TRAC-M- TR-20-004
November 2019

US Army Post Initial Entry Training First-Term Attrition Analysis: Part II

The Research and Analysis Center
700 Dyer Road
Monterey, CA 93943-0692

This study cost the Department of Defense approximately $41,000 expended by TRAC in Fiscal Year 19.
Prepared on 20191106
TRAC Project Code # 060338

DISTRIBUTION STATEMENT: Approved for public release; distribution is unlimited. This determination was made on April 2018

THIS PAGE INTENTIONALLY LEFT BLANK

TRAC-M- TR-20-004
November 2019

US Army Post Initial Entry Training First-Term Attrition Analysis: Part II

Authors

Dr. Samuel E. Buttrey
Dr. Lyn R. Whitaker
MAJ Ta'Lena Fletcher
CPT(P) Aaron Devig
MAJ Gabe Gobea
MAJ Anthony D. Smith

PREPARED BY:

TA'LENA FLETCHER
MAJ, US Army
TRAC-MTRY

APPROVED BY:

BRIAN M. WADE
LTC, US Army
Director, TRAC-MTRY

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** 06 NOV 2019	**3. REPORT TYPE AND DATES COVERED** Technical Report, June 2018-September 2019	
4. TITLE AND SUBTITLE US Army Post Initial Entry Training First-Term Attrition Analysis: Part II			**5. PROJECT NUMBERS** TRAC Project Code 060338
6. AUTHOR(S) Dr. Samuel E. Buttrey, Dr. Lyn R. Whitaker, MAJ Ta'Lena Fletcher, CPT(P) Aaron Devig, MAJ Gabe Gobea, MAJ Anthony D. Smith			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The Research and Analysis Center - Monterey 700 Dyer Road Monterey CA, 93943-0692			**8. PERFORMING ORGANIZATION REPORT NUMBER** TRAC-M-TR-20-004
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) Army Analytics Group (AAG) Army Resilience Directorate (ARD)			**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES Findings of this report are not to be construed as an official Department of the Army (DA) position unless so designated by other authorized documents.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited			**12b. DISTRIBUTION CODE**
13. ABSTRACT (maximum 200 words) This research continues previous work in Army post Initial Entry Training (IET) first-term attrition by including medical data with existing personnel data to predict and understand attrition. We use supervised machine learning models to (1) identify the demographic and medical factors of Army enlisted personnel with highest probability of failure to implement preventative measures and (2) estimate total failures during the first enlistment term to set proper recruiting targets. We use classification and survival analysis techniques within the Person-event Data Environment (PDE) to inform sponsors on attrition trends. We use model results as inputs to an application that displays the predicted probability of success for first term enlistees. The results and application have applicability to other DoD organizations concerned with accession and retention. We find that a soldier's medical history, particularly his Dental Class, PULHES Deployable status, and the duration of the initial contract are significant predictors of whether a soldier will complete his or her first term. Knowledge of the key factors and other influencing variables assists the Army Resiliency Directorate in creation of models to better advise U.S. Army leadership on intervention strategies and preventative measures to preclude the loss of first-term soldiers.			
14. SUBJECT TERMS Supervised Machine Learning, Binary Logistic Regression, Classification Trees, Random Forests, Survival Analysis, Manning, Personnel, Prediction, Attrition			**15. NUMBER OF PAGES** 67 **16. PRICE CODE**
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500

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

THIS PAGE INTENTIONALLY LEFT BLANK

NOTICES

DISCLAIMER

Findings of this report are not to be construed as an official Department of the Army (DA) position unless so designated by other authorized documents.

REPRODUCTION

Reproduction of this document, in whole or part, is prohibited except by permission of the Director, TRAC, ATTN: ATRC, 255 Sedgwick Avenue, Fort Leavenworth, Kansas 66027-2345.

DISTRIBUTION STATEMENT

Approved for public release; distribution is unlimited.

DESTRUCTION NOTICE

When this report is no longer needed, DA organizations will destroy it according to procedures given in AR 380-5, DA Information Security Program. All others will return this report to Director, TRAC, ATTN: ATRC, 255 Sedgwick Avenue, Fort Leavenworth, Kansas 66027-2345.

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This research continues previous work in Army post Initial Entry Training (IET) first-term attrition by including medical data with existing personnel data to predict and understand attrition. We use supervised machine learning models to (1) identify the demographic and medical factors of Army enlisted personnel with highest probability of failure to implement preventative measures and (2) estimate total failures during the first enlistment term to set proper recruiting targets. We used classification and survival analysis techniques within the Person-event Data Environment (PDE) to inform sponsors on attrition trends. We used the model results as inputs to an application that displays the predicted probability of success for first term enlistees. The results and application have applicability to other DoD organizations concerned with accession and retention. We found that a soldier's medical history, particularly his Dental Class, PULHES Deployable status, and the duration of the initial contract are significant predictors of whether a soldier will complete his or her first term. Knowledge of the key factors and other influencing variables assists the Army Resiliency Directorate in creation of models to better advise U.S. Army leadership on intervention strategies and preventative measures to preclude the loss of first-term soldiers.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AAG	Army Analytics Group
AFQT	Armed Forces Qualification Test
ARD	Army Resiliency Directorate
ASVAB	Armed Services Vocational Aptitude Battery
AWD	Army Waiver Database
AUC	Area Under the Curve
BASD	Basic Active Service Date
CMF	Career Management Field
CTS-OCO	Contingency Tracking System – Overseas Contingency Operations
DA	Department of the Army
DCIPS	Defense Casualty Information Processing System
DMDC	Defense Manpower Data Center
DOD	Department of Defense
GAO	Government Accountability Office
GED	General Education Diploma
HOR	Home of Record
HRC	Human Resources Command
IET	Initial Entry Training
MEPCOM	Military Entrance and Processing Command
MOS	Military Occupation Specialty
MTOE	Modified Table of Organization and Equipment
PDE	Person-event Data Environment
PID	Person Identifier

TAPDB	Total Army Personnel Database
TSC	Test Score Category
TDA	Table of Distribution and Allowances
TRAC	The Research and Analysis Center
TRADOC	Training and Doctrine Command
RF	Random Forest
ROC	Receiver Operating Characteristic
USAREC	United States Army Recruiting Command

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The ability to successfully recruit and retain soldiers will be a challenge as long as there is a volunteer Army. Training civilians to become soldiers and then ensuring they complete their initial obligation is an expensive proposition. The goal of this research is to identify, through the development of statistical models, demographic, administrative, and medical factors that influence the probability of first-term active component enlisted soldiers to complete their initial contract obligation. Our work is an extension of work summarized in the Smith, Anglemyer, Speten, and Shingleton (2018) report. Previous work focused on demographic and administrative factors and their relationship with first-term attrition. We extend this work through the addition of medical factors and explore two approaches. Following Speten (2018), we first fit the binary regression models, logistic regression and random forests, to estimate the probability of first-term attrition based on the individual attributes immediately following IET and with the addition of medical attributes. We also employ survival analysis, using survival tree models, which account for the time-varying nature of medical and other soldier attributes, to get a temporal look at attrition. The objective of the modeling efforts are to gain insight and to provide predictive models of attrition at the individual or group level.

We find that in all models a soldier's Dental Class, PULHES Deployable status, and the duration of the initial contract are significant predictors of whether a soldier will complete his or her first term. A Dental Class of four and a PULHES Deployable status of non-deployable tend to increase the chance of attrition. As expected, the chance of first-term attrition increases with contract length. Models are fit using cohort data, where a cohort is defined as all soldiers enlisting in the same Fiscal Year (FY). Data for a cohort includes features measured over the course of their first term. We use FY 2008 – FY 2011 cohorts for model fitting and testing. We find that models using FY 2010 enlistees are most similar to those of FY 2011 and we postulate that this is, in part, due to the missing medical data before FY 2010.

The accuracy rate of the logistic regression model is 83% and provides enough fidelity to warrant consideration of its use by Army planners. We also find that the

survival analysis approach shows promise, because unlike binary regression approaches, survival analysis allows for predicting attrition rates at any time during the first term and accounts for time-varying variables. The analysis from this project enhances the Army Resiliency Directorate's (ARD) ability to recommend intervention strategies and preventative measures to preclude the loss of first-term soldiers. Most importantly, this research provides ARD insight as the agency continues its efforts to improve soldier resiliency and, by extension, first-term attrition rates. Application of our predictive model to the administrative records of current enlistees could provide policy makers with probability estimates of all first-term soldiers and facilitate the creation of intervention programs and prioritized resource strategies built upon a quantitative foundation.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 1. INTRODUCTION

1.1. PURPOSE

This report summarizes the second year of the Post-IET first-term attrition study for active duty (AD) soldiers. It continued the previous year's work summarized in US Army Post-IET First-Term Attrition Analysis (Anglemyer, Smith, & Speten, 2018) report and based primarily on Speten's Operations Research Master's thesis (Speten, 2018). The goal of our research was to extend this work by adding medical factors to demographic and administrative factors available in every soldier's service record. We then used these augmented datasets to inform statistical models to predict a soldiers' probability of failing to complete their initial contractual obligation. We continued to address questions posed by the ARD at the beginning of the study: (i) what are the demographic and medical factors of personnel with highest probability of failure? And (ii) what is the mean number of total failures during the first enlistment term? For this year's work, our focus was to address the following questions: (i) Does the incorporation of medical data provide additional insights to inform understanding of soldier first-term attrition? (ii) Do alternative modeling approaches provide additional insights? (iii) What models and data might be useful to a decision maker concerned with first-term attrition in their organization?

To answer these questions we took two approaches. Following Speten's methodology laid out in his thesis (Speten, 2018), we first fitted the binary regression models, logistic regression and random forests, to estimate the probability of first-term attrition but with the addition of medical attributes. Secondly, we conducted survival analysis, which accounted for the time-varying nature of medical and other soldier attributes, to get a temporal look at attrition. The purpose of the model fits, using either approach, was to gain insight and to give ARD models with which they can predict attrition for an individual, or a group of, based on individual soldier attributes. We scoped the analysis with the constraints, limitations and assumptions described in the next section. Most of these were carried over from last year's work.

1.2. CONSTRAINTS, LIMITATIONS, & ASSUMPTIONS

Constraints - *limit the study team's options to conduct the study*:

- All analysis must be performed in the Person-Event Data Environment (PDE).

- We must finish our research no later than 30 September 2019.

Limitations - ***a study team's inabilities to investigate issues within the sponsor's bounds*****:**

- Analysis was limited to six types of data available through the PDE.
    - Active Duty Army Master and Transaction files, Army Waiver Database (AWD) files, Military Entrance Processing Command (MEPCOM) files, Periodic Health Assessment (PHA) files, Medical Protection System (MEDPRO) files.
- Data available lacked a consistent indicator of attrition and the date of attrition.
- Unit Identification Codes (UIC) were scrambled and duty station zip codes were obscured in the PDE-provided data.
- Even though soldiers were required to have a yearly PHA, the PHA files have a large number of missing values.
- Some Career Management Fields (CMF) codes were re-coded during the course of the study. In our studies, we took CMF codes as they appeared in the dataset and did not adjust for changes in how CMFs were coded.

Assumptions – *study-specific statements that are taken as true in the absence of facts*:

- Data maintained within the PDE were accurate and represented complete first-term soldier information for soldiers enlisted in the seven Fiscal Years (FY) 2005 – 2011.
- No significant changes occurred in a soldier's record that were not accurately captured by the quarterly snapshot dates available.
- Less than 1% of our total population have contractual obligation durations with "odd" values of one, two, seven, or eight years. We assumed the standard enlistment contract was between three to six years. Since the odd values represented such a small percentage of the data, we removed these observations.
- In the creation of the predictor variables, 2.5% of the soldiers had the same odd contractual obligation values, which prevented selection of the soldier record with

the correct end date of their first-term. We assumed these entries were erroneous so we computed the average contractual period from the full population and assigned the odd contractual obligation observations a value of 4 years.

- Soldiers successfully completed their initial obligation if their separation date was within three months of their first-term end date.

1.3. BACKGROUND & LITERATURE REVIEW

Smith et.al. (2018) and Speten (2018) give a general literature overview, which we did not repeat here. In this section, we reviewed some of the Speten (2018) contributions that form the basis of our FY 2019 work. We also give a brief introduction to survival analysis, since these methods were not as well known in military manpower applications.

1.3.1. Previous Research

Because Speten (2018) was the first to perform a modern post-IET attrition study using data available in the PDE, an invaluable part of his contribution was the data science aspect of understanding and manipulating the PDE databases. Speten (2018) was able to form an AD cohort of soldiers who completed IET based on first-term records of all enlisted soldiers who assessed in FY 2005 - FY 2010. He provided a roadmap for how to identify soldiers who completed IET and whether they attrite or not at the end of their first term. Speten (2018) found that the cohort (post-IET first-term) attrition rate was fairly constant with an overall attrition rate of 26%.

For his cohort, Speten (2018) constructed 32 demographic and administrative variables that he used to estimate the probability of attrition. He fitted three binary regression models based on an 80% training set selected at random from his cohort: an additive logistic regression model, a classification tree, and a random forest. The variables used in each are given in Table 1. Unless obvious (e.g. Prior Service) or specified (e.g. Rank (Enlistment)), these variables were computed using all records available in the first term. These groups of variables were the ones that were the most strongly related to attrition in the sense that they gave the best estimates of the probability of attrition for each model.

Variable	Logistic Regression	Classification Tree	Random Forest
AFQT Category	x		
Citizenship Origination	x		
Contract Duration	x	x	x
Days Deployed	x	x	x
Dependents	x		x
Education Tier	x	x	x
Gender	x		
Height (Enlistment)	x		x
Marital Status	x	x	x
Military Occupation Group	x		x
Non-hostile Injuries	x		
Prior Service	x		x
Rank (Enlistment)	x	x	x
Unit Type	x	x	x
Waiver (Admin)	x		
Waiver (Conduct)	x		
Weight (Enlistment)	x		x

Table 1 Variables used, reproduced from Speten (2018)

The performance of the three models on the remaining 20% test set were similar. Because the random forest and classification tree models automatically include interaction terms, they were able to almost match the additive logistic regression model performance using fewer variables. Speten (2018) discussed the pros and cons of these models. In addition, Smith et. al. (2018) provided a web-based application to facilitate the use of these models.

1.3.2. Survival Analysis

Most manpower attrition studies used binary-type regression models. They were used to estimate the probability of attrition in a specific time window -- in our case, between completion of IET and the end of the first term. Without modifying the time window, these models cannot estimate the probabilities of attrition for times before the end of the time window. Nor can they give conditional probabilities of, for example, attrition in the second year, given that the soldier has completed his/her first year. An alternative was to use survival analysis, an approach often used in the bio-medical fields. See Kalbfliesch and Prentice (2002) for an introduction to survival analysis techniques. In survival analysis, the goal is to estimate the survival function, $S(t)$, for $t \geq 0$. The survival function is the probability of "surviving" past time *t*. For post-IET, first-term

attrition, time was taken to be the time (in years) since enlistment. Surviving to time *t* means that the soldier did not attrite at time *t* or before.

As an example, Figure 1 shows the estimated survival functions from a sample of soldiers enlisting in FY 2010 with a four-year enlistment obligation by the variable Dental Class. Survival functions are estimated separately for each group using the non-parametric Kaplan-Meier estimate (Kaplan & Meier, 1958). Here, the Dental Class value was the first Dental Class reported within six months of IET. Dental Class 1, 2, and 3 indicate severity, but Dental Class 4 indicates no dental examination in the prior year. A binary regression model with response "attrite" or "not attrite" in the first term would give the estimated probability of attrition for these soldiers at only the four-year mark. Figure 1 survival functions, however, give the probabilities of survival (1 – probability of attrition) for all *t* up to four years. They show that although the four-year attrition probability for Dental Classes 3 and 4 are about the same, Dental Class 3 attrition tends to occur earlier in the first term. So in this example the time-bases analysis reveals information not available to the logistic regression.

There are a variety of regression-type approaches for estimating survival functions that depend on input variable values. We choose the approach survival trees which marries traditional survival analysis with machine learning. Early work on survival trees dates back to Gordon and Olshen (1985). Survival trees first partition the data into subsets according to variable values and then use the Kaplan-Meier estimator to estimate the survival function for each subset. The algorithm that chooses the subsets was similar to that used in Classification and Regression Trees (CART). Using the predictor variables, the algorithm recursively partitions the data into two subsets, where each split was chosen so that the two resulting subsets give the estimated survival functions that are the most dissimilar. Dissimilarity was measured by the *p-values* for the log-rank test that the survival functions of the two subsets are different. This approach has the advantages that splits are found algorithmically and not by hand, and that in each final subset (or leaf), the survival functions are estimated non-parametrically. Further, a relatively new application of survival trees allows explicit inclusion of time-varying covariates (Fu and Siminoff, 2016). We illustrate these methods in Chapter 3.

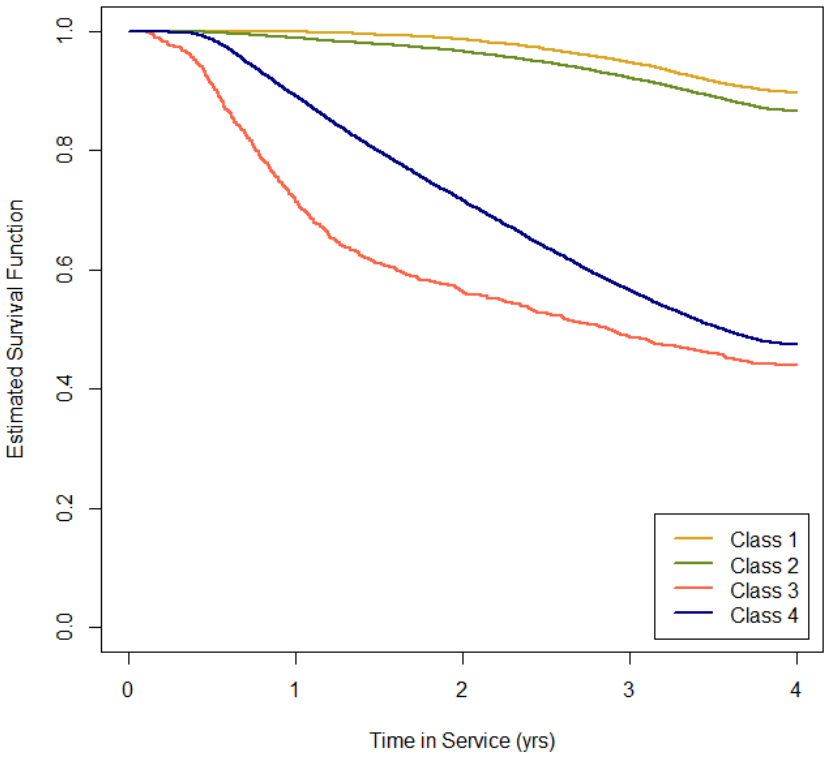


Figure 1 Survival Function, four-year term of enlistment by Dental Class within six months of IET.

To our knowledge, this was the first application of survival trees to military attrition studies. However, while not common, survival analysis has been used to study military attrition; Hawes (1990) studies attrition of male Marines as a function of Armed Forces Qualification Test (AFQT) scores, education credential, and whether the enlistee received a moral waiver. In addition to findings that support earlier non-survival analysis based work, he finds that the relationship between attrition and aptitude scores weaken with time and that difference in attrition rates as a function of the three variables studied was most noticeable in the first four months after enlistment. Rubiano (1993) also used survival analysis in studies of U. S. Coast Guard attrition, but rather than using estimated survival functions, he forecasted monthly attrition using a regression model.

1.4. TECHNICAL APPROACH

All analysis was done in the PDE using the R computing environment (R Core Team, 2019). In this section, we briefly discuss the binary regression and survival tree approaches, how

predictor variables are chosen and constructed for each, and how we chose and assessed the final model fits.

1.4.1. Methods

The first set of methods used a combination of additive logistic regression and random forests, as laid out by Gobea (2019) and discussed further in Chapter 3, to estimate probabilities of post-IET first-term attrition. This approach was patterned after that of Speten (2018). The second set of methods allows for explicit use of time-varying variables using survival trees to estimate the probability of attrition at any time during the first term. This gave a temporal look at the factors that contribute to first-term attrition.

The ultimate goal was to use these estimated attrition probabilities to better understand how the combined demographic and medical factors contribute to attrition, to predict attrition for individuals, and to estimate the expected number of attrites, based on data currently in hand. These goals required several considerations. First, input (predictor) variables must be predictors of, and not consequents, of attrition. Second, with the addition of medical variables, we relied more heavily on the ideas from statistical/machine learning, rather than on traditional statistical methods, to assess model performance. Finally, unlike the demographic variables used by Speten (2018), medical variables based on PHA had a great number of missing values. We mitigated the effect of the missing values in two ways. First, we limited the model fitting to records obtained no earlier than FY 2008. Secondly, we imputed missing values with random forests. For binary regression models imputation was implemented using the R package **`missRanger`** of Mayer (2009). For survival trees we used the methods of Ishwaran, Kogalur, Blackstone, and Laur (2008) implemented in the R package **`randomForestSRC`** (Ishwaran and Kogalur, 2019). The imputation was non-parametric and used all variable values that were present in a record to impute the missing values.

1.4.2. Predictor Variables

Speten (2018) identified variables that were strongly related to attrition. Some, the time-constant variables (e.g. gender and enlistment waivers), could be ascertained using data from the completion of IET or earlier. These could be used to estimate the probability of first-term attrition

for an individual or group of soldiers at any time before their term is completed and as early as the end of IET. Other variables such as number of deployments, marital status, and education could change with time. We call these *time-varying* variables. To accommodate time-varying variables, Speten (2018) used variable values measured at enlistment and the highest or last observed value in the first term. Thus, for example, he used both the rank at enlistment and the maximum rank attained in the first term as input variables. His analyses showed which of these variables were most related to attrition. His models also established the importance of including time-varying as well as time-constant variables in attrition models. However, to use these models to forecast or predict attrition, one must estimate or guess the values of the time-varying variables (e.g. the number of first-term deployments) and this information would not be available until the end of a soldier's first term. Further, many of these variables, such as number of deployments and maximum rank had values that were, in part, a consequence of length of time in service. For example, soldiers who completed their first-term commitment naturally tend on average, to have deployed for more days than similar soldiers who attrite before the end of their first term. Without accounting for the time until attrition, it was difficult to discern how much the values of these variables were a consequence of attrition or a predictor of attrition.

For these reasons, Gobea (2019) estimated the probability of first-term attrition using binary regression models similar to those of Speten (2018). Unlike Speten's, Gobea's demographic and administrative variables utilized values known only at the time of enlistment and for the values of the medical variables this information was gleaned as closely as possible, immediately after IET. Gobea (2019) explains theses distinctions in further detail. Devig (2019), on the other hand used survival analysis to estimate the probability of attrition as a function of time in service. Thus, his response variables were from start of enlistment to attrition or end of first term, whichever occurs first, plus the binary (event) variable indicating whether the soldier has undergone attrition or not. To fit regression-type survival function models, values of time-varying variables must be "predictable." That means, roughly, that the probability of surviving past a time t can only depend on values of time-varying variables up to time t. Predictable means that we cannot look into a soldier's future to estimate his probability of surviving past a certain time.

1.4.3. Test, Validation, and Test Sets

The methods used in this work have their roots in traditional statistics. Under appropriate conditions, large-sample inference results (e.g. standard errors, confidence intervals, hypothesis tests) were available for both logistic regression and survival analysis using Kaplan Meier estimators. We used some of these to guide our choice of models. However, because the datasets were large in number of records and, with the addition of medical variables, in the number of variables available for each record, it was infeasible to rely solely on traditional statistics. Instead, we assessed model performance using independent hold-out sets. These hold-out sets were of two types: validation sets for model selection, and test sets to assess the final model. The remainder of the data, the training set, was used for model fitting. The training, validation, and test datasets were constructed after construction of the cohort dataset.

The training and validation sets were independent sets selected at random with an 80/20 split. For the binary regression models, Gobea (2019) selected training and validation sets from soldiers who enlisted in FY 2008 - FY 2009. Devig (2019), in his final survival analysis model, selected training and validation sets from soldiers who enlisted in FY 2010. The training set was used to fit multiple models. Performance metrics for how well these models predict attrition, based on the validation set, were used to compare and choose among these model fits. This allowed for comparisons based on soldier records that had not been used to fit the models. In addition to validation, some techniques, such as lasso-regularized logistic regression, used cross-validation on the training set. Cross-validation was used to select a model of appropriate complexity among many model fits. With cross-validation, the training set was randomly partitioned into (usually) ten subsets or "folds." Each subset in turn played the role of the hold-out validation set and cross-validated measures of performance were averages of the folds' performance measures.

For both approaches, the test set is chosen to be soldiers who enlist a year later than those in the training and validation sets. Thus Gobea (2019) uses soldiers from FY 2010, and Devig (2019), from FY 2011. This ensures that different records are used for fitting and selecting the final model than are used for assessment. It also gives us a hint about the potential effects of forecasting attrition. We note that this is not a complete assessment of forecast results because the soldiers whose records are used for model training and validation have first terms that overlap with those of the test set soldiers.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 2. DATA PREPARATION

This project uses eight datasets accessed through The Army Analytics Group (AAG) PDE environment. The PDE and its purpose is described by Jensen (2016) and independently assessed by Knapp, Asch, DeMartini, Ruder and Hanley (2018). The PDE was developed and is currently used by the AAG to support analysis projects for senior Army leadership, research data management and model validation for large Army studies. The system is designed as a self-service and collaborative environment, allowing those who need such data to retrieve and analyze the data with some support, and give Department of Defense (DoD) senior leaders actionable information. PDE includes a project management suite that allows users to define a study, invite team members to join the study, specify data sets from a data catalog, conduct analyses and publish results with controlled or open availability (Jensen, 2016, pg 6).

2.1. DATA SOURCES

We use only four of the six data sources used by and described in detail by Speten (2018). The primary dataset is the quarterly snapshots of demographic and administrative information in the Active Duty Military Personnel Master. This is merged with the Active Duty Military Personnel Transaction dataset, which captures the changes in a soldier's record such as enlistment into and separation from the Army. Information about the soldier at enlistment is taken from the MEPCOM -700 file and the U.S. Army Recruiting Command (USAREC) maintained AWD files. We do not use the Defense Casualty Information Processing System (DCIPS) Injury file or the Contingency Tracking System – Overseas Contingency Operations (CTS-OCO) file.

We use two datasets that contain medical information, not used by Speten (2018). These are the PHA datasets of yearly PHA mental and physical assessments. We merge the new and old format PHA datasets. The MEDPROS dataset that tracks immunization, readiness, and deployability status is also used. In particular it provides the Physical / Upper (Extremities) / Lower (Extremities) / Hearing / Eyesight / Psychiatric (PULHES) six-digit code of a soldier's overall health. The MEDPROS dataset is transactional in that its fields are not updated at fixed intervals, but as a soldier's record changes. The DD3349 dataset has too many missing values to be useful and is not used in our work. Speten (2018) constructs a cohort of records for all AD soldiers

enlisting in FY 2005 - FY 2010 who complete IET. For the survival analysis portion of our study, we add soldiers enlisting in FY 2011 to that cohort.

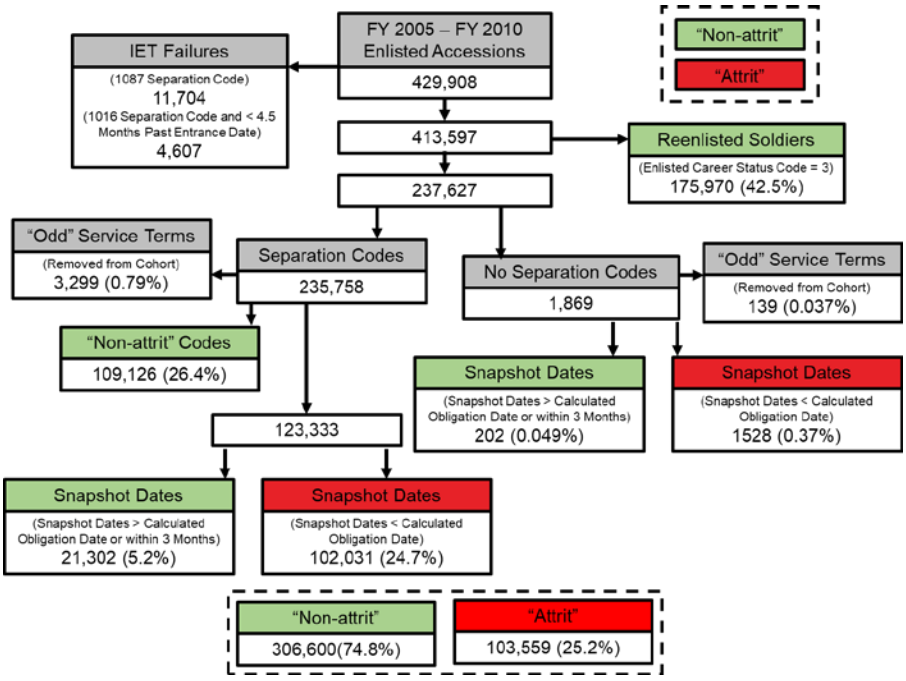
2.2. COHORT CONSTRUCTION

Cohort construction followed that of Speten (2018) very closely, with a slight modification. In addition, we added medical predictor variables and made adjustments for time-varying variables. The dataset used for the binary regression models was based on a cohort that contains 410,159 records and 105 variables. The survival analysis dataset was based on a cohort of 461,964 records and 221 variables. These are described briefly here, and in detail in Gobea (2019) and Devig (2019).

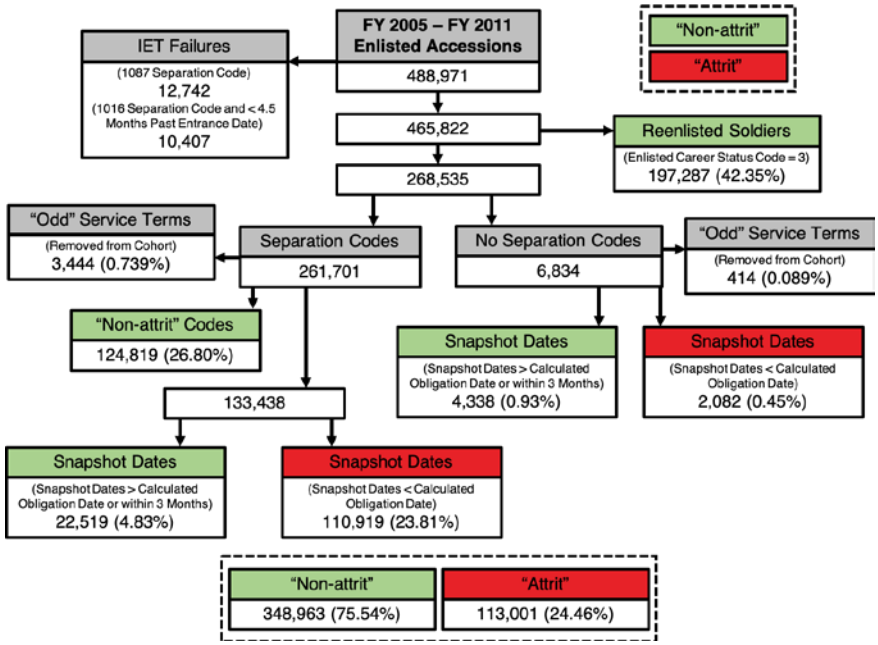
2.2.1. Cohort and Response Variable Construction

Construction of the cohort was tied to the construction of the attrite/not attrite response variable. We followed the methods of Speten (2018) with the modification, described by Devig (2019), that we also exclude soldiers with separation code "1016" with a separation time of less than 4.5 months whom we identify as not completing IET. In Figure 2 we reproduced the flowcharts illustrating how the cohort was constructed from Gobea (2019) and Devig (2019) respectively.

Survival analysis methods require variables that record a time when the event of interest (i.e. attrition) occurs or the time when that observation ends (i.e. end of first term, if no attrition occurs) and an event indicator (attrite or not at that time). Thus, for each record in the cohort, we included a Start Date and End Date. The Start Date was the date of enlistment and the End Date was the date of attrition (for soldiers who attrite) or the end of first-term obligation (for soldiers who complete their first term). These dates were used to compute the start times and end times, measured in years from the Start Date.



(a)



(b)

Figure 2 Flowcharts depicting cohort construction for (a) the binary regression analyses and (b) the survival analyses. Reproduced from Gobea (2019) and Devig (2019) respectively.

2.2.2. Predictor Variables for Binary Regression Models

For the binary regression models, the non-medical predictor variables were selected from Speten's (2018) variables. These were listed in Table 2 and described by Speten (2018).

Variable	Data Source	Data Type	Factor Levels
AFQT Category	Master	Categorical	5
Age at Enlistment (Years)	Master	Numeric	N/A
ASVAB GT Score (Scale 3-150)	MEPCOM	Numeric	N/A
Citizenship Origination	Master	Categorical	3
Citizenship Status (Enlistment)	Master	Binary	2
Contract Duration (Years)	Master	Numeric	N/A
Education Level (Enlistment)	Master	Categorical	4
Education Tier	Master	Categorical	3
Gender	Master	Binary	2
Height at Enlistment (Inches)	MEPCOM	Numeric	N/A
Home of Record Region	Master	Categorical	5
Military Occupation	Master	Categorical	26
Military Occupation Group	Master	Categorical	3
Prior Service	Transaction	Binary	2
Rank (Enlistment)	Master	Categorical	6
Waiver (Admin)	AWD	Binary	2
Waiver (Conduct)	AWD	Binary	2
Waiver (Drug)	AWD	Binary	2
Waiver (Medical)	AWD	Binary	2
Weight at Enlistment (Pounds)	MEPCOM	Numeric	N/A

Table 2 Variable Summary and Data Source Mapping

We added thirty-nine new medical variables from the PHA and MEDPROS datasets to those in Table 2. All are categorical. Of these, 29 were binary (yes/no) variables. These include two reported by a medical provider: the Medical Non-deployable Profile and the Limited Duty Profile. Also included was a PULHES Deployable code computed from the PUHLES six-digit code as "no" if any of the PULHES six digits are 3 (significant limitation) or 4 (severely limited) and "yes" if all six digits are 1 (no limitations) or 2 (some limitation). The rest of the binary variables were self-reported indicators of health issues such as anemia or epilepsy, allergies, tobacco chewing or smoking, and an answer to a question about medical conditions that kept the soldier from performing their duties. The ten multi-level categorical medical variables were: Dental. Hearing, and Vison Readiness Classes, the six variables from the six-digit PULHES code,

and two variables indicating father's and mother's history of chemical dependence. Description of these variables and their dataset field names are given in Gobea (2019) and Devig (2019).

Medical variables such as the presence of some conditions (e.g. allergies to latex or bee stings) were not time-varying, but might be discovered at any time during the first term. These we set to "yes" if there was any record of them in the first term. A "no" for these variables means that the soldier did not have the condition or it had not yet been recorded. Others, such as the presence of back pain and PULHES codes, were time-varying. These we set to their first value that was recorded after IET.

2.2.3. Predictor Variables for Survival Analysis

For the survival analysis, variables were treated as time-varying where possible. Devig (2019) gave a detailed description of these variables. In Table 3, we list those that were treated as time-varying. For each of the variables in Table 3, we captured the date the record changed as well as what it changed to. Keeping track of changes more than doubled the number of fields required for the cohort dataset. The cohort dataset with time-varying variables recorded with one record per soldier is called a "short" format dataset.

Variable
Anemia
Asthma
Back Pain
Cancer
Chronic Pain
Dental Class
Diabetes
Epilepsy
Headaches
Hearing Readiness Class
Heart Murmur
Heart Trouble
Hypertension
Joint Pain
Kidney Disease
Liver Disease
Marital Status Code

Variable
Mental Health Concerns
Pregnancy Status
PULHES–Eyesight
PULHES–Hearing
PULHES–Lower Extremities
PULHES–Physical Capacity/Stamina
PULHES–Psychiatric
PULHES–Upper Extremities
Vision Readiness Class

Table 3 Time-Varying Covariates

To use the data in survival analysis, it was restructured to a "long" format developed by Anderson and Gill (1982). In the long format, each soldier's record was rendered as at least one "pseudo-record". Each pseudo-record corresponds to a half-open time interval (start time, stop time) during which none of the time-varying variables change (where start and stop times were times since Start Date measured in years). An event variable was defined for each pseudo-record; it was set to 1 if the soldier attrites at the end of the pseudo-record time interval and 0 otherwise. The event for the pseudo-record whose end time corresponds to End Date was equivalent to the response variable used for the binary regression models. This well-known trick of reformatting data from short to long form, allowed us to fit models with time-varying variables by using survival analysis methods (and software tools) designed for time-constant variables as long as these methods accommodated left-truncated data.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 3. ANALYSIS AND FINDINGS

3.1. COHORT DATASET OVERVIEW

Devig (2019) and Speten (2018) gave post-IET first-term attrition rates by enlistment year. Devig (2019) rates are reproduced in Table 4. Table 4's reported attrition rate across the entire cohort of approximately 24.5% was a bit lower than that of Speten (2018), likely because the cohort construction differed a bit as explained in Section 2.3.1.

	Fiscal Year						
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Non-attrit	77.28% (50,520)	75.84% (56,293)	74.37% (51,353)	73.96% (50,578)	75.07% (45,705)	76.71% (50,712)	75.60% (43,802)
Attrit	22.72% (14,856)	24.16% (17,929)	25.63% (17,694)	26.04% (17,809)	24.93% (15,175)	23.29% (15,401)	24.40% (14,137)

Table 4 Attrition Rate by Fiscal Year of Enlistment

Devig (2019) and Speten (2018) also gave attrition rates and distributions for demographic and administrative variables. Their results are comparable and are not included here. We do note, though that the Vehicle Mechanics CMF was recoded from 63 to 91. For the FY 2008 and FY 2009 data, CMF 63 for Military Occupation had an attrition rate greater than 90% whereas CMF 91 attrition rates were more comparable to the overall 24% attrition rates. We postulate that the soldiers who had been coded as 63, but attrited, were never updated to the new 93 CMF code. It appears that for these years, the 63 code was a consequence and not a predictor of attrition. We did not delve more deeply into other possible CMF code changes.

Devig (2019) gave overviews of attrition rates and distributions of the new medical variables measured as closely as possible to their values immediately after IET. We highlight a few of his results here. But first it is important to note that even though soldiers were required to have a yearly PHA, their results were not always recorded. Table 5 shows the percentage of soldiers who have complete PHA records by enlistment year. We take two steps to mitigate missing values. First we only used data from soldiers who enlist in FY 2008 or beyond. Secondly, we imputed missing values using random forests. As a consequence of the imputation, the marginal distribution of a variables imputed values is the non-missing variable's distribution.

		Fiscal Year						
		2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
PHA Data?	**Yes**	20.48%	42.44%	64.29%	80.05%	88.31%	93.43%	94.39%
	No	79.52%	57.56%	35.71%	19.95%	11.69%	6.57%	5.61%

Table 5 Physical Health Assessment Data by Fiscal Year

Of the PHA variables, there was a marked difference in attrition rates by Dental Class, and to a lesser extent by Hearing and Vision Readiness Class. Figure 3 shows the attrition rates by Dental Class and enlistment year. Dental Class 3 attrition rates were high across enlistment years, and Dental Class 1 attrition rates tend to be low across enlistment years. However, patterns in attrition rates for the other classes, especially Class 4, change with time. Since Figure 3 only reflects attrition rates of records with non-missing Dental Class, it is probable that the attrition rate changes over time are, in part, an artifact of PHA record keeping practices. For all three variables (Devig, 2019), the attrition rates by class seem to be fairly similar for the last two enlistment years FY 2010 and FY 2011 when fewer soldiers were missing PHA data.

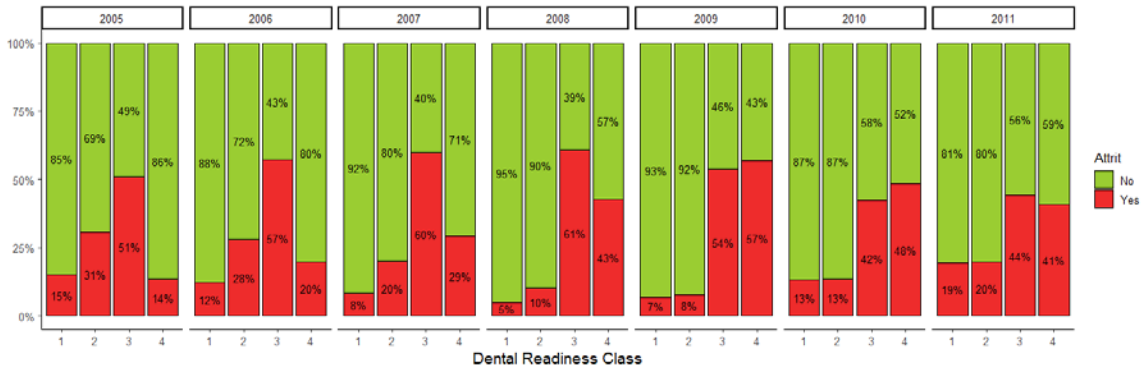


Figure 3 Attrition by Dental Readiness Class following IET Completion

There is also a marked difference in attrition rates by PULHES code where the higher the score, the more severe the limitation, as seen in Figure 4.

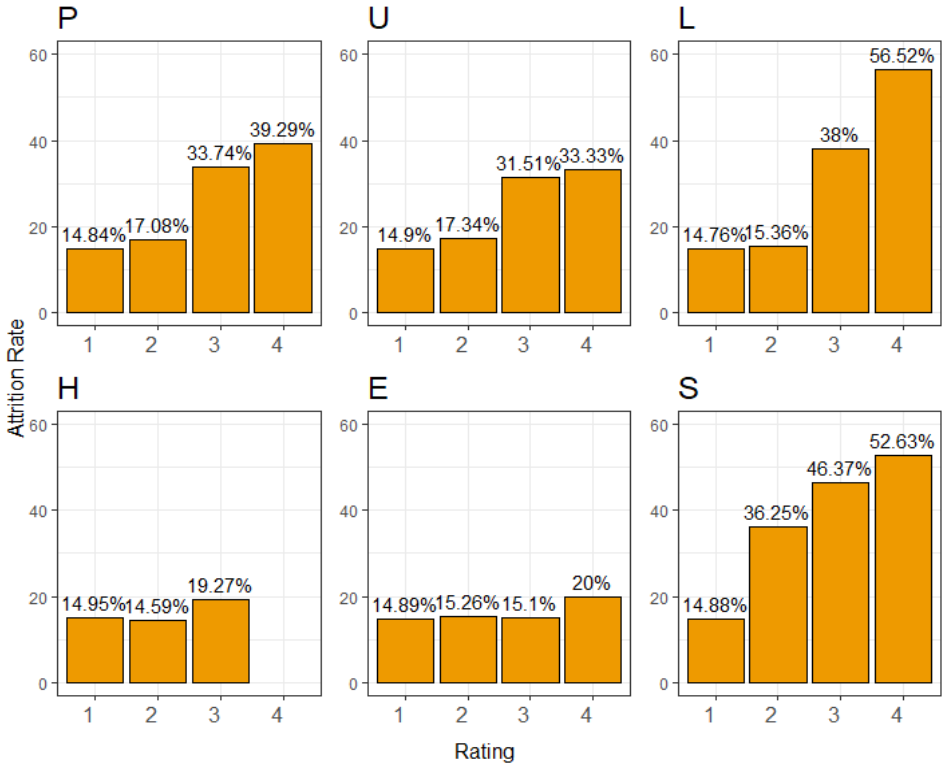


Figure 4 Attrition rates by PULHES code

3.2. BINARY REGRESSION MODELING

In this section, we report the results of Gobea (2019) along with some additional observations. The goal was to fit models to estimate the probability of first-term attrition that can be used as part of a prediction tool, to see the effect of adding medical predictor variables, and to gain an understanding of the importance of these variables. To do this, we used logistic regression and random forest models fit to the 80% training set from the FY 2008 – FY 2009 enlistments (with imputed medical variables). The logistic regression fits were primarily additive. This means they did not account for potential interactions or for potential transformations of the numeric predictors (e.g. height and weight at enlistment). We used additive logistic regression models so that ARD can implement them fairly easily with explicit equations. We also fitted random forest models. Random forests naturally accounts for interactions and transformations of numeric predictors, but we used them as a variable selection device and for exploration rather than for prediction. To understand model performance, especially the effects of the medical variables, we

plotted Receiver Operating Characteristic (ROC) curves computed on the 20% validation set. This helps avoid overfitting, especially when using the random forest models. Our final model fit was evaluated on the FY 2010 test set data.

3.2.1. Insights from Initial Model Fits

Before trying to reduce the number of predictors, we performed a few simple experiments to see if and how medical predictors aid in predicting attrition. We first fitted three additive logistic regression models: a full model with all predictors used by Gobea (2019), a logistic regression model with just the medical predictors, and a logistic regression model without medical predictors. The validation set ROC curves for the three are given in Figure 5. The ROC curves computed on the training set, but not displayed here, were almost identical to those of Figure 5. This was an indication that our models are not over-fit. We see from Figure 5 that the additive model with the medical variables alone performs almost as well as the full model. This underscores the importance of medical variables in our models. Fitting the logistic regression with only the predictable

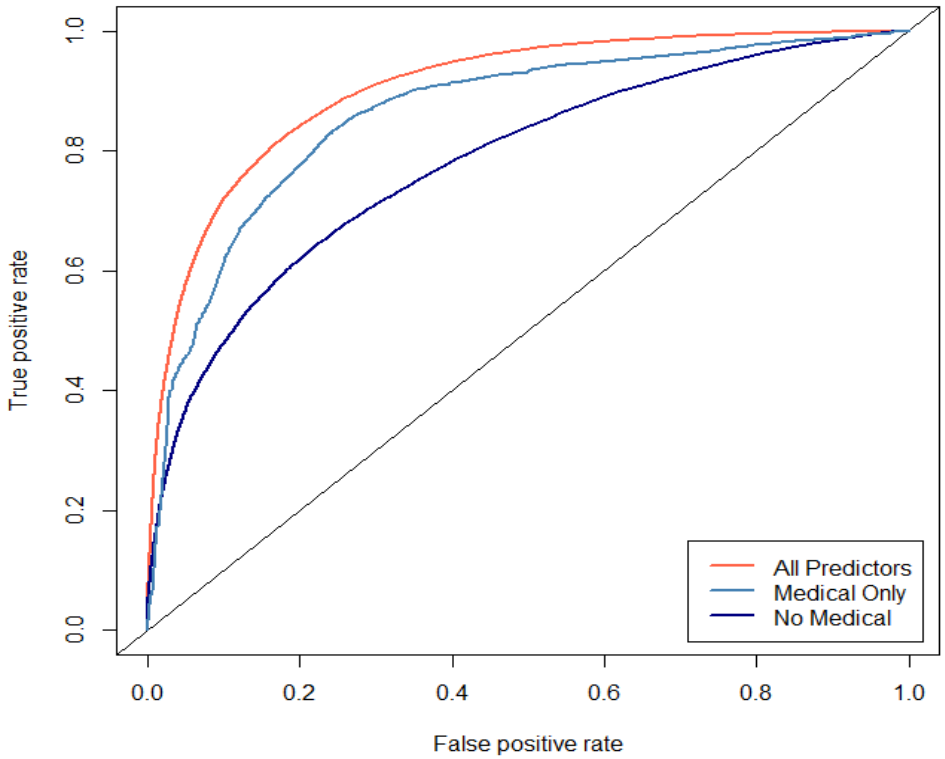


Figure 5 Validation set ROC curves for three additive logistic regression fits.

demographic and administrative variables (analogous to the Speten (2018) models), produced a model fit inferior to those that include the medical variables.

For the second experiment, we fitted a random forest model and compared it to the full additive logistic regression model. The validation set ROC curves of the two models are given in Figure 6. This random forest fit was unexpectedly good. It indicates that there were important interactions among variables that are omitted in the additive logistic regression fit. The random forest without medical variables performs almost the same as the additive logistic regression fit without medical variables. This confirms the Speten (2018) results of little difference between his random forest and additive logistic regression fits, neither of which include medical variables. It appears, though, when medical variables are added to the model, the random forest outperforms the additive logistic regression fit. This hints that the important interactions missing from the additive model are interactions between or with medical variables, but not among the administrative and demographic variables.

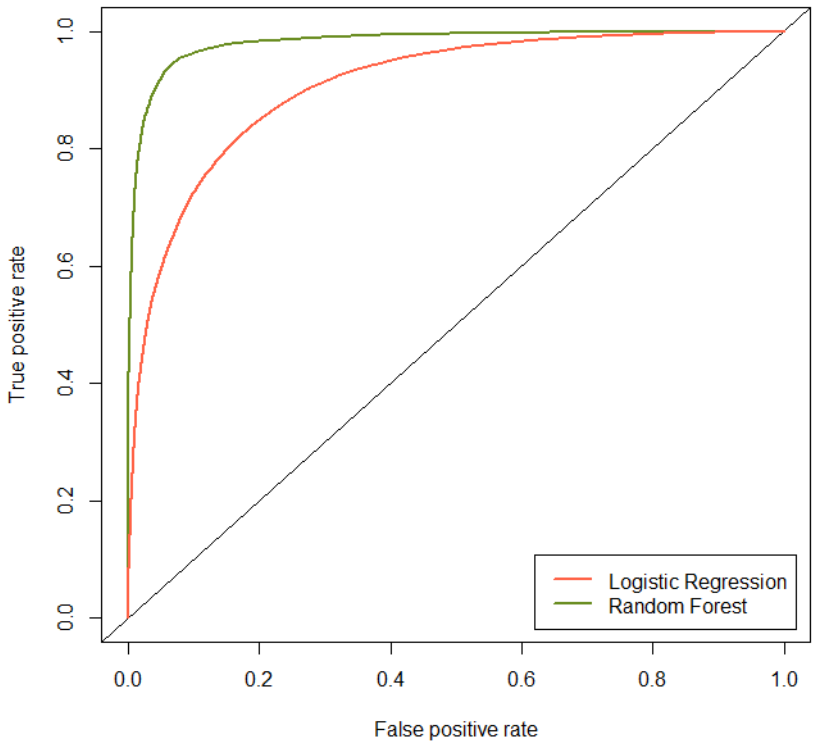


Figure 6 Validation set ROC curves for the additive logistic regression fit and the random forest fit.

3.2.2. Variable Selection

In this section, we identify a smaller, but still large, set of variables that can be used to predict attrition with little loss in accuracy. These results are reported in Gobea (2019). We used the permutation measure of how "important" a variable was for correctly classifying attrition based on the random forest fit. This permutation measure captured how much the random forest's ability to predict attrition is degraded when a variable's values are replaced with noise. High or low variable importance did not imply either a causal relationship or a lack of one between the variable and attrition. It merely measured the variable's contribution to predicting attrition in the presence of all the other variables. This means, for example, that a set of highly correlated variables that were also all strongly related to attrition could all get low importance scores because any single variable in the set was not needed when the rest were included.

We used the random forest variable importance, because it accounted for how important the variable was, including in its interactions, with other predictors. Variable importance based on an additive logistic regression model did not do this. The relative variable importance (relative to the most important, Dental Class), was reported, in order in Table 6. Table 6 only includes those variables whose relative importance was at least 1% of that Dental Class.

Variable	Relative Importance
Dental Class	1.00
PULHES Deployable	0.90
Contract Duration	0.29
PULHES H Field	0.24
Unit Type Citizenship Status	0.18
Marital Status	0.11
Smoker	0.10
Hearing Readiness Class	0.10
Gender	0.09
Medical Non-Deployment Profile	0.09
PULHES E Field	0.09
Military Organization	0.08
Rank (Enlistment)	0.06
Height (Enlistment)	0.06
Age (Enlistment)	0.06
Weight (Enlistment)	0.05
Military Organization Group	0.05
Education Level (Enlistment)	0.04

Variable	Relative Importance
Education Tier (Enlistment)	0.04
Overall Health Profile ….	0.03
AFQT Category	0.03
Prior Service	0.03
Number of Dependents (Enlistment)	0.03
Back Pain	0.02
Chew Tobacco	0.02
Joint Pain	0.01
Home of Record, State	0.01
Waiver (Admin)	0.01
Home of Record, Region	0.01
Limited Duty	0.01
Allergy, Penicillin	0.01
PULHES L Field	0.01
Chronic Pain	0.01

Table 6 Relative random forest variable importance

The two variables Dental Class and PULHES Deployable have, by far, the largest variable importance. We took a conservative approach and only omit variables that do not appear in Table 6. Figure 7 gave the validation set ROC curves for four logistic regression fits. The topmost orange fit is for the additive model using only the variables listed in Table 6. This ROC curve was indistinguishable from the ROC curve for the full model. Thus, removing the variables not listed in Table 6 has a negligible effect on the fit. However, using the Table 6 variables and removing Dental Class and PULHES Deployable had a dramatic effect on the logistic regression model fit as illustrated by the orange dotted ROC curve in Figure 7. The sienna solid ROC curve was for the logistic regression fits with just Dental Class and PULHES Deployable and their interaction. The magenta ROC curve corresponds to a model that included contract duration (TERM) and its interactions to the model with PULHES Deployable and Dental Class. The three variables, Dental Class, PULHES Deployable, and contract length along with their interactions were able to yield a model that performs almost as well as the full logistic regression additive model.

For completeness, we also fitted several alternatives to the additive logistic regression model using the Table 6 variables. To account for non-linear effects of the three numeric variables height, weight, and age at enlistment, we replaced them with their categorical counterparts. We also constructed a variable that counts the number of conditions reported by a soldier; this type of

summary variable has been helpful in other studies. These and experiments adding potential interactions terms did not change model results enough to warrant further investigation at this time. We also did some limited exploration of the random forest fit. The random forest fit with just the Table 6 variables is almost, but not quite, as good as the random forest fit with all the variables. But removing the medical variables gave a random forest fit comparable to the logistic regression fit without medical variables. Finally, Gobea (2019) used a lasso-regularized logistic regression fit to select variables. The results of this fit are similar to that of the additive logistic regression fit with Table 3, but not nearly as informative.

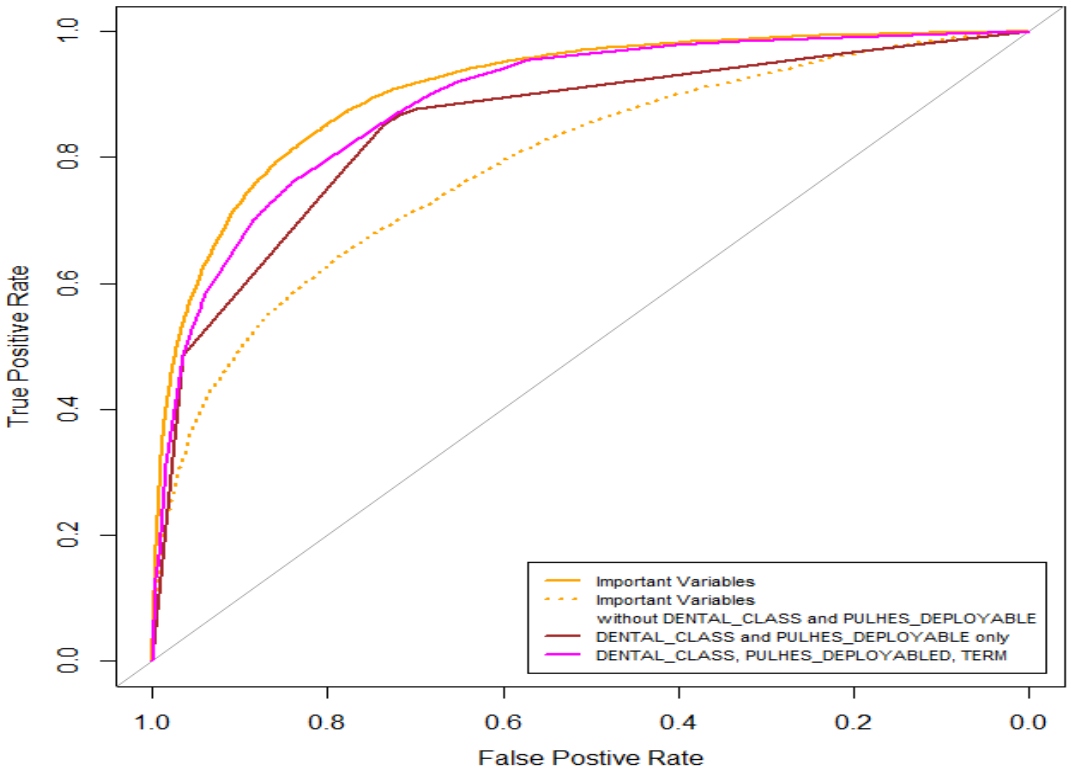


Figure 7 Validation ROC curves for four logistic regression fits using Table 6 variables: All, all without Dental Class and PULHES Deployable, Just Dental Class and PULHES Deployable, Just Dental Class, PULHES Deployable, and contract length (TERM).

For the purposes of this report, we stuck with the additive logistic regression fit. Gobea (2019) gave the coefficients for the final additive logistic regression model fit. Even though it used more variables than are strictly needed, and does not include interactions, it did provide insight and a useable model fit.

3.2.3. Final Model

Gobea (2019) showed how this final model fits on the FY 2010 test set. As expected, the performance degrades. This is due in part to non-stationarity over time, but may also be due to the immaturity of the medical data, particularly PHA. The FY 2010 ROC curves yield an Areas Under the Curve (AUC) measure of 0.82. This was considered respectable for models used to predict attrition for an individual soldier. We do advise some caution in using this model for predicting individual soldier attrition because of the questions surrounding data quality. However, in general, these models gave good estimates of the probability of attrition for groups of soldiers. Gobea (2019) constructed a plot of the average estimated probabilities for 300 subsets of the validation set (with about 100 per group) reproduced in Figure 8. This plot shows that estimated probabilities per group were close to the actual attrition rates in each group. We get a similar plot for the FY 2010 data.

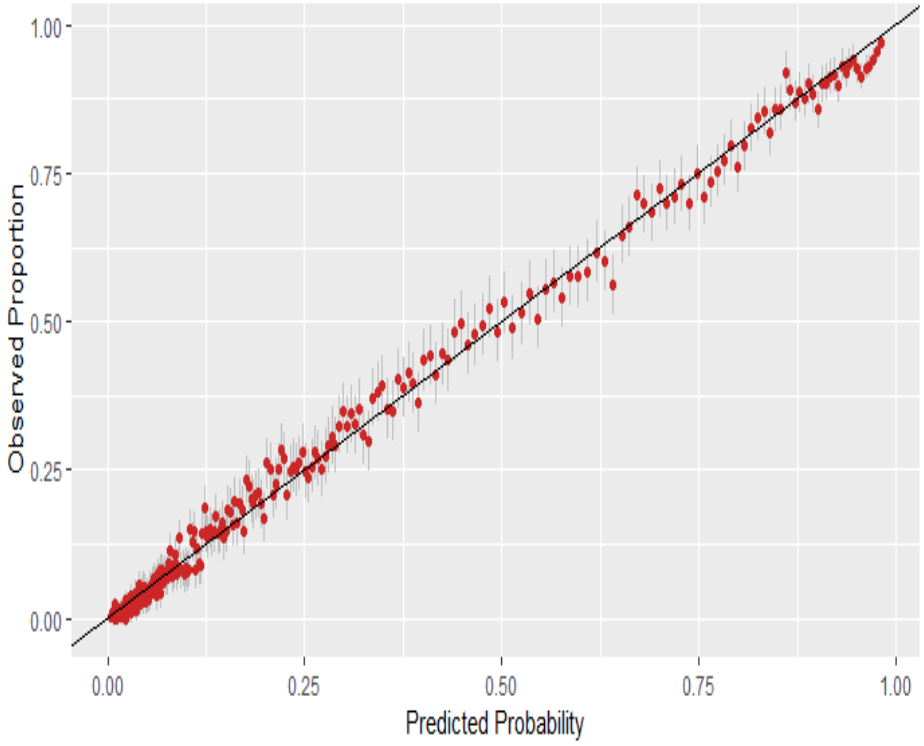


Figure 8 Validation set observed proportion of attrits against average estimated probability of attrition for the group using the additive logistic regression model of Gobea (2019).

Using measures of importance specific to logistic regression models, Gobea (2019) identified the predictor variables: PUHLES Deployable, Dental Class, Contract Duration as having the highest scores, followed by Unit Type, Medical Non-deployable Profile, Hearing

Readiness Class, Gender, Smoker, Education Tier, Marital Status and Military Occupation and then the rest of the variables. For a complete list see Gobea (2019). It was clear that the PHA variable Dental Class, taken post-IET, especially Class 4 was an important predictor of attrition. Our concern was that some of Dental Class 4's relationship to high attrition might be because, for some soldiers, Dental Class 4 (no dental exam in the previous year) was a consequence of attrition and not a predictor. This would mean that the importance of Dental Class was overstated in the models fitted in this work.

3.3. SURVIVAL ANALYSIS

Devig (2019) performed survival analysis primarily using FY 2010 enlistment year records for training and validation and FY 2011 for testing. Like Gobea (2019), he also imputed missing values prior to model fitting. Devig (2019) explicitly constructed predictable time-varying variables. He incorporated them into his model fits by using the long-form data described in Section 2.2.2, and fitting survival tree models that accommodate left-truncation as described by Fu and Simonoff (2016) and implemented in their the R package **`LTRCtrees`** (Fu and Simonoff, 2018). Survival tree methods based on time-constant variables are not new, but survival trees that use time-varying covariates are relatively new.

Survival trees yield a partition of the training data into subsets (or leaves) and then give an estimated survival curve based on the data in each leaf using the Kaplan-Meier estimator. In Figure 9 we show an example of all the survival curves, one per leaf, from a survival tree fit. Taken as a whole, the group of survival functions are difficult to interpret, but can give insights, and Section 3.3.1 shows some examples of such insights. Further, like most algorithmic models, survival trees were most useful for prediction. In particular, in Section 3.3.2 we show that they perform quite well when used to predict the expected number of AD soldiers not attiring over time.

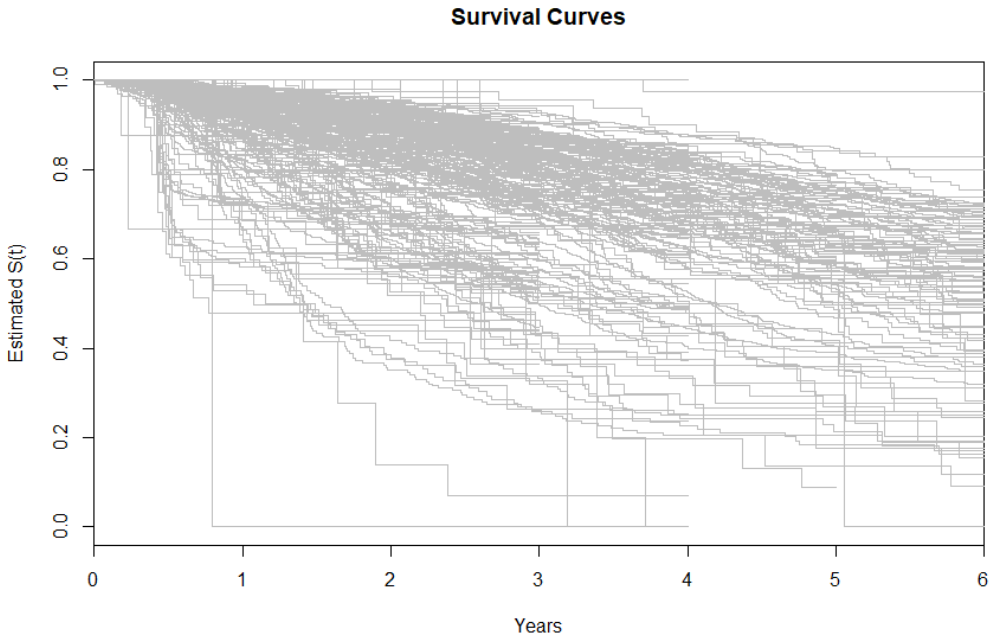


Figure 9 Estimated survival functions for a survival tree based on FY 2008 – FY 2011 enlistments, where t (years) is time since enlistment.

3.3.1. Emerging Insights

In this section, we illustrate how survival analysis, specifically survival trees, might be used to gain insights in ways that traditional binary regression approaches cannot. We do not perform a comprehensive study to characterize differences in survival functions resulting from the survival tree.

We start by noting that the fitted survival functions are step functions. Those with deep steps, as seen in the lower left of Figure 9, are survival functions estimated from leaves that contain only a few observations. For exploratory purposes, we tend to ignore these. In general, survival curves as depicted in Figure 9 are difficult to interpret, but by plotting subsets of the curves, by paying attention to the first few variables used by the tree to split the data, and by plotting average survival curves for individual soldiers in certain groups, we can gain insight.

Our first exploratory survival tree fit was based on enlistment year FY 2008 through enlistment year FY 2011 and includes the enlistment year as a variable. With enlistment year as a variable, this model cannot be used for prediction, but it does help us see if the survival tree model detected changes in attrition patterns over time. The first split for this survival tree partitions the

data into enlistment years FY 2008 – FY 2009 and enlistment years FY 2010 – FY 2011. See Devig (2019) for the complete listing of the tree splits. The average of the survival function over soldiers in each enlistment year and based on the immediate post-IET values of their variables are reproduced from Devig (2019) in Figure 10. Because the curves in Figure 10 were averages over all soldiers in each enlistment year, we do not expect to see large differences. Further, because the survival functions are based on soldier attributes immediately after IET, we expect differences, when they are visible, to appear early in the term. The average survival functions for FY 2010 and FY 2011 give lower attrition rates than do those of FY 2008 and FY 2009, especially in the first two years of the first term. We do not know why there are differences, but we suspect that some of these difference may be due to the maturity of the medical data. Thus, in the final model fitting, Devig (2019) only uses FY 2010 enlistments. The FY 2010 set gave survival function estimates that most closely resemble the FY 2011 estimates and with much less missing PHA data than previous enlistment years.

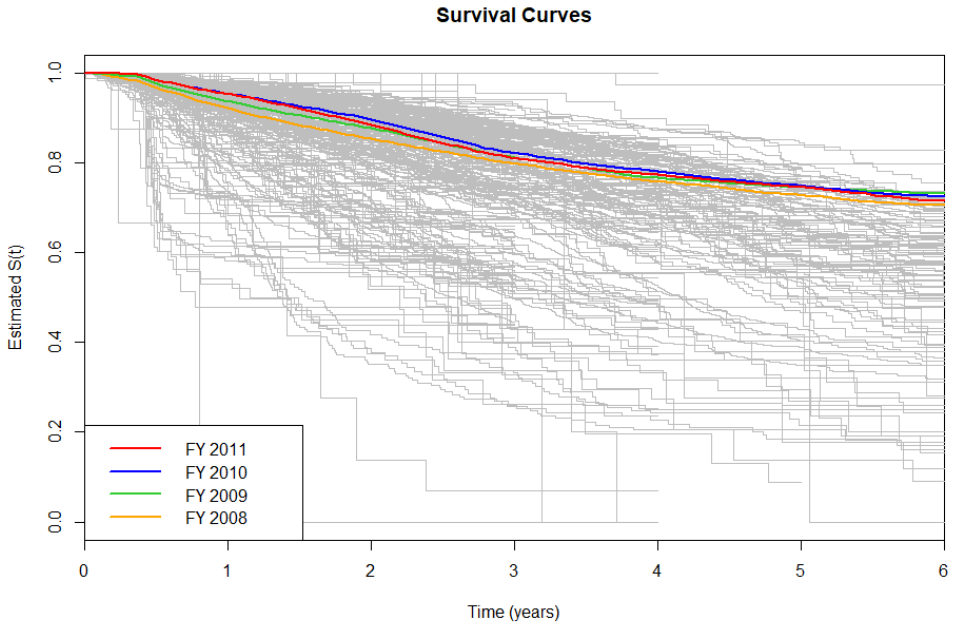


Figure 10 Estimated survival functions from Figure 9 with average survival functions for soldiers from each enlistment year.

In all survival tree fits, contract duration is an early splitting variable. In the survival tree fit using only FY 2010 enlistment data, it was the first splitting variable, separating three-year enlistment contract records from the rest. Survival tree leaves that only contain three-year contract records produce survival functions up to three years. After that the survival functions are

undefined. Figure 11 shows the survival functions estimated based on three year contracts. We mention this because care must be taken when averaging survival functions estimated using survival trees. When averaging survival functions for soldier attrition, we recommend averaging only survival functions of soldiers with the same contract length or only over those times up to the shortest contract length.

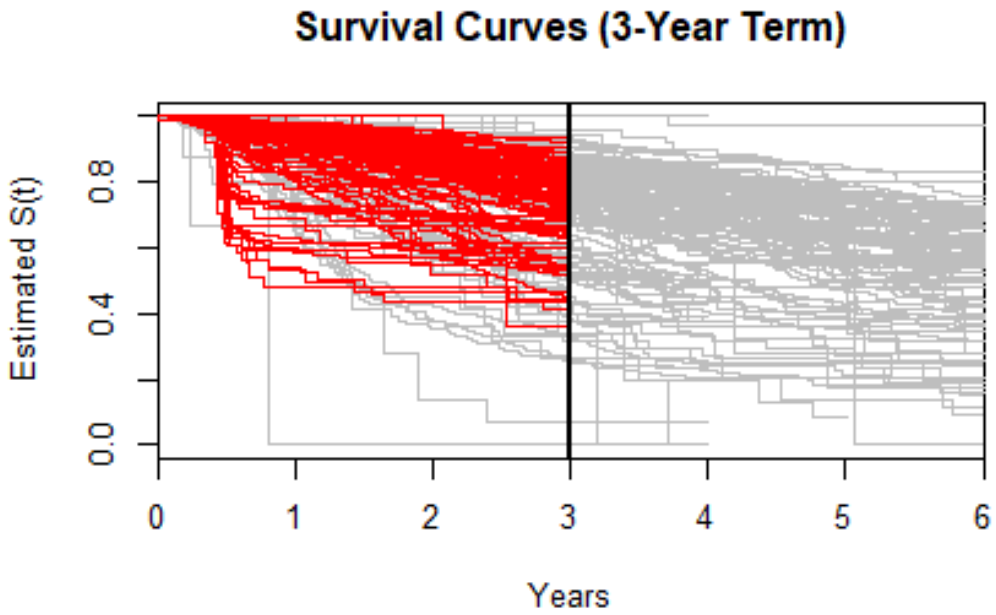


Figure 11 Estimated survival functions from the survival tree based on FY 2010 enlistments with survival functions corresponding to three-year contracts in red.

The effects of contract duration cannot be considered without also considering Military Occupation. Table 7 shows first-term non-attrition rates by Military Occupation (CMF) and by the three categories of Military Occupation Group: Force Sustainment (FS), Operations Support (OS), and Maneuvers, Fires and Effects (MFE). In Table 7, CMFs 63 and 91 have been combined into CMF Vehicle Mechanic (VM).

CMF	Group	Rate
CMF.68	FS	76.9
CMF.42	FS	77.4
CMF.89	FS	77.8
CMF.74	OS	78.4
CMF.92	FS	78.4
CMF.94	FS	78.4
CMF.88	FS	79.2
CMF.VM	FS	80.0
CMF.35	OS	80.4
CMF.31	OS	80.6
CMF.25	OS	81.0
CMF.14	MFE	81.1
CMF.12	OS	82.2
CMF.15	MFE	82.2
CMF.11	MFE	82.3
CMF.19	MFE	82.4
CMF.13	MFE	82.7

Table 7 Completion rate by Military Occupation (CMF) and Military Occupation Group.

We expect soldiers with longer contracts to have greater overall attrition, but attrition in even the first few years differs by contract length. In Figure 12, we see average survival functions based on FY 2010 enlistments by CMF and contract length. All curves in Figure 12 were based on averaging at least 1000 soldier survival functions. Across CMFs, probability of attrition for any time in the first three years was lower for four-year contracts than for three-year contracts. The greatest differences were in the first year. After that, the survival functions are almost parallel, indicating that as a soldier completes year one (and based only on his/her attributes immediately post-IET), the chance of attrition in the next time period is about the same across contract year and CMF groups.

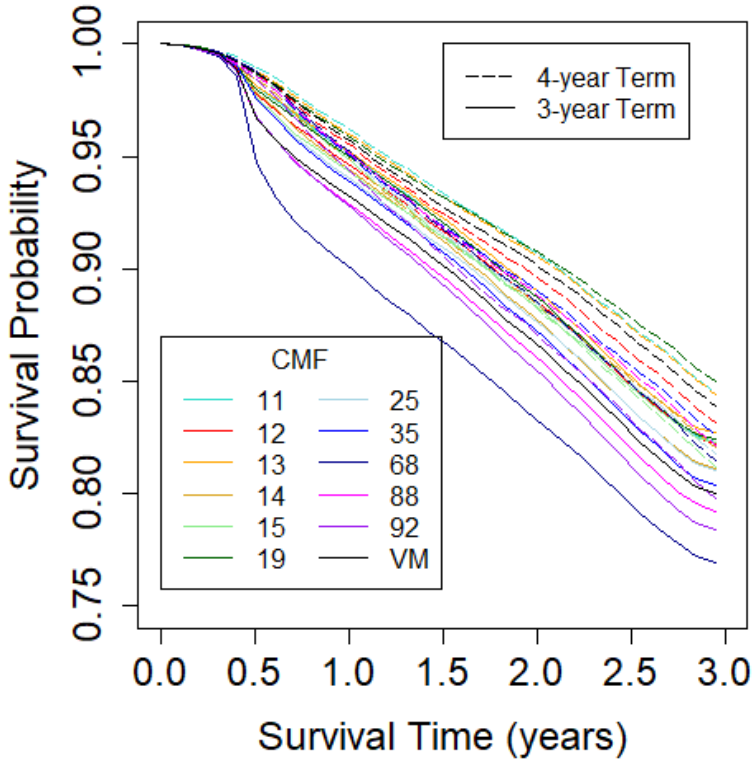


Figure 12 Average survival functions by CMF group and contract length.

This type of analysis can be done for groups of soldiers using other attributes as long as one pays attention to contract length. Plots of average survival functions such as the plot in Figure 12 do not give a sense of the variability of each group. Capturing this variability can be important especially for groups that include a large number of soldiers. To do this in a way that was useful required some thought. One possibility was to plot the upper and lower curves that contain (pointwise), say, 90% of soldiers' estimated probabilities in a group. In addition, the survival functions averaged to give the curves in Figure 12 were based only on immediate post-IET variable values. We have not compared survival functions of soldiers after they complete as certain number of years into their first term. But as we discuss in Section 3.2.2, from the survival tree, in principle, it was straightforward to compute a soldier's conditional survival function past time *s*, given his/her variable values observed up to time *s*, and given that the soldier has completed the first *s* years of his or her term.

3.3.2. Model Performance

Devig (2019) assessed the survival tree model fit based on FY 2010 enlistments using model performance computed with FY 2011 enlistments. We look at how well the FY 2010 model predicted FY 2011 aggregate numbers of soldiers completing the first *t* years into their first term and how well it predicted individual soldier attrition.

First, we used the survival tree model fit to predict the aggregate number of soldiers who were still enlisted at time *t* after enlistment. This was done separately by contract length. For each contract length, the records for FY 2011 enlistees were used to compute the actual number who had not attrited by time *t* years from the date of enlistment. The actual number of enlistees by time is shown in black in Figure 13. The corresponding predicted curves based on the survival trees fitted using FY 2010 data are plotted in red in Figure 13. The predicted numbers of enlistees are based on the individual soldier variable values observed at the end of IET. To gain an appreciation of how well this model worked when predicting for an entire first term, time-varying variable values observed later in the first term are not used. To predict, we first determine into which tree leaf each soldier's record (in the FY 2011 test set) fell. This gave us a predicted survival function estimate for each soldier. These were then summed for each contract length to estimate the expected number of completions. We see from Figure 13, that the predictions were better early in the first term, but that they yield results that were fairly good even when predicting six years into the future.

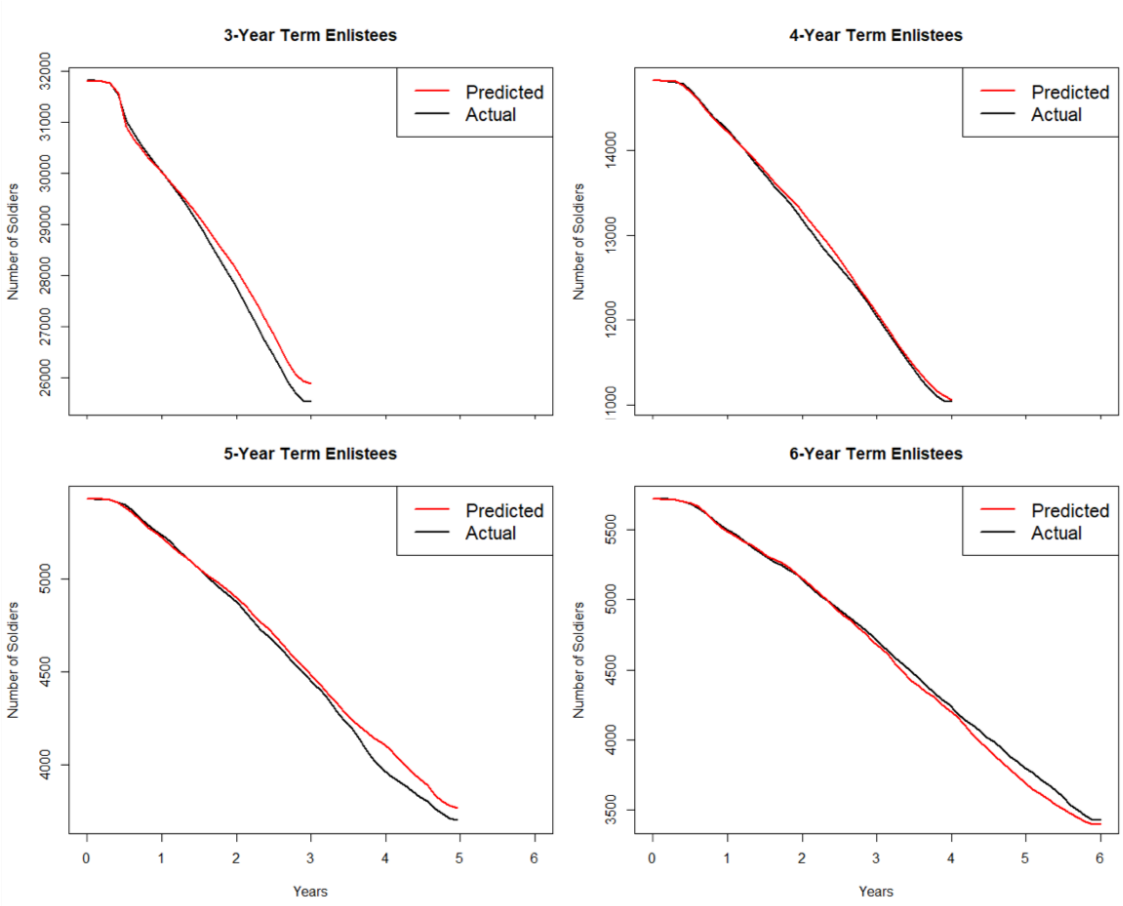


Figure 13 Actual and predicted number of soldiers enlisting in FY 2011 completing the first t years into their first term. Predictions are based on FY 2011 immediate post-IET attributes and the survival tree model fit using FY 2010 enlistees.

Survival functions can be used to predict attrition at any time for individual soldiers. One such approach was to use an estimated survival probability value as "score" at time *t* for risk of attrition in before time *t*. The higher the score, the lower the risk of attrition. The Figure 14 plots are ROC curves to see how well these scores can predict attrition within one, two, and three years for FY 2011 three-year contract enlistees using only the values of time-varying covariates available immediately post-IET. As expected, the greater the time from IET, the more difficult it was to predict attrition for individual soldiers. Further, these ROC curves have AUC's of 0.67, 0.62 and 0.59 for one, two, three year predictions respectively indicating that this approach was not viable for predicting individual soldier attrition.

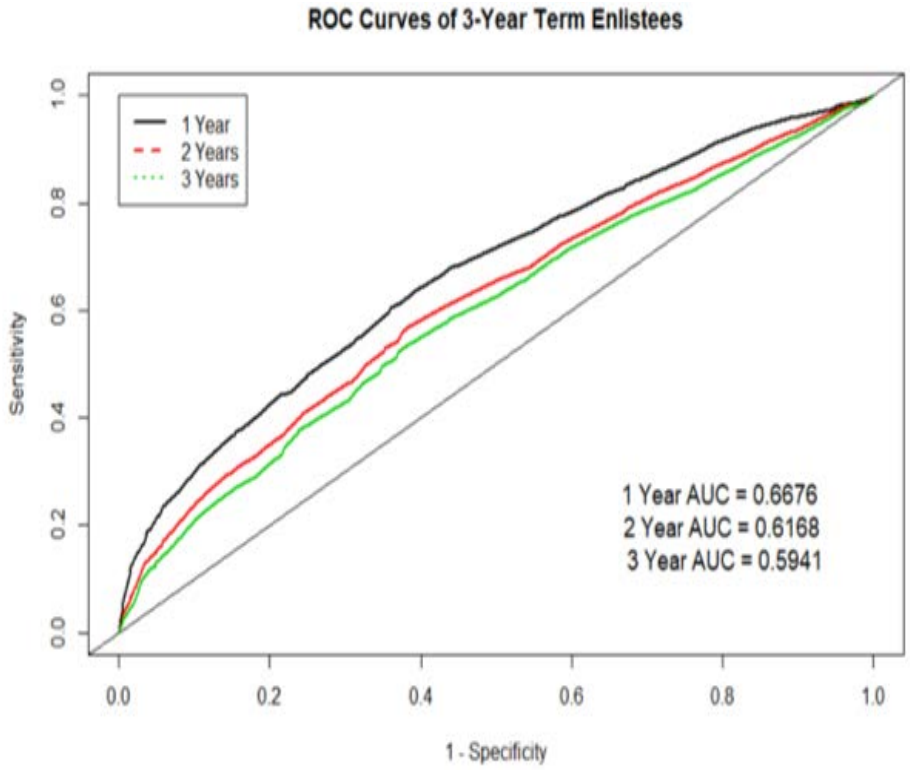


Figure 14 ROC curves for predicting FY 2011 attritions within one, two, and three years of enlistment based on variable values available immediately post-IET.

As mentioned in the previous section, because survival functions provide survival probability estimates for the entire first-term, we used them to get an estimated conditional probabilities of survival for *y* additional years given survival up to time *s* and based on the values of time-varying covariates at time *s*. For a single soldier, this entailed extracting the values of his/her time-varying variables at time *s*. These and the values of his/her time-constant variables are used to determine his/her tree leaf at time *s*. Because his/her time-varying variables may have changed since IET, this leaf may be different than the one we would get had we used the values of his/her variables immediately post-IET. Let $\hat{S}(t)$ be the estimated survival function for the soldier. The probability that the soldier survives an additional *y* years given survival up to time *s* and using what we know about him/her at time *s* is $\hat{S}(s+y)/\hat{S}(s)$.

We have done some experimentation with this approach to try to improve individual estimates, but have not had much success and hence these results are not reported in Devig (2019). We do, however, see potential in using these conditional probabilities to predict aggregate numbers still enlisted in the next *y* years for a *group* of first-term soldiers with a mix of different enlistment dates as would be found in a unit. For such a group, we would compute the time since enlistment

for each soldier, extract the current values of their time-varying variables, and use the survival tree to give an estimated survival function for each one. The survival functions would then give estimates of the conditional probability of surviving *y* more years for each soldier. Summed over the soldiers in the unit, we would have an aggregate estimate of the expected number that will still be enlisted at the end of the year.

3.4. IMPLEMENTATION

MAJ Gabriel Gobea and CPT Aaron Devig demonstrated to ARD as part of a class project, a proof of principle web-enabled decision support tool available in the PDE. Their tool employs the models discussed in this report to predict the likelihood of attrition in an interactive manner with potential to support use cases in recruiting and by organizational leaders.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 4. CONCLUSION

This year's work was a continuation of last year's work studying post-IET first-term attrition reported by Smith et.al. (2018) and based on the Master's thesis of Speten (2018). The new features of this year's work are: we were able to include medical variables from two databases PHA and MEDPROS; as much as possible, we restricted attention to variables that were predictive of attrition and not a consequence of attrition; and we predict attrition for any time in the first term.

To study post-IET first-term attrition, we used two approaches: The first, reported primarily in the Master's thesis of Gobea (2019), used the more traditional binary regression approach, patterned after last year's work of Speten (2018). Using binary regression provided a more direct comparison of models with and models without medical variables. It also gave ARD an additive logistic regression model fit with explicit equations for estimating probabilities that can be implemented fairly easily. Secondly, we used a new method as reported by Devig (2019), survival trees with time-varying variables, which gave a temporal view of attrition. Survival tree models not only gave estimates of probability of attrition for anytime in the first-term (post-IET), but these models are fitted using values of variables that change with time, e.g. rank, marital status, medical variables, etc. They provided insights not possible with binary regression models. The survival tree models performed well in estimating aggregate numbers of attrites as they changed over time in the first term. However, our initial survival tree models did not predict attrition for individuals as well as the simpler logistic regression fits. Survival tree models, though, in general have the greatest potential for use by decision makers.

4.1. ANSWERS TO QUESTIONS POSED BY ARD

In this section, we briefly summarized responses to questions posed by ARD. First we addressed the questions initially posed by ARD and answered by Speten (2018):

(i) What are the demographic and medical factors of personnel with highest probability of failure?

Our findings were comparable to those of Speten (2018). For example, (as reported by Devig, 2019) females tended to have a higher post-IET first-term attrition rate than males, 37.2% versus 22.1% respectively.

Of the medical variables, Dental Class (recorded at the end of IET) had the most extreme differences in attrition rates; Dental Class 3 had an overall attrition rate of 54.3%, but made up only 3.7% of the cohort. Four of the six PULHES codes (at the end of IET) had higher attrition rates for scores of 3 or 4. These were PULHES codes for physical capability, upper extremities, lower extremities, and psychiatric issues. In our models the PULHES Deployable indicator which was "no" if any of the PULHES scores were 3 or 4 was consistently among the two most important variables.

(ii) What is the mean number of total failures during the first enlistment term?

This analysis showed post-IET attrition rates slightly smaller than that of Speten (2018) because we uncovered some additional IET attritions not accounted for in earlier work. We found that FY 2011 enlistments had an attrition rate of 24.40% (14,137 soldiers), which was higher than the attrition rate for FY 2010 enlistees of 23.29% (15,401 soldiers). We also provided the total number of non-failures by number of years since enlistment for FY 2011 in Table 4.

And, for this year's work we address:

(i) Does the incorporation of medical data provide additional insights to inform understanding of soldier first-term attrition?

Yes. Incorporating medical variables did provide extra insight and improved all models that estimated probability of post-IET attrition. Medical variables, especially Dental Class and PULHES Deployable variables were important predictors of attrition in the current data. We noted that when the PHA data were more mature, the strength of effects of these and other medical variables might change.

(ii) Do alternative modeling approaches provide additional insights?

Yes. The survival analysis approach used here and Devig (2019) gave us a temporal view of attrition and made it possible to do a better job of accounting for changes in time-varying covariates. It seemed to be particularly good at predicting aggregate

numbers of non-attriters at times throughout the first term. At present, it does not perform as well as the more simple additive logistic regression model for predicting individual soldier attrition. It is interesting, that with the addition of medical variables and unlike models without medical variables, random forests seem to outperform the simpler additive logistic regression models.

(iii) What models and data might be useful to a decision maker concerned with first-term attrition in their organization?

Current and immediately post-IET values of the approximately twenty medical, administrative, and demographic variables in combination were good predictors of attrition. We expect that as the medical data matures, our models will change and become more productive. And although, we think some of our insights especially for Dental Class 4 are a consequent and not a predictor of attrition, of all the medical variables, Dental Classes 3 and 4 and PULHES Deployable variables are worth keeping track of.

The traditional additive logistic regression models gave good estimates of first-term attrition probabilities. These are easy to use and understand and are effective for predicting first-term attrites based on variables captured immediately post-IET.

We also see promise in using the survival analysis approach. But more work needs to be done in this area. Our current models predict aggregate numbers of attrites well at any time in the first term. In addition, their estimated survival functions can be tailored to predict aggregate numbers of attrites in the next interval of time for a group of soldiers with a mix of enlistment dates and based on current knowledge of these soldier's variables.

4.2. DATA CONCERNS

The PDE was an extremely valuable resource for Army personnel analytics due to the consolidated data tables created from disparate data sources and the collaborative environment. However, the lack of comprehensive data definitions challenges users to fully understand the variables.

As Speten (2018) reported, the universal application of separation codes continues to be a concern nearly four decades after the issue was first raised. Speten (2018) and our work highlights the potential errors in the sole use of the codes to study attrition and we used Speten's (2018) methodology for identifying whether a soldier successfully completed his or her contractual obligation by examining the historical records of each soldier with a small modification to identify which soldiers attrite during IET.

Careful selection of predictor variables was critical for accurately depicting a soldier's demographic, administrative, and medical history. Time-related snapshot and transactional data gave us access to variables that change with time. Using these so that their values are predictive and not a consequence of attrition is difficult. With missing PHA data, we had the greatest difficulty in determining medical variable values immediately following IET. Further, there is a need to have access to a careful catalog of changes to variable coding and use. For our work, this was particularly true of the CMF coding.

Our concerns are that some the effects of for example Dental Class may appear somewhat stronger in our models than they actually are. Models will change somewhat as will their performance in predicting future attrition as the medical datasets mature.

4.3. RECOMMENDATIONS

4.3.1. Implementation

The observations concerning implementation made for last year's efforts continue to be true of this year's work. We tailor these observations to apply to our current work.

First we made two specific recommendations concerning the data: data, especially medical data from the Reception Battalion would give immediate post-IET soldier attributes and help alleviate the data concerns outlined in Section 3.4; having access to UIC and duty station zip codes would aid tailoring analysis to organizations.

All in all, the accuracy rate of our predictive models with the inclusion of medical variables provide enough fidelity to give insight and merits consideration of use by Army planners. Of the models to predict attrition at the end of the first term based on information available at the end of IET, the random forest model was the most accurate, and warrants further analysis, but the additive

logistic regression model was still fairly accurate and easier to manipulate for ARD analysts. General attrition rate findings based on combined demographic and medical predictor variables may help to inform force strength requirements, recruiting goals, and retention efforts. The flexible and repeatable nature of the random forest and additive logistic regression modeling technique provides analysts the ability to react quickly to changes in data availability and shifts in both policies and priorities.

The survival tree models hold great promise and even, as they stand, provide insight and are able to predict aggregate attrition over the entire first term and not just at the end.

Most importantly, this research provides ARD insight as the agency continues its efforts to improve soldier resiliency and, by extension, first-term attrition rates. Application of our predictive models to the administrative and medical records of current enlistees could provide policy makers with probability estimates of all first-term soldiers and facilitate the creation of intervention programs and prioritized resource strategies built upon a quantitative foundation.

A web-based predictive tool available to Army leaders at the lowest unit level would allow human resource professionals or junior Non-Commissioned Officers to engage new soldiers during annual record reviews and monthly professional counseling sessions. Once the attrition probability assessment is completed for each soldier, the appropriate training, administrative actions, or other intervention strategies could be leveraged to best assist the soldier.

4.3.2. Future Work

1. Use UIC and duty station location to aid in making first-term attrition work useable at an organizational level.

   We have not studied UIC and duty location zip codes because they are obscured in the version of the PDE data currently available to us. We are working on gaining an approved protocol from the Naval Postgraduate School Internal Review Board to work with these fields and we are currently working with AAG to get access to this data.

2. Update models and insights using current data.

   The PHA data after FY 2010 are more complete than years prior to FY 2010. In light of our data concerns of Section 3.3, the models described in this report should be

refitted using the most recent enlistment data, as it becomes available. Further, unlike the binary regression methods, the survival analysis methods can use soldier records for soldiers who have not completed their first term.

Further access to records from the Reception Battalion immediately preceding IET might prove helpful in this regard.

3. Develop and implement methods for use and visualization of survival tree results.

   The most important variables for predicting attrition change over the course of a soldier's career. These are particularly challenging to work with. Survival analysis, in general, and survival trees in particular, show promise in capturing these changes and using them to predict and understand attrition. Use of survival trees to time-varying variables is new and we have just scratched the surface of using the trees to capitalize on these variables. Work needs to be done to in the areas of: assessing and visualizing variability of attrition patterns within groups; finding appropriate baseline perhaps average survival functions; updating survival functions for soldiers who have completed a portion of their first term using their most current information; leveraging what we know about a soldiers attribute history across his/her first term to "stitch" together survival functions that capture these changes; combining survival functions across organizations with soldiers who are at different times into their first terms.

   In addition, we were surprised that survival trees did not predict individual soldier attrition as well as the additive logistic regression models. In classification and regression, random forests often outperform single classification and regression trees. That may be the case with survival functions. Survival random forest software is available for time-constant variables. These models should be investigated and adapted to accommodate the time-varying variables.

4. USAREC has expressed interest in augmenting accession data currently available to us in the PDE.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 5. WORKS CITED

Anglemyer, A., Smith, A., & Speten, K. (2018). *US Army Post Initial Entry Training First-Term Attrition Analysis.* Monterey.

Devig, A. (2019). *Predicting U.S. army enlisted attrition after initial entry training using survival analysis.* Monterey: Calhoun.

Fu W., Simonoff, J. (2018, 03 29). *LTRCtrees: Survival Trees to Fit Left-Truncated and Right-Censored and Interval-Censored Survival Data*. Retrieved from CRAN.R-project.org: https://CRAN.R-project.org/package=LTRCtrees.

Fu, W., & Simonoff, J. (2016). LTRCtrees: Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, 352-369. Retrieved from https://doi.org/10.1093/biostatistics/kxw047

Gabe, G. (2019). *Predicting U.S. Army First-term Attrition After Initial Entry Training Part II.* Monterey: Calhoun.

Gordon, L., & Olshen, R. A. (1985). Tree–structured Survival Analysis. *Cancer Treatment Reports 69*, 1065-1069.

Hawes, E. (1990). *An application of survival analysis methods to the study of Marine enlisted attrition.* Monterey: Naval Postgraduate School. Retrieved from http://hdl.handle.net/10945/34851

Ishwaran H, & Kogalur, U. (2019, July 08). *Random Forests for Survival, Regression, and Classification (RF-SRC).* Retrieved from CRAN.R-project.org: https://CRAN.R-project.org/package=randomForestSRC

Ishwaran, H., Kogalur, U., Blackstone, E., & Lauer, M. (2008). Random survival forests. *The Annals of Applied Statistics*, 841-860. Retrieved from Project Euclid: The Annals of Applied Statistics: https://doi.org/10.1214/08-AOAS169

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. In Kaplan, & P. Meier, *Nonparametric estimation from incomplete observations* (pp. 457-481). J.Am. Stat. Assoc.

Speten, K. J. (2018). *Predicting U.S. Army First-Term Attrition After Initial Entry Training.* Monterey: Calhoun.