which Variable is Important	Average Threshold	Included in Modeling
AFQT	18	6.3	Yes
ASVAB GT	18	6.6	No
Age at Enlistment	18	6.7	Yes
Age Group	18	7.1	No
BMI	18	6.4	Yes
BMI Category	18	6.7	No
CMF at Enlistment	18	6.3	Yes
CMF Group	18	6.5	No
Days Deployed	10	18.9	No
Deployment	10	18.8	Yes
HOR Region	18	6.9	No
HOR State	18	6.7	Yes
Hostile Injury	9	19.7	Yes
Hostile Injury Count	8	20.5	No
Non-Hostile Injury	6	20.2	Yes
Non-Hostile Injury Count	6	20.5	No

For modeling, we use each model's important variables, which range from 25 variables for 5T0Y to 49 variables for 3T2Y. Appendix C shows variable inclusion for

each model, where the number indicates the relative importance of the variable based on initial modeling in that specific model, and blanks indicate the variable is not included. But first, we review model parameters and tune for performance and computational resources.

D. MODEL PARAMETERS

We tune four model complexity parameters prior to growing production forests: nodesize, nsplit, ntree, and ntime. Nodesize and ntree both control the complexity of the random forest itself, while nsplit and ntime control the complexity of individual splits. For each parameter, we use both the largest (3T0Y) and smallest (6T5Y) datasets for model tuning.

The nodesize parameter is the minimum allowable number of observations in a terminal node, and the default size is 15 (Ishwaran and Kogalur 2020). For the default, if a node has less than 30 observations, the algorithm will stop splitting on that node. This serves as an indirect way to control tree depth in the forest; a large node size will decrease tree depth, while a small node size will increase tree depth. As discussed previously, we want to grow deep trees in order to limit bias (Ishwaran 2015, p. 76). Figure 19 shows the OOB error rate for various node sizes. For both models, the OOB error starts high, quickly reaches a minimum, and then gradually increases, with 3T0Y having a stronger trend. The 3T0Y model's minimum error is 18.2% at a node size of 4, and the 6T5Y model's minimum is 32.0% at a node size of 15. Both models' errors range roughly 1%, showing minimal difference between the minimum and maximum errors. The results suggest that smaller datasets benefit from shallower trees while larger datasets benefit from deeper trees. For uniformity, we will bias toward the larger dataset and choose a nodesize of 10.

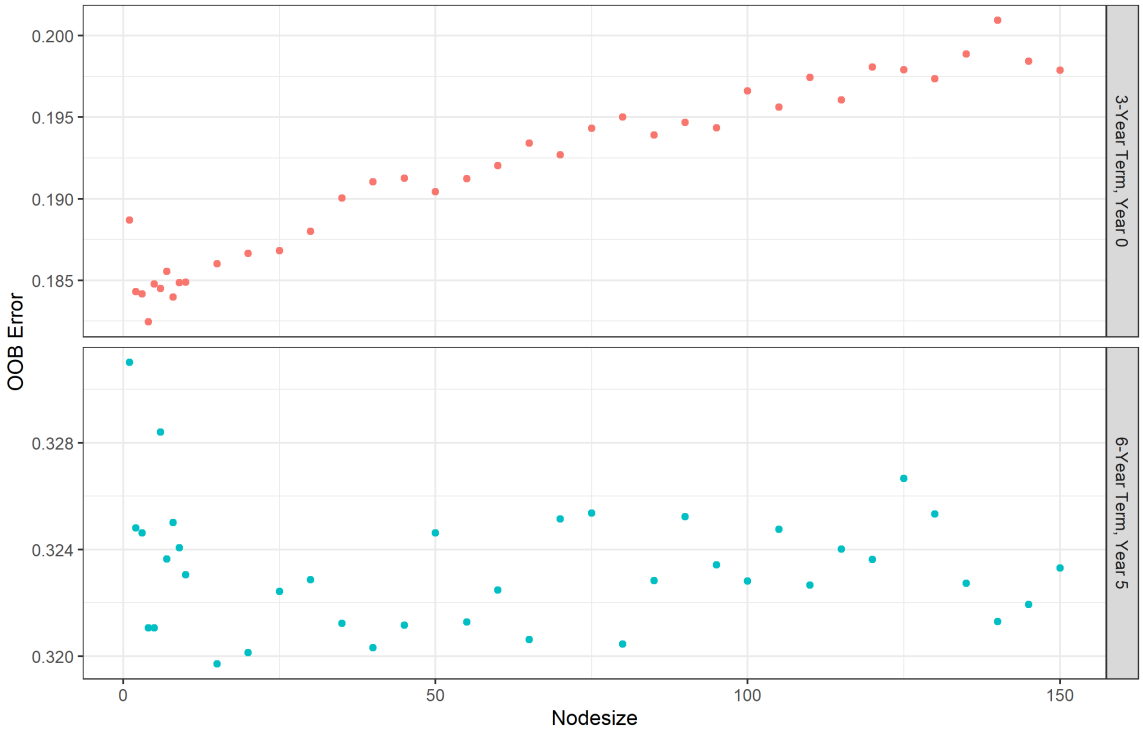

Figure 19. OOB Error Rate for Various Node Sizes

The nsplit parameter is the number of potential split points considered at a node for each selected variable, and the default is 10 (Ishwaran and Kogalur 2020). If no nsplit value is set, the algorithm will deterministically consider all possible split locations. As discussed previously, using a higher split value or deterministic splitting will bias the tree toward continuous and large categorical variables. We evaluated nsplit at three additional settings, one, three, and seven, which correspond to all possible split points for a categorical variable with two, three, and four levels. Table 14 shows the OOB error rates for both models at various nsplit values. We will use an nsplit of three, which minimizes OOB error for both models.

Table 14. OOB Error Rate for Various nsplit Values

nsplit	3T0Y		6T5Y	
	OOB Error	**Rank**	**OOB Error**	**Rank**
10	17.2%	1	32.2%	2
7	17.4%	2	32.7%	4
3	17.1%	3	32.1%	1
1	17.5%	4	32.3%	3

The ntree parameter is the number of trees the algorithm will grow in the forest, and the default is 1000 (Ishwaran and Kogalur 2020). Figure 20 shows the cumulative OOB error rate for forests with increasing numbers of trees, with each dot representing the OOB error for a forest with that corresponding number of trees. Of note, due to the size of the 3-year term dataset, we calculated the OOB error after every 50 trees, while we calculated it after every tree for the 6-year term dataset. For both models, the OOB error starts high then rapidly decreases and levels off, with the 3T0Y model reaching an OOB error minimum of 16.7% at 850 trees, and the 6T5Y model reaching a minimum of 31.6% at 105 trees. For uniformity, we will grow 850 trees in all future modeling.

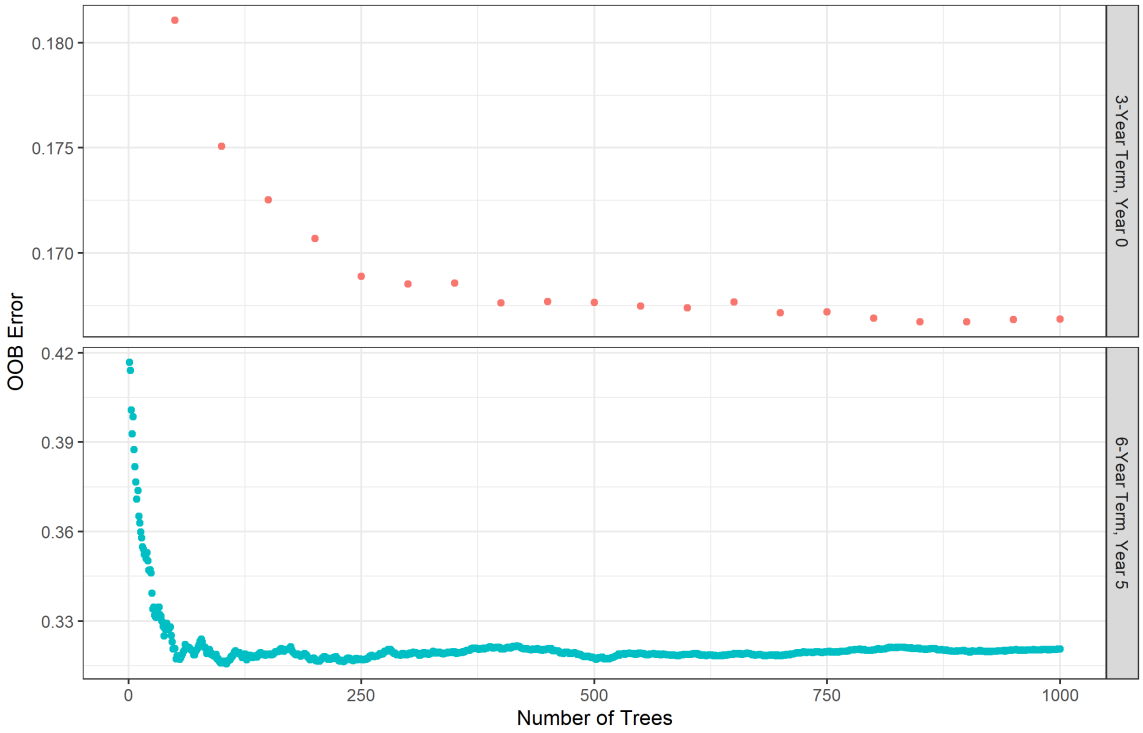

Figure 20. OOB Error Rate for Various Number of Trees

Thus far we have minimized the memory requirements of our models by turning off several unneeded parameters. However, we will need these parameters for analyzing our models and using them for prediction on the test sets. After turning these parameters back on, we attempted to increase the ntime value in order to increase the granularity of the survival analysis, considering only 3T0Y since it is the largest dataset. However, any increase in ntime caused R to crash. Our final model parameters for growing production forests are in Table 15. Other model parameters remain at default settings.

Table 15. Final Model Parameters

Parameter	Setting
samptype	SWR
ntree	850
nodesize	10
nsplit	3
ntime	75

E. MODEL PERFORMANCE AND ANALYSIS

Following model training, we passed the FY 2011 test data to the models for prediction. We use several methods to assess our overall model performance. First, we use Harrell's C-Index, which we also used to assess goodness of fit during model training. Next, we use receiver operating characteristic (ROC) curves and area under the curve (AUC) to assess predictive power for individuals. Third we compare the predicted cohort survival with actual survival. Finally, we build four secondary models, one for each term length, to assess our assumption that approximating TVCs produces a better model than just using values after IET.

1. Harrell's Concordance Index

Table 16 shows the Harrell's C-Index for the primary models. Model 3T0Y has the best performance, with a C-Index of 0.80. This is expected since we tuned model parameters primarily on this dataset. Model 6T2Y has the worst performance, with a C-Index of 0.61. Overall, year-0 for each term length has the best performance, with all subsequent years having lower C-Indices.

Table 16. Harrell's Concordance Index for Primary Models

Term Length	Year of Term					
	0	1	2	3	4	5
3	0.80	0.62	0.62			
4	0.72	0.63	0.62	0.63		
5	0.74	0.62	0.67	0.66	0.65	
6	0.70	0.64	0.61	0.64	0.64	0.63

2. Individual Prediction

We used ROC curves and AUC to assess model performance for individual prediction. Figure 21 shows ROC curves and AUC for all models. The y-axis is the true positive rate, or sensitivity, and the x-axis is the false positive rate, or one minus specificity. The curves are generated by varying the proportion of "attrit" votes in the forest necessary

for the forest to classify a soldier as attrit. A proportion of "1" corresponds with the curve at (0,0), and a proportion of "0" corresponds with the curve at (1,1). An AUC score of 0.8 or higher is considered a good model, while a score of 0.5 or below is considered worthless for prediction.

Our results are on par or slightly below the results achieved by Devig (2019). Considering that Devig's analysis used the future knowledge of non-deployable profiles and dental, hearing, and vision readiness class 4 levels, our results may represent an improvement in performance. As with Harrell's C-Index, model 3T0Y has the strongest performance with an AUC of 0.66. Model 4T1Y has the weakest performance with an AUC of 0.55. Unlike the C-Index, AUC is not consistently highest for the first year of every term. Overall, our model does not show strong performance for prediction of individual attrition.

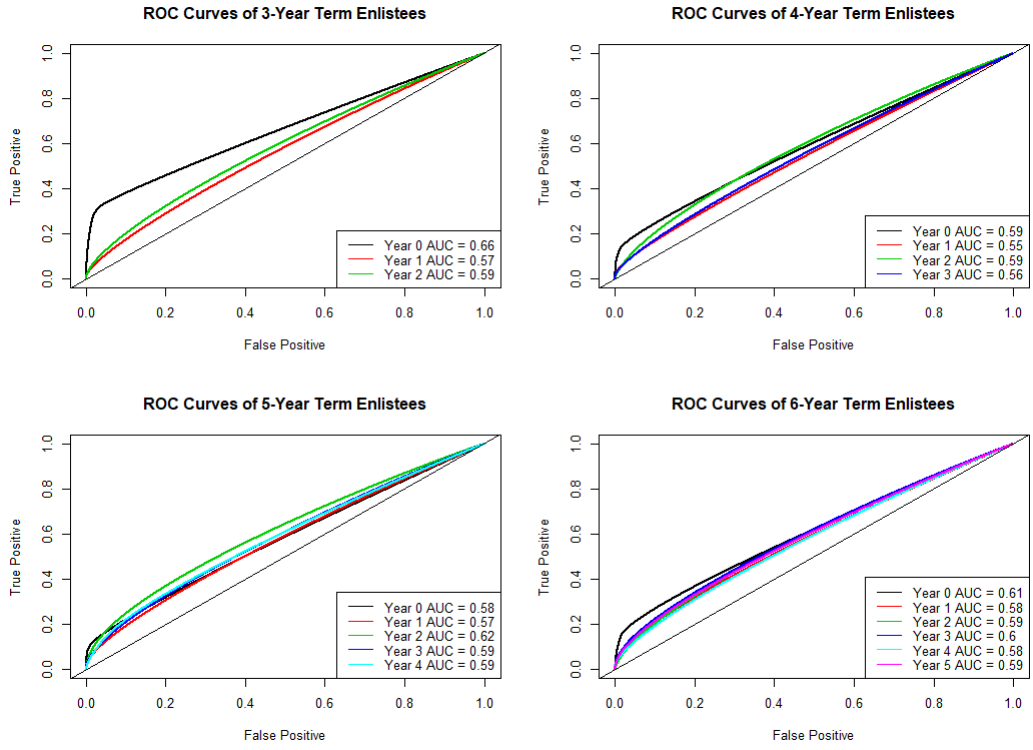

Figure 21. Receiver Operating Characteristics (ROC) Curves for Term Length and Year of Term

3. Cohort Prediction

We used our models' estimated survival probabilities to predict the number of soldiers remaining in a cohort at time, *t*. Figure 22 shows our predicted number (in red) compared to the actual number (in black) for each term length. The flat section in the first part of year 0 is caused by our removal of all IET attrition from the cohort. Consistent with their high C-Index, the year 0 predicted numbers closely match the actual numbers. Predicted errors increase from the start to end of each year, with the largest deviation observed at the end of model 5T3Y. The 6-year term models' predictions most closely match the actual numbers.

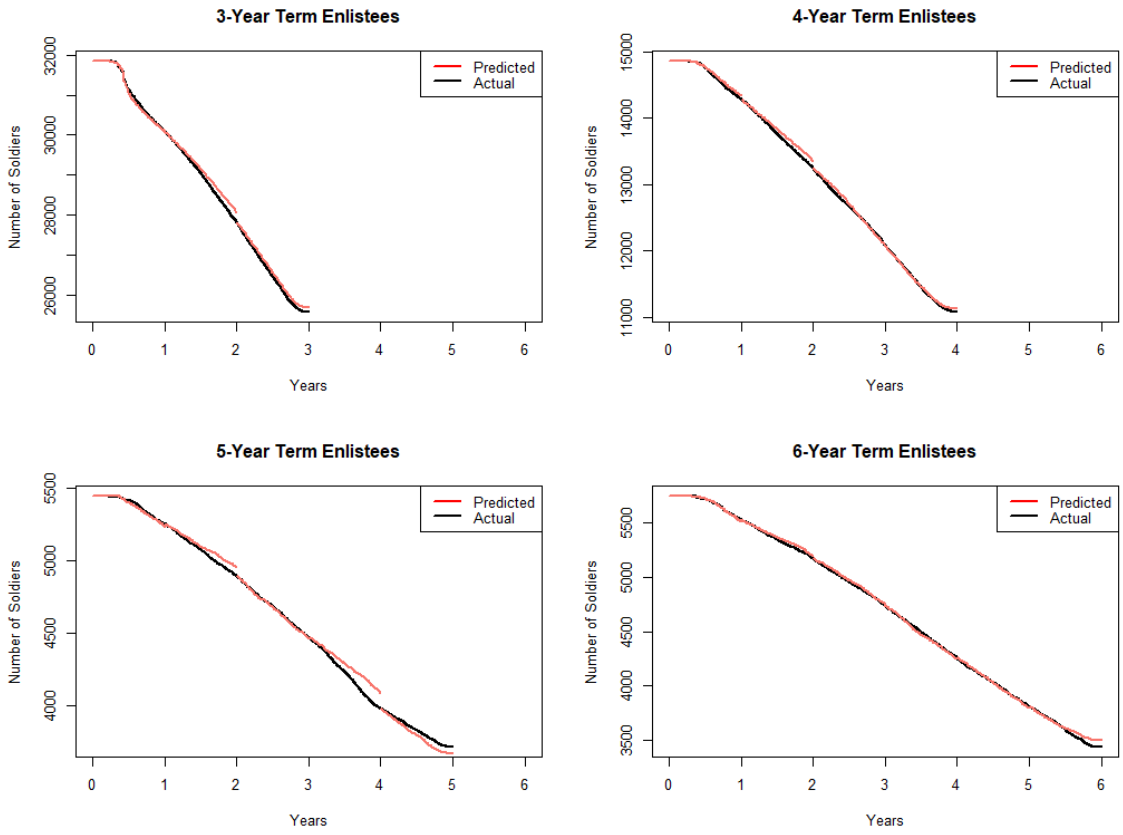

Figure 22. Forecasting FY 2011 Attrition Using FY 2010 Primary Models

As a final check of model fit, we calculated the standard deviation for each model using the square root of the sum of the squared differences between the model projected and observed population at time, *t*, over the number of observations. Table 17 shows the standard deviation for each model. The small standard deviations, especially as compared to the population size for each model, further reinforces the strength of fit for each model.

Table 17. Standard Deviations for Primary Models

Term Length	Year of Term					
	0	1	2	3	4	5
3	57	144	97			
4	36	72	40	18		
5	14	36	6	77	32	
6	8	19	14	15	9	26

4. Secondary Modeling

Our modeling broke each term length into one-year subsets to approximate effects of TVCs, assuming that this would produce improved results. In order to check this assumption, we produced four additional models, one for each term length, which used only data available after IET to predict if and when a soldier would attrit during his or her first term. Of note, we did this only to confirm our assumption. We cannot validly use full FY 2010 data to predict FY 2011 because at the start of the FY 2011 cohort we would only have one year of FY 2010 data, not the three to six years necessary for the modeling.

We followed the same process of variable selection and complexity tuning as discussed for our primary modeling, with two exceptions. First, we automatically excluded the same correlated variables as in primary modeling, rather than comparing the correlated variables and selecting the better predictor. Second, we increased the ntime variable as the term length increased. Table 18 shows parameters for secondary modeling, and Appendix D shows variable inclusion for each of the four secondary models.

Table 18. Secondary Model Parameters

Parameter	Setting
samptype	SWR
ntree	900
nodesize	10
nsplit	7
ntime (3-, 4-, 5-, 6-Year Term)	75, 100, 125, 150

Figure 23 compares the predicted number of soldiers for both our primary and secondary models, with the primary model in red, secondary model in blue, and actual number in black. With the exception of 6-year term enlistees, our primary models outperform the secondary models for all terms, thereby confirming our assumption.

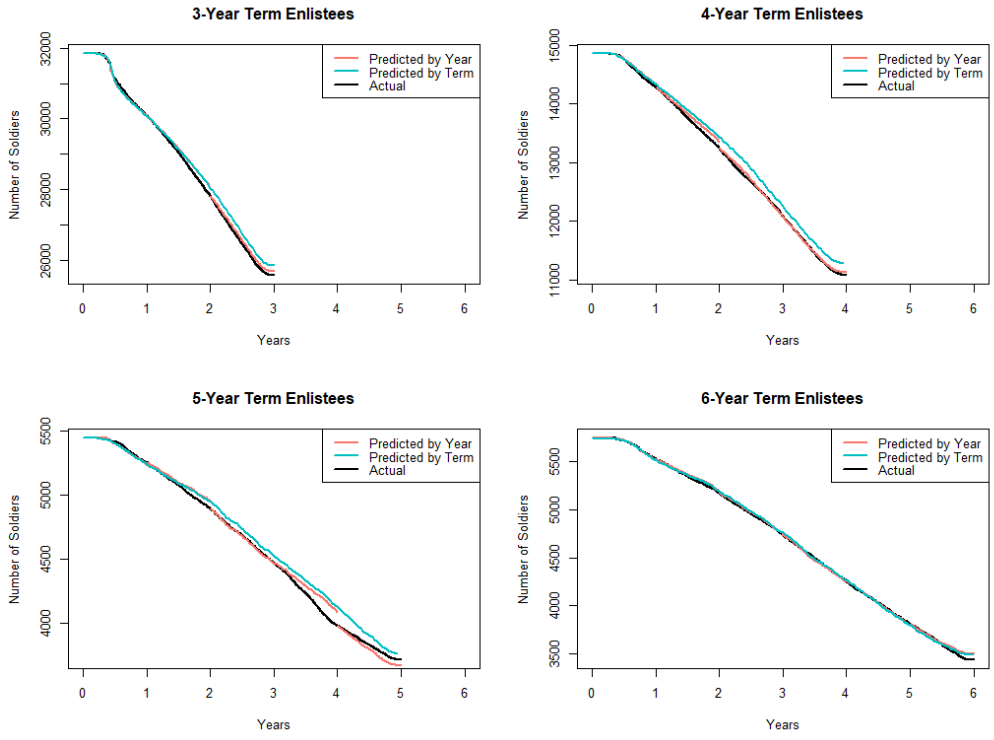

Figure 23. Forecasting FY 2011 Attrition Using FY 2010 Primary and Secondary Models

5. Final Model Terms

Analysis of the final models' terms show that the 10 most important variables are CMF, both at enlistment and after IET, dental readiness class, BMI and age at enlistment, AFQT category, HOR state, PULHES lower extremities after IET, marital status, and gender. Only two of these variables are medical, and four of them are TVCs. Devig (2019) also found that dental readiness was highly important, but he included dental class 4s. It is of special interest that dental readiness remains a strong predictor even after removing class 4s, while both hearing and vision readiness dropped in importance.

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY AND CONCLUSIONS

This chapter provides a summary of conclusions as well as recommendations for future research.

A. CONCLUSIONS

During data analysis we identified an issue with previous work. Dental, hearing, and vision readiness class 4s represent a lack of information, and it is inappropriate to include them in modeling. Additionally, nondeployable and limited duty profiles were reported as constant variables based on their final observation in a soldier's first term. This constitutes future knowledge and it is inappropriate to include them in modeling.

Another issue that arose was a lack of snapshots for time-varying covariates (TVCs). Most TVCs were captured at only three times. This does not provide sufficient information, especially for soldiers enlisted for 4-year terms and greater. Capturing TVCs at additional times would provide greater granularity for modeling and analysis.

We used the CTS-OCO dataset to build deployment and injury histories for soldiers, including these variables as TVCs, both as counts and logical variables. Although soldiers who experienced injuries attrit at significantly different rates from the uninjured, injuries account for very small proportions of the cohort.

Similar to those of Devig (2019), our models do not provide good prediction of attrition for individuals, but perform well for predicting attrition of cohorts. Estimating the impacts of TVCs by taking data snapshots at the beginning of each year improves model performance over only including initial values for these terms. Although our random survival forest models performed similarly to Devig's (2019) survival tree model, his results are better than they likely should be due to inclusion of future knowledge.

Dental readiness remained an important variable for predicting attrition even after removing class 4s, while hearing and vision readiness become less important. Only one PULHES code, lower extremities, was an important predictor in our model. The remaining

top 10 important variables were all demographic variables. Time-varying covariates accounted for four of the top 10 important variables.

B. FUTURE RESEARCH

The Army seeks to reduce IET and post-IET first-term attrition, and increase retention. This thesis analyzed factors that contribute to first-term attrition. Follow-on work for post-IET attrition should consider duty location and accession bonus data. We were unable to access unmasked location data in time for this thesis. Do attrition rates vary by duty location? Does a soldier's relative distance between his or her home of record and first duty station impact attrition? Does receiving a bonus impact attrition rates? Does the total amount of the bonus impact attrition rates?

Due to time constraints, we used the same parameters for growing models on all 18 datasets. However, the datasets vary greatly and a more rigorous tuning of parameters for each dataset will possibly produce improved model performance. TVCs were often only captured at three periods in a soldier's first term. In order to add granularity to the analysis, future research should capture additional, if not all, TVC changes. Additionally, developing techniques to grow random survival forests on large datasets without aggregating the number of attrition points to prevent computational memory issues may also possibly improve model performance.

Tangential research should also focus on IET attrition and retention beyond the first term. Is there a relationship between IET and post-IET attrition? For soldiers who complete their first term, what factors predict whether a soldier reenlists for a second term? The Army must balance recruiting, retention, retirement, and attrition in order to maintain Congressionally approved end-strength levels.

APPENDIX A. FAITH GROUPS

Levels	Level Descriptions
AC	Advent Christian Group
AJ	Jehovah's Witnesses
AS	Seventh Day Adventists
BA	American Baptist Churches in USA
BB	Baptist Church
BC	Southern Baptist Convention
BF	Free Will Baptist Churches, Other
BG	General Association of General Baptists
BN	National Baptist Convention of America
BR	General Association of Regular Baptist Churches
BT	American Baptist Conference
CR	Roman Catholic Church
DL	The Church of Jesus Christ of Latter-Day Saints
EC	Episcopal Church
GC	Christian Church and Churches of Christ
GE	Christian Church (Disciples of Christ)
GX	Church of Christ
HC	Churches of Christ in Christian Union
HN	Church of the Nazarene
II	Islam
JJ	Judaism
KB	Buddhism
KH	Hindu
LE	Evangelical Lutheran Church in America
LL	Lutheran Churches, Other
LM	Lutheran Church, Missouri Synod
MC	Christian Methodist Episcopal Church
ME	African Methodist Episcopal Church
MM	Methodist Churches, Other
MN	Free Methodist Church of North America
MU	United Methodist Church
NC	Christian, no Denominational Preference
NO	No Religious Preference
OE	Eastern Orthodox Churches
Other	All faith group codes with less than 100 observations
PA	Assemblies of God
PC	Church of God in Christ
PD	Full Gospel

Levels	Level Descriptions
PH	Pentecostal Holiness Church
PJ	Pentecostal Church of God
PT	Church of God (Cleveland, TN)
RC	Congregational Churches
RD	Christian Reformed Church in North America
RR	Reformed and Presbyterian Churches, Other
RU	United Church of Christ
TN	Protestant, no Denominational Preference
TO	Protestant, other Churches
UU	Unitarian Universalist
VF	Evangelical Free Church of America
VM	Christian and Missionary Alliance
VV	Evangelical Churches, Other
XX	Unclassified Religions
ZA	Atheist

APPENDIX B. HOME OF RECORD STATES/TERRITORIES

Levels	Level Descriptions
AK	Alaska
AL	Alabama
AZ	Arizona
AR	Arkansas
CA	California
CO	Colorado
CT	Connecticut
DE	Delaware
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
OH	Ohio
OK	Oklahoma
OR	Oregon
PA	Pennsylvania
RI	Rhode Island

Levels	Level Descriptions
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
DC	District of Columbia
GU	Guam
PR	Puerto Rico
VI	U.S. Virgin Islands
WW	Unknown

APPENDIX C. PRIMARY MODEL VARIABLE INCLUSION

Term Length	3			4				5					6					
Year of Term	0	1	2	0	1	2	3	0	1	2	3	4	0	1	2	3	4	5
Variable																		
AFQT Category	5	8	11	13	7	11	11	8	4	5	12	9	8	9	12	13	10	15
Age at Enlistment	11	7	17	14	6	14	26	6	8	10	16	22	9	11	14	19	13	16
Anemia	44	33	32		38	26	36								41			
Asthma	35	23	24	30	32	31	24		14	34	36				33	39	28	23
Back Pain	8	10	8	5	13	7	5	13	17	20	3	4	7	24	1	14	3	6
BMI	17	18	16	16	16	9	12	9	10	12	13	8	10	10	6	10	8	7
Chronic Pain	37	15	2	34	31	2	1		25	4	1	3		28	15	6	2	1
CMF after IET	2	11	20	8	10	17	17	4	9	9	10	10	5	4	9	11	11	9
CMF at Enlistment	1	14	19	7	14	19	19	3	5	8	14	16	4	5	8	12	12	14
Dental Readiness	3	1	13	1	1	6	20	1	1	7	9	11	1	2	16	17	17	12
Deployment			29			35	27			18	21	19			13	16	19	17
Diabetes			49			47	47											
Education Category	27	9	35	22	17	15	42			22	35	32	22	20	37	15	37	
Epilepsy					43						26							
Gender	6	3	14	2	2	1	13	2	2	2	11	18	3	3	2	21	20	22
Headaches	18	13	3	17	9	3	3		26	25	15	23		13	10	28	6	5
Hearing Reading	38	35	25		35	38	30			30	30	38		32	38	24	36	
Heart Murmur	43	37	41		28	44	43			38				31				
Heart Trouble	33	5	10	31	27	22	28		21	6	25	34			39	5	34	36
Hispanic	16	27	27	21	20	20	33	15	19	21	29	28	15	22	23	33	31	24
HOR State	15	12	22	9	12	18	18	7	7	11	17	13	6	8	11	18	14	10
Hostile Injury		45	42			42	35				40	25			42	40	38	
Hypertension	40	30	31	33	36	28	31			36	37	19			30	32	21	8
Joint Pain	7	19	9	3	21	10	6	16	15	3	2	6	14	19	3	8	4	4
Liver Disease			48															
Marital Status	13	6	15	19	5	8	14	12	11	14	19	12	11	6	4	20	15	11
Mental Health	20	4	4	29	4	4	8	18		17	4	1		17	18	1	9	3
Non-Hostile Injury			39			41	38				33					42	39	
Number of Dependents	22	22	30	23	26	27	39	19	22	24	28	31	17	12	17	30	26	19
P after IET	9	20	7	4	8	25	9	23	28	31	27	15	21	25	7	22	16	26
U after IET	28	42	12		42	21	7			35	20	5			25	9	7	32
L after IET	29	39	6		33	16	4			32	6	2		30	28	3	5	2
H after IET	34	36	26	32	37	40	16					37			36	29	40	
E after IET	12	16	23	6	11	23	15	5	6	13	18	14	13	7	19	27	23	25
S after IET	42	40	5	27	34	45	2			37	5	30			31	2	1	13
P at Enlistment	23	31	37	11	25	36	34	24	24	23	32	36	24	26	24	25	32	29
U at Enlistment	31	44	47		44	46	32											
L at Enlistment	26	41	44		39	43	44				39				40	38	41	35
H at Enlistment	36	38	43	35	40	39	45								35	41		
E at Enlistment	21	21	28	20	22	24	22	10	16	16	22	27	18	16	22	26	22	21
S at Enlistment	41	34	45	36	41	37												
Pregnancy	32	26	21	24	18	5	10		12	1	34	17		29	34	4	18	33
Prior Service	25	2	1	15	3	12	28	25	3	15	8	21	2	1	5	7	29	28
Race Code	14	17	18	10	15	13	23	11	13	19	23	24	12	14	21	31	24	18

Term Length	3			4				5					6					
Year of Term	0	1	2	0	1	2	3	0	1	2	3	4	0	1	2	3	4	5
Variable																		
US Citizen. Origination	30	32	40	26	29	32	41	17	29	27	38	33	23	27	29	36	35	27
US Citizen. Status	39	43	46		45		46											
Vision Readiness	10	24	34	12	19	30	21	14	18	26	24	26	16	15	20	23	25	20
Waiver Admin	4	29	36	28	30	33	40	21	27	33	41	29		21	32	37	33	31
Waiver Conduct	24	25	32	25	23	29	25	22	20	29	7	7	20	23	26	34	30	34
Waiver Medical	19	28	38	18	24	34	37	20	23	28	31	35	19	18	27	35	26	30

APPENDIX D. SECONDARY MODEL VARIABLE INCLUSION

Term Length	3	4	5	6
Variable				
AFQT Category	6	7	5	10
Age at Enlistment	7	4	3	7
Anemia	36	37	31	40
Asthma	42	40		43
Back Pain	29	25	28	19
BMI	12	5	9	3
Chronic Pain	25	19	13	12
CMF after IET	2	8	4	5
CMF at Enlistment	3	9	6	4
Dental Readiness	1	1	1	1
Education Category	23	16	25	18
Epilepsy		45		
Gender	8	3	2	6
Headaches	28	22	29	21
Hearing Reading	37	35	26	32
Heart Murmur	40	38	40	37
Heart Trouble	31	32	27	36
Hispanic	22	29	24	28
HOR State	10	12	10	9
Hypertension	39	39	37	39
Joint Pain	33	28	16	17
Kidney Disease	45			
Marital Status	19	17	17	20
Mental Health	11	6	14	22
Number of Dependents	13	15	19	13
P after IET	5	2	11	8
U after IET	18	34	32	16
L after IET	9	18	15	11
H after IET	21	33	35	25
E after IET	16	10	7	15
S after IET	30	13	36	27
P at Enlistment	34	26	30	34
U at Enlistment	43	43		38
L at Enlistment	26	42	39	41
H at Enlistment	32	31	38	35
E at Enlistment	24	24	22	24
S at Enlistment	41	41	41	42
Pregnancy	38	23	21	33
Prior Service	4	11	8	2
Race Code	15	14	12	14
US Citizen. Origination	20	20	23	26
US Citizen. Status	44	44		
Vision Readiness	14	21	20	23
Waiver Admin	17	36	33	31
Waiver Conduct	35	27	18	29
Waiver Medical	27	30	34	30

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32, https://doi.org/10.1023/A:1010933404324.

Cammack J (2020) Predicting Army post-IET attrition using logistic regression and time-varying covariates. Master's thesis, Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA. http://hdl.handle.net/10945/65485.

Centers for Disease Control and Prevention (2021) Defining adult overweight and obesity. Accessed June 3, 2021, https://www.cdc.gov/obesity/adult/defining.html.

Devig A (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.

Dickstein C (2020) Army hits 2020 recruiting, retention goals amid pandemic, but top officials say more diversity needed. Stars and Stripes. Accessed March 29, 2021, https://www.stripes.com/news/us/army-hits-2020-recruiting-retention-goals-amid-pandemic-but-top-officials-say-more-diversity-needed-1.648068.

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

Ishwaran H, Kogalur U (2020) Fast unified random forests for survival, regression, and classification (RF-SRC). R package version 2.9.3, https://cran.r-project.org/package=randomForestSRC.

Ishwaran H (2015) The effect of splitting on random forests. *Machine Learning* 99(1):75-118, https://doi.org/10.1007/s10994-014-5451-2.

Ishwaran H, Kogalur U, Blackstone E, Lauer M (2008) Random survival forests. *Annals of Applied Statistics* 2(3), https://doi.org/10.1214/08-AOAS169.

Ishwaran H, Kogalur U, Gorodeski E, Minn A, Lauer M (2010) High-dimensional variable selection for survival data. *Journal of the American Statistical Association* 105(489):205-217, https://doi.org/10.1198/jasa.2009.tm08622.

Kaplan E L, Meier P (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282):457-481.

Power R (2019) U.S. military basic training attrition. The Balance Careers. Accessed April 24, 2021, https://www.thebalancecareers.com/united-states-military-basic-training-attrition-4052608.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schmidt M, Wright M, Ziegler A (2016) On the use of Harrell's C for clinical risk prediction via random survival forests. Accessed June 21, 2021, https://arxiv.org/pdf/1507.03092.pdf.

South T (2019) Rising costs, dwindling recruit numbers, increasing demands may bring back the military draft. *Military Times*. Accessed March 29, 2021, https://www.militarytimes.com/news/your-military/2019/11/19/rising-costs-dwindling-recruit-numbers-increasing-demands-may-bring-back-the-draft/.

Speten K (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.

Sprent P, Smeeton NC (2007) *Applied Nonparametric Statistical Methods*, 4th ed. (Chapman & Hall/CRC, New York, NY).

U.S. Army Recruiting Command (2021) Support Army Recruiting: FAQ. Accessed April 24, 2021, https://recruiting.army.mil/How-Can-I-Help/Support-Army-Recruiting/FAQ/.

Vie L, Griffith K, Scheier L, Lester P, Seligman M (2013) The Person-Event Data Environment: leveraging big data for studies of psychological strengths in soldiers. *Front. Psychol*. 4(934), https://doi.org/10.3389/fpsyg.2013.00934.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California