NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

SURVIVAL ANALYSIS OF ARMY ENLISTED DEFENSE LANGUAGE INSTITUTE GRADUATE ATTRITION FACTORS

by

Philip J. Lukanich

March 2023

Thesis Advisor: Ruriko Yoshida
Second Reader: Candice Farney,
The Research Analysis Center-Monterey

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2023	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
SURVIVAL ANALYSIS OF ARMY ENLISTED DEFENSE LANGUAGE INSTITUTE GRADUATE ATTRITION FACTORS

5. FUNDING NUMBERS

6. AUTHOR(S) Philip J. Lukanich

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)

Defense Language Institute (DLI) is the Department of Defense's (DOD) multi-service language school and on average hosts 2,648 students annually, at a cost of $323K per student. In an increasingly challenging recruiting environment, failure to recognize influential DLI graduate attrition factors places a heavier burden on recruiting efforts and the graduate's follow-on command for retention. Utilizing the Person-Event Data Environment (PDE), this analysis looked at active duty enlisted Army service members who joined the military between January 1, 2010, and December 31, 2012. After cleaning and joining four different datasets, there were 1,469 unique records. Kaplan-Meier, Cox Proportional Hazard, and Random Survival Forest models were used to identify which factors most contributed to DLI graduate attrition, and at which point in a soldier's career they left the service. Though slightly different, all models showed consistent results. The most significant variables contributing to attrition were age, ethnicity, marital status, language difficulty, education level, and AFQT percentile. Over half (52.1%) of observations have left the Army, with 68.1% of this attrition taking place between 4 and 6.5 years of service. This analysis provides an improved understanding of when and why DLI graduate attrition occurs, supporting DOD decision makers in the development and adjustment of future policies focused on retention.

14. SUBJECT TERMS Defense Language Institute, DLI, Person-Event Data Environment, PDE	**15. NUMBER OF PAGES** 71
	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

SURVIVAL ANALYSIS OF ARMY ENLISTED DEFENSE LANGUAGE INSTITUTE GRADUATE ATTRITION FACTORS

Philip J. Lukanich
Lieutenant Commander, United States Navy
BS, United States Naval Academy, 2011
MBA, Western Governors University, 2019

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2023

Approved by: Ruriko Yoshida
Advisor

Candice Farney
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Defense Language Institute (DLI) is the Department of Defense's (DOD) multi-service language school and on average hosts 2,648 students annually, at a cost of $323K per student. In an increasingly challenging recruiting environment, failure to recognize influential DLI graduate attrition factors places a heavier burden on recruiting efforts and the graduate's follow-on command for retention. Utilizing the Person-Event Data Environment (PDE), this analysis looked at active duty enlisted Army service members, who joined the military between January 1, 2010, and December 31, 2012. After cleaning and joining four different datasets, there were 1,469 unique records. Kaplan-Meier, Cox Proportional Hazard, and Random Survival Forest models were used to identify which factors most contributed to DLI graduate attrition, and at which point in a soldier's career they left the service. Though slightly different, all models showed consistent results. The most significant variables contributing to attrition were age, ethnicity, marital status, language difficulty, education level, and AFQT percentile. Over half (52.1%) of observations have left the Army, with 68.1% of this attrition taking place between 4 and 6.5 years of service. This analysis provides an improved understanding of when and why DLI graduate attrition occurs, supporting DOD decision makers in the development and adjustment of future policies focused on retention.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AFMG Armed Forces Mental Group

AFQT Armed Forces Qualification Test

AIC Akaike Information Criterion

CHF Cumulative Hazard Function

CPH Cox Proportional Hazard

DOD Department of Defense

DEERS Defense Enrollment Eligibility Reporting System

DLAB Defense Language Aptitude Battery

DLIFLC Defense Language Institute Foreign Language Center

DLPT Defense Language Proficiency Test

DMDC Defense Manpower Data Center

FY fiscal year

HR Hazard Ratio

ILR Interagency Language Roundtable

KM Kaplan Meier Method

MEPCOM Military Entrance Processing Command

MOS Military Occupational Specialty

NPS Naval Postgraduate School

OOB Out-of-Bag

OR Operations Research

PDE Person-Event Data Environment

PID Personal Identifier

PH Proportional Hazards

RSF Random Survival Forest

ROC Receiver Operating Characteristic

SQL Structured Query Language

YOS Years of Service

Executive Summary

The Defense Language Institute Foreign Language Center (DLIFLC), located in Monterey, California, is responsible for training over 2,500 military and Department of Defense (DOD) annually. The culturally based foreign language training provided by the DLIFLC strives to give the DOD a competitive advantage in the joint operational environment and protect United States national security.

Language programs typically range from 26 to 64 weeks and on average, it costs $323,000 per student to attend the DLIFLC. However, some students are recycled, change languages, or need additional enhancement training, adding significant costs and time to their training pipeline. The DOD time and resources involved in training DLIFLC graduates are necessary and produce highly capable linguists but also make retention a high priority. In an increasingly difficult recruiting environment, it has become vital to understand when and why DLIFLC graduate attrition occurs.

This research aims to identify key factors associated with Army Enlisted DLIFLC graduate attrition over time. The dataset is comprised of enlisted Army linguists who joined the military between January 1, 2010 and December 31, 2012, and has 1,469 observations. By highlighting significant factors associated with attrition, the DOD and the DLIFLC can explore ways to improve DLIFLC graduate retention.

Survival analysis is commonly applied when evaluating time-to-event data and was an appropriate choice for analyzing DLIFLC graduate attrition. In this case, time-to-event is measured by the time from a service member's initial service entry date, until the time they separate from military service. Three separate survival models, Kaplan-Meier, Cox Proportional Hazard, and Random Survival Forest were used to identify and compare significant factors causing attrition, and at which point in a service member's career they were most likely to attrite. While long-term attrition was a focal point of this analysis, the 4-–6.5 Years of Service (YOS) time period was also of interest because this is when most first term enlisted contracts expire and a service member's first opportunity to voluntarily leave the military.

The overall Kaplan-Meier (KM) model had a DLIFLC graduate median survival time of 10.6 YOS. Understandably, attrition does not begin until around the 4 YOS mark because of initial training commitments. However, between 4—6.5 YOS survival probability decreases by 29.3%, indicating a significant amount of attrition is experienced during this time frame. The most significant variable influencing attrition in the KM model is age. DLIFLC graduates under 30 have a much higher probability of attrition than those 30 or older. The KM model also identified AFQT percentile and ethnicity as additional variables, which may impact attrition rates. Final survival probability for the KM model was 47.9%.

The Cox Proportional Hazard (CPH) model used had slightly better median survival time than KM of 12.1 YOS. Survival probability between 4-–6.5 YOS is slightly better as well, decreasing 23.6%. The benefit of the CPH model is that it takes into consideration other independent variables. As a result, the CPH model was able to highlight greater differences between variable factor levels and their influence on DLIFLC graduate attrition. Age, education level, language category, and marital status all had significant differences between factor level survival curves. The final survival probability for the CPH model was 49.6%.

The final model used was Random Survival Forest (RSF). This is an ensemble model, based on the Random Forest classification algorithm, and uses decision trees to estimate survival probabilities. This model estimated 1,469 survival curves and for comparison to KM and CPH models, an average of these estimates was taken. The final RSF survival probability was slightly lower at 46.3% than KM and CPH models. The RSF model also identified variables of importance, consistent with the previous two models. The five most important variables were age, followed by education level, rank, AFQT percentile, and ethnicity.

The fundamental goal of this research was to provide analysis and results that can assist the DOD in identifying factors influencing attrition and high attrition time periods in a DLIFLC graduate's career. These insights are provided as support tool in an effort to aid decision-makers in developing future policy and improving DLIFLC graduate retention.

Acknowledgments

I would first like to thank the Operations Research (OR) department faculty. I could not have completed this thesis without the foundational knowledge and guidance that was imparted to me throughout my time at Naval Postgraduate School (NPS). I am leaving NPS a better Naval Officer and I attribute this to the patience, dedication, and genuine passion the OR department shares in developing its students.

I would also like to thank Major Candice Farney. Candice, from the beginning, your help framing the problem and navigating the data procurement process was invaluable. I felt more confident in my analysis because I could always count on you to provide constructive feedback or a different perspective, always improving upon my existing work. I cannot thank you enough for your time and support.

Professor Yoshida, I am unbelievably grateful for your willingness to advise me on this thesis. The number of advisees that you guide is a testament to your passion and ability to teach and empower your students, I consider myself lucky to be one of them. As a student, I was amazed at the depth of your knowledge, and as your advisee I could not imagine conducting this research without being able to rely on your expertise and advice. Thank you for everything!

Finally, I would like to thank my wife, Joanna, and our daughters, Frances and Louise. None of this would be possible without your love, support, and encouragement. You serve as a constant reminder of how blessed I am, bringing me joy and laughter every day. I love you all very much!

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

This research aims to identify key factors associated with Army enlisted Defense Language Institute Foreign Language Center (DLIFLC) graduate attrition over time. By highlighting significant factors associated with attrition, the Department of Defense (DOD) and DLIFLC can explore ways to improve DLIFLC graduates retention.

1.1 Background

DLIFLC is multi-service school, located in Monterey, California, responsible for providing "culturally based foreign language education, training, evaluation, and degrees for the DOD globally. In an effort to afford a comprehensive understanding of the joint operational environment, a competitive edge to our warfighters, and safeguard the national security of the United States" (Defense Language Institute Foreign Language Center 2022, p. 7). The faculty trains more than 2,500 military personnel and select DOD employees, in over a dozen languages annually. As well as providing instruction and sustainment foreign language training for thousands more through its extension, distance-learning, and online programs.

DLIFLC strives to meet unique agency language requirements. Languages are grouped into four categories (I, II, III, or IV) based on difficulty. Prospective students must meet minimum Defense Language Aptitude Battery (DLAB) scores, which vary based on language category for program acceptance. Language programs range from 26 to 64 weeks, focusing on four functional skills: listening, speaking, reading, and writing.

Language proficiency is measured by the Defense Language Proficiency Test (DLPT). DLIFLC tests, listening comprehension (L), reading comprehension (R), and speaking proficiency (S). DLPT proficiency levels are, 0 (No Proficiency), 1 (Elementary), 2 (Limited Working), 3 (General Professional), 4 (Advanced Professional), 5 (Functional Native). The "+" designation indicates proficiency substantially exceeds on skill level but does not fully meet the criteria for the next level (Interagency Language Roundtable 2023).

In addition to satisfactory completion of DLIFLC requirements, students must achieve the

following DLPT scores in order to graduate and re-test annually thereafter: Basic programs L2/R2/S1+, Intermediate programs L2+/R2+/S2, and Advanced programs L3/R3/S2 (Defense Language Institute Foreign Language Center 2022).

1.2 Objective

On average, it costs $323,000 per student to attend DLIFLC. Some students are recycled, change languages, or need additional enhancement training, adding significant costs and time to their training pipeline. Young enlisted students on their initial service contract account for the largest share of DLIFLC students. Attrition is a significant loss in DOD time, resources, and money. In an increasingly difficult recruiting environment, it has become vital to understand why and when attrition occurs. This analysis focused on identifying significant factors contributing to attrition and at what point in a DLIFLC graduate's career they left the service. By highlighting these elements and better understanding DLIFLC graduate attrition, this research provides valuable insight to the DOD and aid decision makers in their retention efforts.

1.3 Structure

The following chapters provide the framework used to gather data, conduct analysis, obtain results, and address the research objective. Chapter 2 includes a literature review and analysis methods. Chapter 3 describes and familiarizes the reader with the dataset. Chapter 4 presents analysis, results, and interpretation. Chapter 5 discusses conclusions, recommendations, and proposed follow-up research.

CHAPTER 2:
Literature Review and Methods

This chapter provides a review of previous research conducted using survival analysis to analyze military attrition. While these studies are not focused on DLIFLC graduates, they do investigate important factors surrounding service member attrition. The majority of DLIFLC attrition studies has focused on military members currently enrolled in language study. Although it does not utilize survival analysis, Hinson (2005), explores individual success following DLIFLC graduation and provides meaningful insight. It is unknown if any studies utilize survival analysis to approach the DLIFLC graduate attrition problem.

2.1 Military Retention

As multiple service branches faced the possible reality of missing their fiscal year (FY) 22 recruiting goals, in July 2022, the Department of the Army released a memorandum addressing the current Armed Services recruiting challenges. Stating, "America's military faces the most challenging recruiting environment since the All-Volunteer Force was established in 1973, driven in part by the post-COVID labor market, intense competition with the private sector, and a declining number of young Americans interested in uniformed service. Currently, only 23% of 17- to 24-year-old Americans are fully qualified to serve" (Department of the Army 2022, p. 1). These challenges, coupled with the recent rise in minimum DLPT score requirements, make retention of the DOD's most qualified and proficient linguists even more crucial. Despite efforts such as Critical Skill Retention Bonuses, 35M (Army Human Intelligence) and 35P (Army Cryptologic Linguist) Military Occupational Specialty (MOS) codes are among those in high demand and considered critical jobs; often difficult to fill with new recruits due to the high Armed Forces Qualification Test (AFQT) and DLPT requirements. The reliance on recruiting high quality candidates, is no longer enough to fill the nation's need for trained linguists and efforts must instead be applied to retention.

2.2 Literature Review

Devig (2019) used survival analysis to investigate Army enlisted attrition following initial entry training. This study used the Person-Event Data Environment (PDE) to gather information on individual demographics, job assignments/deployment history, medical status, waivers, and initial service entry data. The data set was comprised of enlisted soldiers ranks of Private (E-1) to Staff Sergeant (E-6), who enlisted in the Army between FY 2005 and FY 2011. Survival analysis, specifically survival trees, were used to "develop a predictive survival model to forecast the probability a soldier will either leave the service through attrition within the first t years into their first term or will continue to serve in the Army past their initial first term obligation" (Devig 2019, p. xv). While the models developed did not estimate an individual's probability of attrition, they performed well at predicting the aggregate number of soldiers remaining in that Army at time $t > 0$.

Survival analysis was also used by Hawes (1990) to study U.S. Marine enlisted attrition. The author uses data on Marine enlisted accessions from October 1, 1983, to August 31, 1988. The study uses the Armed Forces Mental Group (AFMG), a score based on the AFQT, education level, and whether the individual received a moral waiver. These covariates were examined to determine any effects or association with premature attrition exist. While results were consistent with previous military attrition studies, Hawes did note two significant findings. First, attrition in "alternate high school credential holders varied significantly according to credential type. Second, the relationship between aptitude and attrition appeared to have weakened in recent years" (Hawes 1990, p. iii).

Rubiano and Enrique (1993) applied survival analysis techniques in the study of U.S. Coast Guard enlisted attrition. Using data from FY 1983 to FY 1990, this study investigated U.S. Coast Guard attrition using the following predictor variables: sex, marital status, race, pay-grade, and rating. Their goal was to develop a more accurate model, in order to predict monthly attrition. Rubiano and Enrique concluded males and married individuals had a higher survival probabilities. Additionally, enlisted ranks of E-6 to E-9 had lower probabilities of attrition than E-1 to E-5, and Asians had the highest probability of survival. A regression model was then used to a provide monthly attrition forecast. Also noted were high rates of attrition observed at four years of service and 20 years of service.

Hinson (2005) utilized DLIFLC data from 1997–2000 combined with Defense Manpower

Data Center data. "This study used classification trees and logistic regression to investigate how military, academic and personal characteristics influence first-term success after successfully completing DLIFLC. The author defined success as completing a first term enlistment contract and maintenance of language proficiency. Motivation came from DLIFLC management's interest in the difference in success for individuals that graduated DLIFLC via the different training pipelines. These different training paths included completing the program as originally assigned, recycling, relanguaging, or taking DLPT enhancement training. 63% of students graduated and only 45% of those that graduated were successful post-DLIFLC. The analysis identified service affiliation, contract lengths, citizenship, AFQT scores and gender were common factors in predicting success. Individuals in the Army had the worst odds of success. Males had higher odds of success than females. Contract lengths were very influential in determining success. Individuals with contracts greater than four years had lower odds of success" (Hinson 2005, p. v).

2.3 Methods

"Survival analysis is collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs" (Kleinbaum and Klein 2005). In many cases, the outcome measured is failure or death, resulting in survival analysis being used heavily in medical studies. However, an event can also refer to other outcomes of interest with a known time of occurrence, such as military separation.

2.3.1 Definitions

There are several commonly used definitions in survival analysis but can vary slightly in terms of how or what they reference. Following each general definition below, is a description unique to this analysis.

Exposure: The incident that starts the clock for the variable of interest. Exposure in this research refers to a DLIFLC graduate's initial service entry date.
Time: The number of units being measure (i.e., years, months, or weeks) from exposure until an event occurs, usually referred to as survival time. Time in this analysis is measured in Years of Service (YOS).
Event: The incident that stops the clock for the variable of interest. Service separation date

is the event that stops a DLIFLC graduate's clock.
Censoring: Happens when some information about an individual's survival time is known, but the exact time-to-event is unknown. There are three types:

- **Left-Censored**: The time-to-event is less than the observation time;
- **Interval-Censored**: The time-to-event is bracketed by a time interval;
- **Right-Censored**: The time-to-event is greater than the observation time.

Survival Function: The survival function, $S(t)$, is the probability that an observation's survival time, T, is greater than a specified time t (Kleinbaum and Klein 2005). The survival function will be used to plot survival curves over time.

$$S(t) = Pr(T > t)$$

Hazard Function: "The hazard function, *h(t)*, gives the instantaneous potential for an event to occur given survival time up to time t. In contrast to the survivor function, which focuses on not failing, *h(t)*, focuses on failing (i.e., the higher the average hazard, the lower probability of survival)" (Kleinbaum and Klein 2005).

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

2.3.2 Models Used

Kaplan Meier Method (KM): Is a nonparametric method that estimates a survival function with censored data. The dataset in this study is right censored because some observations are still currently on Active Duty and do not experience an event. For this reason KM is an appropriate method for evaluating probability of failure, in this case probability of attrition. Using the following information, we can calculate the KM curve.

t_j: ordered service separation times since a fixed time t_0;
n_j: number of DLIFLC graduates still remaining at time t_j;
m_j: number of attrites between time t_{j-1} and t_j;
$R(t_j)$: Risk set, DLIFLC graduates who have survived (remain on Active Duty) at least to

time t_j.

The general KM formula for calculating the probability of failure at time t_j is

$$S(t_j) = S(t_{j-1}) * Pr(T > (t_j)|T \geq (t_{j-1})),$$

where

$$Pr(T > (t_j)|T \geq (t_{j-1})) = \frac{n_j - m_j}{n_j}$$

for $n_j \neq 0$. The curve begins with the entire population and shows the percentage surviving over time. Horizontal lines represent time periods with no events occurring. For smaller datasets, the KM curve may resemble a step function, while larger datasets will appear continuous.

Log-Rank Test: Determines equivalence or not of two or more KM curves, the statistical criteria measured allows for overall comparison between the KM curves (Kleinbaum and Klein 2005). Additionally, the log-rank statistic can be computed without calculating variance as an approximation using as a "classic chi-square that sums over each group being compared, the square of the observed (O_i) minus expected value (E_i), divided by the expected value" (Kleinbaum and Klein 2005).

$$\chi^2 \approx \sum_{i}^{\text{\# of groups}} \frac{(O_i - E_i)^2}{E_i}$$

If the number of groups is ≥ 2, the log-rank statistic involves variances and covariances of O_i-E_i but has roughly a large sample chi-square distribution with, # of groups -1 degrees of freedom (Kleinbaum and Klein 2005).

Cox Proportional Hazard (CPH): Is a semiparametric, regression-based model that predicts the hazard rate for an observation as a function of both time and the observations corresponding predictor values. A key reason for its popularity is that, "even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves can be obtained for a wide variety of data situations" (Kleinbaum and Klein 2005). The model is usually written in terms of the

hazard model formula but can also be converted to a corresponding survival function, which provides a basis for determining survival curves. Unlike KM survival curves, CPH survival curves are adjusted for covariates. The formulas are

$$\text{Hazard Function Formula} : h(t,\boldsymbol{X}) = h_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i\right)$$

and

$$\text{Survival Function Formula} : S(t,\boldsymbol{X}) = S_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i\right),$$

where

$h_0(t)$: is the baseline hazard;
$S_0(t)$: is the baseline survival function;
X_i : predictor variables (X_1,X_2,$X_3 \ldots X_p$); and
β_i : predictor variable coefficient, estimated by model's algorithm.

This formula has an underlying assumption of time independence. This means that the predictor variables do not depend on the Hazard Function of time t and vice versa.

CPH model outputs are similar to regression model outputs, including regression coefficients, standard errors, and p-values, from significance tests in which the null hypothesis $H_0 : \beta_i = 0$ is tested against the alternative hypothesis $H_1 : \beta_i \neq 0$ where β_i is the coefficient of the explanatory variable X_i. An additional output is the hazard ratio, which indicates the variables' effect on the hazard rate, after being adjusted for other variables in the model (Kleinbaum and Klein 2005).

Random Survival Forest (RSF): Is an extension of the Random Forest ensemble methods used for classification and regression analysis, capable of dealing with time-to event and censored data. Random forest model is a nonparametric ensemble model which can be applied to regression. For details on the process of random forest model, see James et al. (2013). A Cumulative Hazard Function (CHF) is then calculated for each tree in the random forest model and averaged to obtain the ensemble CHF (Ishwaran et al. 2008). The CHF estimate for each tree node p at time t, is estimated by is the Nelson-Aalen estimator

$$H_p(t) = \sum_{t \leq T} \frac{d_t}{Y_t},$$

where, d_t and Y_t are the number of people who attrite and the total number of individuals at risk of leaving the Army at time t (Nelson 1969).

The ensemble CHF is calculated, as the average over all CHFs from all L decision trees. Allowing for the cumulative hazard prediction of a new observation x,

$$H(t|x) = \frac{1}{L} \sum_{i=1}^{L} H_i(t|x),$$

where, $H_i(t|x)$ denotes the CHF of the tree grown from the i-th bootstrap sample.

The predictive performance of a RSF model is most commonly evaluated using the concordance index (C-index) for survival analysis, also known as "Harrell's C" (Harrell et al., 1982; Ishwaran et al., 2008). "The C-index estimates the probability that in a randomly selected pair of cases, the case that fails had a worse predicted outcome. Unlike other measures of survival performance, the C-index does not depend on a single fixed time for evaluation and specifically accounts for censoring" (Ishwaran et al. 2008). The C-Index can be interpreted in a similar manner as a Receiver Operating Characteristic (ROC) curve. For example, a $C = .5$, signifies the prediction is no better than random chance, while $C = 1$, indicates the model perfectly predicts an outcome.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:
Data Overview

This chapter provides insight into the data procurement and cleaning process required to generate the dataset used for analysis. Developed by the Army Analytics Group, this research utilized the Person-Event Data Environment (PDE), to request and access service member data. The PDE safely stores service member data and allows users to perform queries and analysis, through a remote desktop equipped with several statistical tools. For this research, we used Toad for Oracle to conduct Structured Query Language (SQL) queries across databases and R for cleaning and analysis.

3.1 Data Procurement and Cleaning

This analysis focused on active duty enlisted Army, DLIFLC graduates, entering service between January 1, 2010, and December 31, 2012. This time frame was chosen to ensure completion of initial contract service obligations. After receiving authorization and creating a project within the PDE, initial data requests were submitted through PDE's database catalog. The required data was sourced from Defense Manpower Data Center (DMDC), Defense Enrollment Eligibility Reporting System (DEERS), and Military Entrance Processing Command (MEPCOM) catalogs. Data requests included; Army Active Duty Military Personnel, Army Workforce Transaction Language files, Contingency Tracking System—Overseas Contingency Operations, and Active Duty Military Personnel Transaction files. Once the data was procured through the PDE, a Personal Identifier (PID), unique to the requested data, was assigned to each service member record. This is a standard PDE procedure performed to further protect service member privacy and was used to join multiple different datasets.

We used Toad for Oracle to perform SQL queries on necessary data files. The Army Active Duty Military Personnel file was filtered by Army linguist MOS (35M = Army Human Intelligence Collector, 35P = Army Cryptologic Linguist, and 35X = Intelligence Senior Sergeant/Chief Intelligence Sergeant) and their rank. Using the unique PID for all Army linguists, the next query was performed on the MEPCOM data file, providing

gender, age, education level, marital status, AFQT percentile, ethnicity, and initial service entry dates. Subsequent searches, using the unique PID, were performed on the Army Workforce Transaction Language file, filtering DLIFLC graduates, language studied, DLPT proficiency in speaking, listening, and reading. Deployment history was pulled from the Contingency Tracking System Overseas Contingency Operation file and the Active Duty Military Transaction file provided a service separation date. Ultimately, four Excel files were created through this process, later being cleaned and joined in R. After joining, we were left with 1,469 unique records of Active Duty Army service members, who graduated DLIFLC between January 1, 2010, and December 31, 2012. The merged dataset was relatively clean, only missing four Education Level observations, which were imputed with the median Education Level.

Table 3.1 provides a list of all dataset variables, description, and type. Deployment history was created as a binary variable, identifying a deployment with 1 and no deployments with 0. Two additional variables were created for analysis purposes, YOS and Status. YOS was calculated using the difference between Service Entry Date and Separation Date. Status is a logical variable, indicating a service member is still active duty. The majority of variables were transformed into categorical variables for analysis.

Table 3.1. Variable types in the dataset. n is the sample size of the data.

Variable	Description	Type	Factor Levels
ID	Unique ID	Character	n
SVC_ENTRY	Service Entry Date	Date	NA
SVC_SEPARATION	Separation Date	Date	NA
GENDER	Gender	Categorical	2
AGE	Age	Numeric	NA
RANK	Rank	Categorical	2
MRTL_STAT	Marital Status	Categorical	4
ETH	Ethnicity	Categorical	5
EDU_LVL	Education Level	Categorical	4
AFQT_PCTL	AFQT Percentile	Numeric	NA
LANG	Language Category	Categorical	4
DLPT_SPEAK	DLPT Speaking Score	Categorical	4
DLPT_LISTEN	DLPT Listening Score	Categorical	6
DLPT_READ	DLPT Reading Score	Categorical	6
DEPLOYED	Deployed or Not	Binary	2
YOS	Years of Service	Numeric	NA
STATUS	In Service or Not	Logical	2

3.2 Explanatory Variables and Descriptive Statistics

The original PDE data was coded and needed to be transformed into an interpretable format, using the PDE data dictionary. The DLPT scores were the service member's most recently recorded DLPT. Most predictor variables were condensed into a more manageable number of factor levels. Table 3.2 provides a summary of the categorical predictor variables used and their factor levels, followed by a description of how factor levels were created and statistics for each variable.

Table 3.2. Description of Categorical Predictors.

Variable	Factor Levels
GENDER	F - Female, M - Male
RANK	Junior Enlisted (E4-Below), Junior NCO (E5-E6)
MRTL_STAT	Married, Not Married, Divorced, Other
ETH	Asian, Hispanic, Native or Pacific Island, Other, None
EDU_LVL	HS Equiv, Some College, College Ed, Graduate Ed
LANG	CAT I, CAT II, CAT III, CAT IV
DLPT_SPEAK	1/1+, 2/2+, 3/3+, Unknown
DLPT_LISTEN	0, 1/1+, 2/2+, 3/3+, 4, Unknown
DLPT_READ	0, 1/1+, 2/2+, 3/3+, 4, Unknown

RANK: Ranks ranged from E01 to E06 and were factored into E4 and below (Junior_Enlisted) and E05-E06 (Junior_NCO). Junior_Enlisted ranks accounted for 73.2% of observations, while Junior_NCO made up 26.8%.

MRTL_STAT: Marital status originally encompassed married, not married, divorced, annulled, and legally separated. Married graduates made up the largest portion of observations, accounting for 49.4%. The remainder of the dataset is comprised of Not_Married = 45.4%, Divorced = 5%. Annulled and legally separate had the fewest observations and were grouped together into the level Other, accounting for 0.2% of observations.

ETH: There were 18 different ethnic categories. The data dictionary was used to create factor levels. Asian, includes Asian Indian, Chines, Filipino, Japanese, Korean, Vietnamese, and other Asian descent. Hispanic, includes Mexican, Puerto Rican, Cuban, Latin with Hispanic descent, and other Hispanic descent. Native or Pacific Island, includes Aleut, US or Canadian Indian tribes, Micronesian, and Polynesian. Other and None were kept the same as the original data. The dataset composition for each level was Asian = 7.8%, Hispanic = 8.9%, Native or Pacific_Island = 0.7%, Other = 75%, and None = 7.6%.

EDU_LVL: There were 15 different education levels. HS_Equiv includes Secondary School Credential, Test Based Equivalency, Occupational Program Certificate, Home Study Diploma, Adult Education Diploma, ARNG Challenge Program, Other Non-traditional

High School Credential, and High School Diploma. Some_College includes Completed 1 semester of college or Associate Degree. College_Ed, includes Baccalaureate Degree. Graduate_Ed, includes Master's Degree, Post Master's Degree, First Professional Degree, Doctorate Degree. The composition of each education level was HS_Equiv = 56.4%, Some_College = 16.7%, College_Ed = 24.8%, and Graduate_Ed = 2.1%.

LANG: There were 24 different languages. These languages were grouped according to DLIFLC's category criteria. CAT I, includes French and Spanish. CAT II, includes Indonesian. CAT III, includes Serbian-Croatian, Hebrew, Armenian, Persian (Iranian and Afghan dialects), Russian, Tagalog, and Urdu. CAT IV, includes Arabic (Sudanese, Modern, Algerian, Egypt, Gulf, Iraqi, Syrian, Peninsula, and Yemeni dialects), Chinese, Japanese, Korean, and Pushtu. CAT IV and CAT III languages made up the bulk of observations, accounting for 60% and 27%, respectively. CAT I languages accounted for 11.6% and CAT II languages 1.4%.

DLPT_SPEAK,**DLPT_LISTEN**,**DLPT_READ**: Figure 3.1 shows the distribution of DLPT scores. Raw scores were converted to Interagency Language Roundtable (ILR) proficiency levels. Some service members did not have data for their last DLPT and was recorded as Unknown.

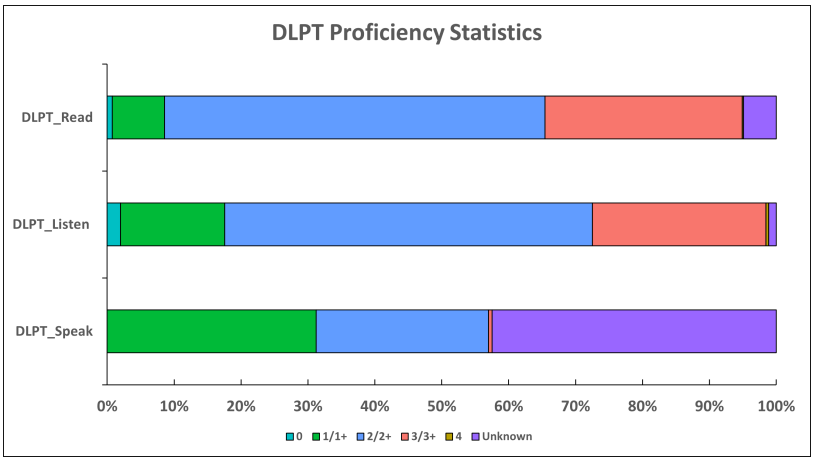

Figure 3.1. DLPT test scores are indicated as 0 = cyan, 1/1+ = green, 2/2+ = blue, 3/3+ = red, 4 = gold, and Unknown = purple. DLIFLC DLPT_Speak scores were 1/1+ (31.2%), 2/2+ (25.7%), 3/3+ (0.5%), and Unknown (42.4%). DLPT_Listen scores were 0 (2%), 1/1+ (15.6%), 2/2+ (55%), 3/3+ (26%, 4 (0.4%), and Unknown (1.2%). DLPT_Read scores were 0 (0.7%), 1/1+ (7.8%), 2/2+ (56.8%), 3/3+ (29.5%), 4 (0.2%), and Unknown (5%).

Three additional, non-categorical variables were also used in the models, AGE, AFQT, and DEPLOYED. These variables were left as numerical when fitting the overall KM, CPH, and RSF models. However, in order to provide a better visualization, they were factored when producing survival curves in the following ways.

AGE: Ages ranged from 20 to 53 and were factored into levels containing service members in their Twenties, Thirties, and Forties_Above. The average age of observations was 31 and the composition by factor level was Twenties = 38.9%, Thirties = 55.6%, and Forties_Above = 5.5%.

AFQT_PCTL: The AFQT percentiles ranged from 23 to 99. These were factored into levels of Below_70, 70–79, 80–89, and 90_Above. 90_Above made up the largest portion of observations with 63%, followed by 80–90 (22.1%), 70–79 (8%), and Below_70 (6.9%).

DEPLOYED: Deployment = 1 and No Deployments = 0. 40.3% of observations had at least one deployment, while 59.7% had no recorded deployments.

CHAPTER 4:
Analysis and Results

This chapter presents analysis and interpretation of the results discussed in Chapter 2. R's `survival` package (Therneau 2021) was used to fit KM and CPH models and the `survminer` package (Kassambara et al. 2021) to produce survival plots. The RSF model was fit using the `ranger` package (Wright and Ziegler 2017). The best CPH model was determined using the `MASS` package's `stepAIC` function (Venables and Ripley 2002). This method determines the best fit model by comparing different possible models, and selecting the one that explains the greatest amount of variation, using the fewest variables. The CPH model identified the variables AGE, MRTL_STAT, ETH, EDU_LVL, and LANG. While the KM model used all predictor variables, for continuity purposes this chapter presents interpretation for the corresponding CPH model variables. Additional KM survival curves for the remaining variables can be found in the Appendix. In addition to median survival times, factor level survival times, and statistically significant differences, special attention was given to survival probabilities between 4–6.5 YOS, due to the fact that most initial service entry contracts expire during this time frame.

4.1 Kaplan-Meier Model Results

Figure 4.1 is the overall KM model. The survival curve shown in red is the probability of an observation experiencing an event, in this case an event is considered attrition. Since this dataset is comprised of DLIFLC graduates, who spent the early part of their careers in basic training and language school, it makes sense we see very little attrition until 4 YOS. However, between 4–6.5 YOS, there is a 29.3% drop in survival probability and the final probability of survival for DLIFLC graduates is 47.9%. The median survival time for graduates was 10.6 years, indicated by the dashed line. Of those in the dataset that did attrite, 68.1% were between 4–6.5 YOS.

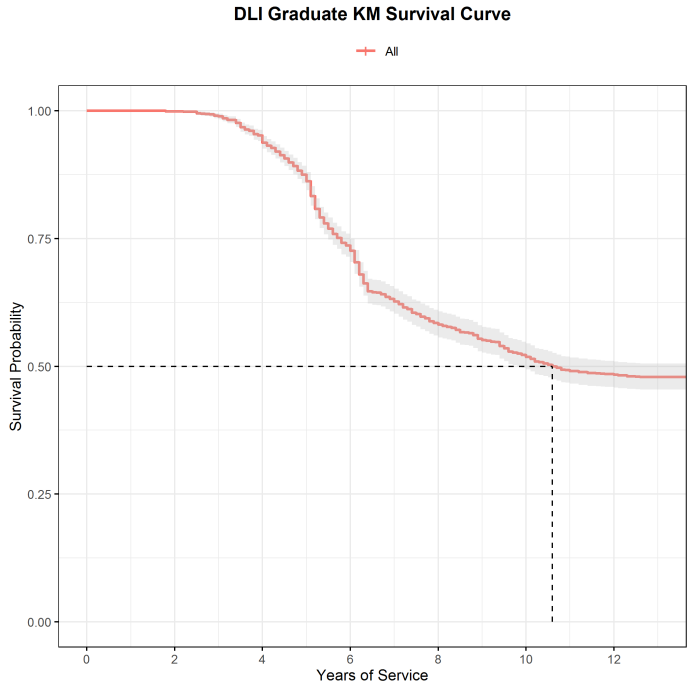

Figure 4.1. The red curve signifies DLIFLC graduate survival probabilities, with a confidence interval shown in gray. The dashed line indicates the median survival time, which occurs at 10.6 YOS. At 4 YOS the survival probability is 93.8% and drops to 64.5% by 6.5 YOS, a decrease of 29.3% in 2.5 years. After 6.5 YOS survival probability continues to decrease at a slightly lower rate, leveling out at 47.9%. Altogether, 68.1% of events (attrition) occur between 4–6.5 YOS.

4.1.1 KM Survival Curves by Age

Figure 4.2 shows survival times by age group. The key takeaway from looking at DLIFLC graduates by age, is that if the Army can retain graduates through their twenties, there is a significantly higher probability of long-term retention. A large difference can be witnessed between observations in their twenties compared to those thirty and older. The survival probability for observations in their twenties decreased 54.1% between 4–6.5 YOS and had a median survival probability of 5.6 YOS, indicated by the dashed line. The final survival probability for DLIFLC graduates in their twenties was 17.2%. Observations in their thirties experienced a 13.7% drop between 4–6.5 YOS and those forty or older decreased 12.4% during this time frame. Graduates thirty and over do not reach the median threshold of 50%, having final survival probabilities of 67.7% (thirties) and 65.4% (forties or above).

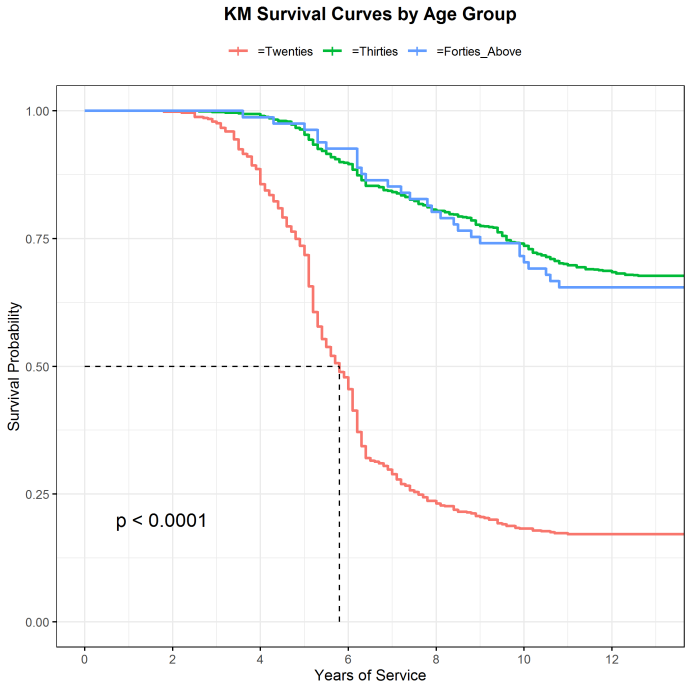

Figure 4.2. Age groups are indicated as twenties = red, thirties = green, and forties or older = blue. When comparing survival curves, the null hypothesis (H_0) is: Survival probabilities are equal across all groups. The alternative hypothesis (H_a) is: At least one group has a survival probability different from the other groups. A p-value is a measurement to validate our null hypothesis. In this case, a p-value $<$ .0001 indicates there is statistical significance between age groups and we reject H_0. Observations in their twenties, saw survival probability decrease from 85.6% at 4 YOS to 31.5% at 6.5 YOS, a drop of 54.1%. The dashed lines show what time a survival probability reaches 50% (median survival probability) and only occurred for observations in their twenties at 5.6 YOS. While thirties experienced a 13.7% and forties or older 12.4% during the 4–6.5 YOS time frame. Final survival probabilities by age group were twenties = 17.2%, thirties = 67.7%, and forties or above = 65.4%.

4.1.2 KM Survival Curves by Ethnicity

Figure 4.3 shows survival times by ethnicity. The KM model does not indicate any significant difference in the relationship between ethnicity and attrition. Asian, Hispanic and Other have relatively similar survival probabilities. Between 4–6.5 YOS survival probabilities for Asian, Hispanic, and Other dropped 27.2%, 27.7%, and 30.8%, respectively. Although they had

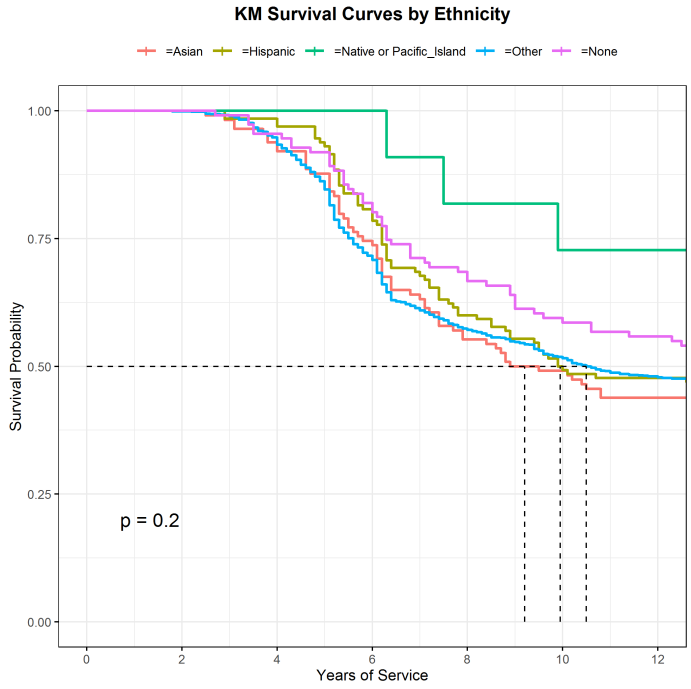

Figure 4.3. Ethnicity's are indicated as Asian = red, Hispanic = gold, Native or Pacific_Islander = green, Other = blue, and None = pink. Using p-value as a measurement to validate our null hypothesis, a p-value of .2 is not statistically significant Therefore, we fail to reject the null hypothesis (H_0) that: Survival probabilities are equal across ethnicities. The dashed lines show what time survival probability reaches 50% (median survival probability). Asian, Hispanic, and Other had median survival probabilities of 9.2 YOS, 9.95 YOS, and 10.5 YOS, respectively. While not statistically significant, from 4–6.5 YOS we see the large drops in probability of survival among Asian, Hispanic, and Other ethnicities, with decreases of 27.2%, 27.7% and 30.8%, respectively. Although they had the smallest sample sizes, ethnicities of None and Native or Pacific_Islander's have the best chance at remaining in the service. From 4–6.5 YOS None dropped 21.6%, while Native or Pacific_Islander dropped 9.1%. Final survival probabilities by ethnicity were, Asian = 43.9%, Hispanic = 46.9%, Native or Pacific_Islander = 72.7%, Other = 47.6%, and None = 54.1%.

smaller sample sizes, Native or Pacific_Islander and None had the least dramatic drops in survival probability between 4–6.5 YOS, decreasing 9.1% and 21.6%. These two groups also have the highest final survival probabilities, with Native or Pacific_Islander ending at 72.7% and None = 54.1%. Asian had the lowest final survival probability of 43.9%, followed

by Hispanic = 46.9%, and Other = 47.6%.

4.1.3 KM Survival Curves by Language Category

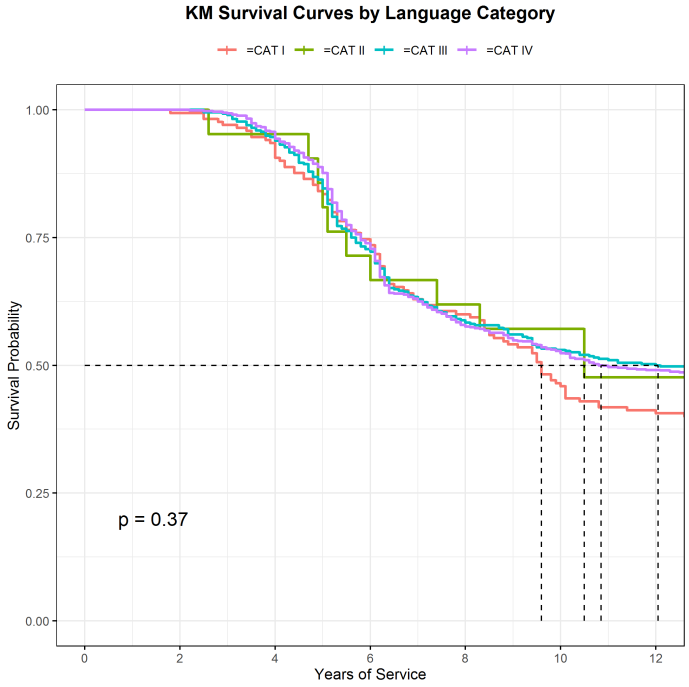

Figure 4.4. Language categories are indicated as CAT I = red, CAT II = green, CAT III = blue, and CAT IV = purple. Using p-value as a measurement to validate our null hypothesis, a p-value of .37 is not statistically significant Therefore, we fail to reject the null hypothesis (H_0) that: Survival probabilities are equal across language categories. The dashed lines show what time survival probability reaches 50% (median survival probability). CAT I had the lowest median YOS at 9.6, followed by CAT II = 10.5 YOS, CAT IV = 10.9 YOS, and CAT III = 12.1 YOS. While CAT III languages have a slightly longer median YOS, all of the survival probabilities decrease around the same rate, further validating our null hypothesis that language category does not have a significant impact on graduate attrition. The largest decrease in survival probabilities occur between 4–6.5 YOS. With CAT IV languages experiencing a 30.2% drop, CAT III a 29% drop, CAT II a 28.5% drop and CAT I a 25.3%. Final survival probabilities were CAT I = 40.0%, CAT II = 47.6%, CAT III = 49.7%, and CAT IV = 48.6%.

Figure 4.4 shows survival times by language category. Survival probability across language categories is relatively consistent. Although long term, more difficult languages have a

slightly higher survival probability and the longest median years of service. CAT III languages had a median service length of 12.1 YOS and CAT IV = 10.9 YOS, compared to CAT I = 9.6 YOS and CAT II = 10.5 YOS. Between 4–6.5 YOS, CAT IV language graduates experience the largest drop in survival probability, decreasing 30.2%, followed by CAT III = 29%, CAT II = 28.5%, and CAT I = 25.3%.

4.1.4 KM Survival Curves by Education Level

Figure 4.5 survival curves by education level. Median survival times, indicated by dashed lines, ranged from 9.5 YOS for observations with graduate level education, followed by high school education = 10.1 YOS, college graduates = 11.1 YOS and some college = 12.2 YOS. Few DLIFLC graduates have a graduate level degree but we witnessed a 41.9% drop in survival probability between 4–6.5 YOS for these observations. High school, some college and college education levels experienced drops of 30.2%, 25.2%, and 29.2%, respectively, during the same 4–6.5 YOS period.

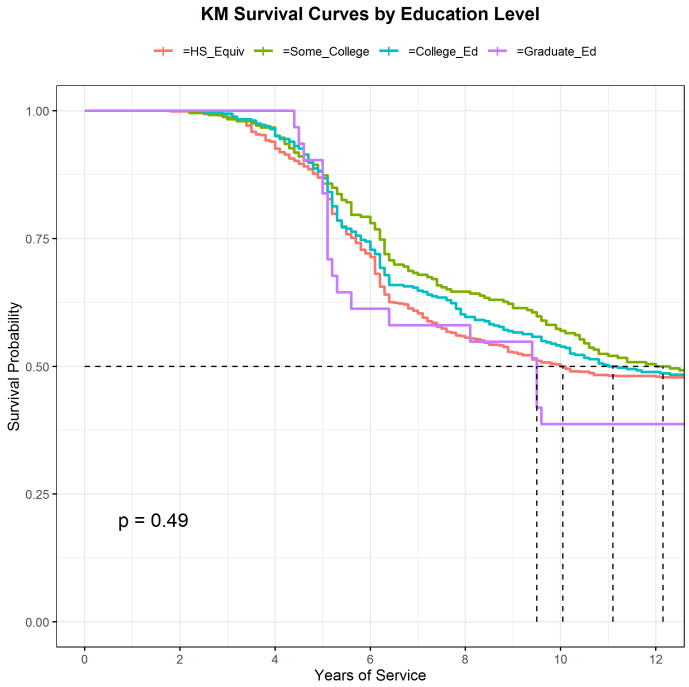

Figure 4.5. Education levels are indicated as HS_Equiv = red, Some_College = green, College_Ed = blue, and Graduate_Ed = purple. Using p-value as a measurement to validate our null hypothesis, a p-value of .49 is not statistically significant Therefore, we fail to reject the null hypothesis (H_0) that: Survival probabilities are equal across education levels. The dashed lines show what time survival probability reaches 50% (median survival probability). Graduate_Ed had the lowest median YOS at 9.5, followed by HS_Equiv = 10.1 YOS, College_Ed = 11.1 YOS, and Some_College = 12.2 YOS. Although very few DLIFLC graduates have a graduate level education, we see a drastic drop in their survival probability just past the 4 YOS mark. Survival probability for DLIFLC graduates with a Master's or Doctorate degree drops from 100% to 58.1% between 4–6.5 YOS. Of the remaining education levels, graduates with some college have the best survival probability, decreasing 25.2% during 4–6.5 YOS. While high school educated graduates probability of survival drops 30.2% and college educated graduates drop 29.2%. Final survival probabilities by education level were HS_Equiv = 47.7%, Some_College = 49.2%, College_Ed = 48.4%, and Graduate_Ed = 41.9%.

4.1.5 KM Survival Curves by Marital Status

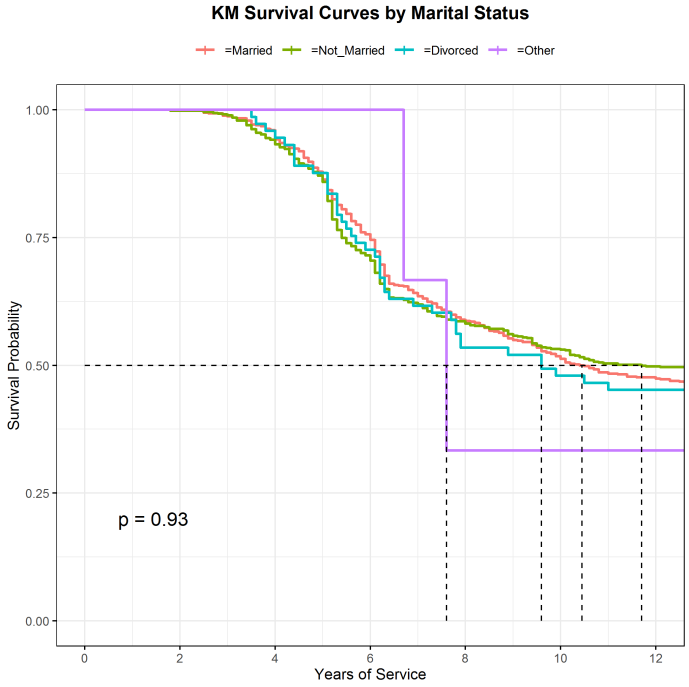

Figure 4.6. Marital status is indicated as Married = red, Not_Married = green, Divorced = blue, and Other = purple. Using p-value as a measurement to validate our null hypothesis, a p-value of .93 is not statistically significant Therefore, we fail to reject the null hypothesis (H_0) that: Survival probabilities are equal across marital status. The dashed lines show what time survival probability reaches 50% (median survival probability). The shortest median time in service was Other with 7.6 YOS, followed by Divorced = 9.6 YOS, Married = 10.4 YOS, and Not_Married = 11.7 YOS. Between 4–6.5 YOS survival probability decreased among Married by 28.5%, Not_Married = 30.2%, Divorced = 31.5% and Other had no decrease during this time frame but ultimately two out of the three observations leave service. Final survival probability by marital stats was Married = 46.7%, Not_Married = 49.7%, Divorced = 45.3%, and Other = 33.3%.

Figure 4.6 shows survival times by marital status. Having only three observations, the level Other has the appearance of a step function. The majority of observations in this model were either Married or Not_Married, with Not_Married service members having the longest median time in service at 11.7 years. There is no indication that marital status has a significant impact on DLIFLC graduate attrition. Looking at Married, Not_Married

and Divorced observations their survival probabilities decreased 28.5%, 30.2%, and 31.5%, respectively, between 4–6.5 YOS. While final survival probabilities for these three groups were Married = 46.7%, Not_Married = 49.7%, and Divorced = 45.3%.

4.2 Cox Proportional Hazard Model

In contrast to the overall KM model, survival times for the overall CPH model are slightly better and presented in figure 4.7. A significant decline in survival probability between 4–6.5 YOS is still present, dropping 23.6% but less drastic than the KM model. In addition to survival curves, summary statistics for the CPH model are presented in Table 4.1.

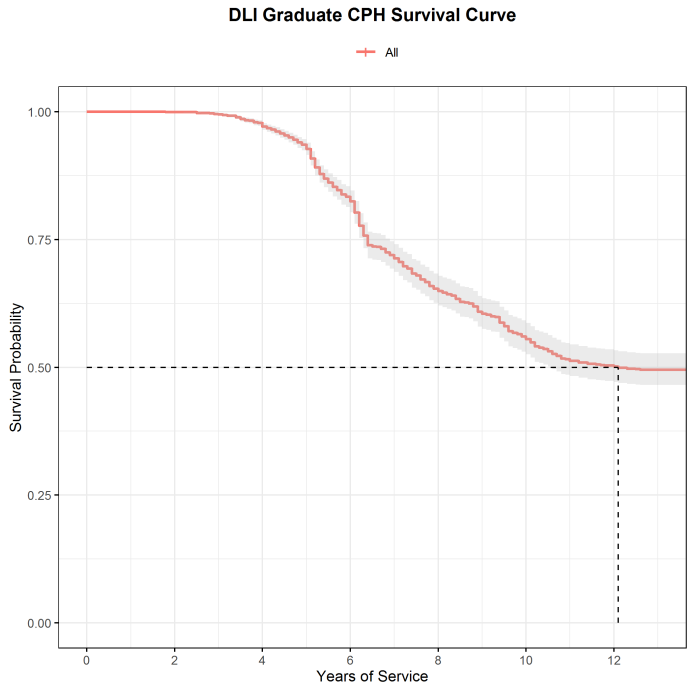

Figure 4.7. The red curve signifies DLIFLC survival probabilities, with a confidence interval shown in gray. The dashed line indicates the median survival time, which occurs at 12.1 YOS. At 4 YOS the survival probability is 97.1% and drops to 73.5% by 6.5 YOS, a decrease of 23.6%. After 6.5 YOS probability continues to decrease at a slightly lower rate, leveling out at 49.6%. Altogether, 68.1% of events occur (attrition) between 4–6.5 YOS.

For feature selections via CPH p-values in Table 4.1 are computed from the following

hypotheses:

$$H_0: \text{ the coefficient for the given predictor } = 0,$$
$$H_1: \text{ the coefficient for the given predictor } \neq 0.$$

Some statistics of interest from the CPH model are shown in Table 4.1.

Table 4.1. CPH Significance Codes: 0 - ***, .001 - **, .01 - *, .05 - ·

Variable	Hazard Ratio	Lower 95% CI	Upper 95% CI	p-value
AGE	.73	.71	.75	< .0001***
Married (reference)	-	-	-	-
Not_Married	.78	.67	.91	.001**
Divorced	.79	.57	1.01	.17
Other	.38	.10	1.55	.18
Asian (reference)	-	-	-	-
Hispanic	.70	.50	.99	.04·
Native or Pacific_Islander	.54	.17	1.73	.30
Other	.86	.66	1.11	.24
None	.51	.35	.74	< .0001***
HS_Equiv (reference)	-	-	-	-
Some_College	1.4	1.14	1.72	.001**
College_Ed	3.22	2.64	3.93	< .0001***
Graduate_Ed	6.29	3.92	10.10	< .0001***
CAT I (reference)	-	-	-	-
CAT II	1.10	.59	2.05	.77
CAT III	.74	.58	.94	.02·
CAT IV	.85	.68	1.06	.14

4.2.1 CPH Survival Curves by Age

Age was identified as a highly significant predictor. Figure 4.8 shows survival times by age group. CPH survival times for graduates in their twenties decreased 56.3% between 4–6.5

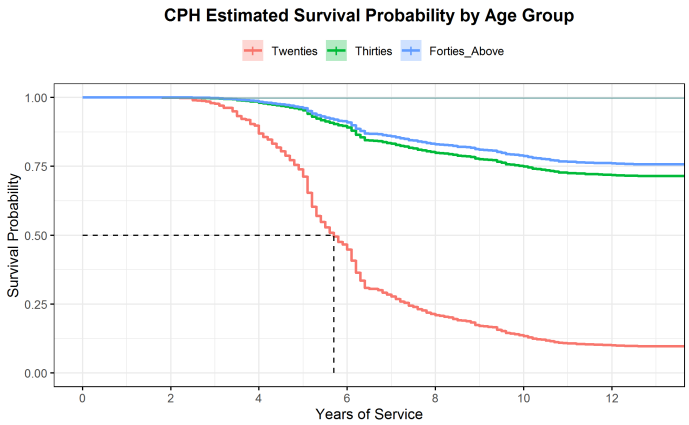

Figure 4.8. Age groups are indicated as twenties = red, thirties = green, and forties or older = blue. When comparing CPH survival curves, two statistics of interest were the p-value and Hazard Ratio (HR). The null hypothesis (H_0) is: Survival probabilities are equal across all groups. The alternative hypothesis (H_a) is: At least one group has a survival probability different from the other groups. A p-value is a measurement to validate our null hypothesis. In this case, a p-value $< .0001$ indicates there is statistical significance between age groups and we reject H_0. HR is a measurement for risk of failure in relation to the reference group Twenties, in this case a HR of .73 signifies each year a DLIFLC graduate remains in service, their risk of attrition decreases by a factor of .73. Observations in their twenties saw survival probability decrease from 86.9% at 4 YOS to 30.6% at 6.5 YOS, a drop of 56.3%. The dashed lines show what time a survival probability reaches 50% (median survival probability) and only occurred for twenties at 5.7 YOS. While thirties experienced a 13.7% decrease and forties or older a 11.5% decrease during the 4–6.5 YOS time frame. Final survival probabilities by age group were twenties = 9.7%, thirties = 71.5%, and forties or above = 75.7%.

YOS and had a final probability of 9.7%, both of which are lower than the KM model. While CPH survival probabilities are slightly higher than the KM model for observations thirty and older. Between 4–6.5 YOS, graduates in their thirties experienced a drop of 13.7% and those forty and above fell 11.5%. Final survival probability for thirties was 71.5% and forties or above was 75.7%.

4.2.2 CPH Survival Curves by Ethnicity

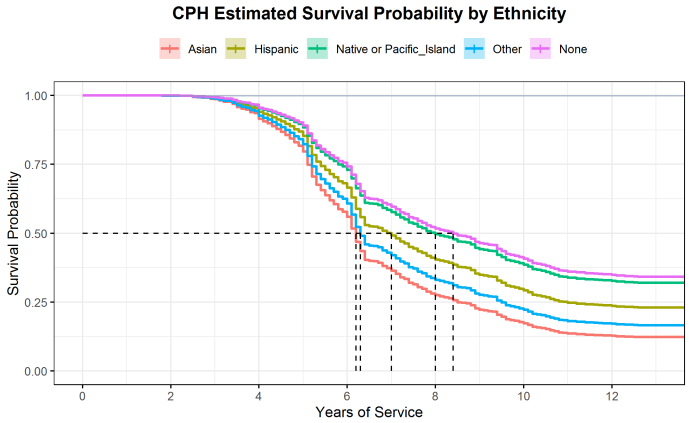

Figure 4.9. Ethnicity's are indicated as Asian = red, Hispanic = gold, Native or Pacific_Islander = green, Other = blue, and None = pink. When comparing CPH survival curves, two statistics of interest were the p-value and HR. The null hypothesis (H_0) is: Survival probabilities are equal across all ethnicities. The alternative hypothesis (H_a) is: At least one ethnicity group has a survival probability different from the other groups. A p-value is a measurement to validate our null hypothesis. Two ethnicities were identified as statistically significant, Hispanic had a p-value = .04 and None had a p-value $<$ .0001 and we reject H_0. HR is a measurement for risk of failure in relation to the reference group Asian. Looking at statistically significant ethnic categories in this case, Hispanic had a HR of .70 and None had a HR of .51. These HRs signify each year a DLIFLC graduate remains in service, Hispanic graduates' risk of attrition decreases by a factor of .70, while None's risk of attrition decreases by a factor of .51. The dashed lines show what time survival probability reaches 50% (median survival probability). Median survival probabilities for each ethnicity were Asian = 6.2 YOS, Hispanic = 7.0 YOS, Native or Pacific_Islander's = 8.0 YOS, Other = 6.3 YOS, and None = 8.4 YOS. All groups experience significant drops in survival probability between 4–6.5 YOS. The largest drop was experienced by Asian falling 51.6%, followed by Other = 47.1%, Hispanic = 41.4%, Native or Pacific_Islander = 34.6%, and None = 33.1%. Final survival probabilities by ethnicity were, Asian = 12.3%, Hispanic = 23.0%, Native or Pacific_Islander = 32.0%, Other = 16.6%, and None = 34.2%.

Survival curves shown in Figure 4.9 vary drastically from the KM model. We see a decrease in the median survival times, identified by the dashed line, for all ethnicity groups. Only

Native or Pacific_Islander and None reach eight years of service or more, while median service time for Asian, Hispanic, and Other were 6.2 YOS, 7.0 YOS, and 6.3 YOS, respectively. All ethnic groups experienced relatively large drops compared to the KM model regarding survival probability between 4–6.5 YOS. The largest decrease occurs among Asians, falling 51.6%, followed by Other = 47.1%, Hispanic = 41.4%, Native or Pacific_Islander = 34.6%, and None = 33.1%. Also significantly lower than KM model estimates were the final CPH survival probabilities. By ethnicity final survival probabilities were Asian = 12.3%, Hispanic = 23.0%, Native or Pacific_Islander = 32.0%, Other = 16.6%, and None = 34.2%.

4.2.3 CPH Survival Curves by Language Category

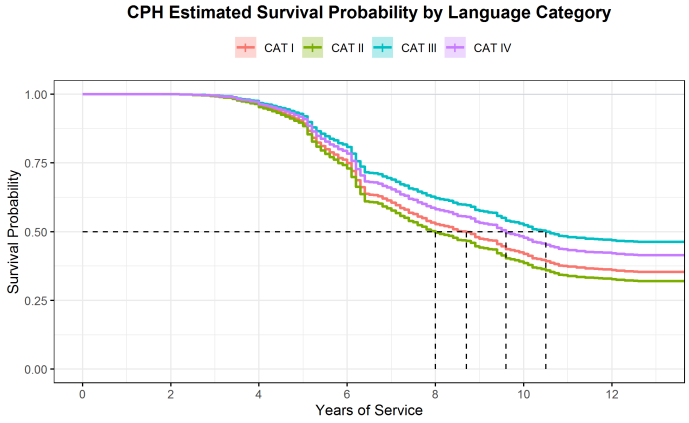

Figure 4.10. Language categories are indicated as CAT I = red, CAT II = green, CAT III = blue, and CAT IV = purple. When comparing CPH survival curves, two statistics of interest were the p-value and HR. The null hypothesis (H_0) is: Survival probabilities are equal across all ethnicities. The alternative hypothesis (H_a) is: At least one language category group has a survival probability different from the other groups. A p-value is a measurement to validate our null hypothesis. CAT III was the only statistically significant, language category with a p-value = .02 and we reject H_0. HR is a measurement for risk of failure in relation to the reference group CAT I. Looking at CAT III graduates, the only statistically significant group, had a HR of .74. Signifying each year a DLIFLC graduate remains in service, CAT III graduates risk of attrition decreases by a factor of .74. The dashed lines show what time survival probability reaches 50% (median survival probability). Median survival probabilities were CAT I = 8.7 YOS, CAT II = 8.0 YOS, CAT III = 10.3 YOS, and CAT IV = 9.6 YOS. All groups experience significant drops in survival probability between 4–6.5 YOS. The largest drop was experienced by CAT II falling 34.6%, followed by CAT I = 32.3%, CAT IV = 26.3% and CAT III = 25.5%. Final survival probabilities by language category were, CAT I = 35.4%, CAT II = 32.0%, CAT III = 46.3%, and CAT IV = 41.4%.

Consistent with the KM model, we can see in Figure 4.10, graduates who studied more difficult CAT III and CAT IV languages remain in service longer. However, in this model, CAT II languages now have the lowest median survival time at 8 YOS, followed by CAT I = 8.7 YOS, CAT IV = 9.6 YOS, and CAT III = 10.3 YOS. Although significant, decreases in survival probability between 4–6.5 YOS varied only slightly from KM model estimates,

CPH estimates were CAT I = 32.3%, CAT II = 34.6%, CAT III = 25.5%, and CAT IV = 26.3%. Final survival probabilities were CAT I = 35.4%, CAT II = 32.0%, CAT III = 46.3%, and CAT IV = 41.4%.

4.2.4 CPH Survival Curves by Education Level

Figure 4.11 shows survival times by education. Some_College, College_Ed, and Graduate_Ed were all highly significant factors in this model and we can see a significant difference in median survival times between the three compared to KM estimates. HS_Equiv graduates have the highest probability of survival, only experiencing a 6.7% decrease between 4–6.5 YOS and a final survival probability of 83.4%. DLIFLC graduates with some college also have a relatively good probability of survival, decreasing 9.5% between 4–6.5 YOS and finishing with a probability of 77.6%. College and graduate level educated DLIFLC graduates are the most likely to leave service. College graduates had a drop of 20.1%, while graduate educated service members fell 34.6% between 4–6.5 YOS. Final survival probability for college and graduate education levels were 55.8% and 32.0%, respectively.

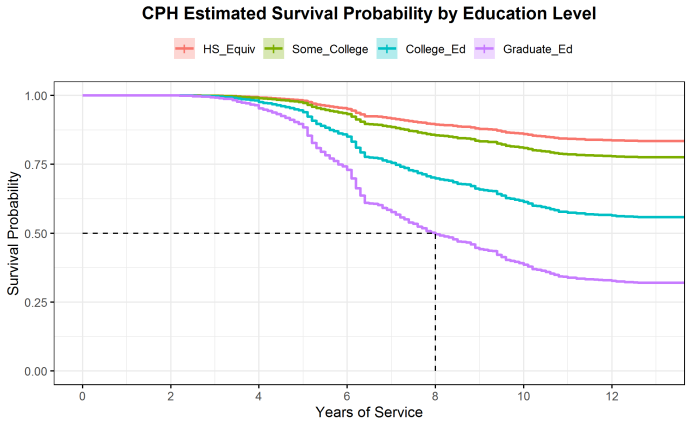

Figure 4.11. Education levels are indicated as HS_Equiv = red, Some_College = green, College_Ed = blue, and Graduate_Ed = purple. When comparing CPH survival curves, two statistics of interest were the p-value and HR. The null hypothesis (H_0) is: Survival probabilities are equal across all education levels. The alternative hypothesis (H_a) is: At least one education level group has a survival probability different from the other groups. A p-value is a measurement to validate our null hypothesis. Three education levels were identified as statistically significant, Some_College had a p-value = .001, while College_Ed and Graduate_Ed had a p-values $<$ .0001 and we reject H_0. HR is a measurement for risk of failure in relation to the reference group HS_Ed. Looking at statistically significant education levels in this case, Some_College had a HR of 1.4, College_Ed had a HR of 3.2, and Graduate_Ed had a HR of 6.32. These HRs signify each year a DLIFLC graduate remains in service, the graduate's risk of attrition based off their education level of Some_College, College_Ed, Graduate_Ed increases by a factor of 1.4, 3.2, and 6.32, respectively. The dashed lines show what time survival probability reaches 50% (median survival probability). Median survival probabilities for HS_Equiv, Some_College, and College_Ed were not reached, while Graduate_Ed had a median of 8.0 YOS. All groups experience significant drops in survival probability between 4–6.5 YOS. The largest drop was experienced by Graduate_Ed falling 34.6%, followed by College_Ed = 20.1%, Some_College = 9.5%, and HS_Eq = 6.7%. Final survival probabilities were, HS_Equiv = 83.4%, Some_College = 77.6%, College_Ed = 55.8%, and Graduate_Ed = 32.0%.

4.2.5 CPH Survival Curves by Marital Status

Figure 4.12 shows, married graduates have the lowest median survival at just 5.6 years of service. While non-married and divorced graduates have almost identical survival curves.

Survival probabilities between 4–6.5 YOS decrease significantly faster than KM estimates. The CPH model survival probabilities for this time period decrease 61.6% for married, followed by Divorced = 55.6%, Not_Married = 55.2%, Other = 35.3%. Final survival probabilities were also significantly lower in the CPH model, with Married = 4.8%, Not_Married = 9.2%, Divorced = 8.9%, and Other = 31.1%.

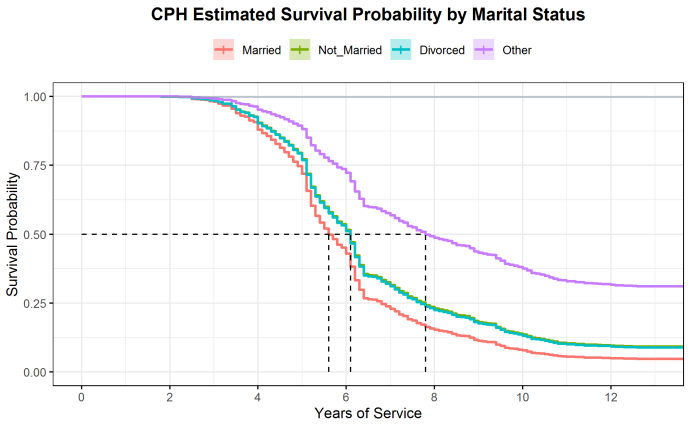

Figure 4.12. Marital status is indicated as Married = red, Not_Married = green, Divorced = blue, and Other = purple. When comparing CPH survival curves, two statistics of interest were the p-value and HR. The null hypothesis (H_0) is: Survival probabilities are equal across all marital status groups. The alternative hypothesis (H_a) is: At least one group has a survival probability different from the other groups. A p-value is a measurement to validate our null hypothesis. Not_Married graduates were statistically significant with a p-value $<$ .0001 and we reject H_0. HR is a measurement for risk of failure in relation to the reference group Married. Looking at statistically significant, in this case Not_Married graduates, had a HR of .78. Signifying each year a DLIFLC graduate remains in service, a Not_Married graduate's risk of attrition decreases by a factor of .78. The dashed lines show what time survival probability reaches 50% (median survival probability). Median survival probabilities for Married = 5.6 YOS, Not_Married = 6.1 YOS, Divorced = 6.1 YOS, and Other = 7.8 YOS. All groups experience significant drops in survival probability between 4–6.5 YOS. The largest drop was experienced by Married graduates falling 61.6%, followed by Divorced = 55.6%, Not_Married = 55.2%, and Other = 35.3%. Final survival probabilities were, Married = 4.8%, Not_Married = 9.2%, Divorced = 8.9%, and Other = 31.1%.

4.3 Random Survival Forest Model

The last model we looked at was RSF. This model was fit similar to the KM model, in that it used all predictor variables. Because RSF is an ensemble method, results are presented in Figure 4.13 as an average of all estimated survival curves, allowing for comparison between the overall KM and CPH survival curves. Most notably, again we see a sharp drop in survival probability between four and six years of service and a steady decline after six years of service until about 50% survival probability. The decrease between 4–6.5 YOS of 20.3%

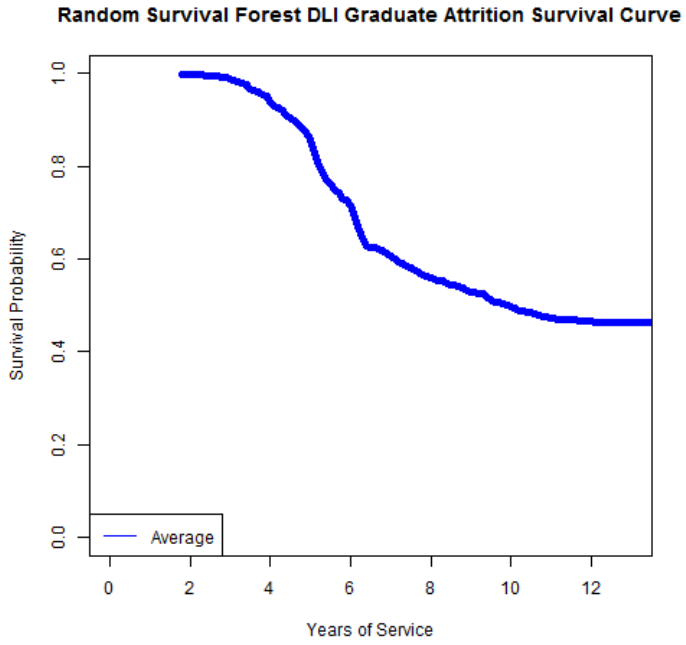

Figure 4.13. The survival curve shown in blue, is the average of 1,469 survival curves estimated by the RSF model. The decrease in survival probability between 4–6.5 YOS is 20.3% and slightly less than the drops in KM (29.3%) and CPH (23.6%) witnessed during the same time period. However, the RSF model average survival probability at the 4 YOS mark was 87.5%, which is 6.3% and 9.6% lower than KM and CPH, respectively, at 4 YOS. The average RSF final survival probability was 46.3%.

was slightly less than KM and CPH drops of 29.3% and 23.6%, respectively. However at 4 YOS, the RSF model's average survival probability had already dropped to 87.5%, which is 6.3% lower than the KM model and 9.2% lower than the CPH model at the same point in time. Final RSF survival probability was 46.3%.

Figure 4.14 shows which variables in the RSF model were identified as most important. Below we can see that age was clearly the most important factor contributing to DLIFLC graduate attrition. Several other variables such as education level, ethnicity, and marital status, identified as significant in the CPH model, register as being important as well.

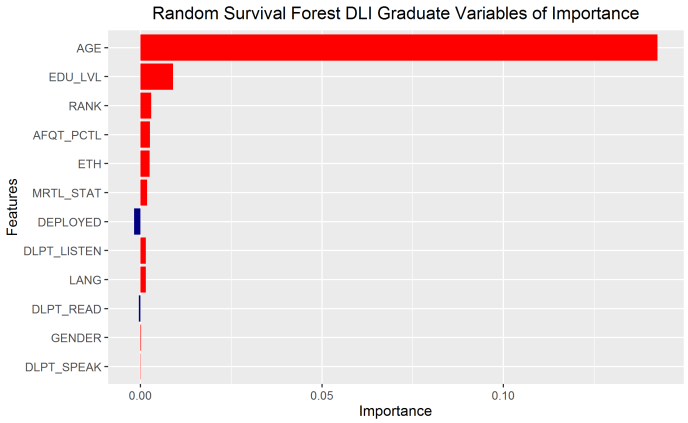

Figure 4.14. RSF Variables of Importance. The value of the bar chart for each feature indicates how strong it is correlated to DLIFLC graduate attrition. RSF shows that age is the most important feature followed by education, rank, AFQT percentile, and ethnicity.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
Conclusion

5.1 Summary

The time, cost, and difficulty associated with training Army linguists were the motivating factors in investigating DLIFLC graduate attrition. Using survival analysis, this study focused on identifying significant factors contributing to attrition and at which point graduates left military service. Utilizing KM, CPH, and RSF models to conduct survival analysis, several insights have been gained about DLIFLC graduate attrition.

Over half (52.1%) of Army enlisted DLIFLC graduates that entered the service between January 1, 2010, and December 31, 2012 have left the Army. 68.1% of this attrition occurs between 4–6.5 YOS and coincides with the end of most first term enlistment contracts. Army linguists are currently offered some of the highest enlistment bonuses, upon joining the service. While this may attract new linguists, current active-duty incentives like language pay and reenlistment bonuses do not appear as effective in retaining linguists. Considering just the average cost of DLIFLC, the DOD spent approximately $168 million to train these linguists, while getting a return of perhaps one operational duty assignment before they left the service.

The most significant factors related to differences in graduate attrition were age, education level, AFQT percentile, and language category. The difference in attrition between age groups strongly indicates, retaining graduates until their thirties would result in better long term retention. Another approach to addressing retention when looking at age is exploring additional options for enlistment waivers and perhaps increasing age limitations for potential linguists, with the understanding they have a higher probability of remaining in the service after DLIFLC. There was also a clear distinction between education level and attrition. As education level increased, so did the probability of attrition. College and graduate level educated DLIFLC graduates are more likely to attrite than those with high school education or some college. An assumption could be made regarding education, that the combination of advanced education and language training, provides higher paying civilian

or other government agency job opportunities. In contrast, the more difficult languages, CAT III and CAT IV, had the highest survival probabilities and longer median times in service. In terms of ethnicity, Asians experience attrition at the highest rate, while Native or Pacific_Islander have the highest probability of survival. Marital status did not seem to have a significant impact on attrition, although married DLIFLC graduates do have a slightly lower survival probability that non married and divorced graduates. Though not included in the CPH model, the KM and RSF models also identified AFQT percentile being a significant contributor to DLIFLC attrition. Highlighting that those who scored low (Below_70) and high (90_Above) were most likely to leave the Army.

Ultimately, this research provides a baseline understanding of some influential factors impacting DLIFLC graduate attrition. Used as a framework for additional research and increased scope of work, these results and analysis methods can help the DOD reevaluate policies currently in place and help determine if additional incentives could improve retention, specifically past a graduate's first contract.

5.2 Future Work

There are several opportunities for future work regarding DLIFLC graduate attrition. Linguist retention challenges are not unique to the Army and the scope of this analysis could be expanded to include other service branches. Comparison among the different service branches could provide valuable insight into differences in retention and what service branches do differently, that may positively or negatively influence retention.

In addition to analyzing all service branches, one recommendation on how to expand upon this problem would be to increase the service date range; providing more observations over a longer period of time. Another recommendation for continued research would be including other explanatory variables. The PDE database catalogs are vast and have an enormous amount of data available to analysts. For example, service member duty station, home of record, initial service contract length, and reenlistment bonuses could also provide insight into DLIFLC graduate attrition.

One final area of interest would be to look at DLIFLC graduate attrition, in relation to overall service branch attrition. If not already being done, this could include surveying linguists

leaving the military and determine additional influential factors and post military career plans. Research of this nature could help determine if factors impacting linguist retention are unique to linguists, or are consistent throughout the service branches.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX: Outputs from Survival Analyses

A.0.1 Gender KM Survival Curves

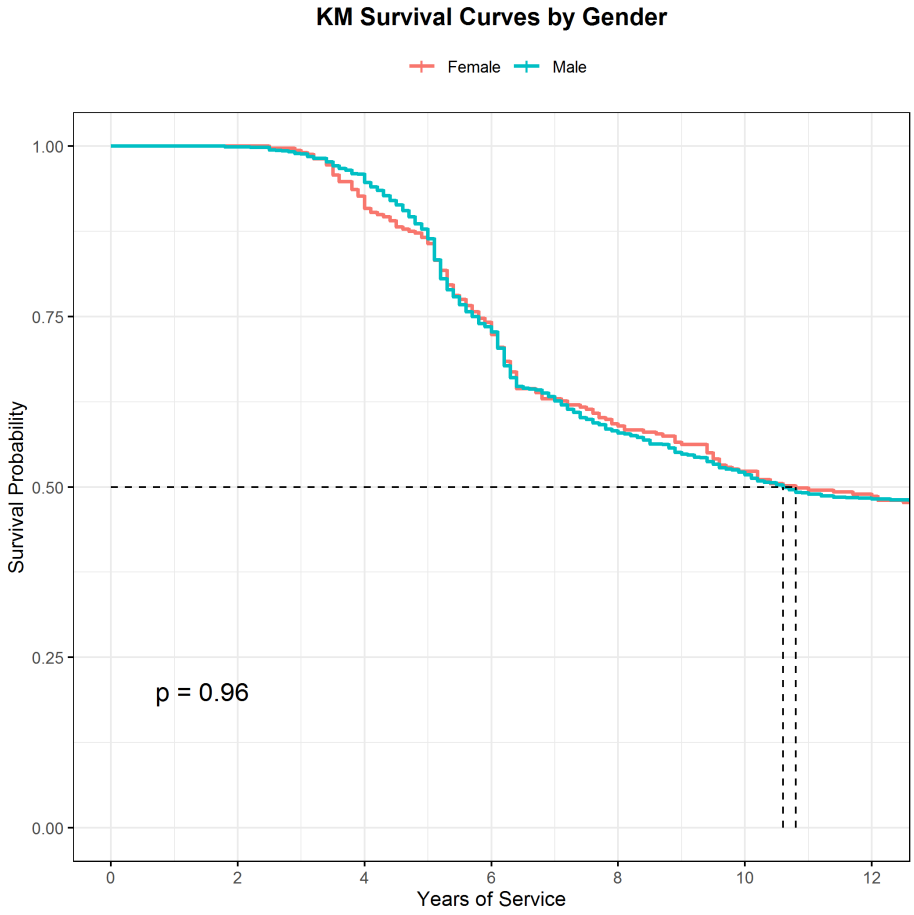

Figure A.1. Gender is indicated as Female = red and Male = blue. Gender does not appear to have an impact on DLIFLC graduate attrition. Female have a slightly smaller decrease in survival probability between 4–6.5 YOS of 26.5%, compared to Males who drop 30.1%. However, median survival probabilities for female and male were very close at 10.7 YOS and 10.4 YOS, respectively. Additionally, final survival probabilities by gender were extremely close with females at 47.7% and males at 48.0%.

A.0.2 Deployment KM Survival Curves

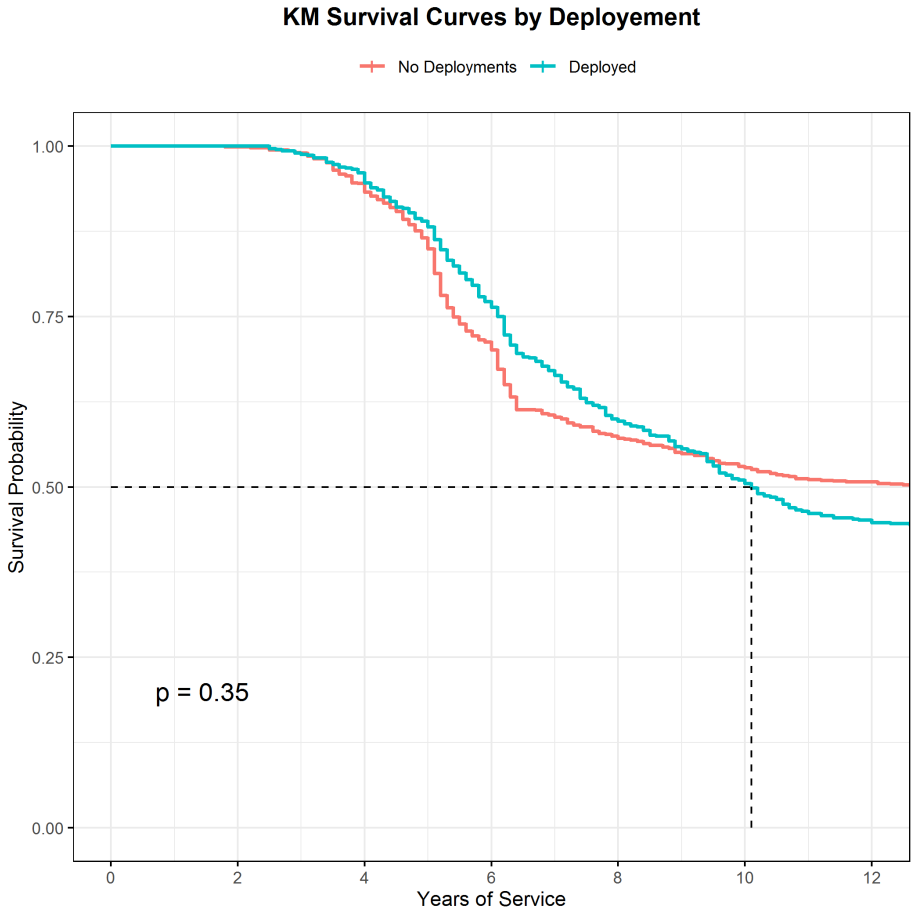

Figure A.2. Deployment status is indicated as No Deployment = red and Deployment = blue. Initially, observations that had at least one deployment appear to have a higher survival probability. Between 4–6.5 YOS, graduates who had deployed had a decrease of 25.2%, while those with no deployments dropped 32%. However, around 9.5 YOS survival probability for the two groups cross, which could be a possible indication of operational tempo impacting linguist attrition. Ultimately, linguists with no deployments have a slightly higher survival probability at 50.3% compared to those who had deployed at 44.4%.

A.0.3 Rank KM Survival Curves

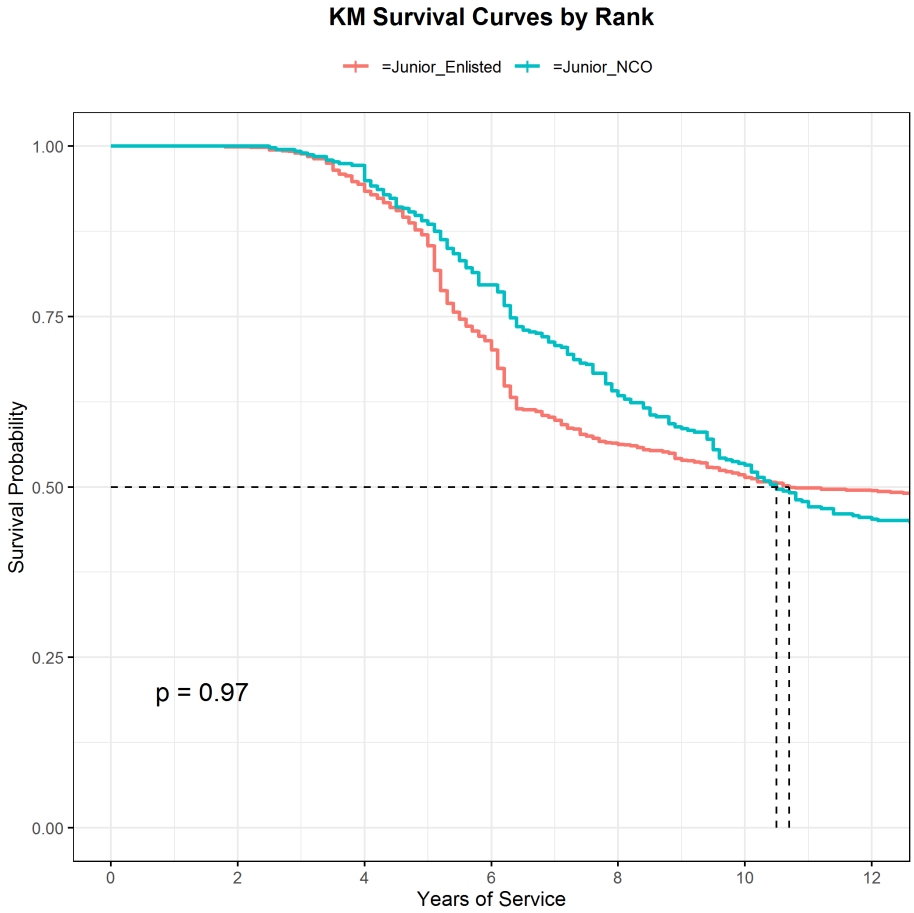

Figure A.3. Rank is indicated as Junior_Enlisted = red and Junior_NCO = blue. While not statistically significant, Junior_Enlisted DLIFLC graduates experienced a higher decrease in survival probability between 4–6.5 YOS, dropping 32.1% compared to Junior_NCO at 21.9%. However median survival times are very similar with Junior_Enlisted averaging 10.6 YOS, while Junior_NCO average 10.4 YOS. Final survival probability for Junior_Enlisted was 49.1%, while Junior_NCO was 44.8%

A.0.4 DLPT_Speak KM Survival Curves

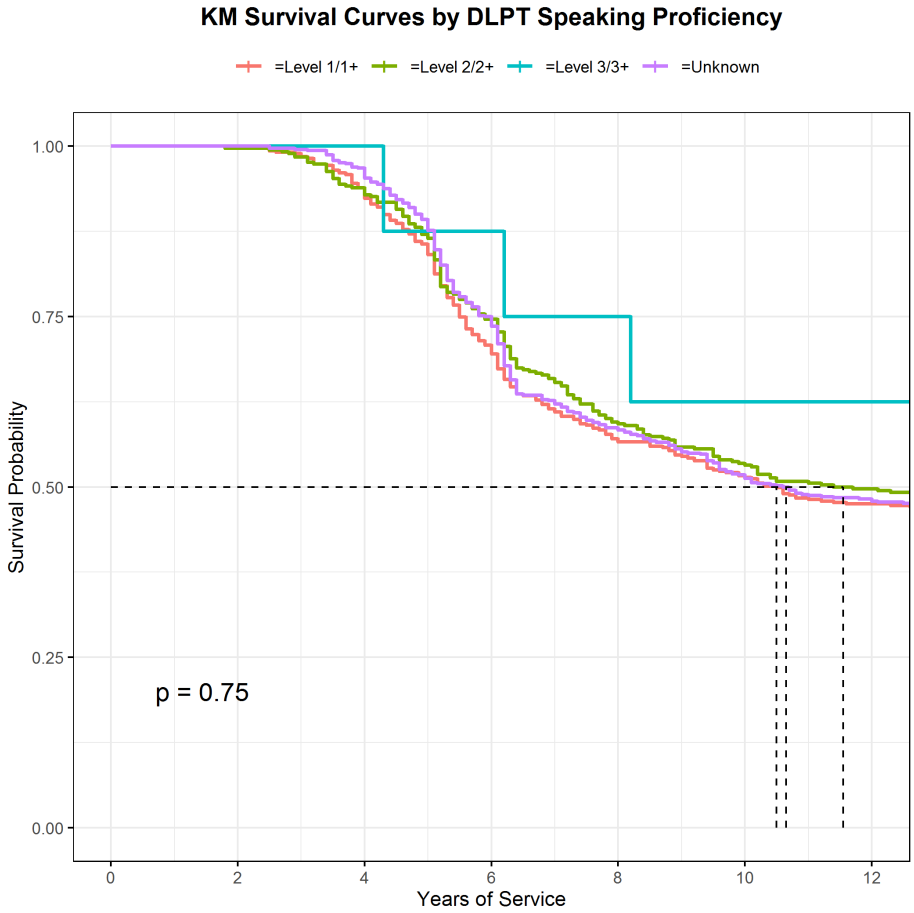

Figure A.4. DLPT scores are indicated as 1/1+ = red, 2/2+ = green, 3/3+ = blue, and Unknown = purple. The speaking proficiency is the DLPT score from a graduate's most recent test. There is no significant difference between speaking proficiency and survival probability. The following were the percentage decline in survival probability between 4–6.5 YOS. 1/1+ = 29%, 2/2+ = 25.7%, 3/3+ = 25.0%, and Unknown = 31.9%. The final survival probability for proficiency levels was 1/1+ = 47.1%, 2/2+ = 49.2%, 3/3+ = 62.5%, and Unknown = 47.6%.

A.0.5 DLPT_Listen KM Survival Curves

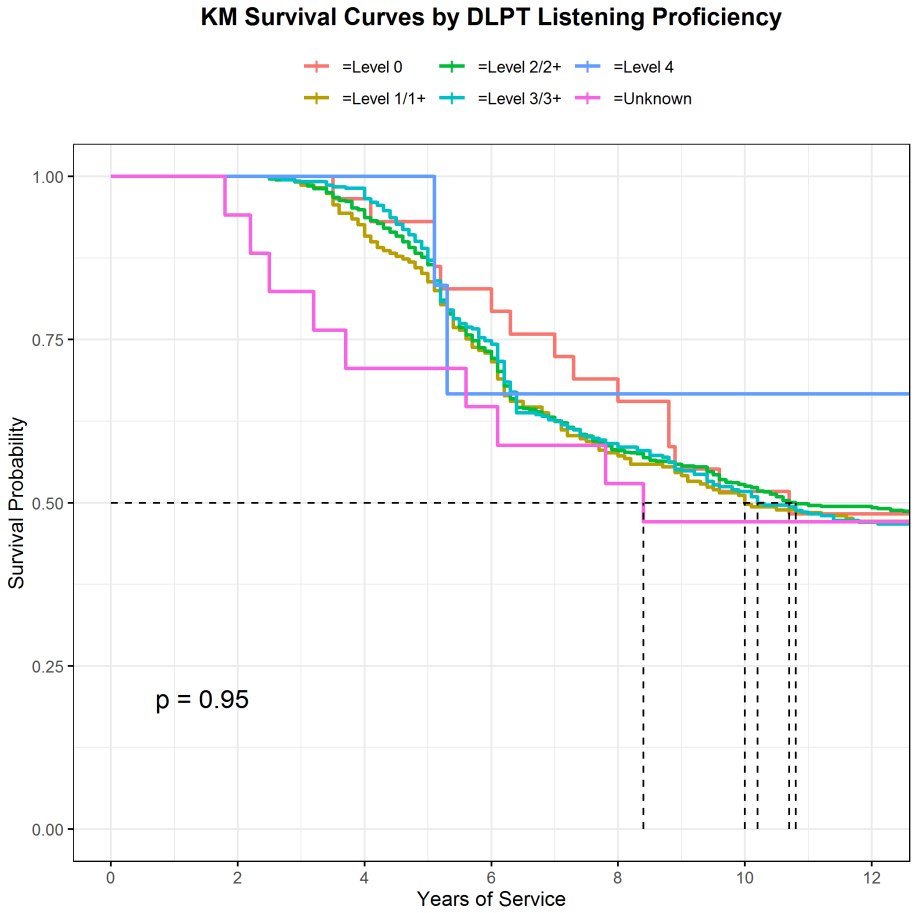

Figure A.5. DLPT scores are indicated as 0 = red, 1/1+ = gold, 2/2+ = green, 3/3+ = cyan, 4 = blue, and Unknown = pink. The listening proficiency is the DLPT score from a graduate's most recent test. While not significant, there is more variation among listening survival curves than speaking curves. DLIFLC graduate scores of 0 and 4 are rare and the reason they resemble a step function. However, the bulk of DLPT scores are made up of 1/1+, 2/2+, and 3/3+ and these scores have relatively similar survival probabilities. The following were the percentage decline in survival probability between 4–6.5 YOS. 0 = 20.7%, 1/1+ = 26.2%, 2/2+ = 29.3%, 3/3+ = 32.8%, 4 = 33.3%, and Unknown = 11.8%. The final survival probabilities were 0 = 48.3% ,1/1+ = 47.2%, 2/2+ = 48.6%, 3/3+ = 46.7%, 4 = 66.7%, and Unknown = 47.1%.

A.0.6 DLPT_Read KM Survival Curves

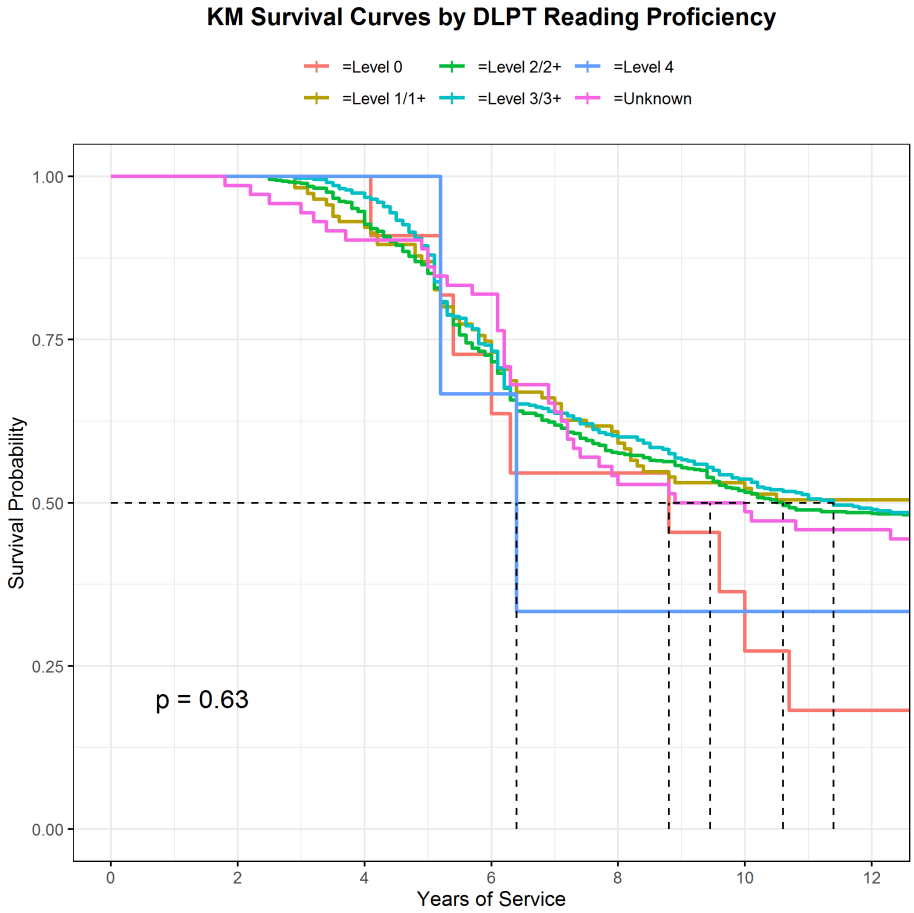

Figure A.6. DLPT scores are indicated as 0 = red, 1/1+ = gold, 2/2+ = green, 3/3+ = cyan, 4 = blue, and Unknown = pink. The reading proficiency is the DLPT score from a graduate's most recent test. The following were the percentage decline in survival probability between 4–6.5 YOS. 0 = 45.5%, 1/1+ = 25.2%, 2/2+ = 29.0%, 3/3+ = 31.7%, 4 = 33.3%, and Unknown = 11.8%. The final survival probabilities were 0 = 18.2%, 1/1+ = 50.4%, 2/2+ = 48.1%, 3/3+ = 48.3%, 4 = 33.3%, and Unknown = 44.4%.

A.0.7 KM Survival Curves by AFQT Percentile

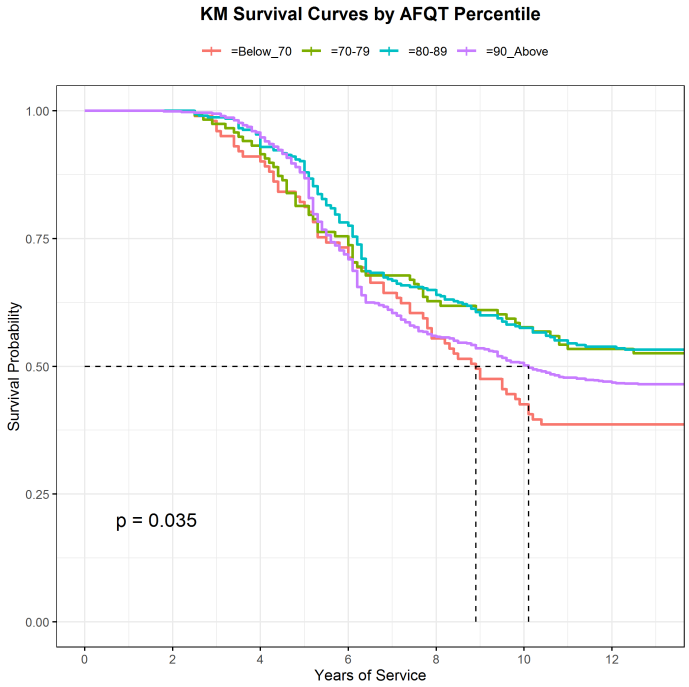

Figure A.7. AFQT percentiles are indicated as 70_Below = red, 70–79 = green, 80–89 = blue, and 90_Above = purple. The null hypothesis (H_0) is: Survival probabilities are equal across all groups. A p-value is a measurement to validate our null hypothesis. In this case, a p-value = .035 indicates a statistical significance between AFQT percentile groups and we reject H_0. It appears that DLIFLC graduates who had low (Below_70) or high (Above_90) AFQT scores have the highest probabilities of attrition. Between 4–6.5 YOS, all the AFQT percentile groups experience similar declines in survival probability. 70_Below decreased 23.8%, 70–79 = 23.7%, 80–89 = 24.6%, and 90_Above dropped by 32.2%. However, there is a larger difference between final survival probabilities. Below_70 has the lowest survival probability of 38.6%, followed by 90_Above = 46.5%, while 70–79 and 80–89 had final probabilities of 52.5% and 53.2%, respectively. DLIFLC graduates who scored below the 70th percentile or above the 90th percentile have the highest probability of attrition.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

Defense Language Institute Foreign Language Center (2022) General catalog 2021-2022. Accessed October 17, 2022, https://www.dliflc.edu/resources/publications/general-catalog/.

Department of the Army (2022) A call to service to overcome recruiting and retention challenges. https://www.army.mil/article/258577.

Devig A (2019) Predicting U.S. Army Enlisted Attrition After Initial Entry Training Using Survival Analysis. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/62725.

Hawes E (1990) An Application of Survival Analysis Methods to the Study of Marine Enlisted Attrition. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/34851.

Hinson W (2005) A Statistical Analysis of Individual Success After Successful Completion of Defense Language Institute Foreign Language Center Training. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/1934.

Interagency Language Roundtable (2023) Descriptions of proficiency levels. Technical report, https://www.govtilr.org/Skills/ILRscale1.htm.

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

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning* (Springer Texts in Statistics, New York, NY).

Kassambara A, Kosinski M, Biecek P (2021) survminer. https://cran.r-project.org/web/packages/survminer/index.html.

Kleinbaum DG, Klein M (2005) *Survival Analysis A Self-Learning Text* (Springer Science+Business Media, Inc., New York).

Nelson W (1969) Hazard plotting for incomplete failure data. *Journal of Quality Technology* 1(1):27–52, URL http://dx.doi.org/10.1080/00224065.1969.11980344.

Rubiano O, Enrique L (1993) An Analysis of the Coast Guard Enlisted Attrition. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/26352.

Therneau T (2021) survival: A package for survival analysis in R. R package version 3.2-13, accessed December 07, 2021, https://cran.r-project.org/package=survival.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S* (New York: Springer), fourth edition, URL https://www.stats.ox.ac.uk/pub/MASS4/, iSBN 0-387-95457-0.

Wright M, Ziegler A (2017) ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77(1), https://doi.org/10.18637/jss.v077.i01.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California