2022-12

PERFORMANCE OF RANDOM SURVIVAL FORESTS WITH TIME-VARYING COVARIATES IN PREDICTION OF U.S. ARMY ENLISTED ATTRITION COMPARED TO TRADITIONAL MANPOWER ANALYSIS METHODS

Rooney, Connor A.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PERFORMANCE OF RANDOM SURVIVAL FORESTS WITH TIME-VARYING COVARIATES IN PREDICTION OF U.S. ARMY ENLISTED ATTRITION COMPARED TO TRADITIONAL MANPOWER ANALYSIS METHODS

by

Connor A. Rooney

December 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*

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** December 2022	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
PERFORMANCE OF RANDOM SURVIVAL FORESTS WITH TIME-VARYING COVARIATES IN PREDICTION OF U.S. ARMY ENLISTED ATTRITION COMPARED TO TRADITIONAL MANPOWER ANALYSIS METHODS

5. FUNDING NUMBERS

6. AUTHOR(S) Connor A. Rooney

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 importance of identifying qualified candidates and properly forecasting future manpower strength will always be critical to military recruiting and organization. The ability to assess the cross-section of covariates of a cohort of enlistees and forecast manpower strength would allow for improved planning and allocation decisions. We leverage an innovative method of survival analysis—random survival forests (RSF) with time-varying covariates (T-VC)—to predict Army first-term post-Initial Entry Training attrition rates. Using random survival forests with time-varying covariates (TV-RSF) is an emerging method of survival analysis that has not been used in a military manpower setting. Using a Brier Score we compare TV-RSF with three other methods. We illustrate that using a single tree rather than the computationally intensive TV-RSF may suffice for predicting future year attrition. We also illustrate that TV-RSFs outperform traditional classification methods (logistic regression, random forests) that only account for yearly changes in T-VCs.

14. SUBJECT TERMS
survival analysis, attrition, random survival forest, RSF, time-varying covariates, T-VC, forecasting attrition, army, random survival forest with time-varying covariates, TV-RSF, person-event data environment, PDE, left-truncated right-censured, LTRC random forest, statistical learning, machine learning, DOD manpower, forecast manpower, survival trees, logistic regression, random forests

15. NUMBER OF PAGES
79

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.

PERFORMANCE OF RANDOM SURVIVAL FORESTS WITH TIME-VARYING COVARIATES IN PREDICTION OF U.S. ARMY ENLISTED ATTRITION COMPARED TO TRADITIONAL MANPOWER ANALYSIS METHODS

Connor A. Rooney
Ensign, United States Navy
BS, United States Naval Academy, 2021

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 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 importance of identifying qualified candidates and properly forecasting future manpower strength will always be critical to military recruiting and organization. The ability to assess the cross-section of covariates of a cohort of enlistees and forecast manpower strength would allow for improved planning and allocation decisions. We leverage an innovative method of survival analysis—random survival forests (RSF) with time-varying covariates (T-VC)—to predict Army first-term post-Initial Entry Training attrition rates. Using random survival forests with time-varying covariates (TV-RSF) is an emerging method of survival analysis that has not been used in a military manpower setting. Using a Brier Score we compare TV-RSF with three other methods. We illustrate that using a single tree rather than the computationally intensive TV-RSF may suffice for predicting future year attrition. We also illustrate that TV-RSFs outperform traditional classification methods (logistic regression, random forests) that only account for yearly changes in T-VCs.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AAG	Army Analytics Group
ACT-TRAN	Active Duty Military Personnel Transaction
ACT-MAST	Active Duty Military Personnel Master
AFQT	Armed Forces Qualifications Test
API	application programming interface
AWD	Army Waiver Database
BMI	body mass index
BS	Brier Score
CDC	Center for Disease Control
CMF	career management field
CPH	Cox Proportional Hazards
CPU	central processing unit
CRAN	Comprehensive R Archive Network
FY	fiscal year
IBS	integrated brier score
IET	Initial Entry Training
IMD	Invasive Meningococcal Disease
HOR	home of records
LTRC	left-truncated, right-censured
MEPCOM	Military Entrance Processing Command
MEDPROS	Medical Protection System
PDE	Person-Event Data Environment
PHA	Physical Health Assessment

PID	personally identification numbers
PII	personally identifiable information
PULHES	physical capacity/stamina (P), upper extremities (U), lower extremities (L), hearing and ears (H), eyes (E), and psychiatric (S)
RAM	random-access memory
RSF	random survival forest
RFL	Research-Facilitation Laboratory
RODBC	Oracle Database Connectivity (for R)
SQL	Structured Query Language
T-CC	time-constant covariate
TOAD	Tools for Oracle Application Development
T-VC	time-varying covariate
TV-RSF	time-varying random survival forest

EXECUTIVE SUMMARY

In 2019, only 45,000 U.S. Army enlistees were recruited of the 62,000 goal, a goal that has been considerably reduced from over 80,000 recruits in previous years (South 2019). Of the soldiers the Army are able to recruit, an average of 29.7% attrite before their first-term service obligation (Marrone 2020). According to Marrone (2020), in a RAND Corporation research study, first-term attrition costs the Army an annual amount ranging from $580 million to $652 million fiscal year (FY) 2022. We address the first-term retention problem facing the Army and compare more traditional analysis methods with an innovative method of survival analysis (Yao, Frydman, Larocque, Simonoff 2022a)—random survival forests with time-varying covariates (TV-RSF)—as a tool for manpower estimation and forecasting.

We study the performance of inclusion of time-varying covariates (T-VC) into manpower modeling, using random forests adapted to survival analysis over the use of traditional manpower attrition modeling. Yao et al. (2022a) formally propose a method of TV-RSF in their research along with an R package LTRCforests (Yao, Frydman, Larocque, Simonoff 2022b) to support the model fits in our research. We utilize the previous work of Speten (2018) who lays the groundwork for other research with his scoping of the problem, definition of attrition, cohort selection to only include records of soldiers who have completed Initial Entry Training, and extensive code extracting administrative and demographic data being utilized. The models we compare are conditional inference TV-RSF, left-truncated, right-censored (LTRC) survival trees, and the traditional manpower classifiers: random forests and logistic regression. The survival analysis methods that accommodate T-VCs use pseudo-records to capture changes in T-VCs and the LTRC methodology while random forests and logistic regression utilize Cammack's (2020) method of accounting for year changes in T-VCs. For these two traditional manpower methods, three separate models are tuned and trained with snapshots of the surviving subject's data at times two, three and four years. We measure these models' estimation power through the use of a modified Brier Score (BS) analytic (Brier 1950).

This research leverages the data and resources from the Person-Event Data Environment (PDE): "a consolidated data repository that contains unclassified but sensitive manpower, training, financial, health, and medical records covering U.S. Army personnel…" (Vie, Griffith, Scheier, Lester, Seligman 2013) Numerous analytical and statistical resources are available within the PDE's virtual environment. The only statistical resources utilized for our work are the PDE desktop's R software environment (R Core Team 2017) and the PDE's RStudio server (RStudio Team 2020). One of the major limitations in this research are the computational restrictions of the PDE's RStudio server where the computationally intensive TV-RSF model fits takes over 320 hours and over 90% of the available 128GB of RAM.

With our computational limitations, a smaller cohort of only a training set of FY2010 enlistees and test set of FY2011 enlistees, with only four-year terms of first-term military obligation, are used in our analysis. The variables selected for this research are determined through consideration of past research and decreasing computational run-time for our model fits. We use Lazzarevich's (2022) research to select our variables because the goal of his work is to identify variables—including T-VCs—important in predicting attrition. We reconstruct or collapse variables to more accurately capture their effect on attrition or to better fit our modeling methods.

We find our survival analysis models (TV-RSF and LTRC survival trees) outperform the traditional manpower methods at predicting first-term post-Initial Entry Training attrition. This is due to the more effective capture of T-VCs in our survival analysis models than the method used by Cammack (2020) which only incorporates annual T-VC values for traditional classification models. The survival analysis methods also produce more useful results than the classification models with the estimated cohort survival functions giving senior leaders insights on which groups of soldiers they can expect to attrite. The ability to forecast manpower strength is compatible with that of survival analysis methods, not of the traditional classification methods.

For new-year data, the TV-RSF only marginally outperforms the LTRC survival tree. With first-term attrition costing the Army $652 million annually, the small increase in prediction power with the TV-RSF may very well be worth the additional computational

time (Marrone 2020). We speculate that differences in policy, or economic conditions may have decreased the effectiveness of the TV-RSF in predicting first-term post-IET attrition in FY2011 data when trained with FY2010 data. This is indicated by how well the TV-RSF fits to the FY2010 data with a comparatively low training error to LTRC survival trees and the fact that cohort year is an important variable when included in previous models (Devig 2019). With the inclusion of these underlying policy or economic variables, the difference between new-year prediction accuracy could be greater between our two survival analysis methods.

References

Brier G (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review. 78 (1): 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

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.

Lazzarevich N (2022) Predicting U.S. Army enlisted attrition after initial entry training using random survival forests. Master's thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA. http://hdl.handle.net/10945/69666.

Marrone J (2020) Predicting 36-month attrition in the U.S. military: a comparison across service branches. RAND Corporation, https://www.rand.org/pubs/research_reports/RR4258.html. Also available in print form.

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/.

RStudio Team (2020) RStudio: integrated development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

South T (2019) Rising costs, dwindling recruit numbers, increasing demands may bring back the military draft. Military Times. Accessed August 21, 2022, https://www.militarytimes.com/news/your-military/2019/11/19/rising-costsdwindling-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.

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.

Yao W, Frydman H, Larocque D, Simonoff JS (2022b) LTRCforests: Ensemble methods for survival data with time-varying covariates. R package version 0.5.5, https://cran.r-project.org/web/packages/LTRCforests/.

Yao W, Frydman H, Larocque D, Simonoff JS (2022a) Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research,* https://doi.org/10.1177/09622802221111549.

ACKNOWLEDGMENTS

I would like to thank Dr. Lyn Whitaker for being my advisor for this thesis. Her mentorship, expertise, and kindness since the beginning of my time at NPS were invaluable. Thank you to Dr. Sam Buttrey for your guidance during this process. Thank you to my friends and family for all of your support during my time at NPS, especially Jacqui who made writing this thesis so much more enjoyable. Thank you to my professors and instructors at USNA who led me this point, especially CAPT David Ruth and LCDR Kai Seglem—without your help and motivation, I would not have attended NPS. Thank you to all the professors, instructors and entire OR who made my time at NPS so rewarding.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

According to Army Secretary Christine Wormuth, fiscal year (FY) 2022 is the "Army's most challenging recruiting year since the start of the all-volunteer force, [which] will only achieve 75% of [its proposed goal]" (Harrison 2022). In 2019, only 45,000 enlistees were recruited of the 62,000 goal, a goal that has been considerably reduced from over 80,000 recruits in previous years (South 2019). Of the soldiers the Army are able to recruit, an average of 29.7% attrite before their first-term service obligation—the highest of all the military services by a margin of 6.1% (Navy) to 11.1% (Marine Corps) (Marrone 2020). According to Marrone (2020), in a RAND Corporation research study, first-term attrition costs the Army an annual rate ranging from $580 million to $652 million FY2022. Confronted with the fiscally limited environment of recent years, the Army must work to retain new recruits to reconcile prevailing and planned manpower and force goals; the ability to assess cross-sections of recruits to predict or forecast manpower strength will allow for improved planning and allocation decisions.

A. PURPOSE OF RESEARCH

The purpose of this research is multi-faceted, addressing both the first-term retention problem facing the Army and assessing an innovative method of survival analysis (Yao, Frydman, Larocque, Simonoff 2022a)—random survival forests with time-varying covariates (TV-RSF)—as a tool for manpower estimation and forecasting. With manpower problems often relying on data collected over time, time-varying covariates (T-VC) are often present; T-VCs are variables (also referred to as "covariates") that have values that change over time like the variable *number of dependents* increasing each time a soldier's child is born. Potentially being important predictors, changes in T-VCs are not usually or easily captured by traditional manpower modeling methods, leading to decreased prediction power in models that omit them. Further, traditional manpower analysis utilizes classification models to estimate, for example, attrition rates at the end of a fixed time period. In contrast, survival analysis—typically used in biomedical research—applies survival (retention) functions to manpower analysis; these functions allow analysts to

easily estimate cohort attrition rates across time. In addition, using T-VCs with survival analysis gives analysts a way to estimate future attrition rates among a cohort using the latest information about the individuals in that cohort. The TV-RSF is an emerging method never before used to estimate and forecast with manpower data; we compare this method to previously used, more traditional methods applied to the problem of assessing Army attrition. We illustrate how to use TV-RSF and when it is a superior method.

B. RELATED WORK

This research continues the work of Speten (2018), Gobea (2019), Devig (2019), Cammack (2020), and Lazzarevich (2022). Their focus is addressing first-term Army attrition of soldiers who have completed Initial Entry Training (IET). Their work is based on Army manpower and medical data spanning FY2005 through FY2017. We use a subset of this data; all soldiers with a four-year term of obligation who enlist in FY2010 and in FY2011. We rely heavily, with some modifications, on their data pre-processing, variable construction, and lessons learned. In this section we highlight pertinent features of their work. We also introduce the work of Yao et al. (2022a) who generalizes random survival forests (RSF) to incorporate T-VCs.

1. Traditional Manpower Attrition Modeling

The first iterations of research on this problem come from Speten (2018) and Gobea (2019) who utilize logistic regression models to estimate first-term attrition rates without accounting for T-VCs. Gobea (2019) uses the first value for each T-VC which is then treated as a time-constant covariate (T-CC). For example, only marital status, number of dependents, etc., recorded at enlistment are used. Speten (2018) lays the groundwork for other research with his scoping of the problem, definition of attrition, cohort selection to only include records of soldiers who have completed IET, and extensive code extracting administrative and demographic data being utilized; refer to Speten's (2018) work for an understanding of his efforts. The most important variables for predicting attrition by Speten's (2018) models are (1) *deployment history* (in part, due to an error in how deployment history is captured as T-CCs), (2) *initial contract length*, and (3) *marital status*. With the inclusion of the two medical datasets pre-processed by Devig (2019) and Speten's

(2018) six administrative and demographic datasets, Gobea's (2019) most important variables are (1) *PULHES* (physical capacity/stamina (P), upper extremities (U), lower extremities (L), hearing and ears (H), eyes (E), and psychiatric (S)) *non-deployable profile/ codes*, (2) *dental-class*, and (3) *initial contract length*). While traditional manpower modeling most commonly utilizes logistic regression, other classification methods like classification trees and random forests are also used.

Cammack (2020) also utilizes the traditional manpower modeling method of a logistic regression classification model but unlike Speten (2018) and Gobea (2019), attempts to include T-VCs. Applying the dataset constructed by Devig (2019) and used by Gobea (2019), Cammack (2020) fits a sequence of conditional logistic regression models trained with snapshots of surviving subjects' covariates at the beginning of each year, and estimating the conditional attrition rate for the year among soldiers surviving the beginning of the year. Thus, each model in the sequence treats the T-VCs as time constant. Training a model for each year in a cohort's contract term length allows for some of the changes in the T-VCs to be captured. Her method is superior to the method used by Speten (2018) and Gobea (2019) with its allowance for limited inclusion of T-VCs but unlike the survival analysis modeling methods used by this research it does not observe every change made by a T-VC. Cammack (2020) inspires the research of Lazzarevich (2022) who also fits a sequence of conditional models capturing the time-varying values of covariates at the beginning of each year. Like Cammack (2020), each model treats T-VCs as time constant. Unlike Cammack (2020), Lazzarevich (2022) fits RSFs which have more flexibility and tend to predict better than traditional logistic regression models.

2. Survival Analysis in Manpower Attrition Modeling

Unlike the traditional manpower modeling approach of logistic regression which classifies attrition or retention, survival analysis applied to manpower data estimates a survival function, *S(t)*, the probability of surviving (or retention) beyond time *t.* See James, Witten, Hastie, and Tibshirani (2021) for an introduction to survival analysis in the context of other statistical learning methods including logistic regression, trees, and random forests. We adopt survival analysis terminology where "surviving" to time *t* means that the

soldier did not attrite prior to time *t,* where *t* is time in service (in years) measured from date of enlistment. For each subject's survival function, the complement 1 – *S(t)* gives the probability of attrition at or before time *t*. Rarely used in a manpower setting, survival analysis is more commonly used in biomedical literature. As an example, Shen, Bouée, Aris, Corrine, and Ekkehard (2022) in their research on long-term mortality of Invasive Meningococcal Disease (IMD), estimate survival functions for those who survived IMD and the general healthy U.S. population as controls in Figure 1. The survival curves start and end with the same probabilities, but convey much more information between the start and end term; traditional manpower classification methods do not have the innate ability to capture these relationships.

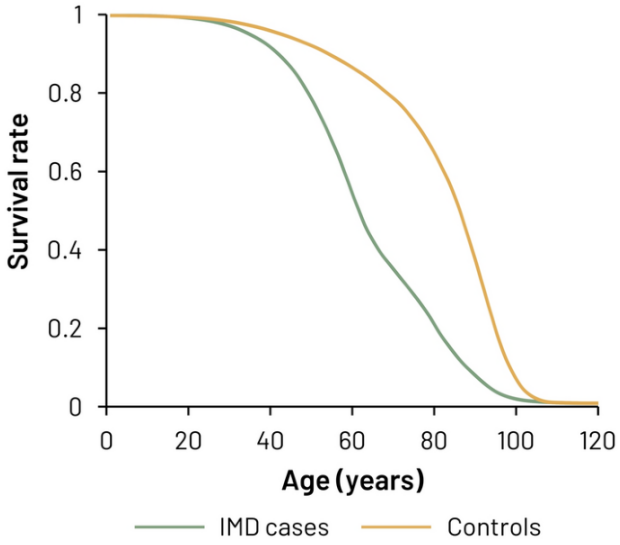


Figure 1. Extrapolated Lifetime Survival Curves and Life Expectancy by Age with IMD. Source: Shen et al. (2022).

The survival analysis method used in most biomedical literature to incorporate T-VCs is the Cox Proportional Hazards model (CPH). James et al. (2021) have an accessible introduction to CPH, in addition to T-VCs. This method presumes that hazards for different values of covariates are proportional which does not apply to the large military datasets used in this research. An alternative to CPH are algorithmic models like survival trees (Wei and Simonoff 2017) and tree ensembles (James et al. 2021) which partition the data using

the covariates and then produce survival functions from those subsets. These models capture T-VCs by treating each subject's whole record as an amalgamation of many pseudo-records. These pseudo-records represent time intervals when the subjects' T-VCs are constant. Individually a pseudo-record is treated as an incomplete observation, right censored and/or left truncated (see e.g., James et al. (2021) for definitions). A pseudo-record is right censored if it is not the last pseudo-record for a subject or left truncated if it is not the first pseudo-record for a subject. Each time a subject's T-VC changes, a pseudo-record is right censored and another pseudo-record that had been left truncated carries on the subject's covariates until the next potential change, attrition, or contract completion. Survival analysis methods that accommodate left-truncated and right-censured (LTRC) data can be adapted for use with T-VCs through the use of pseudo-records. For a review of survival analysis and survival modeling methods, as they apply to our problem, refer to the research of Devig (2019) and Lazzarevich (2022).

Devig (2019) is the first to utilize survival analysis and fully account for T-VCs on this problem with his use of LTRC survival trees (Wei and Simonoff 2017). Devig (2019) expands Speten's (2018) administrative and demographic datasets to include medical data, and refines the response variable of attrition through careful evaluation of service codes in the data to allow for more effective analysis. The most important variables found in Devig's (2019) research were: (1) *dental class*, (2) *vision class,* and (3) *PULHES codes*. We see compared to Speten's (2018) work that all of the important variables are both from medical datasets and T-VCs. We note that Devig (2019) did not use deployment history data due to concerns over data quality.

The natural progression from trees is to use random forests, a method which almost always outperforms trees; Devig's (2019) use of LTRC survival trees motivates Lazzarevich's (2022) chosen modeling methodology of RSFs, but the software to fit TV-RSF was not yet available to him. Lazzarevich (2022) trains his models to partially account for T-VCs with the approach used by Cammack (2020). His goal is to identify important variables in this problem. Gobea (2019) and Devig (2019) show that a class four *dental class* or *vision class* (these factor levels are defined in Table 7) is important for attrition, but Lazzarevich (2022) points out that these soldiers have already made the decision to

attrite. Thus, from the soldier's perspective, the act of attriting is a predictor of a class four *dental/vision class* rather than the other way around. Lazzarevich (2022) treats the fours in these two variables as missing values (coded as NA) and imputes them from existing data. From a higher-level administrative perspective, we believe these class four levels are important predictors of attrition and do not look into the future. Because Lazzarevich (2022) computes a measure of variable importance for each model in his sequence of models, we see from his work how the importance of T-VCs change with time. Some covariates, for example *prior service*, tend to be important for predicting attrition early in the first term, whereas others such as *back pain*, tend to be more important later in the first-term (see Table 7 for variable definitions). We use Lazzarevich (2022) variable importance results to aid variable selection of T-VCs for our model. Like Devig (2019), many of Lazzarevich's (2022) most important variables include medical T-VC covariates.

3. Random Survival Forests with Time-Varying Covariates

The greatest impediment to the modeling approach taken by Lazzarevich (2022) is the inability to leverage LTRC pseudo-records to account for T-VCs in his RSFs. Yao et al. (2022a) formally propose this method of TV-RSF in their research along with an R package LTRCforests (Yao, Frydman, Larocque, Simonoff 2022b) to support the model fits. Their research compares the semi-parametric CPH model to multiple TV-RSF models, including the conditional inference TV-RSF we use in our research. In their simulation studies, when data is generated under the assumptions of the CPH model, CPH outperforms the RSF methods; however, when the assumptions are not met, the ensemble RSF methods outperform the CPH model with the conditional inference TV-RSF performing better than the other forests. We use the conditional inference TV-RSF in our research. While our research utilizes the approach of bootstrapping subjects for bagged trees—as opposed to bootstrapping individual pseudo-records—Yao et al. (2022a) find that the bootstrapping methods have no differences in levels of performance. Yao et al. (2022a) also identify the tuning parameters for their proposed model which include (1) mtry, (2) ntree, and (3) alpha. The parameter mtry represents the size of the sample of covariates to be evaluated at each node for conditional inference. The parameter ntree represents the number of bagged trees in the random forest ensemble. The parameter alpha represents the significance value of

splitting criteria in the conditional inference paradigm of trees, where a larger alpha value induces greater depth within each bagged tree. For an in-depth discussion of the proposed forests for T-VC data refer to Yao et al. (2022a).

C. THESIS OUTLINE

This thesis studies the performance of inclusion of T-VCs into manpower modeling, using random forests adapted to survival analysis over the use of traditional manpower attrition modeling. The models we compare are conditional inference TV-RSF, LTRC survival trees, and the traditional manpower classifiers: random forests and logistic regression. The survival analysis methods that accommodate T-VCs use pseudo-records to capture changes in T-VCs and the LTRC methodology while random forests and logistic regression utilize Cammack's (2020) method of accounting for T-VCs—for these two traditional manpower methods, three separate models are tuned and trained with snapshots of the surviving subject's data at times two, three and four years. We measure these models' estimation power through the use of a modified Brier Score (BS) analytic explored in Chapter III (Brier 1950). This thesis follows a progression structure where previous research is discussed, variables are enumerated, models are tuned and fit with equivalent given information, and compared. This structure fills five chapters. Chapter II explores the data with sections describing the data environment, datasets used, cohort selection, and variables constructed. Chapter III examines our measure of comparison, modeling methodologies, model tuning, and final fits for our four models. Chapter IV discusses the final results of our models with the comparison of the accuracy estimations through BS. Chapter V summarizes our findings and outlines the considerations and our recommendations for future research on this topic.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA

This chapter discusses the data utilized in this research. First, we examine the environment where our datasets and analytic tools are housed. We then explain the datasets used from our environment and how they are extracted. Our chosen research cohort is accounted with a description of our rationale. We then discuss methodology where the unique survival analysis data format, traditional manpower attrition modeling data format, variable selection criteria, and variables used are explained.

A. PERSON-EVENT DATA ENVIRONMENT

This research leverages the data and resources from the Person-Event Data Environment (PDE): "a consolidated data repository that contains unclassified but sensitive manpower, training, financial, health, and medical records covering U.S. Army personnel…" (Vie, Griffith, Scheier, Lester, Seligman 2013) This cloud-based resource—created and maintained by the Army Analytics Group (AAG) and its Research-Facilitation Laboratory (RFL)—is accessible for research through the use of personal identification numbers (PID) to relate the many databases in its repository. This anonymized PID also allows for the removal of all personally identifiable information (PII) of the subjects such as personnel locations, unit information, and social security numbers. The PDE stores the schema/databases in a structured query language (SQL) server with access available through a Windows 10 virtual machine.

Accessing the sensitive proprietary data within the PDE requires careful authorization. Approval for PDE account creation must be given by the AAG before the process for project admission (data access) is initiated; this process can take several months to complete. Upon admittance to a PDE project, only specific datasets are available within the scope of the project's research. Requests for additional data access are granted at the discretion of the AAG. To protect sensitive information, the PDE's closed system environment does not allow internet or local machine access for its users; requests to export analytical products—e.g., charts, plots, and code—are scrutinized for protected data. All

data analytics must be performed with the resources provided in the closed-virtual environment.

Numerous analytical and statistical resources are available within the PDE's virtual environment. The only statistical resources utilized for our work are the PDE desktop's R software environment (R Core Team 2017) and the PDE's RStudio server (RStudio Team 2020). The PDE's RStudio server has access to all the R packages maintained on the Comprehensive R Archive Network (CRAN) with additional packages installed at the discretion of the AAG. While the desktop R software is utilized at the beginning of this research, the RStudio server's superior computational resources and access to R-packages on the CRAN make it our principal statistical tool. As critical as the PDE's RStudio server is to our research, it was not without significant limitations.

The PDE's RStudio server has access to eight central processing units (CPU) with a combined 128GB of random-access memory (RAM). The multiple cores allow us to leverage parallelization when fitting our models. This method of parallelization decreases runtime by approximately eight-fold; our most computationally intensive model—TV-RSF—takes over 40 hours to fit with a computational time of over 320 hours and over 90% of the available 128GB of RAM being utilized. With strained memory in a shared server, fitting larger models is infeasible on the PDE for this research. Due to the computational runtime of RSFs increasing at a rate of $n \times \log(n)$ with n being the number of observations (pseudo-records), the most significant limitation to this research is the bounded size of our subject cohort.

B. DATASETS USED

This research works with some, but not all of the same databases utilized by prior work on this topic. Of the six administrative and demographic datasets used by Devig (2019) and Lazzarevich (2022) only four are used in this research. The two medical datasets used by their work are still included in this research. These datasets stored on the PDE's SQL server are accessed through Oracle Database Connectivity (for R) (RODBC) (Ripley and Lapsley 2021)—an application programming interface (API)/R package—that allows for querying and execution of SQL commands from R scripts. Tools for Oracle Application

Development (TOAD) are also leveraged to browse schemas and visualize tables. While the data used in this research are already transferred from the SQL server into R data frames by Speten (2018) and Devig (2019), we replicate these efforts for fundamental understanding of the data format.

Datasets are constructed using two collection methods: (1) snapshot, and (2) subject development. The snapshot datasets give a snapshot and date of all subject's covariates at certain times (e.g., upon entry, the last day of every month, etc.). The subject development datasets are transactional and include new observations of all a subject's covariates at each time a covariate changes (with the date of the change). Each covariate in the six datasets outlined in Table 1 are connected by a subject's unique PID with a timestamp associated to each change in a subject's covariates.

Table 1. Description of Administrative, Demographic, Medical Datasets.

Dataset	Description
Active Duty Military Personnel Master (ACT-MAST)	Soldier administrative data: marital status, career management field (CMF), rank, terms of service, and service dates. *Collected Quarterly.*
Active Duty Military Personnel Transaction (ACT-TRAN)	Soldier administrative data: determines "attrition status" via reenlistment/separation codes. *Subject Development Collection.*
Military Entrance Processing Command (MEPCOM)	Soldier demographic data: dependent, home or record (HOR) race, ethnicity, and Armed Forces Qualification Test (AFQT) information. *Collected Upon Entry.*
Army Waiver Database (AWD)	Soldier administrative data: administrative and conduct waivers. *Collected Upon Entry*
Physical Health Assessment (PHA)	Soldier medical data: recorded physical/mental conditions. *Collected Annually*
Medical Protection System (MEDPROS)	Soldier medical data: medical/deployment readiness (includes PULHES). *Subject Development Collection.*

C. COHORT DESCRIPTION

Lazzarevich (2022) reduces the cohort created by Devig (2019) by utilizing only subjects starting their service in FY2010 as the training set and subjects starting their service in FY2011 as the test set. These years are chosen for their minimal missing values compared to previous years. As discussed in Section A of this Chapter, our biggest limitation in this research is the computational demand needed to fit a TV-RSF with restricted computational supply in the PDE's RStudio server environment. To further reduce the size of the cohort used by Lazzarevich (2022) only subjects with four-year term contracts are evaluated. Even though out of the four available contract terms (3 – 6 years) three-year contracts are more frequent in the data as seen in Table 2, four-year term contract subjects are selected to enable an additional year of assessment. In choosing only four-year term subjects for our research we reduce the number of subjects in our initial cohort from 124,052 to 34,231.

Table 2. Distribution of Subjects in Data with Differing Year Term Commitments by Service Start Data.

	3-Year Term	4-Year Term	5-Year Term	6-Year Term	Total Subjects
Service Start 2010	53.6%	29.3%	8.1%	9.0%	66,113
Service Start 2011	55.0%	25.7%	9.4%	9.9%	57,939
Service Start 2010/11	54.3%	27.6%	8.7%	9.4%	124,052
Total Subjects	67,336	34,231	10,773	11,712	

Of the 34,231 subjects in our cohort 4.3% are missing at least one covariate value. Most of the missing values are in medical covariates. With IET not being the final training in the Army training pipeline, we rationalize that many of the missing medical data values are due to no collections being made before finishing all training which often lasts up to 52 weeks. To minimize missing data, we remove subjects who survived IET, but attrite in the first year; 2,957 subjects attrited in the first year post-IET and are thus removed from our final cohort. After truncating this year, only 226 subjects have incomplete records; these are also removed. A further 522 subjects are removed for being incompatible with

the TV-RSF method with some of their pseudo-records having negligible time differences between truncation and censoring. The final cohort for our research contains 30,526 subjects with an attrition rate of 22.2%. 80,709 records are in the 2010 (training set) cohort and 52,480 records are in the 2011 (test set) cohort with 21.6% and 22.9% attrition rates respectfully.

D. METHODOLOGY

This section discusses the methodologies and rationale behind data formatting and variable selection, and defines our covariates. The differences between the unique survival analysis and traditional manpower method data formats are explained. The variable selection process is depicted utilizing the variable importance findings from Lazzarevich (2022). The construction methodologies, definitions, and time-varying attribute for each covariate are described.

1. Unique Survival Analysis Data Format

When our data is initially transferred from the PDE's SQL database to R, each subject has one record or row with each of their covariates, changes of covariates, times of changes in covariates, and status of attrition; we call this format the "wide format." To produce the LTRC pseudo-records needed to evaluate T-VCs in survival analysis, the data frame must be transformed into the "long format"; where the single row for each subject in the wide format is expanded to many rows, one for each time a covariate changes values. Each pseudo-record has three additional variables indicating times (in years): *tstart, tstop,* and *status*. The *tstart* variable indicates the time when a variable in the record is changed to a new value, the *tstop* variable indicates the time when a variable in the record is about to change values, and the variable *status* indicates 1 if the subject attrites at *tstop* or 0 if the subject attrites past *tstop*. A more in-depth discussion and example of the "long format" of survival data can be found in Devig (2019). This "long format" forces the number of observations fed into training a model to be much larger than the number of subjects, which in turn dramatically increases the run-time for our TV-RSF model compared to other models.

2. Traditional Manpower Attrition Modeling Data Format

Random forests and logistic regression are not able to inherently handle T-VCs like our survival methods of TV-RSF and LTRC survival trees; to utilize the method leveraged by Cammack (2020) and Lazzarevich (2022) further data preparation is conducted to treat the T-VCs like T-CCs. This is done by taking three annual snapshots of each subjects covariates at the beginning of years one, two and three. While more snapshots can be produced to refine the accuracy of the overall model, it would be impractical to tune and fit more than three models for each classification method. The fact that the TV-RSF and LTRC survival trees are able to handle T-VCs lend to their appeal in survival analysis, especially with data with many or important T-VCs.

3. Variable Selection

The variables selected for this research are determined through consideration of past research and decreasing computational run-time for our model fits. The most similar model to our primary model—TV-RSF—from previous research was utilized by Lazzarevich (2022). We use Lazzarevich's (2022) research to select our variables because the goal of his work is to identify variables—including T-VCs—important in predicting attrition. The variable importance for each of Lazzarevich's (2022) annual snapshots is computed as seen in Table 3 which reports the rankings of variable importance for each year's model. While Lazzarevich (2022) documents the importance for each of his variables across all four term lengths (3 – 6 years) we only examine the four-year term contract variable importances for the models that predict attrition in the second, third, and fourth year.

We use the same set of data and variables to compare our four attrition models. Our primary goal is to examine the effectiveness of TV-RSFs compared to the other three methods. We are incentivized to include more T-VCs to amplify the TV-RSFs ability to predict from T-VCs, but to maintain a balance of T-VCs and T-CCs. In an effort to reduce the computational run-time of fitting our models, we omit variables identified as unimportant by Lazzarevich (2022), or incompatible with our method. Some of the variables selected are reconstructed from Lazzarevich's (2022) variables to more

accurately capture their effect on attrition or collapsed to better fit our methods; these reconstructions/collapses will be discussed in detail in the following section.

Table 3. Variable Selection from Variable Importance Rankings of Four-Year Term Contract Subjects. Adapted from Lazzarevich (2022).

Term Length	**4**			**Included in Research**	**Variable Reconstructed/ Collapsed**	**Time-Varying Covariate**
Year of Term	**1**	**2**	**3**			
Variable						
AFQT Category	7	11	11	Yes	Yes*	No
Age at Enlistment	6	14	26	Yes	No	No
Anemia	38	26	36	No	-	Yes
Asthma	32	31	24	No	-	Yes
Back Pain	13	7	5	Yes	No	Yes
BMI at Enlistment	16	9	12	Yes	No	No
Chronic Pain	31	2	1	Yes	No	Yes
CMF Code after IET	10	17	17	Yes	No	Yes
CMF Code at Enlistment	14	19	19	Yes	No	No
Dental Readiness	1	6	20	Yes	Yes*	Yes
Deployment	45+	35	27	No	-	Yes
Diabetes	45+	45+	45+	No	-	Yes
Educational Tier Code at Enlistment	17	15	42	Yes	No	No
Epilepsy	43	45+	45+	No	-	Yes
Gender	2	1	13	Yes	No	No
Headaches	9	3	3	Yes	No	Yes
Hearing Readiness Class	35	38	30	Yes	Yes*	Yes
Heart Murmur	28	44	43	No	-	Yes
Heart Trouble	27	22	28	No	-	Yes
Hispanic	20	20	33	Yes	No	No
HOR State	12	18	18	Yes	Yes*	No
Hostile Injury	45+	42	35	No	-	Yes
Hypertension	36	28	31	No	-	Yes
Joint Pain	21	10	6	Yes	No	Yes
Liver Disease	45+	45+	45+	No	-	Yes
Marital Status Code	5	8	14	No	-	Yes
Mental Health Concern	4	4	8	Yes	No	Yes
Non-Hostile Injury	45+	41	38	No	-	Yes

Term Length	4			Included in Research	Variable Reconstructed/ Collapsed	Time-Varying Covariate
Year of Term	**1**	**2**	**3**			
Variable						
Number of Dependents at Enlistment	26	27	39	Yes	No	No
P – PULHES after IET	8	25	9	Yes	No	Yes
U – PULHES after IET	42	21	7	Yes	Yes*	Yes
L – PULHES after IET	33	16	4	Yes	Yes*	Yes
H – PULHES after IET	37	40	16	Yes	No	Yes
E – PULHES after IET	11	23	15	Yes	No	Yes
S – PULHES after IET	34	45+	2	Yes	Yes*	Yes
P – PULHES at Enlistment	25	36	34	Yes	No	No
U – PULHES at Enlistment	44	45+	32	Yes	No	No
L – PULHES at Enlistment	39	43	44	Yes	No	No
H – PULHES at Enlistment	40	39	45+	Yes	No	No
E – PULHES at Enlistment	22	24	22	Yes	No	No
S – PULHES at Enlistment	41	37	45+	Yes	Yes*	No
Pregnancy	18	5	10	No	-	Yes
Prior Service	3	12	28	Yes	No	No
Race Code	15	13	23	Yes	No	No
US Citizen. Origination	29	32	41	No	-	No
US Citizen. Status	45+	45+	45+	No	-	No
Vision Readiness	19	30	21	No	-	Yes
Waiver Admin	30	33	40	Yes	No	No
Waiver Conduct	23	29	25	Yes	No	No

*Variables reconstructed/collapsed are discussed in Chapter 2, Section D, Subsection 2 – *Variables Used*.

4. Variables Used

Out of the variables selected in Table 3, this research utilizes two distinct types of variables: (1) T-CCs, and (2) T-VCs. The T-CCs have fixed values over time; T-CCs include many of the administrative, all of the demographic, and some of the medical data collected at the date of enlistment. The T-VCs change throughout the term of an enlistee's contract with most of the medical data and some of the administrative data adjusting to the subject's experience one-year post-IET. Both T-CCs and T-VCs have categorical data but only the T-CCs have numerical data. Each of the three numerical covariates are discrete and the 30 categorical covariates have range of 2 to 29 levels. There is a combination of 130 factor levels among the categorical variables.

We see in Table 3 there are a total of 19 T-CCs and 14 T-VCs with additional columns identifying the variables as either "Constructed" or "Collapsed." The constructed variables are created out of other covariates, while collapsed variables are factors (categorical variables) in which the number of levels is reduced by combining unimportant or sparse levels with neighboring levels. Some factors are first collapsed in the research conducted by Devig (2019) and/or further collapsed by Lazzarevich's (2022) work. Their rationale is to lower the total number of levels in each factor and combine factors with similar definitions (Devig 2019). Our research also calls to further collapse factor levels but for a reason specific to our modeling methods. The function used to fit our TV-RSFs—ltrccif from the LTRCforests R package (Yao et al. 2022b)—is unable to properly handle sparse levels when there is a chance that a random sample used for tuning validation does not include some level. The function incorrectly deletes the level from the factor which then produces an "unknown variable error" when predicting with data containing the deleted level. To mitigate the occurrence of this error, factors in our test and training sets with levels with fewer than 50 pseudo-records are collapsed into neighboring levels. The methods of variable construction, reasons for which factors are collapsed, and descriptions of the variables are outlined later in this section.

Table 4. Summary of Variables. Adapted from Lazzarevich (2022).

Variable	Type	Constructed	Collapsed	Factor Levels	Time-Varying
AFQT Category Code	Categorical	No	Yes	5	No
Age at Enlistment	Numeric	Yes	No	-	No
Back Pain	Binary	No	No	2	Yes
BMI at Enlistment	Numeric	Yes	No	-	No
Chronic Pain	Binary	No	No	2	Yes
CMF Code after IET	Categorical	No	Yes	18	Yes
CMF Code at Enlistment	Categorical	No	Yes	18	No
Dental Class	Categorical	Yes	No	5	Yes
Dependents (Number at Enlistment)	Numeric	No	No	-	No
Educational Tier Code at Enlistment	Categorical	No	No	3	No
Gender	Binary	No	No	2	No
Headaches	Binary	No	No	2	Yes
Hearing Readiness Class	Categorical	Yes	No	5	Yes
Hispanic	Binary	Yes	No	2	No
HOR State/Territory	Categorical	No	Yes	29	No
Joint Pain	Binary	No	No	2	Yes
Mental Health Concern	Binary	No	No	2	Yes
Prior Service	Binary	No	No	2	No
*PULHES after IET**	Categorical	No	Yes/No**	2-3**	Yes
*PULHES at Enlistment**	Categorical	No	Yes/No**	2-4**	No
Race Code	Categorical	No	No	4	No
Waiver (Administrative)	Binary	No	No	2	No
Waiver (Conduct)	Binary	No	No	2	No

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

** Some of the PULHES variables have been collapsed for sparse levels causing errors in the modeling.

a. Time-Constant Covariates

Out of our 16 categorical T-CCs, six are collapsed factors and only one is a constructed variable. The *Armed Forces Qualification Test (AFQT) category code* variable

is initially collapsed by Devig (2019) to six levels is further collapsed to five levels in this research to mitigate errors in model fitting. *Career management field (CMF) code at enlistment* is also initially collapsed by Devig (2019) but further collapsed by Lazzarevich (2022) for reasons discussed in their research. *Home of record (HOR) state/territory*, and the *U/L/S – PULHES at enlistment* variables are all collapsed to mitigate errors in model fitting. The *hispanic* binary variable is constructed by Lazzarevich (2022) using ethic codes from the Military Entrance Processing Command (MEPCOM) dataset. Two of our three numerical T-CCs are constructed by prior researchers. Devig (2019) constructs *age at enlistment* utilizing enlistment dates and Lazzarevich (2022) constructs *body mass index (BMI) at enlistment* utilizing a Center for Disease Control (CDC) formula (CDC 2021).

Table 5. Time-Constant Categorical Covariates. Adapted from Lazzarevich (2022).

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	1-30%
CMF Code at Enlistment	Assigned occupational code.	Multiple (18 levels)	See Appendix A
Educational Tier Code at Enlistment	Indicates high school completion or equivalent.	1	High school diploma
		2	GED or equivalent
		3	No high school diploma, GED, or equivalent
Gender	Gender of enlisted.	M	Male
		F	Female
Hispanic	Whether enlisted is Hispanic or not.	N	Not Hispanic
		Y	Ethnic code of AK, AL, AM, AN, or AO (is Hispanic)
HOR State/ Territory	Home or record state code.	Multiple (29 levels)	See Appendix B
Prior Service	Whether enlisted has prior serve or not.	0	Has no prior service
		1	Has prior service
		1	High level of fitness

Variable	Description	Levels	Level Description
P – PULHES at Enlistment	A qualifier of an enlistee's physical profile and stamina observed at enlistment.	2	Possess a medical condition that limits some activities
		3	Possess a medical condition that requires significant limitations
U – PULHES at Enlistment	A qualifier of an enlistee's upper extremities observed at enlistment.	1	High level of fitness
		2+	Possess a medical condition that limits some activities or requires significant limitations
L – PULHES at Enlistment	A qualifier of an enlistee's lower extremities observed at enlistment.	1	High level of fitness
		2+	Possess a medical condition that limits some activities
H – PULHES at Enlistment	A qualifier of an enlistee's hearing observed at enlistment.	1	High level of fitness
		2	Possess a medical condition that limits some activities
		3	Possess a medical condition that requires significant limitations
E – PULHES at Enlistment	A qualifier of an enlistee's eyesight observed at enlistment.	1	High level of fitness
		2	Possess a medical condition that limits some activities
		3	Possess a medical condition that requires significant limitations
S – PULHES at Enlistment	A qualifier of an enlistee's psychiatric and emotional profile observed at enlistment.	1	High level of fitness
		2+	Possess a medical condition that limits some activities or 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
		N	Did not receive a waiver

Variable	Description	Levels	Level Description
Waiver (Administrative)	If enlistee received an administrative waiver for enlistment.	Y	Received a waiver
Waiver (Conduct)	If enlistee received a conduct waiver for enlistment.	N	Did not receive a waiver
		Y	Received a waiver

Table 6. Time-Constant Numerical Covariates. Adapted from Lazzarevich (2022).

Variable	Description
Age at Enlistment	Recruit's age at time of enlistment. Constructed using birth data and date of enlistment.
BMI at Enlistment	BMI as defined by the CDC (2021).
Dependents (Number at Enlistment)	Number of dependents at the time of enlistment.

b. Time-Varying Covariates

From our 14 T-VCs two are constructed and two are collapsed. The factors *dental class* and *hearing readiness class* are given an extra level "0" that represents subjects that have not yet been evaluated for dental or hearing readiness class. *CMF code after IET* is collapsed for the same reason as *CMF code at enlistment. S – PULHES after IET* is collapsed to mitigate errors in model fitting with not enough observations in a level.

Table 7. Time-Varying Covariates. Adapted from Lazzarevich (2022).

Variable	Description	Levels	Level Description
Back Pain	Medical condition as recorded in the enlisted PHA.	N	Has not been diagnosed with the condition
		Y	Has been diagnosed with the condition
Chronic Pain		N	Has not been diagnosed with the condition

Variable	Description	Levels	Level Description
	Medical condition as recorded in the enlisted PHA.	Y	Has been diagnosed with the condition
CMF Code after IET	Assigned occupational code that may change.	Multiple	See Appendix A
Dental Class	Dental Readiness determined by level of treatment needed.	0	First dental evaluation post-IET has not occurred yet
		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
Headaches	Medical condition as recorded in the enlisted PHA.	N	Has not been diagnosed with the condition
		Y	Has been diagnosed with the condition
Hearing Readiness Class	Hearing readiness determined by level of treatment needed.	0	First hearing test post-IET has not occurred yet
		1	Hearing test current; no hearing issues
		2	Hearing test: minor issues
		3	Minor hearing issues; need evaluation by audiologist
		4	No hearing test within past 13 months; requires immediate hearing test
Joint Pain	Medical condition as recorded in the enlisted PHA.	N	Has not been diagnosed with the condition
		Y	Has been diagnosed with the condition
Mental Health Concern	Medical condition as recorded in the enlisted PHA.	N	Has not been diagnosed with the condition
		Y	Has been diagnosed with the condition
P – PULHES after IET	A qualifier of an enlistee's physical profile and stamina 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

Variable	Description	Levels	Level Description
			limitations, or military duty must be drastically limited
U – PULHES after IET	A qualifier of an enlistee's upper extremities 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, or military duty must be drastically limited
L – PULHES after IET	A qualifier of an enlistee's lower extremities 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, or military duty must be drastically limited
H – PULHES after IET	A qualifier of an enlistee's hearing 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, or military duty must be drastically limited
E – PULHES after IET	A qualifier of an enlistee's eyesight 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, or military duty must be drastically limited
S – PULHES after IET	A qualifier of an enlistee's psychiatric and emotional profile following the completion of IET.	1	High level of fitness
		2+	Possess a medical condition that limits some activities, requires significant limitations, or military duty must be drastically limited

THIS PAGE INTENTIONALLY LEFT BLANK

III. MODELING

This chapter discusses the measurement of comparison, issues encountered while fitting the TV-RSF model, and how we selected tuning parameters for all models. We take extra care in our discussion of fitting a TV-RSF model, to impact lessons learned to future military manpower analysts who might want to use this method. After tuning parameters are identified for each model, final models are fit. Comparisons are made in Chapter IV.

A. BRIER SCORE

The measures of effectiveness used both in model tuning and comparison between models are based on the BS. It quantifies, with respect to time, how well subjects' estimated survival functions at time t predict their survival at time t. Specifically let N be the number of subjects, $S_s(t)$ be the estimated probability that subject s survives past time t given the value of the subject's covariates at time t, and let $O_s(t)$ be the indicator function taking value 1 if subject s is observed to survive past t, and 0 if the subject s attrites in the time interval $(1, t]$ for $s = 1, \ldots, N$ (recall, we only use records of soldiers surviving year one of their first term). Then as a function of $t \in (1, 4]$:

$$BS(t) = \frac{1}{N}\sum_{s=1}^{N}(S_s(t) - O_s(t))^2 \tag{1}$$

1. Survival Analysis (Integrated) Brier Score

With survival analysis methods estimating survival functions for each subject, the $BS(t)$ for subjects can be easily calculated for time t from one-year post-IET to the end of their contract. To compare and tune our TV-RSF and survival analysis methods, we compute an integrated brier score (IBS) across time. This IBS gives a single number to compare models. For a four-year enlistment term where t is time in service measured in years, IBS is:

$$IBS = \int_1^4 BS(t)dt \tag{2}$$

2. Modified Brier Score for Classification Methods

Unlike survival analysis methods, classification attrition models estimate a single probability of attrition for the entire study period. Using Cammack's (2020) method of accounting for T-VCs we fit our logistic regressions and random forests as a sequence of three conditional models. Each model estimates the conditional probability of attrition in a single year given a subject's covariate values at the start of the year and given that the subject has not attrited in previous years. To make this more concrete, let Y be a random variable representing a subject's time of attrition or end of term—whichever occurs first. Further let $H(t)$ represent that subject's covariate values at time t (we have dropped the subscript s, denoting subject s, for convenience). Then the three models give estimates of the conditional probabilities:

$$P(Y > y+1 \mid Y > y, H(y)) \text{ for } y = 1, 2, 3 \tag{3}$$

From equations (1) and (3) we see that we can only compute $BS(t)$ for $t = 2, 3, 4$ from the sequence of logistic regression and random forest fits. The first of these sequential models ($y = 1$) estimates probabilities of surviving past $y = 2$ which can be substituted directed into equation (1) to compute $BS(2)$. Computing $BS(3)$, and $BS(4)$ requires estimates of probabilities of surviving past years three and four respectively given only survival past year one. These are derived by multiplication since:

$$P(Y > 3 \mid Y > 1) = P(Y > 3 \mid Y > 2)P(Y > 2 \mid Y > 1),$$

and

$$P(Y > 4 \mid Y > 3) = P(Y > 4 \mid Y > 3)P(Y > 3 \mid Y > 2)P(Y > 2 \mid Y > 1).$$

B. LEFT-TRUNCATED, RIGHT-CENSURED CONDITIONAL INFERENCE RANDOM SURVIVAL FORESTS WITH TIME-VARYING COVARIATES

1. Fitting the Model

Utilizing the ltrccif function from the LTRCforests package (Yao et al. 2022b) on R we tune and fit our TV-RSF model using IBS as our measure of effectiveness. The inputs to the ltrccif function include the parameters and certain objects like the survival formula

and data. The data to fit the ltrccif object has to be in the "long format" which we discuss in detail in Chapter II. The inputs and methods for the model are defined below.

- ntree—this value determines the number of trees to grow for the forest. In an ensemble method like random forests, the number of trees is limited by computational time. When tuning this parameter, increasing the number of trees will never decrease expected model proficiency but may severely increase the runtime to fit the model which we encounter as a major issue. The default value for this parameter is 100 (Yao et al. 2022b) but in our tuning this value varies from 1 to 400.

- mtry—this parameter is the size of the random sample of covariates to be considered at each node to be selected as the splitting variable in a tree. The default for this value is a tuning function within the LTRCforests package (Yao et al. 2022b) which we did not use as it is not supported within the PDE. Our tuning for this parameter ranges from 4 to 16.

- bootstrap—this parameter allows from a selection of four bootstrapping protocols: (1) by.sub, (2) by.root, (3) by.user, and (4) none. The default value which we use to fit our TV-RSFs is "by.sub" which bootstraps each tree by random subjects. The protocols "by.root" bootstraps trees by pseudo-records, "by.user" bootstraps by a defined array created by the user, and "none" does not use bootstrapping at all but rather uses each subject in the dataset once to build a tree.

- samptype—this selection determines the type of bootstrapping that takes place. The two choices are: (1) "swor" which stands for sampling without replacement, and (2) "swr" which stands for sampling with replacement. The default is to sample without replacement but our models each sample with replacement in, that is, using traditional bootstrapping.

- alpha—in conditional inference forests, the stopping criterion for splitting nodes is indicated by a nonparametric test's p-value (the parameter –

alpha) which has the null hypothesis that each split from the node have the same survival function; the node becomes terminal when the probability that the two branches off the node are the same is less than alpha. Increasing the model's parameter alpha decreases the threshold for splitting nodes (rejecting the null by being less than alpha) and thus increases the depth of the tree. Increasing and decreasing alpha, overfit and underfit the model respectively. The alphas considered in our tuning range from 0.05 to 0.2. The default value of alpha is 0.05.

2. Computational Issues and Parallelization

There are significant issues with fitting these ltrccif models on the PDE. The first issue identified is the considerable amount of time it takes to fit a ltrccif object; our final model takes over 350 computational hours to fit while our tuning models take an average of five hours to fit depending on the parameters. A major progression in transcending this problem is the use of parallelization. Leveraging the entirety of computational resources in the PDE's RStudio server, the eight physical cores in parallel drastically reduce the model fitting time. The bagged trees in the ensemble forest are distributed among the eight cores enabling the model to be fit at nearly eight times the speed. With the PDE's RStudio server being a shared resource with other researchers, another issue arises when fitting larger ltrccif models engages the majority of the server's memory. Future iterations of this research should request additional, appropriated resources in the RStudio server to avoid afflicting other researchers on the PDE's RStudio server.

3. Tuning the Model Parameters

To reduce computational time and limit memory usage on the RStudio server we use a small subset of our training data to tune our TV-RSF model. From our training data (2010 cohort) of 17,289 unique subjects we partition half of the pseudo-records (8,570 subjects) into a tuning training set and the other half of the pseudo-records (8,719 subjects) into a tuning test set; we call these datasets our "full-tuning training/validation datasets." We further reduce the size of our data for tuning by randomly sampling 1,000 subjects from each dataset of our full-tuning datasets; we call these datasets our "reduced-tuning

training/validation datasets." These two sets of tuning datasets are used for tuning all of our models.

After sub-setting our data to create training and validation datasets, we tune our parameters to minimize validation IBS. Included in the LTRCforests package (Yao et al. 2022b) we also utilize functions: (1) predictProb, and (2) sbrier_ltrc. The predictProb function constructs the estimated survival function curve for each subject in a selected dataset with our selected ltrccif model. The sbrier_ltrc function returns either a list of BSs evaluated at requested times or the IBS for a prediction. The necessary inputs for the sbrier_ltrc function are the predictProb object and a survival object of the LTRC pseudo-subject observations.

The process for tuning the TV-RSF model includes five series of tuning single parameters to narrow down the most effective combination taking into consideration dependencies among results of parameter settings between parameters. Due to the long run-time and memory restrictions on the PDE's RStudio server, the first four tuning series utilize the reduced-tuning datasets to train and validate the models. The last tuning series utilizes the full-tuning datasets for reasons discussed below.

a. First Tuning Series: ntree

With a larger than necessary ntree not decrementing the accuracy of the model, our goal in this initial tuning series is to find the minimum number of trees without reducing the effectiveness of the model. The number of trees in this ensemble method is linearly related to the run-time of the model fit; reducing the effective number of trees to utilize for future tuning models significantly decreases the overall time and computational resources to discover the final parameters for this model. We utilize the default value of alpha of 0.05 and set mtry to six which is the closest integer to the square root of the number of covariates.

We see in Figure 2 the training and validation IBS for these tuning models. Validation IBS has marginal decreasing reductions from 1 tree to 150 trees before a slight increase at 250 and 500 trees. Moving forward from the first tuning series we select ntree

to be 100. Even though the plot suggests that 50 trees could be enough, our memory and time limitations can afford up to 100 trees.

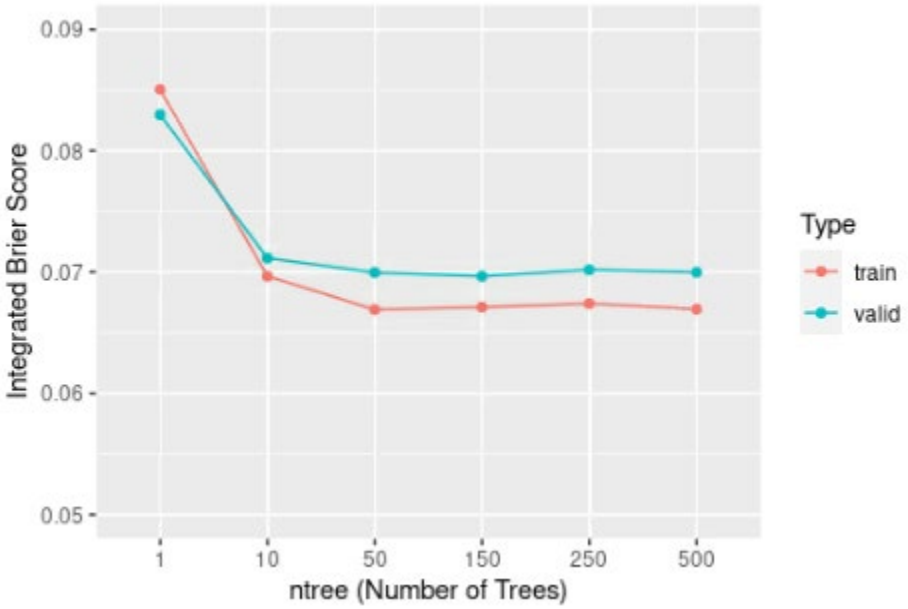


Figure 2. TV-RSF Integrated Brier Score vs. ntree (alpha = 0.05, mtry = 6).

b. Second Tuning Series: alpha

After determining the minimum number of trees needed for effective tuning, we compare different values of alpha—the depth of trees in the forest. As seen in Figure 3, we identify the IBS for alpha values of 0.05, 0.1, 0.15, and 0.2. We see that the validation IBS remains similar across the values while the training IBS decreases significantly from 0.05 to 0.2. This training and validation IBS discrepancy indicates potential for model overtraining. The lowest validation IBS out of these models was an alpha of 0.15 so we continued our tuning process with 0.15 as our alpha.

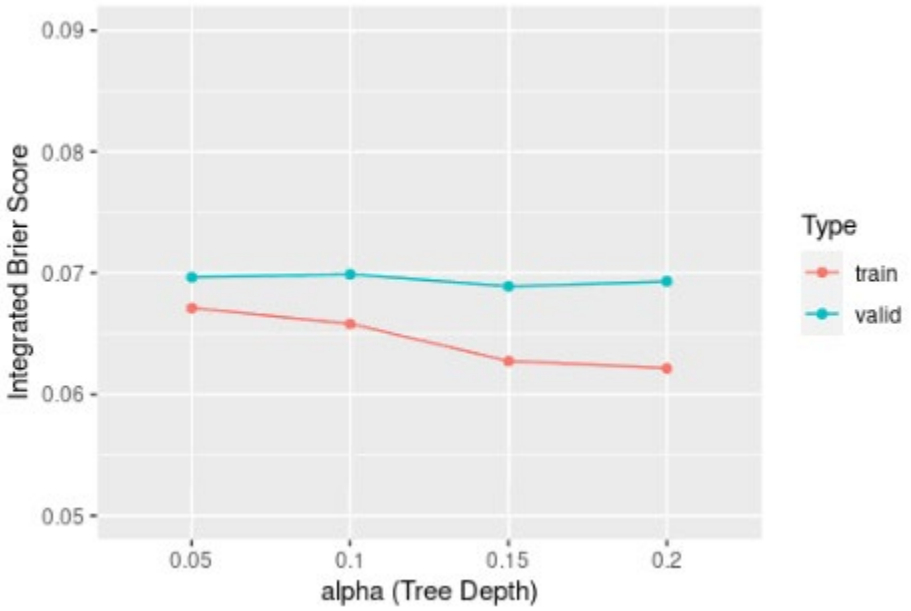


Figure 3. TV-RSF Integrated Brier Score vs. alpha (ntree = 100, mtry = 6).

c. Third Tuning Series: mtry

With the number of trees set to 100 and an alpha at 0.15 we next need to discover the best value for mtry—the size of the sample of covariates to be available to use at each node in the trees. Figure 4 shows our training and validation IBS for the seven values of mtry tested: 4, 6, 8, 10, 12, 14, and 16. We see than both training and validation IBS have decreasing marginal reductions with increases in mtry. While training IBS continues to decrease with each subsequent value through mtry = 16, the validation IBS begins to decrease from mtry values 12 to 14. This indicates that mtry values beyond 12 were over-trained to the test dataset which created too much variance in the model thus increasing the validation IBS. The lowest level of validation IBS was at an mtry of 12, therefore we continued our tuning with an mtry of 12.

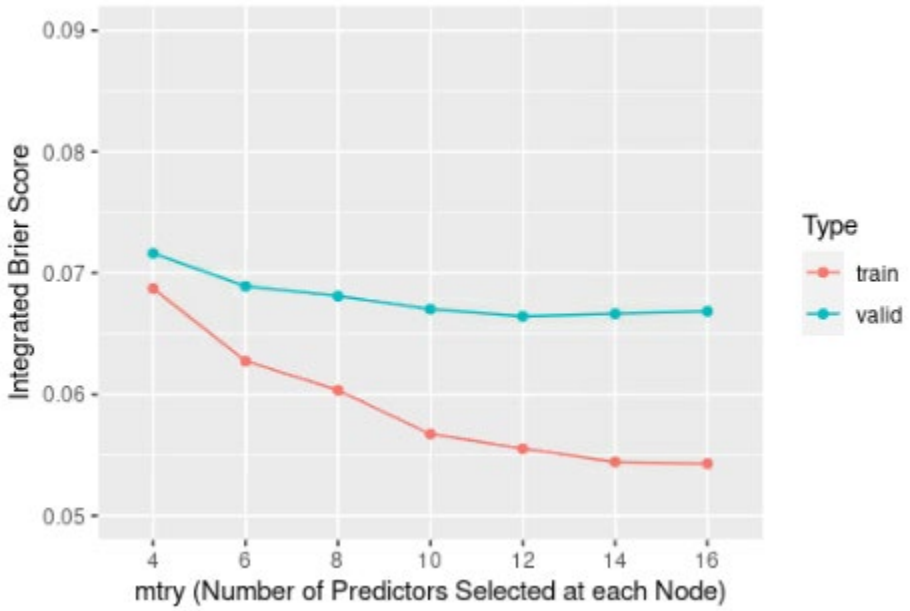


Figure 4. TV-RSF Integrated Brier Score vs. mtry (ntree = 100, alpha = 0.15).

d. Fourth Tuning Series: alpha With Updated mtry

With a new value for mtry, we re-tune our alpha parameter to ensure that we still had an effective value for our model; we wanted to ensure there was not a substantial change in the validation IBS values due to any potential interaction between alpha and mtry. Seen in Figure 5, the validation IBS does not significantly change between alpha values so we feel comfortable maintaining this value through to our final model parameters.

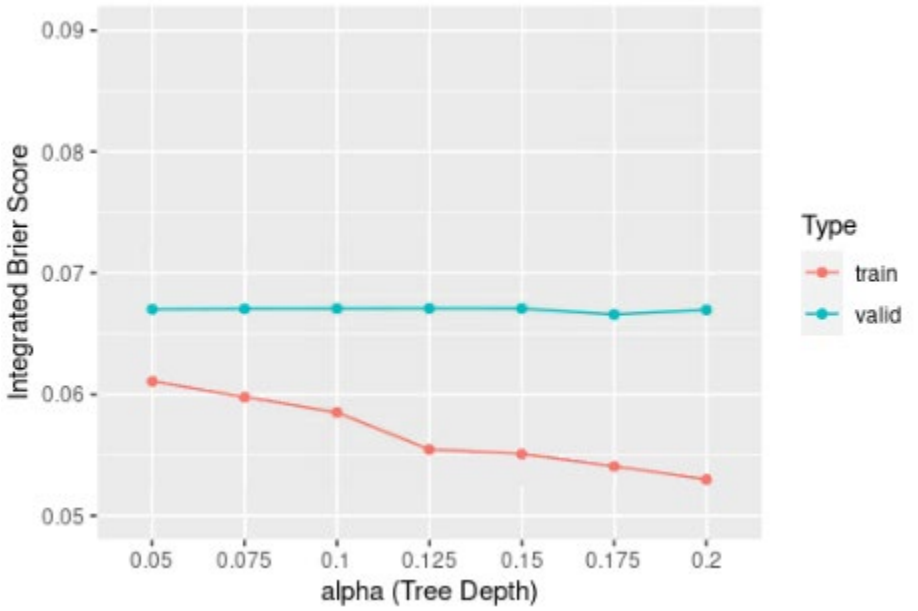


Figure 5. TV-RSF Integrated Brier Score vs. alpha (ntree = 100, mtry = 12).

e. Fifth Tuning Series: Validate alpha with Varying ntree

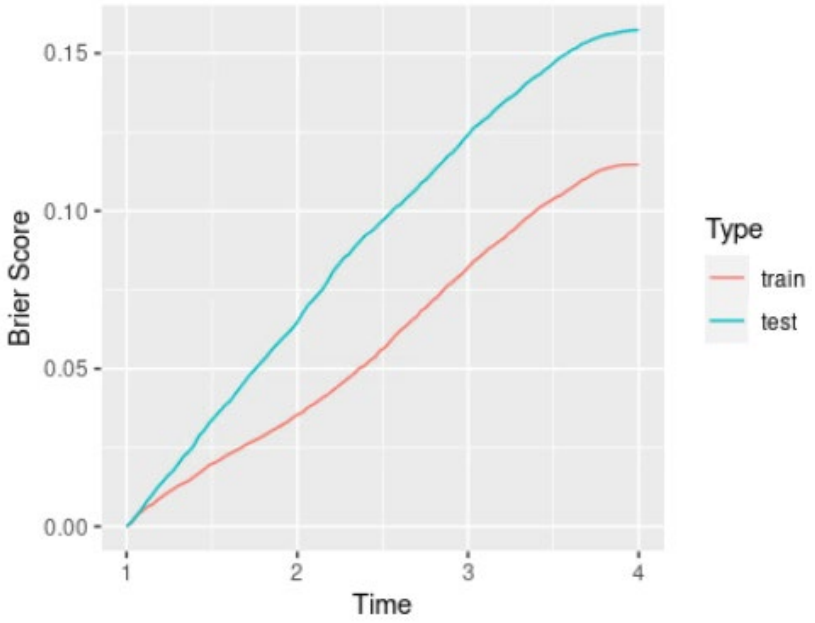
Our final tuning series is conducted out of concern for the number of trees limiting the effectiveness of a larger alpha value; if our trees are deeper/overfitting with a larger alpha we may be able to mitigate the variability in the model results by increasing the size of the ensemble (increase the number of trees). We fit four models, with all combinations of alpha = 0.05/0.2 and ntree = 100/400. If our concern is founded, there will be a significant decrease in validation IBS from model: ntree = 100, alpha = 0.05 to model ntree = 400, alpha = 0.2. For this tuning series we use the full-tuning datasets to train and validate the models to ensure that the bias from a smaller dataset is not the reason for any differences between the IBSs. Using the full-tuning training dataset increases the fitting time for 4–5 hours in the previous tuning models to over 35 hours for each of the four models fit. Table 8 demonstrates that our alpha value is not influenced by the number of trees in the forest and that our previous tuning parameters are correct. This tuning series leads us to confirm our final model parameters.

Table 8. Integrated Brier Scores for Interactions of alpha and ntree.

Test Integrated Brier Score	alpha = 0.05	alpha = 0.2
ntree = 100	0.0635	0.0579
ntree = 400	0.0633	0.0580
Validation Integrated Brier Score	**alpha = 0.05**	**alpha = 0.2**
ntree = 100	0.0702	0.0699
ntree = 400	0.0701	0.0697

4. Final Model

For our final model we find our parameters to be: mtry = 12, alpha = 0.15, and ntree = 250. The number of trees is a function of maximizing the amount of memory used without overflowing the system; fitting our final model takes over 350 computational hours, and nearly 45 hours of run-time. Previous attempts at fitting the model with greater numbers of trees crashes the system with the PDE's RStudio server's memory reaching its limit. Even after the model fit, additional errors arise when trying to predict with our final model; the predictProb function is extremely computationally demanding and the 2011-test data has to be partitioned into nine subsets in order to predict attrition in bite-sized chunks. After partitioned test-set predictions are used to calculate nine sets of BSs and IBSs, they are averaged with respect to the weights of the initial partitioned test sets. This average value of IBS and averaged BS values represent the final test and training IBS and BS values for this model; these values are seen in Figure 6.



The test IBS for this final model was 0.0934 and the training IBS was 0.05715. The training set is the FY2010 cohort, and the test set is the FY2011 cohort.

Figure 6. Final TV-RSF Brier Scores.

In Figure 6 we can see the large difference between the test and training BSs. This difference, greater than the difference between our tuning training and validation IBSs, indicates possible overfitting of the TV-RSF model to the 2010 data; with the test data being from a different year, some of the differences from the training set are not captured. We also see that there is a steady increase in BS as time approaches the end of the subjects' contract term of four years. This shows that with subjects attriting and leaving the cohort we become less accurate in our ability to estimate the remaining subjects' survival functions. With our tuning models' validation IBS nearing 0.065, we gain insight on the test set with a significantly higher IBS of 0.093; the 2011 cohort – test set – is not perfectly represented by the 2010 cohort – training set.

C. LEFT-TRUNCATED, RIGHT CENSURED SURVIVAL TREE WITH TIME-VARYING COVARIATES

1. Fitting the Model

In fitting the LTRC survival tree model, we utilize the same function ltrccif from the LTRCforests package (Yao et al. 2022b) but with some modifications. While there is a package specifically designed for LTRC survival trees, LTRCtrees (Fu, Simonoff, Wenbo 2021), the PDE does not support some of the necessary dependencies for use of the package. Using the same function as fitting the TV-RSF model also allows for use of the predictProb and sbrier_ltrc functions. Different from an RSF, an LTRC survival tree only has one tree and does not rely on bootstrapping—every subject is used in fitting a single tree. This results in the LTRC survival tree only using a single parameter: alpha—depth of the tree. Each other parameter in the function ltrccif is set to a constant value in order to modify the RSF into an LTRC survival tree.

- ntree—with the model no longer being a forest or an ensemble method, this parameter is set to one to only fit a single LTRC survival tree.
- mtry—with only one tree being fit, each of the covariates are available to split on at each node. This parameter is set to 33—the number of covariates in formula.
- bootstrap—this value is set to "none"; no bootstrapping takes place in an LTRC survival tree.
- samptype—we did not bootstrap samples to create this tree, this is set to "swor" or sampling without replacement. The entire training dataset is used to fit the single tree.
- sampfrac—this parameter is used when the samptype is set to "swor" and represents the proportion of subjects to be used for each tree. The default is 0.632 but we set this value to one to ensure each subject is drawn for the model fitting.

- alpha—the depth of the tree is the only parameter tuned for this LTRC survival tree model. The default value is 0.05.

2. Tuning the Model Parameters

With computational time no longer being a factor in the tuning process for this model, we use the full-tuning datasets; each tuning model needs less than one hour to fit. With only one parameter to tune, many models are fit to pinpoint the best value for alpha. We fit 10 models from alpha values 0.025 to 0.25 with a step of 0.025 between models' alpha parameters. Figure 7 shows the training and validation IBSs for these tuning models. The trends in Figure 7 are surprising in that an alpha value near 0.05 led to lower IBS; values of alpha greater than 0.05 demonstrated overfitting evident with the decrease in training IBS and increase in validation IBS. This variance caused by overfitting is attributed to the greater depth of the tree with unnecessary splits. We decide to select an alpha of 0.05 for our final model.

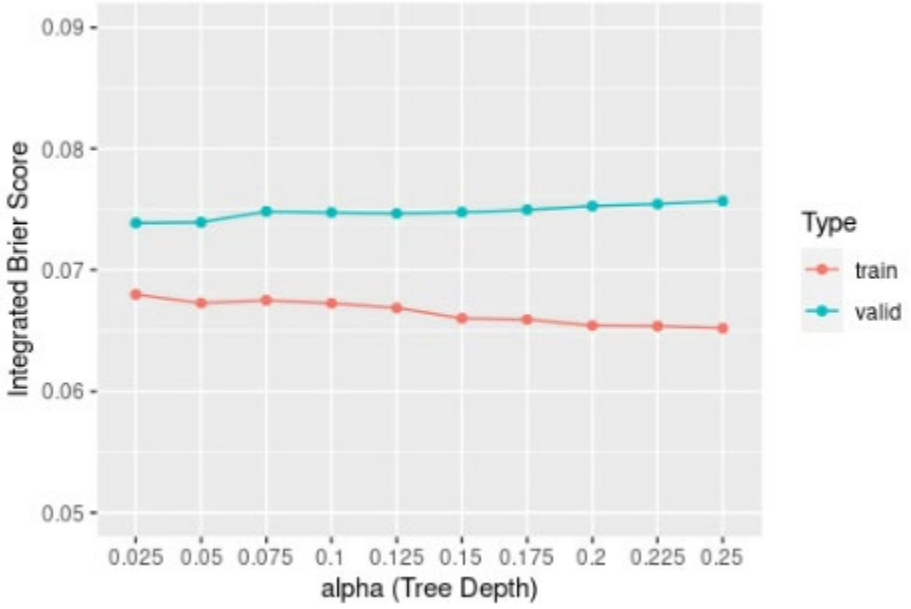
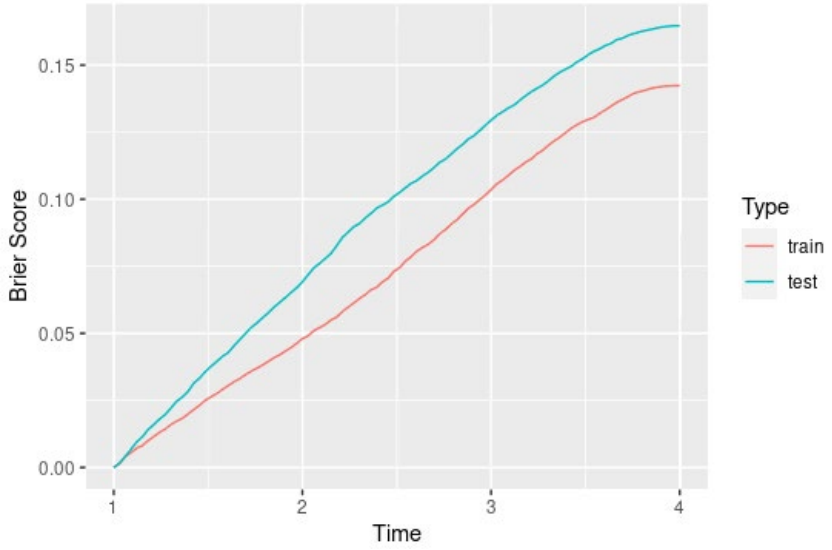


Figure 7. LTRC Survival Tree Integrated Brier Score vs. alpha.

3. Final Model

The final LTRC survival tree model has an alpha of 0.05. With the magnitude of the test set overloading the PDE's RStudio server when attempting to predict, we again utilize the method of creating nine subsets to predict with our final LTRC survival tree model. From these predictions—in predictProb object format—we calculate our BSs and IBSs for both our training and test sets. Figure 8 depicts the BSs as a function of time for both the training and test sets.



The test IBS for this final model was 0.0969 and the training IBS was 0.6877.

Figure 8. Final LTRC Survival Tree Brier Scores.

As with the TV-RSF, the magnitude between the training and test BSs is notable but the difference between the training and test IBS is not as large. We also see that as time increases, we are less accurate in estimating with our survival functions from the model.

D. RANDOM FORESTS

1. Fitting the Model

We fit our random forests models using the ranger function from the ranger package (Wright and Ziegler 2017). We tune three models fit with: (1) snapshot data of covariates

at year one, (2) snapshot data of covariates at year two with only subjects surviving past year two, and (3) snapshot data of covariates at year three with only subjects surviving past year three. The parameters for this random forest method which we tune with the ranger function (Wright and Ziegler 2017) are as follows:

- ntree—similar to TV-RSF, the number of trees is a parameter that will not decrease effectiveness of the model with being larger. We set our ntree value to 1,000 for tuning model fits and 10,000 for final model fits; these values are confirmed in the tuning process to be well within the magnitude to maximize accuracy.

- mtry—same parameter seen in both TV-RSF and LTRC survival tree tuning. We evaluate all 33 possible values of mtry for each of our three random forest models.

- important—this parameter in the ranger function determines the variable importance metric. We set this value to 'impurity' which utilizes Gini index for classification.

- probability—we set this value to 'true' which turns the model into probability forest as in Malley, Kruppa, Dasgupta, Malley, and Ziegler (2012). These forests produce probabilities for classification which we extract for our brier scores and cohort estimations.

- Exclusion of Unimportant Variables—after finding our best mtry parameter value utilizing each of the 33 covariates in our formula, the ten most important are selected to fit another model with the same mtry value for comparison.

2. Tuning the Model Parameters

All three models are tuned by first testing each of the 33 mtry values at a ntree value of 1,000. After the optimal mtry is found, 10,000 trees are used to fit the same model to

compare to 1,000 to ensure there is not a significant difference. The top-ten variables in terms of Gini index are then identified to rerun the model with only those variables.

a. Random Forest Model Tuning: Years One to Two

As seen in Figure 9, all 33 mtry values are tested and compared with BS. We see that the mtry which minimizes validation BS is mtry = 4. The difference between the best BS and worst BS for these models is very small compared to the tuning in previous modeling methods. We do not see evidence of overfitting with the training BS also increasing with the validation BS. After the models are tuned with 1,000 trees the mtry = 4 model is refit with 10,000 trees; no significant difference is noted in either validation or training BS, showing that 1,000 trees reasonable. We then find the top-ten most important variables using the Gini index and refit the model with an mtry of 4. This limited covariate model has a higher validation BS of 0.055 than the model including all of the variables with a validation BS of 0.054.

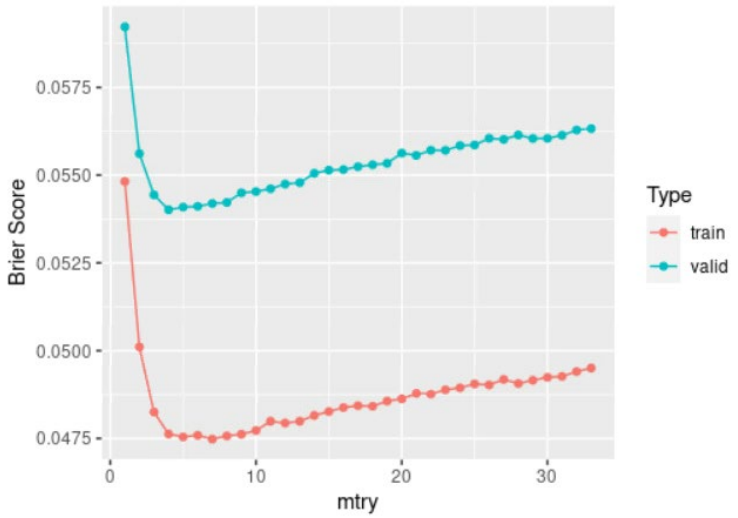


Figure 9. Years One to Two Random Forest mtry Tuning Brier Scores

b. Random Forest Model Tuning: Years Two to Three

We see in Figure 10 that the mtry which minimizes validation BS is mtry = 3. Like the years one to two model, the difference between the best BS and worst BS for these

models is very small compared to the tuning in previous modeling methods. We also do not see evidence of overfitting with any of the models. After the models are tuned with 1,000 trees the mtry = 3 model is refit with 10,000 trees; no significant difference is noted in either validation or training BS. The limited covariate model found with the top-ten most important variables from the mtry = 3 model has a worse validation BS of 0.077 compared to the model including all of the variables with a validation BS of 0.074.

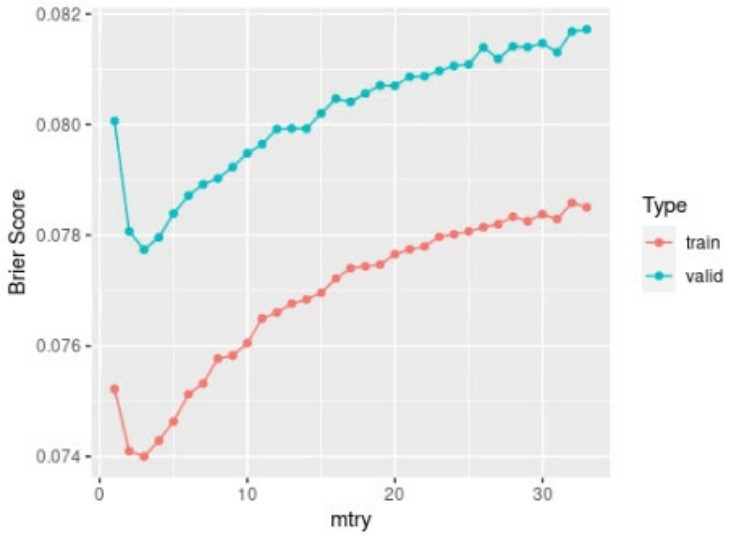


Figure 10. Years Two to Three Random Forest mtry Tuning Brier Scores

c. Random Forest Model Tuning: Years Three to Four

Figure 11 demonstrates that the mtry which minimizes validation BS is mtry = 4. We also do not see evidence of overfitting with any of the models. No significant difference is noted in either validation or training BS when comparing the 10,000-tree model to the 1,000-tree model. The limited covariate model found with the top-ten most important variables from the mtry = 4 model has a worse validation BS of 0.033 compared to the model including all of the variables with a validation BS of 0.031.

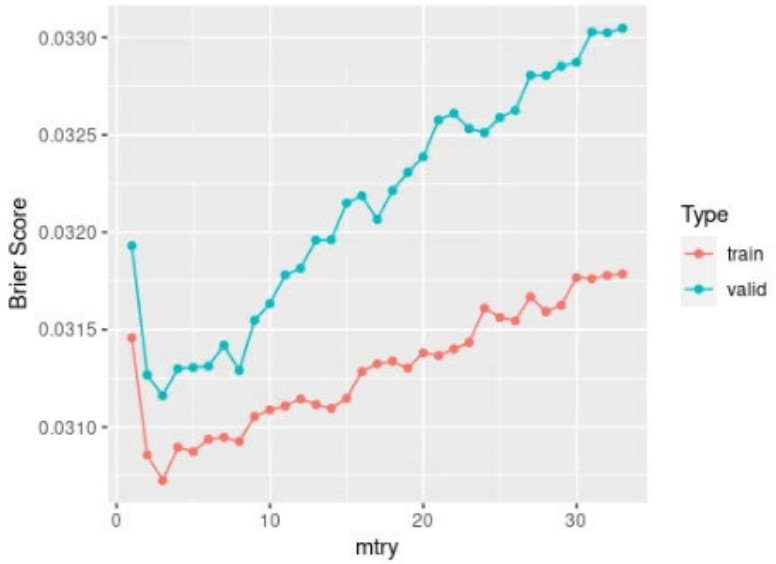


Figure 11. Years Three to Four Random Forest mtry Tuning Brier Scores

E. LOGISTIC REGRESSION

We fit our logistic regression model utilizing the glm function found in base R (R Core Team 2017). We create three models fit with: (1) snapshot data of covariates at year one, (2) snapshot data of covariates at year two with only subjects surviving past year two, and (3) snapshot data of covariates at year three with only subjects surviving past year three. With no tunable parameters, the logistic regression model is not tuned. The final model is trained with all of the 33 covariates and BS are calculated utilizing the same method as random forests.

IV. RESULTS

Figure 12 demonstrates our final model comparisons of BS at the three selected evaluation times of attriting by year two, three and four. We include BS evaluated on the FY2010 training set used to train the four models and on the FY2011 test set. We see that our TV-RSF model performs the best with a lower test BS for each of the three evaluation times. Following closely behind TV-RSF, the LTRC survival tree model performs marginally worse, with a difference of test BS of less than 0.03 for each selected evaluation time. We also see that the training BS for TV-RSF is markedly lower than the training BS for the LTRC survival tree. This is as expected since an ensemble (a random forest) typically outperforms a single model fit (a tree).

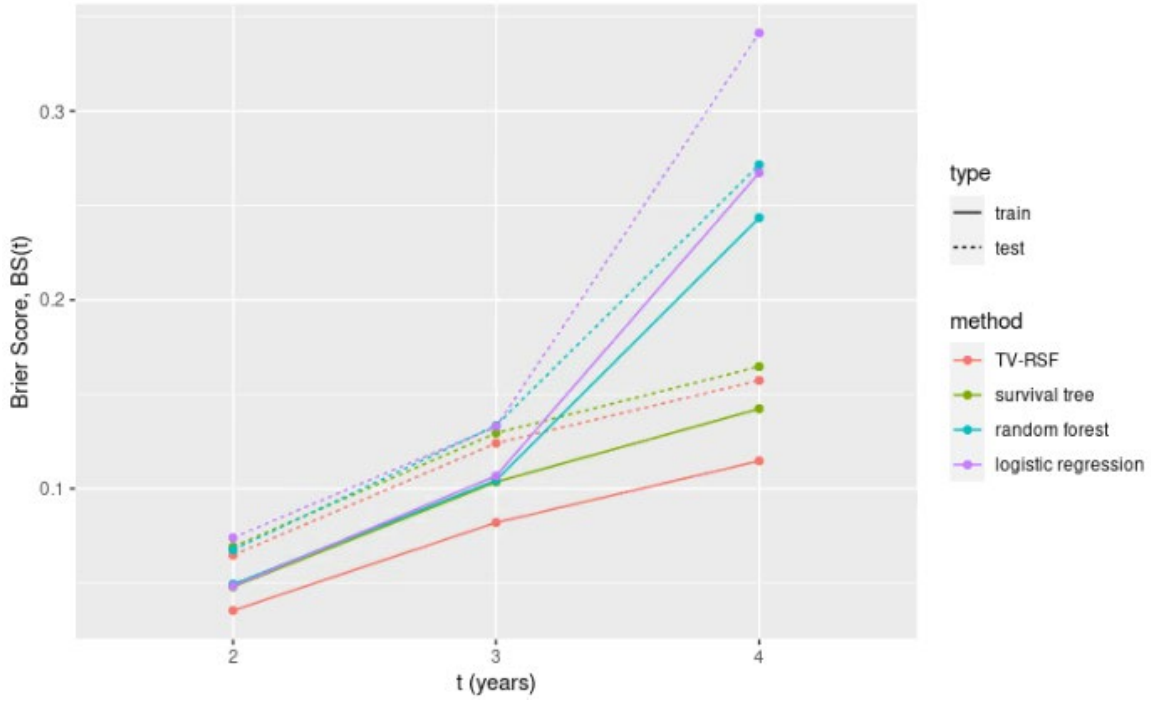


Figure 12. Final Model Training and Test Brier Score Comparisons

We suspect that the advantage TV-RSF has over an LTRC survival tree does not carry over to the FY2011 test set because the FY2011 policies, economic conditions, and recruits are fundamentally different for the FY2010 cohort. This is consistent with

observations of Speten (2018) and Devig (2019) who fit models with cohort year as a predictor and identify cohort year as one of the most important variables in predicting attrition. In a sense the more complex/flexible TV-RSF model is over-fitting the FY2010 data. It predicts FY2010 attrition so well that is does not generalize as well to a new year's cohort. This result suggests that using the much less computationally intensive LTRC survival tree might be nearly as good as using a TV-RSF when analysts expect non-stationarity in the next year. However, in the case that the next year is similar to the training year, a TV-RSF may be worth the extra computational burden. It also suggests that in the current environment where underlying policies, economic conditions, and recruites are changing dramatically that these attrition models must be updated regularly, perhaps even more frequently than annually.

Random forest outperforms both the LTRC survival tree and logistic regression models by a small amount in terms of test BS at time two, but cohort attrition estimation degrades at time three and four compared to survival forests with a significant spike in test BS at time four. In estimating cohort attrition, logistic regression performs the poorest at all three of the selected evaluation times. Similarly, the training BSs for years two and three are about the same for logistic regression, random forests, and the LTRC survival tree, but the training BS of year four is lowest for the LTRC survival tree followed by random forests and logistic regression. It seems that keeping close track of how T-VCs change with time (more frequently than annual changes) using survival analysis is more important in the later years of the first term than in early years.

Finally, we note that survival analysis with T-VC models (the LTRC survival tree or TV-RSF) provide the analyst with greater flexibility than trying to shoe-horn T-VCs into T-CC models like Cammack (2020) and Lazzarevich (2022). For example, say we have a group of first-term soldiers who are at the two-year mark in their contract. We know the values of their covariates at year two. Both the LTRC survival tree and TV-RSF models easily estimate their probabilities of surviving t for $t > 2$ given their covariate values at year two. Using the Cammack (2020) approach, to estimate the probability of surviving four years given survival to year two with covariate values at year two requires fitting a new classification attrition model to the training set. The new classification model would

be based on a snapshot of training records at year two with the response variable of attrition or retention at year four. Further, the survival analysis methods can be applied to estimate probabilities of surviving the next year for a group of soldiers, say in a particular unit, whose time in service so far differs and is measured in fractions of years rather than in one-year increments.

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY AND CONCLUSIONS

This chapter provides a summary of our work, conclusions and recommendations for future research.

A. SUMMMARY

We compare the performance of four methods for estimating attrition rates for a subset of Army post-IET first-term soldiers. Two methods, the LTRC survival tree and TV-RSF, accommodate T-VCs and estimate the entire survival function rather than an attrition rate for a single fixed time window. To our knowledge, our work is the first use of TV-RSF with military manpower data. Thus, to aid future military analysis, for TV-RSFs, we also give details concerning training, tuning parameter selection, and T-VC variable selection. The other two methods train a sequence of classification models (logistic regression and random forests) estimating a year's attrition rate given values of soldiers' covariates at the beginning of the year. These methods, particularly logistic regression, are the ones most often seen in military manpower analysis. We use BS computed on the FY2010 training cohort and on the FY2011 test cohort to compare the four methods.

B. CONCLUSIONS

Survival Analysis has benefits over traditional manpower methods. We find our survival analysis models (TV-RSF and LTRC survival trees) outperform the traditional manpower methods at predicting first-term post-IET attrition. This is due to the more effective capture of T-VCs in our survival analysis models than the method used by Cammack (2020) to only incorporate annual T-VC values for traditional classification models. The survival analysis methods also produce more useful results than the classification models with the estimated survival functions giving senior leaders insights on which groups of soldiers they can expect to attrite and attrition rates over the entire first term conditioned on what has been observed for a solder or group of soldiers so far.

For new-year data, the TV-RSF only marginally outperforms the LTRC survival tree. With first-term attrition costing the Army up to $652 million annually, the small

increase in prediction power with the TV-RSF may very well be worth the additional computational time (Marrone 2020). We see that differences in policy or economic conditions may have decreased the effectiveness of the TV-RSF in predicting first-term post-IET attrition in FY2011 data when trained with FY2010 data. This is indicated by how well the TV-RSF fits to the FY2010 data with a comparatively low training error to LTRC survival trees and the fact that cohort year is an important variable when included in previous models (Devig 2019). With the inclusion of these underlying policy or economic variables, the difference between new-year prediction accuracy could be greater between our two survival analysis methods.

We also note that the computational time required to train a survival model with T-VCs increases with the number of pseudo-records. Datasets with only T-CCs have one record per subject. The number of pseudo-records per subject increases, however, with the number of T-VCs, even when the number of subjects remains the same. Because training TV-RSFs is computationally expensive, pre-screening of T-VCs is important for training these models. As we illustrate, variables that include T-VCs many be pre-screened by training a sequence of traditional classification models. These can be trained rapidly. We use the sequence of random forests trained by Lazzaravich (2022) because random forest packages almost always provide measures of variable importance.

Finally, although we focus on Army first-term attrition, these methods are general. They may be used for attrition studies of civilians and for other services and other manpower applications such as estimating promotion, retention, or continuation rates.

C. FUTURE RESEARCH

Because survival analysis with T-VCs are rarely used in military manpower studies, there are many directions for future research; we list three.

Follow-on work for first-term post-IET attrition should incorporate variables like policy changes in the Army, and economic changes that influence the attrition rates of soldiers year over year. With the inclusion of these variables and a larger training set for TV-RSF to span multiple years to incorporate the effect of these variables, the effectiveness of the model in estimating cohort attrition rates for future years could improve.

Due to computational restraints in the PDE's RStudio server, only a subset of the available variables is used, the number of trees in the TV-RSF is reduced, and a smaller cohort with only one training year is selected. In future research, analysts should seek much larger computational resources to facilitate the demand of the TV-RSF model fitting with the inclusion of additional variables, larger cohort and number of trees in the ensemble. This research could produce improved insights for the Army's first-term attrition problem.

Finally, we use a "goodness of fit" statistic, BS, as a measure of performance based on full cohorts, and we have only alluded to how TV-RSF might be used in a military manpower setting. Future work should examine tangible military benefits of various applications of survival functions estimated from TV-RSFs. For example, how much improvement in end-strength forecasts do we get using the better fitting TV-RSF models? Is there a benefit to such forecasting at the unit level or for a particular MOS? How much added benefit is there to updating these forecasts in time as the T-VCs for a group of soldiers is updated? And, if we were to design a software application capitalizing on the strengths of TV-RSF, what should it look like?

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. CAREER MANAGEMENT FIELDS

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

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. HOME OF RECORD STATES/TERRITORIES

Levels	Level Description
AK	Alaska
AR+MS+AL+LA	Arkansas / Mississippi / Alabama / Louisiana
AZ+NM	Arizona / New Mexico
CA	California
CO	Colorado
CT+RI	Connecticut / Rhode Island
FL	Florida
GA+SC	Georgia / South Carolina
GU+PR+VI+AS	Guam / Puerto Rico / U.S. Virgin Islands / American Samoa
HI	Hawaii
IL	Illinois
IN+OH	Indiana / Ohio
MA	Massachusetts
MD+DC+DE	Maryland / District of Columbia / Delaware
ME	Maine
MI	Michigan
MN+WI	Minnesota / Wisconsin
NE+KS+IA+MO	Nebraska / Kansas / Iowa / Missouri
NH+VT	New Hampshire / Vermont
NJ	New Jersey
NY	New York
OR+WA	Oregon / Washington
PA	Pennsylvania
SD+ND	South Dakota / North Dakota
TN+KY+WV	Tennessee / Kentucky / West Virginia
TX+OK	Texas / Oklahoma
UT+NV	Utah / Nevada
VA+NC	Virginia / North Carolina
WY+ID+MT	Wyoming / Idaho / Montana

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Brier G (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review. 78 (1): 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

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.

Fu W, Simonoff J, (2017) Survival trees for left-truncated and right-censored data, with applications to time-varying covariate data. Biostatistics 18, 352–369 https://doi.org/10.1093/biostatistics/kwx047.

Fu W, Simonoff J, Wenbo J (2021) LTRCtrees: survival trees to fit left-truncated and right censored and interval-censored survival data. R package version 1.1.1, https://cran.r-project.org/web/packages/LTRCtrees/.

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.

Harrison D (2022) Secretary of the Army Christine Wormuth's remarks to the 2022 AUSA opening ceremony (October 10, 2022) (As Prepared). https://www.army.mil/article/260969/secretary_of_the_army_christine_wormuths_remarks_to_the_2022_ausa_opening_ceremony_october_10_2022as_prepared.

James G, Witten D, Hastie T, Tibshirani R (2021) *An Introduction to Statistical Learning: with Applications in R* (Spring Texts for Statistics, (New York, NY).

Lazzarevich N (2022) Predicting U.S. Army enlisted attrition after initial entry training using random survival forests. Master's thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA. http://hdl.handle.net/10945/69666.

Malley J, Kruppa J, Dasgupta A, Malley K, Ziegler A (2012) Probability machines: consistent probability estimation using nonparametric learning machines. Methods Inf Med 51, 74–81 https://doi.org/10.3414/ME00010052.

Marrone J (2020) Predicting 36-month attrition in the U.S. military: a comparison across service branches. RAND Corporation, https://www.rand.org/pubs/research_reports/RR4258.html. Also available in print form.

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/.

Ripley B, Lapsley M (2021) RODBC: ODBC database access. R package version 1.3-19, https://cran.r-project.org/web/packages/RODBC/.

RStudio Team (2020) RStudio: integrated development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Shen J, Bouée S, Aris E, Corrine E, Ekkehard B (2022) Long-term mortality and state financial support in invasive meningococcal sisease—real-world data analysis using the French National Claims Database (SNIIRAM). Infect Dis Ther 11, 249–262. https://doi.org/10.1007/s40121-021-00546-z.

South T (2019) Rising costs, dwindling recruit numbers, increasing demands may bring back the military draft. Military Times. Accessed August 21, 2022, https://www.militarytimes.com/news/your-military/2019/11/19/rising-costsdwindling-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.

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.

Yao W, Frydman H, Larocque D, Simonoff JS (2022b) LTRCforests: Ensemble methods for survival data with time-varying covariates. R package version 0.5.5, https://cran.r-project.org/web/packages/LTRCforests/.

Yao W, Frydman H, Larocque D, Simonoff JS (2022a) Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research,* https://doi.org/10.1177/09622802221111549.

Wright M, Ziegler A (2017) ranger: A fast implementation of random forests for high dimensional data in C++ and R. R package version 0.14.1, https://cran.r-project.org/web/packages/ranger/.

INITIAL DISTRIBUTION LIST

1. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California