Modelling attrition in the Eastern Cape public health system using multilevel survival analysis and machine learning methods

Author: Cailin Perrie (PRRCAI001)

Minor Dissertation presented in partial fulfilment of the requirements for the degree of

Master of Science (Data Science)

in the Faculty of Statistical Sciences at the University of Cape Town

Supervisor: Sheetal Silal

Co-supervisor: Sebnem Er

October 13, 2023

Plagiarism Declaration

I, Cailin Perrie, know the meaning of plagiarism and declare that all of the work in the minor dissertation, save for that which is properly acknowledged, is my own.

Signed by candidate

Signed

13/10/2023

Date

Abstract

The size of South Africa's public health workforce is influenced by many factors including, but not limited to, inter-facility transfers, emigration, voluntary exits, illness, death and retirement. Understanding the rate at which public health workers exit or move within the public health system (*i.e.* the attrition rate), is essential for adequately formulating effective workforce policies and strategies. South Africa's public health system budget currently accounts for an annual 5% attrition rate for health facilities in general. This rate does not consider fluctuations in attrition rates between cadres, across facilities, or across districts. Presently, there are no guidelines or models for predicting attrition within the Eastern Cape (EC) public health care system from an individual, cadre, facility, or district level. As a result, staffing levels are determined entirely by the discretion of facility or departmental managers.

The purpose of this investigation was, therefore, to explore and utilize human resource (HR) data within South Africa's public healthcare system, with specific focus on the EC province, to predict attrition rates within and across cadres, health facilities, and districts. The study places a large focus on using the findings of the study to improve budgeting and health care staffing levels. The study thus aims to develop predictive models that are capable of handling data that is hierarchical in nature, use these models to identify level specific factors that both negatively and positively impact annual attrition rates, and compare predictive models to determine the most effective model for predicting attrition rates in the EC public health sector. The study further aims to perform a historical data analysis on the HR data to identify areas of high concern regarding attrition.

Based on a preliminary and historical exploratory data analysis (EDA) of the EC province's public heath HR data, the annual attrition rates between 2010 and 2020 have consistently exceeded this budgeted 5%, with the annual attrition rate in some years reaching as high as 15.65%. The preliminary analysis further indicated that attrition rates are subject to high variation when computed at different levels (*i.e.* cadre and facility level groupings) as well as across different years. Consequently, the Eastern Cape Department of Health (EC DOH) have been historically and holistically under budgeting for attrition. Additionally, by catering for attrition at a provincial level only, the department has been neglecting the effects that within and between-group variation in attrition has on budget formulation.

The historical EDA further identified several cadres that consistently experienced high levels of attrition namely, the `Medical Services`, `Nursing`, and `Primary Health Care` cadres. The job titles that fall within these cadres (*i.e.* Medical Specialists, Clinic Specialists, and Nurses) are considered

critical to the functioning of any health facility as they are responsible for providing medical care to patients. The historically high attrition levels obtained in these cadres are, therefore, alarming as they suggest that the EC province can expect to consistently see the same or a degrading level of patient care in the years to come.

The findings from the historical EDA, and the potential risks associated with over or under-budgeting for attrition, suggest that there is a financial incentive for the EC DOH to develop models capable of accurately predicting future attrition rates within and between multiple levels within the EC province. The application of both statistical and machine learning (ML) modelling techniques were thus explored in this investigation, however, only one statistical modelling method (multilevel discrete-time event models) and three ML modelling methods (multi-layer perceptron neural networks, generalized linear mixed-model trees, and tree-based mixed effect models) were explored. This was due to their potential ability to handle and, effectively model, the complex multilevel and longitudinal HR data available for use in this study.

Unfortunately, all multilevel machine learning models explored failed to converge, resulted in excessive computational time forcing an abort, or simply resulted in poor model performance when evaluated on unseen data. Based on these findings, and within the limitations of the study scope, it is accepted that these three modelling methods are unable to outperform traditional multilevel statistical methods at this time. The multilevel discrete-time event models, however, are able to handle the complex data used in this investigation. Based on model performance metrics, the best multilevel discrete-time event model developed in this investigation is considered feasible for use in attrition prediction for the EC DOH. The model is further capable of being used to determine time-indicator and healthcare worker level variables influencing attrition. Overall, the insights gained from this investigation can be used to help guide intervention planning, optimize HR capacity planning processes and, in turn, improve overall budgeting for the EC health system.

The findings and limitations of this investigation, however, open up opportunities for future work both as improvements to, or extensions of, the data preparation processes as well as model formulations and optimizations. Such follow-up work may include the exploration of different attrition definitions and the impact that has on the investigations findings, exploring methods for reducing HR healthcare data integrity issues, and provisioning or implementing re-sampling techniques, different cadre grouping strategies, or virtual machines to improve the performance of the machine learning models proposed.

Contents

List of Tables

List of Figures

Nomenclature

AAH: Aalen's Additive Hazard

AFT: Accelerated Failure Time

AIC: Akaike Information Criterion

ANN: Artificial Neural Network

BH: Baseline Hazard

$CHAI$: Clinton Health Access Initiative

CPH: Cox Proportional Hazard

CV: Coefficient of Variation

DOH: Department of Health

DT: Decision Tree

EC: Eastern Cape

EDA: Exploratory Data Analysis

FM: Fuzzy Matching

$GLMM$: Generalised Linear Mixed-Model

$HGLM$: Hierarchical Generalised Linear Model

HLM: Hierarchical Linear Model

HR: Human Resources

ICC: Intra-Class Correlation

LR: Logistic Regression

LRT: Likelihood Ratio Test

LVQ: Learning Vector Quantization

$MDTE$: Multilevel Discrete Time-Event

ML: Machine Learning

MLP: Multilayer Perceptron

MSE: Mean Squared Error

NN: Neural Network

$PEDA$: Preliminary Exploratory Data Analysis

PPH: Parametric Proportional Hazard

PWE: Piecewise Exponential

RF: Random Forests

SME: Subject Matter Expert

$TBME$: Tree-Boosted Mixed Effects

VM: Virtual Machine

XGB: Extreme Gradient Boosted

1 Introduction

1.1 Problem context

The size of South Africa's public health workforce is influenced by many factors including, but not limited to, inter-facility transfers, emigration, voluntary exits (*e.g.* movement to other sectors of employment), illness, death and retirement. Understanding the rate at which public health workers exit or move within the public health system - otherwise known as the attrition rate - is essential for adequately formulating effective workforce policies and strategies. South Africa's public health system budget currently accounts for an annual 5% attrition rate for health facilities in general (Castro-Leal et al., 2000). This rate does not consider fluctuations in attrition rates between cadres, across facilities, or across districts.

Based on a preliminary, exploratory data analysis of the Eastern Cape (EC) province's public heath human resource (HR) data, discussed in Section 4.2.1, the annual attrition rates between 2010 and 2021 have consistently exceeded this budgeted 5%, with the annual attrition rate in some years reaching as high as 15.65%. The preliminary analysis further indicated that attrition rates are subject to high variation when computed at different levels (*i.e.* cadre, facility, district).

According to Castro Lopes et al. (2017), high workforce attrition leads to i) a large loss of public resources due to the additional expenditure on education and training of health workers, ii) worsening working conditions for the remaining workforce due to the increase in workload as a result of the reduction in capacity, iii) a decrease in the projected supply of health workers that South Africa requires in order to meet the growing population need for health care, and iv) an inability to meet health care demands due to inadequate staffing levels.

The preliminary analysis, in addition to research findings, suggest that there is a financial incentive for South Africa to reassess its existing policies regarding the use and computation of public health workforce annual attrition rates in budgeting considerations.

1.2 Problem statement

Presently, there are no guidelines or models for predicting attrition within the EC public health care system from an individual, cadre, facility, or district level. As a result, staffing levels are determined entirely by the discretion of facility or departmental managers.

This study proposes to perform an historical data analysis on the HR data to identify areas of high concern regarding attrition rates, as well as develop statistical and machine learning models for prediction of future attrition events within and across the EC public health system. The historical analysis findings and models developed aim to assist the EC Department of Health (EC DOH) in the strategic planning process by providing insights into historical areas of concern with regards to attrition, factors influencing attrition, and expected future attrition rates across cadres, facilities and districts in the EC. The insights gained from this investigation are to be used to improve budgeting and health care staffing levels and identify areas that require further investigation or intervention to reduce high attrition.

This investigation is to be conducted for, and in collaboration with, the Clinton Health Access Initiative (CHAI) South Africa.

1.3 Scope and objectives

The following objectives are pursued in this project:

I. To *conduct* a thorough review of literature pertaining to:
   i. Attrition, and
   ii. Existing methods for analysing and predicting future workforce attrition, in particular, statistical and machine learning models.

II. To *conduct* a thorough review of methodology pertaining to:
   i. Basics of survival analysis and survival data,
   ii. Univariate, multivariate, and multilevel-multivariate survival analysis and statistical modelling,
   iii. Hazard and event prediction using multilevel-multivariate statistical modelling, and
   iv. Hazard and event prediction using supervised machine learning algorithms.

III. To *conduct* a preliminary exploratory data analysis and data preprocessing to:

   i. Obtain and explore the raw data provided by CHAI, and

   ii. Perform data preprocessing and cleaning to transform data into a format suitable for model development.

IV. To *conduct* an historical exploratory data analysis to:

   i. Gain an understanding of historical attrition rates within and across the EC public health system, with the primary aim of gaining insight into possible trends or areas of concern with respect to workforce attrition.

V. To *develop* statistical and machine learning models capable of:

   i. Handling the complexity of the data utilised in this investigation,

   ii. Accurately predicting future attrition events for possible groupings (*i.e.* cadres, facilities, districts), and

   iii. Identifying influential factors impacting attrition.

VI. To *determine* the most effective model for attrition prediction, within the context of the problem statement:

   i. Compare the performance and limitations of the developed statistical and machine learning models for predicting attrition rates in the public health sector,

   ii. Identify the most appropriate model for future use, for the problem at hand, and

   iii. utilise the best proposed model to gain insight into predicted future attrition rates and areas of concern.

VII. To *recommend* sensible follow-up work related to the work in this project, which may be pursued in the future.

1.4 Research methodology

The methodology followed in this study, in order to achieve the study aim and objectives set out in Section 1.2-1.3, is as follows:

I. *Consult* and *review* existing literature on the definition of, and mathematical formulations for, attrition so as to inform the choice of attrition definition and mathematical formulation used in this investigation.

II. *Consult* and *review* existing literature on the methods that have proven to be successful for the prediction of attrition-like events, specifically for datasets

that may contain longitudinal and multilevel data. This will inform modelling methods that are to be explored in the context of this investigation.

III. *Consult* and *review* existing literature on the methodology and implementation of the modelling methods identified for exploration.

IV. *Perform* preliminary data analysis to understand the datasets available for use in this investigation and the complexities thereof, so as to inform the process of preparing, preprocessing, wrangling, and cleaning the datasets into a form suitable for the development of the modelling methods explored.

V. *Perform* historical attrition analysis on the cleaned and preprocessed dataset, so as to answer pertinent questions from stakeholders regarding average historical attrition rates and possible areas of concern. This will determine any underlying trends in the data, possible influential factors with regards to attrition, as well as indicate any correlation that may be present in the dataset variables.

VI. *Determine*, from the type of data available, the modelling methods to be applied for attrition prediction and influencing factor determination.

VII. *Develop* baseline statistical and machine learning models and evaluate performance using goodness-of-fit tests and evaluation metrics. Thereafter, iteratively *develop* more complex models and assess performance against the previously developed baseline models.

VIII. *Compare* the statistical and machine learning prediction models developed, to determine the most effective model for predicting attrition rates in the public health sector, for the cleaned and preprocessed dataset utilised in the investigation.

IX. *Discuss* parallels between model findings, historical exploratory data analysis findings, and findings obtained from the thorough review of literature and methodology.

X. *Discuss* limitations of the investigation and possible future work.

1.5 Dataset considerations

The HR data, utilised in this investigation, is schematically represented in Figure 1.

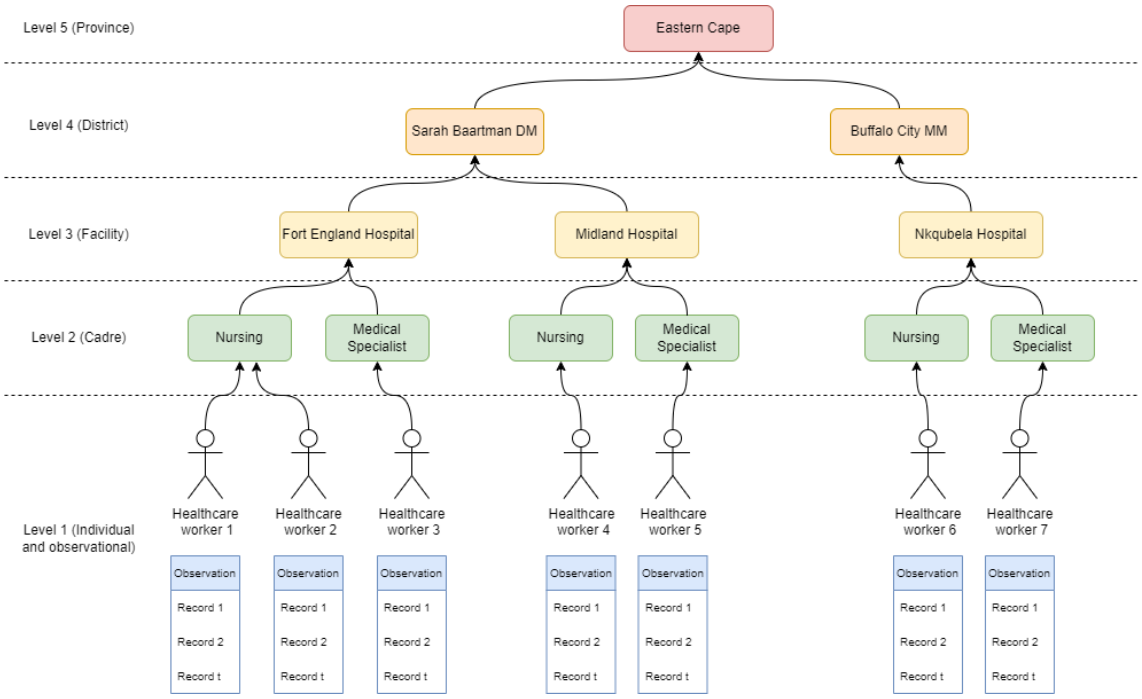

Figure 1: Schematic representation of the nature of the data utilised in this investigation.

The schematic representation of the data, described in Figure 1, implies that a form of grouping structure exists in the dataset, whereby different level units are nested in one (and only one) higher level unit. For example, `Healthcare worker 1`, who currently works as a nurse in `Fort England Hospital` in the `Sarah Baartman DM` district cannot simultaneously work as a nurse in a different hospital and district. This type of data grouping is synonymous with a multilevel data structure. Figure 1 further indicates that the dataset structure is longitudinal in nature. This is evident by the capturing of repeated observations, for the same healthcare workers, over an extended period of time (*i.e.* time period `1` to `t`). Consequently, this investigation aims to explore literature and modelling methods that are capable of handling data that is longitudinal and, potentially multilevel, in nature.

2 Literature review

In this chapter, literature pertaining to the definition, mathematical formulation, and potential causes of attrition is explored. This is followed by a discussion of two overarching methods, statistics and machine learning, that have been used in applied research to predict attrition and attrition rates, as well as determine factors influencing attrition.

2.1 Attrition

According to Castro Lopes et al. (2017), *attrition* is broadly defined as any exit from the workforce. Consequently, *attrition rate* is the pace at which employees leave an organisation. According to Eysenbach et al. (2005), there are four types of attrition, namely:

1. Voluntary attrition: occurs when an employee makes a decision to leave their organisation on their own accord.
2. Involuntary attrition: occurs when an organisation chooses to part ways with an employee and, consequently, relieves the employee of their job duties and responsibilities. Examples include termination due to poor performance, and/or behavioural problems.
3. Internal attrition: occurs when an employee moves between roles, positions, and/or departments within their organisation. This type of movement is considered a form of attrition because as an employee leaves their current position, that role becomes vacant, leading to position-based turnover. This is usually a positive form of attrition as it often indicates employee growth in the form of promotions.
4. Demographic-specific attrition: occurs when an entire group of specific employees leave an organisation at the same time. The group may consist of employees of the same age, ethnicity, gender, qualification, or position.

As discussed in Section 1.1, understanding the rate at which public health workers exit the public health system (*i.e.* attrition rate) is essential in adequately formulating effective workforce policies and strategies. Castro Lopes et al. (2017) conducted a rapid review[1] of studies published between 2005 and 2017 on attrition rates of health workers to gain an understanding of, i) the definition of attrition used in applied research regarding healthcare workforce attrition, and ii) to establish popular

[1]Rapid Review: is a review that simplifies or omits components of the systematic review process in order to produce information that is critical for answering questions, or making decisions, that are time sensitive Castro Lopes et al. (2017).

methods for calculating healthcare workforce attrition.

The objectives of the 51 studies included in the review ranged from i) forecasting the needs of future workforce availability based on current workforce availability, ii) internal and external migration of the workforce, iii) attrition within specific training/health programs, and iv) retention of health workers (factors and levels) (Castro Lopes et al., 2017). 57% of the studies took place at the national level, using national data from census, council registers or the department of health. 41% of the studies took place at the sub-national level and thus focused on states, provinces, districts, rural/remote areas, health facilities and education institutions (Castro Lopes et al., 2017). Moreover, the most featured cadres in the studies included doctors, nurses (registered and enrolled nurses, licensed practical nurses, nurse assistants), midwives, and community health workers. Additional, but not as prevalent, cadres noted in the studies included clinical officers, lab technicians, pharmaceutical staff, and healthcare aides.

The rapid review, conducted by Castro Lopes et al. (2017), identified that only half of the studies provided a full definition of attrition. The word attrition was also frequently used interchangeably with the terms *drop-outs*, *turnover*, *brain-drain*, *losses*, *premature departure*, and *separation* (Castro Lopes et al., 2017; Eysenbach et al., 2005). In many studies, the definition of attrition was also extended to include exiting from the workforce due to retirement, death, and migration (Eysenbach et al., 2005). Consequently, the review suggests that there is a diversity of definitions of healthcare workforce attrition in applied research. Additionally, the types of attrition modelled or investigated were not defined or clarified.

The rapid review further identified the lack of consistency regarding the computation of attrition and attrition rate. Attrition rate estimates were provided for different periods of time, ranging from 3 months to 12 years, using different calculations and data collection systems (Castro Lopes et al., 2017). Despite this inconsistency, however, the annual attrition rate was the most common and the only comparable measure.

The mathematical formulation for this annual attrition rate is defined as follows:

$$A_{rate}(n) = \frac{Z_n}{\left(\frac{X_i + X_n}{2}\right)} \times 100 \tag{1}$$

where $A_{rate}(n)$ is the attrition rate for the system in year n, Z the number of employees that left the system by the end of year n, and X the number of employees in the system at the start of year i and end of year n respectively. In other words, the annual attrition rate is the number of employees that leave the system in year n,

divided by the average number of employees in the system in year n (Castro Lopes et al., 2017; Eysenbach et al., 2005). A few of the studies noted the possibility of utilizing Formula 1 to determine attrition rates per health facility or per cadre of workers per facility. Consequently, healthcare workforce attrition rate estimates seem to be influenced by the purpose and type of study.

The research outcomes from the rapid review suggests that the lack of internationally comparable definitions and guidelines for measuring attrition from the healthcare workforce makes it very difficult for countries to identify the main causes of attrition and to develop and test strategies for reducing it (Castro Lopes et al., 2017). Methods and measures for computing healthcare workforce attrition is, therefore, determined by the data available for use in determining attrition rates and the outcomes of importance or relevance to the decision makers.

2.2 Methods for analysing workforce attrition

As identified in Section 2.2, the term *attrition* is often used interchangeably with the terms *drop-outs*, *turnover*, *brain-drain*, *losses*, *premature departure*, and *separation*. As a result, the literature review search domain was extended to include studies that focused on analysing one or more of these terms as the outcome of interest.

Two overarching methods for analysing attrition/turnover/drop-out rates were identified from the review of literature; namely i) the use of statistical models for multilevel survival analysis, and ii) supervised machine learning models for outcome prediction.

2.2.1 Statistical models for multilevel survival analysis

Two types of statistical models for multilevel survival analysis have been used in applied research to predict attrition/turnover/drop-out and ultimately determine attrition/turnover/drop-out rates. These include Multilevel Cox Proportional Hazards models and Multilevel Discrete-Time Event models. Both of these models cater for data that is hierarchical (multilevel) and longitudinal in nature, contains both categorical and continuous explanatory variables, and whose outcome variable of interest is a time-to-event variable. Examples of data with a multilevel structure include:

- Students that are nested in a school, which is nested in a community;
- Patients that are nested in a hospital, which is nested in a district.

In such instances, subjects who are nested within the same higher level unit within the hierarchy are likely to have outcomes that are correlated with one another,

which traditional univariate and multivariate survival analysis methods are unable to handle.

A review of literature pertaining to the application of these methods for attrition/turnover/drop-out prediction is explored in Sections 2.2.1.1-2.2.1.2.

2.2.1.1 Multilevel Cox proportional hazards model for attrition prediction

According to Finch et al. (2009), a Multilevel Cox Proportional Hazards model catering for survival data was utilised to identify factors influencing the rate of student movement, both within and out of state charter school systems. This model catered for the multilevel nature of the data by incorporating individual and facility level predictor variables in the model formulation.

The response variable (outcome variable of interest) was the time until a student exited a charter school or until they were censored (*i.e.* the study period ended with the student remaining in a charter school). Enrollment was assessed twice a year over a period of six years. Consequently, a time-indicator variable was engineered as the number of periods that a student was enrolled in a school. Individuals who survived in the system until they completed the highest grade available at the school and did not exit the system by the end of the study were labelled as *non-leavers*, while those who exited the system before completing the highest grade available were labelled as *leavers* (Finch et al., 2009).

The Cox model with mixed effects (multilevel Cox proportional hazards survival model) allows for the inclusion of both categorical and continuous independent variables at all levels of the data hierarchy. Consequently, the individual-level variables used in data analysis included; gender, race (2 level factor variable), free/reduced lunch status (2 level factor variable), special education status (2 level factor variable), and eligibility for Title I funded programs (2 level factor variable). The school level variables included the student-teacher ratio, the average experience (in years) and the average attendance rate (in days) (Finch et al., 2009). Since the variables available for use in the study were obtained at both the 2 different levels in the school system (*i.e.* the school and individual levels), a model capable of handling multilevel data was required to ensure that the clustering of individuals, within their respective schools, was accurately catered for in the calculations of the standard errors (*i.e.* avoided the bias in standard errors) (Finch et al., 2009; Singer et al., 2003).

Parameter estimation, for the purpose of this study, was conducted using the profile likelihood method (Finch et al., 2009). A hazard ratio for the variables (within subjects and between schools) was then determined from the parameter estimate. The hazard ratio for the categorical variables represented the relative likelihood

of leaving, in one category of predictor variable, relative to a reference category. Conversely, the hazard ratio, for continuous predictor variables, represented the change in the likelihood of leaving the school prematurely for each 1 unit increase in the predictor variable.

The findings of this investigation suggest that an exit event from the school system is influenced by several variables, namely an individual's initial test scores when first entering the a school, an individuals race group, and whether or not the individual participated in Title I funded programs (Finch et al., 2009). It was further identified, that individuals were less likely to exit the system before fully completing their academic levels at schools with more experienced teachers(Finch et al., 2009).

2.2.1.2 Multilevel discrete time event models for attrition prediction

The second type of statistical model, prevalent in literature and applied research, for analysing multilevel survival data is the Multilevel Discrete-Time Event models. A research paper that made use of this model type to predict attrition and determine factors influencing attrition (or like terms) is explored.

Guillory (2008) conducted a study with the aim of investigating college student retention. Due to the longitudinal and hierarchical nature of retention data, it is rather difficult to identify factors influencing these retention rates with traditionally simple statistical methods (Barber et al., 2000). As a result, this study aimed to explore the application of multilevel statistical methods to student retention problems, with a primary focus on identifying their efficacy with respect to determining factors influencing retention. Consequently, a multilevel discrete-time hazard model was developed to analyze individual and school-level factors that effected the risk of a student leaving a university. According to this study, the student attrition rate, in tertiary education, is defined to be the percentage of students who exit the system before completing a semester at a university and do not return to the university the next semester (Guillory, 2008).

Three main objectives of the study included exploring the likelihood that a student left a university during a year, exploring what individual level factors were most influential in a student's decision to leave a university during a year, and exploring the extent at which the type of school a student attended effected the risk of a student leaving a university during a year.

Gender, ethnicity and school-type were used to model the timing of students leaving a university from a cohort of first-time freshmen over a five year period (Guillory, 2008). The discrete-time intervals used in the analysis spanned 1 year. The variables used in the study are described below:

- Outcome variable: A discrete variable that indicated whether or not an indi-

vidual experienced an attrition event, when assessed at the beginning of the year (*i.e.* was enrolled, or not enrolled).

- Student-level variables: Gender, ethnicity, age, high school GPA, residency, and duration of enrolment.
- School-level variables: school type (*i.e.* public or private)

This study utilised the random-shape baseline hazard model to determine the effect of the predictor variables (both level one and level two) on student retention in tertiary education. The addition of random effects to the baseline hazard allows for additional flexibility (Barber et al., 2000). The data used in this study was obtained from the *National Longitudinal Survey of Youth 1997*, conducted by the U.S. Department of Labour (Guillory, 2008). This survey contained data about the labour market and consisted of 8,984 records comprising of both individual and school level factor variables (Guillory, 2008).

Results indicated school type, a level two variable, was significant. This suggests that the risk associated with attrition differs depending on the school in which a student resides. It was further identified that students enrolled in private universities had a higher risk of exiting the system and not returning the next year. One of the key findings of this study was that students who left the system early, did so because they failed to integrate or adapt to the university environment. A students ethnicity and gender were also identified to influence the risk of an individual prematurely exiting a university (Guillory, 2008).

2.2.2 Supervised machine learning models for attrition prediction

Three types of supervised machine learning (ML) models have been used in applied research to predict attrition and ultimately determine attrition rates. These include Neural Networks (NN), Generalised linear mixed-model (GLMM) trees, and Extreme Gradient Boosted (XGB) trees with mixed effects.

A review of literature pertaining to the application of these methods for attrition prediction is explored in Sections 2.2.2.1-2.2.2.3.

2.2.2.1 Neural network paradigms for employee turnover prediction

Two neural network (NN) paradigms, multilayer perceptron (MLP) and learning vector quantization (LVQ), were used to investigate voluntary employee attrition amongst a sample of 577 healthcare workers, in a specific hospital. The study aimed to assess whether or not artificial neural networks offered greater predictive accuracy over traditional logistic regression methods (Somers, 1999).

According to Somers (1999), accurately predicting employee turnover has proven to be a difficult problem, and calls for new methods and new directions have been an ongoing theme in employee attrition research dating back to the 1980's. Consequently, this study was based on the premise that NNs offered a potential improved solution to employee attrition prediction.

The data presented in this study was analyzed using three modelling techniques, namely, logistic regression (LR), MLP NN, and LVQ NN. The LR model was used as a baseline model to assess the performance of the NN models developed. The MLP that provided the best results took the form of a 4-3-1 architecture. This architecture describes a NN with 4 neurons in the input layer representing each input variable, three in the hidden layer, and 1 in the output layer representing the probability of a turnover event occuring. Seeing as the outcome variable of interest was a probability of an event occuring, the activation function and loss function chosen was the Sigmoid Function and Mean Squared Error (MSE), respectively (Somers, 1999). The LVQ model architecture included 4 neurons in the input layer, 50 neurons in the hidden layer, and 2 neurons in the output layer representing *leavers* and *stayers*. Since the LVQ model aimed to predict a discrete outcome, the activation and loss function chosen for use was Tanh and cross-entropy, respectively (Somers, 1999).

According to the evaluation statistics, the LR model yielded an overall correct classification accuracy of 76%, with the MLP and LVQ models obtaining a classification accuracy of 88% and 84%, respectively. Although the LVQ network obtained an overall correct classification rate that was lower than the MLP model, this model was able to correctly predict 87% of *stayers* and 77% of *leavers*, and was therefore the only model able to accurately classify both classes when predicted on the test dataset. This was a significant improvement on the LR model which was only able to correctly identify 1% of *leavers* (Somers, 1999). One major implication of this study concerns the effective management of employee attrition as results from the LVQ model suggest that aggregate turnover levels can be estimated far more accurately than traditionally applied methods with a very small set of classifiers.

According to Somers (1999), the prediction of an employee turnover event can also be defined from the angle of a survival event, whereby the outcome variable of interest can be defined as a death event or defined as the probability of experiencing a hazard event. Consequently, the application of NNs can, and have, also been extended to cater for continuous and discrete-time survival prediction (Kvamme and Borgan, 2019; Gensheimer and Narasimhan, 2019) where data is in longitudinal format (*i.e.* repeated measures for each individual in the sample or population). Although these models have also proven to be effective in the prediction of such events, discrete-time survival NNs do not seem to have been previously applied to employee turnover problems. Moreover, to my knowledge, there is also no research

indicating the efficacy of these models for multilevel survival data.

2.2.2.2 Generalised linear mixed-model trees for attrition prediction

The second type of supervised machine learning model, prevalent in literature and applied research, for analysing not only survival data but also multilevel data is the flexible decision-tree method known as the Generalised Linear Mixed-Model (GLMM) Tree. One research paper, making use of this method for event prediction, is explored. Although the outcome of interest in this study is not an attrition event, the study looked at determining the risk of an individual experiencing one of two events, whilst nested in a specific facility. This draws parallels to the employee attrition problem that aims to predict a discrete event based on individuals who are nested within one or more higher level units, and is therefore explored in this literature review.

Fokkema et al. (2021) conducted a study with the aim of investigating baseline patient characteristics that could be used to predict treatment outcomes for individual patients. GLMM trees were applied to a dataset containing 3256 unique individuals who were receiving treatment at one of several mental-health facilities in the UK. The discrete event treatment outcomes were regressed on 18 predictor variables, which were demographic, case, and characteristic level variables (Fokkema et al., 2021). A baseline model containing only time-indicator variables was first developed and acted as a baseline model for further comparison. Additional comparative models were also developed with the aim of assessing the performance of GLMM trees over more traditionally applied techniques such as Generalised Linear Mixed Models (GLMM) and Random Forests (RF). In order to effectively assess model performance, all models developed were trained on a training dataset and tested on a test dataset, with an 80:20 training-test dataset split (Fokkema et al., 2021).

Results indicated that the GLMM trees yielded modest predictive accuracy with the cross validated multiple R [2] values of 0.18 and 0.25. Furthermore, it was found that the predictive accuracy for the GLMM and random forest models did not differ significantly from the predictive accuracy obtained for the GLMM tree model. However, the GLMM tree required far fewer predictor variables to obtain this similar accuracy. This study further identified that GLMM trees can also be utilised to determine complex nested structures in datasets with more than 2 nested structures (Fokkema et al., 2021). One big limitation of this ML modelling technique, identified in this paper, is that these models can become increasingly complex when the dataset contains many predictor variables of type factor, that contain many levels. In such cases, it becomes increasingly difficult for the model to detect effective splits, resulting in the models failure to converge. This is further exacerbated in cases when

[2]Multiple R: the absolute value of the correlation coefficient (Fokkema et al., 2021).

these predictor variables are also nested in a large number of levels (Fokkema et al., 2021).

2.2.2.3 Extreme gradient boosted trees for attrition event prediction

In the study entitled, *Predicting Employee Attrition using XGBoost Machine Learning Approach*, by Jain and Nayyar (2018) a novel model for predicting employee attrition using Extreme Gradient Boosted (XGB) Trees, a highly robust machine learning method, is proposed. Additional comparative models were also developed with the aim of assessing the performance of XGB trees over more traditionally applied techniques such as LR or basic decision trees (DT). In order to effectively assess model performance, all models developed were trained on a training dataset and tested on a test dataset, with a 75:25 training-test dataset split.

The performance of the proposed method was validated using six evaluation metrics: accuracy, precision, recall, F1-score, receiver operating characteristic curve, and area under the curve. The XGB classifier displayed the best classification performance amongst the three algorithms used in this study. The XGB Tree model had the highest specificity (*i.e.* lowest Type-I error) and highest recall (i.e lowest Type-II error) when compared to the LR and DT models. It is important to note, however, that the highest recall obtained was 16.98% (Jain and Nayyar, 2018). The model was thus able to predict a high correct number of non-events, but often predicted non-events as attrition events. These results, however, when compared to past attrition prediction models are considered an improvement.

Although these models have proven to be rather successful for attrition prediction, limited research has been undertaken to apply these methods to multilevel or longitudinal survival data (Jain and Nayyar, 2018) or similar problems with such complex data. Upon further investigation, a novel methodology, combining tree-boosting with mixed-effects models, has been proposed as an improvement on XGB Trees (Liu et al., 2022). These Tree-Boosted Mixed Effects (TBME) models are designed to cater for survival data that is multilevel and longitudinal, with the intention of predicting an outcome event of interest or outcome probability. This methodology, however, has only been applied to very small samples of data with only two nesting levels (Liu et al., 2022).

2.3 Chapter summary

The review of literature focused on two of the study's main objectives. Firstly, it explored the myriad of attrition definitions, mathematical formulations, and potential causes of attrition that currently exist in applied research (Section 2.1). Secondly, it explored modeling techniques that have proven to be successful for the prediction of attrition-like events in applied research (Section 2.2.1-2.2.2).

Based on this literature review, there is a lack of internationally comparable definitions and guidelines for defining and measuring attrition. This has, historically, made it very difficult for researchers to identify not only the most appropriate definition and formulation of attrition, but also made it difficult to determine the main causes of attrition and to develop and test strategies for reducing it (Section 2.1). Consequently, a key finding from the review of literature is that methods and measures for computing healthcare workforce attrition is determined by the data available for use and the outcomes of importance or relevance to the decision makers (Section 2.1). In the context of this investigation, attrition rates need to be factored into the strategic planning process for effective budgeting and workforce planning. Most of this planning is performed annually. Consequently, annual attrition rates are of most concern for this investigation.

The review of literature also focused on the application of statistical and machine learning methods for predicting and analysing attrition events (or like terms). The statistical methods explored were able to effectively determine the likelihood of a future attrition event as well as determine factors influencing attrition, for data that was both longitudinal and hierarchical (Section 2.2.1). These studies further indicate that HR data often utilised for attrition prediction, comprises of a multilevel structure, and if assessed overtime, is also longitudinal. Methods capable of handling this type of data must therefore be given consideration for attrition prediction problems (Section 2.2.1).

Moreover, the ML algorithms explored in this review of literature are considered to outperform traditional statistical methods when it comes to highly complex and possibly non-linear data (Section 2.2.2). Although these models have proven to be rather successful for the prediction of attrition events (or like terms), limited research has been undertaken to apply these methods to multilevel or longitudinal survival data or similar problems with such complex data. The possibility of their application to large, multilevel and longitudinal (survival) datasets, however, is proposed (Section 2.2.2).

3 Methodology

Based on the thorough review of literature, discussed in Chapter 2, there are a multitude of different statistical and ML methods that have been applied to attrition-like prediction problems. In some cases, these models had to make use of HR data that was both longitudinal and multilevel, and contained an outcome variable of interest that was a time-to-event variable. Time-to-event outcome variables are often considered an indication that the dataset also contains survival data. Consequently, it was important to gain an understanding of what survival data and survival analysis is and how it influences the modelling methods identified during the review of literature.

Consequently, this chapter includes an exploration of the theory behind survival analysis and explores univariate, multivariate, and multilevel-multivariate survival analysis and statistical modelling techniques. The chapter further explores hazard and event prediction using multilevel-multivariate statistical modelling, as well as supervised machine learning algorithms.

3.1 Survival analysis

Survival analysis is one of the most common statistical techniques employed to assess the *time to an event* of interest, such as time from birth until death, time until relapse of a disease, or time from entry into a clinical trial until tumor response (Kleinbaum et al., 2012).

Survival analysis is, therefore, a series of statistical methods that aim to analyze data that includes an outcome variable of interest that is a time to event variable (*i.e.* the outcome variable has both an event and a time value associated with it). This differs from a typical regression problem where the outcome variable is often continuous (*e.g.* housing price) or a classification problem where the outcome variable is categorical (*e.g.* Class I or Class II) (Aalen et al., 2008).

According to Aalen et al. (2008), the characteristics of survival data include:

i) Appropriate definition of the time of origin for each study subject (*i.e.* time since entry into study),

ii) appropriate definition of the end event or failure event (*i.e.* death, relapse, tumor response time),

iii) study subjects should be comparable at their time of origin (*i.e.* every enrolled individual is to be followed from a baseline date until the end event or termination of study), and

iv) survival data can never be negative. Time is a positive value that may be in hours, days, weeks, months, or years from the time of origin until an event occurs.

There are numerous tests and models used in survival data analysis. They all aim to identify and test how explanatory variables predict an outcome variable that measures the time until an event. Additionally, they are all based on concepts that are central in any time-to-event analysis, namely censoring, survival functions, the hazard function, and cumulative hazards. These concepts are explored in Sections 3.1.1-3.1.2.

3.1.1 Censoring

As identified in Section 3.1, survival analysis involves the consideration of the time between a fixed starting point (*e.g.* diagnosis of cancer) and a terminating event (*e.g.* death). The key feature that distinguishes survival data from other types is that the event may not have necessarily occurred in all individuals by the time the study terminates, and consequently, the full survival times for these individuals are unknown (Klein and Moeschberger, 2003). In such instances, these individuals are said to be "censored." There are 3 types of censoring, namely:

- Right-censoring: occurs when the individual's true survival time is greater than or equal to their observed survival time (Klein and Moeschberger, 2003).
- Left-censoring: occurs when the individual's observed survival time is less than or equal to their true survival time (Klein and Moeschberger, 2003).
- Interval-censoring: occurs when an individuals true survival time falls within an interval of two events (Klein and Moeschberger, 2003).

These concepts are illustrated by means of Figure 2.

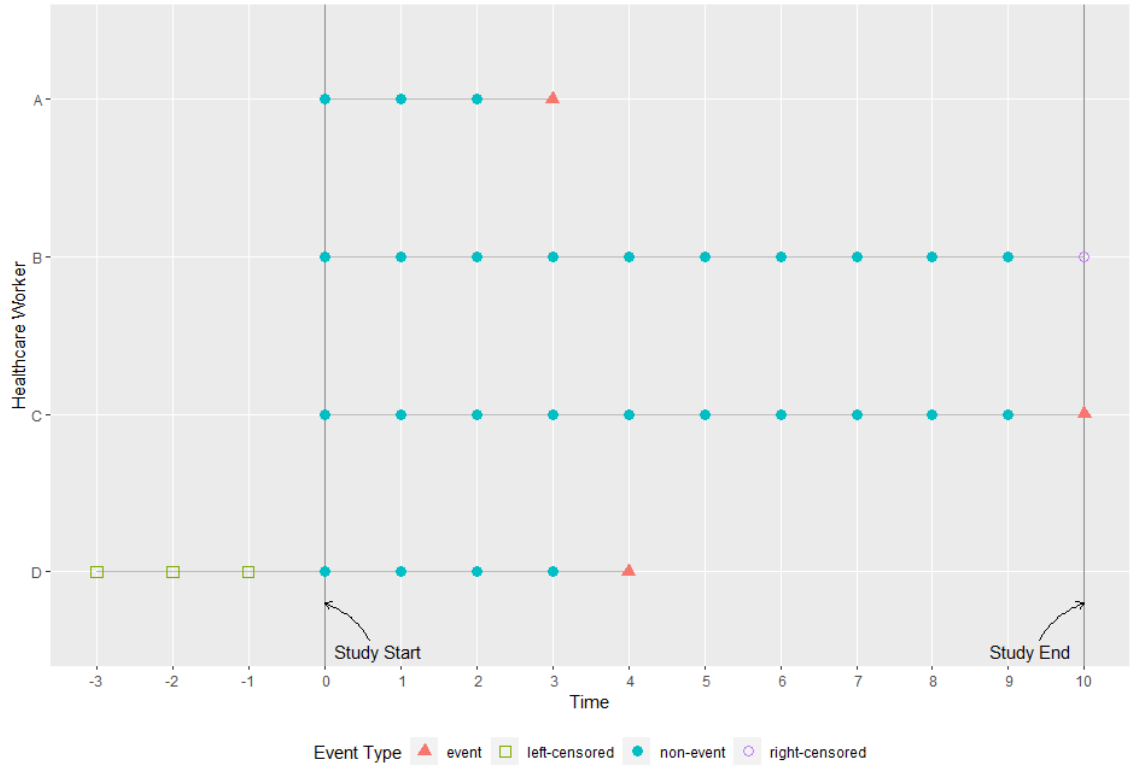

Figure 2: Schematic example of censored and non-censored observations in survival data.

As per Figure 2, healthcare worker A enters the system at the start of the study period and experiences an attrition event before the study ends. Consequently, this healthcare worker requires no censoring as their exact survival time (time until event) is known. Healthcare worker C also enters the system at the study start but manages to survive the entire duration of the study before experiencing an event. Since this healthcare worker experiences an attrition event in the last period of the study, their survival time is known, and thus no censoring is required. This, however, is not the case for healthcare workers B or D. Although healthcare worker B enters the system at the study start, they have not yet experienced an attrition event upon the completion of the study. Consequently, they survive up until at least the end of the study, but their exact survival time is unknown. For this reason, they would need to be right-censored. On the other hand, healthcare worker D, experiences an attrition event during the study period, but existed in the system prior to the start of the study. Consequently, their true survival time is longer than their observed survival time. They, therefore, need to be left-censored.

According to Kleinbaum et al. (2012), right-censoring is the most widely used form of censoring in survival analysis.

3.1.2 Survival, hazard, and cumulative hazard functions

The most important quantitative terms in any survival analysis are the survival function and the hazard function.

The **survival function**, denoted by $S(t)$, is often defined by the hazard function, and is mathematically expressed as:

$$S(t) = 1 - F(t) = Pr(T > 1) \tag{2}$$

where T is a non-negative, random variable denoting the time until a failure event, and t the specific value of interest for random variable T (Kleinbaum et al., 2012). $F(t)$ is Ts cumulative distribution function. The survival function, $S(t)$, gives the probability that a individual survives beyond time t. In other words, it is the probability that there is no failure event prior to time t (Aalen et al., 2008).

According to Kleinbaum et al. (2012), all survival functions follow two principles:

1. The function is equal to one at $t = 0$.
2. The function decreases towards zero as t tends to infinity.

The **hazard function**, denoted by $h(t)$ - also known as the conditional failure rate, the intensity function, and the force of mortality (Kleinbaum et al., 2012) - is mathematically expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{3}$$

where T is a non-negative, random variable denoting the time until a failure event, and t the specific value of interest for random variable T. $f(t)$ is Ts probability density function. The hazard function, $h(t)$, is the instantaneous rate of failure, with units $\frac{1}{t}$. It is the (limiting) probability that the failure event occurs in a given interval, conditional upon the subject having survived to the beginning of that interval, divided by the width of the interval (Aalen et al., 2008).

According to Aalen et al. (2008), the hazard function follows three principles:

1. The hazard rate can vary from zero (*i.e.* no risk) to infinity (*i.e.* the certainty of failure at that instant).

2. Over time, the hazard rate can increase, decrease, remain constant, or take on more serpentine shapes.
3. There is a one-to-one relationship between the probability of survival past a certain time and the amount of risk that has been accumulated up to that time. The hazard rate measures the rate at which risk is accumulated.

The **cumulative hazard function**, denoted by $H(t)$, is mathematically expressed as:

$$H(t) = \int_0^t h(u)\,du \tag{4}$$

and thus

$$H(t) = \int_0^t \frac{f(u)}{S(u)}\,du = -\int_0^t \frac{1}{S(u)}\left\{\frac{d}{du}S(u)\right\}du = -ln\{S(t)\} \tag{5}$$

where t is the specific value of interest for random variable T (Kleinbaum et al., 2012). The cumulative hazard function, $H(t)$, measures the total amount of risk that has been accumulated up until time t. According to Equation 5, there is an inverse relationship between the accumulated risk and the probability of survival.

Based on Equation 5, the survival function, the cumulative distribution function, and the the probability density function can be expressed as follows:

$$S(t) = \exp\{-H(t)\} \tag{6}$$

$$F(t) = 1 - \exp\{-H(t)\} \tag{7}$$

$$f(t) = h(t)\exp\{-H(t)\} \tag{8}$$

This survival probability and, consequently, the hazard and cumulative hazard probabilities, can be estimated parametrically, semi-parametrically, and non-parametrically using a variety of different methods. This can be performed on a sample of data points or on the entire population set. These methods are explored in the remainder of this chapter.

3.2 Univariate survival analysis and statistical modelling

Two univariate analyses that can be performed on survival data are the Kaplan–Meier method and the Log-Rank test.

3.2.1 Kaplan-Meier method

The survival probability can be estimated, non-parametrically, from observed survival times - both censored and uncensored - using the Kaplan-Meier (or product-limit) method (Kaplan and Meier, 1958).

Using the Kaplan-Meier method, the survival function can be estimated as follows:

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right) \tag{9}$$

where n_j is the number of alive participants just before time t_j, d_j is the number of events at t_j, $t_0 = 0$ and $S(0) = 1$ (Kaplan and Meier, 1958). The value of $S(t)$ is constant between events and, therefore, the estimated survival probability is a step function that changes value only at the time of each event. This estimator thus allows each participant to contribute information to the calculation for as long as they are known to be event free (Kaplan and Meier, 1958). If no censoring is exhibited, the estimator would reduce to the ratio of the number of individuals considered event free at time t divided by the number of people who entered the study. This is graphed, and the curve referred to as the Kaplan-Meier survival curve. The Kaplan–Meier curves indicate the outcome of interest, censoring, and number of subjects at risk or survival probability (Kaplan and Meier, 1958; Goel et al., 2010).

According to Kaplan and Meier (1958), the use of the Kaplan–Meier approach relies on the assumption that:

- censoring is independent of the likelihood of developing the event of interest, and
- survival probabilities are comparable in participants who are recruited early and later on into the study.

The Kaplan-Meier method can also be used to construct survival curves for different participant groups (Goel et al., 2010). In case of comparison of these groups, the aforementioned assumptions are also required to be satisfied for each group.

The main limitation of Kaplan–Meier estimate is that it cannot be used for multivariate analysis as it only studies the effect of one factor at a time (Goel et al., 2010).

3.2.2 Log-rank test

The log-rank test is a non-parametric test for comparing survival in two or more groups of participants. In other words, it is a hypothesis test to compare the survival distributions of two groups (Kleinbaum et al., 2012).

The log-rank test compares the observed number of events to the expected number of events for each group by computing the test statistic as follows:

$$\chi^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} \tag{10}$$

where O_i is the observed number of events for treatment group i, E_i is the expected number of events for treatment group i, g is the number of groups, and χ^2 is the test statistic (Kleinbaum et al., 2012; Peto et al., 1977). This value is then compared to a chi-squared distribution with $(g-1)$ degrees of freedom. In this way, a *p-value* may be computed to calculate the statistical significance of the differences between the complete survival curves. In essence, the log-rank test compares the observed number of events in each group to what would be expected if the null hypothesis were true (*i.e.* if the survival curves were identical) (Peto et al., 1977; Collett, 2015).

Other non-parametric tests may be used to compare groups in terms of survival, however, the log-rank test is considered the most widely used (Collett, 2015).

3.3 Multivariate survival analysis and statistical modelling

Both the Kaplan-Meier and log-rank methods, discussed in Section 3.2, are examples of univariate analyses - they describe survival with respect to a single factor under investigation and ignore the impact of any additional factors. In industry, it is more common to encounter a situation where several known variables, attributes, or covariates potentially impact the survival of a participant (Lawless, 2011). In such a case, it is often desirable to adjust for the impact of multiple factors when investigating survival. In addition, whilst the log-rank test provides a *p-value* for the differences between groups, it offers no estimate of the actual effect size; in other words, it offers a statistical, but not a clinical, assessment of the factor's impact (Peto et al., 1977).

The use of statistical models improves on these methods by i) allowing survival to be assessed with respect to several factors simultaneously, and ii) offers estimates of the strength of effect for each constituent factor (Cox, 1972). Consequently, statistical models are important and frequently used tools which, when constructed appropriately, offer valuable insights into the survival analysis process.

Several statistical methods have been proposed for modelling survival analysis data. In all cases, the models assume that all survival times are independent of one another and that censoring only occurs as right-censoring and is non-informative[3] (Cox and Oakes, 2018). The methods used to summarize multivariate survival data, that will be discussed in Sections 3.3.1-3.3.4, may be divided into two broad categories, namely semi-proportional and proportional hazard models and accelerated failure time models.

3.3.1 The Cox ('Semi-Parametric') proportional hazards model

The Cox proportional hazards (CPH) model implements survival regression - a technique that regresses covariates [4] against survival duration - to provide insight into how specific covariates influence survival duration (Cox, 1972). This model is classified as a semi-parametric model because it incorporates both parametric and non-parametric elements in its formulation (Cox, 1972; Cox and Oakes, 2018).

Mathematically, the CPH model is written as:

$$h(t) = h_0(t) \times \exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_p\} \tag{11}$$

where the hazard function $h(t)$ (as described in Section 3.1.2) is influenced by a set of p covariates $(x_1, x_2, \ldots, x_p)$ whose impact is measured by the size of their respective coefficients $(\beta_1, \beta_2, \ldots, \beta_p)$ (Cox, 1972). The term $h_0(t)$ is the baseline hazard which is estimated non-parametrically. It represents the hazard if all of the variables are equal to zero (Cox, 1972).

The CPH is, consequently, a multiple linear regression of the logarithm of the hazard on the variables, x_i, where the baseline hazard, h_0, is an intercept term that changes over time. This model, therefore, allows the hazard rate to fluctuate, as opposed to adhering to a fixed pattern (as typically seen with parametric models) (Christensen, 1987).

The model is, however, dependent on the proportional hazards assumption: the hazard of the event, in any group, is a constant multiple of the hazard in any other group (Cox and Oakes, 2018). This implies that the hazard ratios, $\exp\{b_i\}$, between groups remain constant. A value of b_i greater than zero (*i.e.* a hazard ratio greater than one) indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases (Cox and Oakes, 2018). This

[3] Non-informative: censoring is independent or unrelated to the likelihood of developing the event of interest (Cox and Oakes, 2018)

[4] covariate: is any variable that is measurable and considered to have a statistical relationship with the dependent variable (Christensen, 1987).

assumption, therefore, implies that the covariates act multiplicatively on the hazard at any point in time.

3.3.2 Parametric proportional hazard models

Parametric proportional hazard (PPH) models are a class of models similar in both concept and interpretation as the CPH model (Section 3.3.1) (Cox and Oakes, 2018). Hazard ratios have the same interpretation and the proportionality of hazards is still assumed. The key difference between the two models, however, is that for PPH models the hazard is assumed to follow a specific statistical distribution (Lin and Wei, 1989). No such constraint is enforced for CPH models since the baseline hazard function is estimated non-parametrically, and consequently, the survival times are not assumed to follow a particular statistical distribution (Cox, 1972).

As described in Section 3.1.2, there is a direct link between survival and hazard. Consequently, the choice of hazard distribution determines the distribution that the survival times are assumed to follow. Examples include Exponential, Weibull, and Gompertz PPH models (Lawless, 2014).

The main limitation of PPH models is this need to specify the hazard distribution (Lawless, 2011). Identifying and, thereafter verifying, the distribution that most appropriately mirrors that of the actual survival times is often difficult to do. However, when a suitable distribution can be found, the PPH model has been documented to be more informative than a CPH model and yield slightly more precise estimates of survival (Cox and Oakes, 2018).

3.3.3 Accelerated failure time models

The accelerated failure time (AFT) model is another type of model that may be used for the analysis of survival time data. It does not assume proportionality of hazards and can, therefore, be used as the alternative to CPH and PPH models if the constant hazards assumption is violated (Cox and Oakes, 2018; Lawless, 2011; Wei, 1992).

According to Wei (1992), the AFT model is mathematically formulated as:

$$S(t) = S_0(\phi t) \tag{12}$$

where $S(t)$ is the survival function, S_0 the baseline survival function, and ϕ is an *acceleration factor* that is dependent on the covariates as described by the formula

$$\phi = \exp\{(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)\} \tag{13}$$

where $(x_1, x_2, \ldots, x_p)$ are a set of p covariates, and $(\beta_1, \beta_2, \ldots, \beta_p)$ the respective coefficients. The effect of a covariate is to stretch or shrink the survival curve along the time axis by a constant relative amount ϕ (Wei, 1992).

The AFT model is commonly rewritten as being log-linear with respect to time (Lawless, 2011; Wei, 1992), resulting in the following mathematical formulation

$$log(T) = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon) \tag{14}$$

where ε is a measure of residual variability in survival times. Under this AFT model specification, the survival times can be considered to be multiplied by a constant effect and the exponentiated coefficients, $\exp\{\beta_i\}$, referred to as time ratios (Wei, 1992). A time ratio above one suggests that the covariate prolongs the time to the event.

AFT models require that survival times are assumed to follow a specific distribution, such as the *Log-Normal*, *Generalised Gamma*, *Log-Logistic*, and *Weibull*. It is also assumed that covariate effects are constant and multiplicative on the timescale (that the covariate impacts survival by a constant factor). They thus hold similar limitations to PPH models (Lawless, 2011, 2014).

3.3.4 Other approaches

In addition to the aforementioned statistical models for survival analysis, there are a few alternative approaches that can be considered for specific use cases. Such alternatives include *Aalen's Additive Hazard Model (AAH)*, and most recently, *ML* approaches such as classification trees and artificial neural networks (explored in Section 3.5).

Aalen's additive hazard model is a method for modelling the relationship between survival and covariates, but does so by assuming that the covariates act additively and not multiplicatively on the baseline hazard (Aalen, 1989; Lawless, 2011). Unlike the CPH and PPH models, the covariate effects are not constrained to be constant. Consequently, covariate impact is allowed to vary freely over time (Aalen, 1989). In this method, estimating the baseline hazard non-parametrically is not straightforward and, consequently, the cumulative baseline hazard is used (Aalen, 1989).

The AAH model is rather flexible, however, the coefficients are not easy to understand as they change repeatedly over time and offer no single quantifiable effect size. Additionally, Aalen plots are effectively the only useful method for determining the effect sizes. Based on the aforementioned reasons, the AAH model is not widely adopted (Aalen, 1989; Lawless, 2011).

3.4 Multilevel, multivariate survival analysis and statistical modelling

Data with a multilevel structure occur frequently across a wide range of disciplines including education, public health, and health services research. In multilevel data, a level one unit (*e.g.* employees) is nested in only one level two unit (*e.g.* hospitals). This pattern can be extended to more than two levels (*e.g.* hospitals nested in regions) (Singer et al., 2003; Snijders and Bosker, 2011). Further levels of nesting (or clustering) are possible.

Conventional statistical methods for modelling survival analysis data (such as CPH, PPH, and ART models) assume that all subjects and, consequently, survival times are independent of one another (Aalen, 1989; Cox, 1972; Lawless, 2011; Wei, 1992). However, in cases where survival data are hierarchical in nature, individuals nested in the same higher level grouping as other individuals are likely to experience outcomes that are correlated with one another. The presence of this within-cluster homogeneity violates the assumption of independent observations (Snijders and Bosker, 2011). This correlation can arise or be introduced into the system by unmeasured variables at the cluster/grouping level (*e.g.* hospital culture) or subject level (*e.g.* salary or overtime hours worked when subjects are clustered within families) (Singer et al., 2003; Snijders and Bosker, 2011).

Multilevel, multivariate regression models are able to cater for the clustering effects on the outcome variable of interest and, therefore, allow for the statistical analysis of survival data that have a multilevel structure (Goldstein, 2011). In such models, the outcome variable is evaluated at the lowest level of the multilevel structure. The predictor variables, however, can be evaluated on units at any level in the multilevel structure (Barber et al., 2000; Goldstein, 2011).

Three families of regression models for the analysis of multilevel and multivariate survival data will be discussed in Sections 3.4.2.1-3.4.2.4. A precursor to this analysis will include a brief summary on two hierarchical generalised linear models (HGLM) as these models form the basis for two of the multilevel statistical models explored here.

3.4.1 Hierarchical generalised linear models

Hierarchical linear models (HLMs) and HGLMs are well known models used for the analysis of multilevel data (Singer et al., 2003; Snijders and Bosker, 2011). HLMs are used when the outcomes are continuous (*i.e.* time until attrition event), whereas HGLMs are used when the outcomes are discrete (binary: event vs non event, or integer count: years until attrition event) (Snijders and Bosker, 2011). For the purpose

of this analysis, two HGLM models are explored, namely multilevel logistic regression models for binary outcomes, and random coefficient models. The discussion assumes the data have two levels.

For the purpose of this discussion:

- Let Y_{ij} denote the binary response variable (1: event, 0: non-event or censored observation) for the i^{th} subject, nested within the j^{th} cluster
- Let $X_{1ij}, \ldots, X_{pij}$ denote p explanatory variables that are measured on this individual (*e.g.* healthcare worker characteristics)
- Let $Z_{1j}, \ldots, Z_{qj}$ denote q explanatory variables that are measured on the j^{th} cluster (*e.g.* hospital characteristics)

3.4.1.1 Multilevel logistic regression models for binary outcomes

According to Singer et al. (2003); Snijders and Bosker (2011), a random intercept logistic regression model incorporates a single random effect, allowing the intercept to vary randomly across clusters. The mathematical formulation for this model is defined below:

$$logit(Pr(Y_{ij} = 1)) = \alpha_{0j} + \alpha_1 X_{1ij} + \cdots + \alpha_p X_{pij} + \beta_1 Z_{1j} + \cdots + \beta_q Z_{qj} \qquad (15)$$

where it is assumed that the random effects follow a normal distribution. This model allows the probability of the occurrence of the outcome for a reference subject to vary across clusters. However, the effects of the individual explanatory variables are constrained to be equal across clusters (Singer et al., 2003).

3.4.1.2 Random coefficients model

The random coefficients model for analysing multilevel data incorporates a higher level of complexity into the model formulation when compared to the logistic regression models (Section 3.4.1.1).

In this model, the regression coefficients for the subject-level covariates are allowed to vary across clusters (Singer et al., 2003). The simplest example is a model with a random intercept and a random slope for a covariate X_{1ij} (as previously defined in Section 3.4.1). According to Singer et al. (2003); Snijders and Bosker (2011), the mathematical formulation for the aforementioned, simple random coefficients model is defined by Equation 16 below:

$$logit(Pr(Y_{ij} = 1)) = \alpha_{0j} + \alpha_{1j}X_{1ij} + \cdots + \alpha_p X_{pij} + \beta_1 Z_{1j} + \cdots + \beta_q Z_{qj} \quad (16)$$

where X_{1ij} has a school-specific regression coefficient (or slope α_{1j}).

It is important to note that not all regression coefficients for the subject-level covariates are required to vary across clusters. It is possible to have a random coefficients model in which a subset of the regression coefficients for the subject-level covariates are constrained to be fixed across clusters, and the rest allowed to vary across the clusters (Snijders and Bosker, 2011).

All three of these models introduce random effects into the model formulation. Random effects are incorporated in order to account for within-cluster homogeneity in outcomes (which is prevalent in multilevel data).

3.4.2 Statistical models for multilevel survival analysis

There is a significant amount of research into methods for analysing multilevel data (Snijders and Bosker, 2011), however, many of these references omit methods for the analysis of multilevel survival data or simply provide a cursory discussion of multilevel survival analysis. In applied research, time-to-event outcomes occur frequently (Austin et al., 2010), and the context of the data available for investigation are hierarchical in nature. There is thus a growing need for models that account for the analysis of multilevel survival data.

Rabe-Hesketh and Skrondal (2008) propose three models for the analysis of such data, namely i) frailty models, which are Cox proportional hazard models with mixed effects, ii) piecewise exponential (PWE) survival models with mixed effects, and iii) discrete time survival models with mixed effects. These models will be explored in the following subsections.

3.4.2.1 Frailty models: CPH regression model with mixed effects

As identified in Section 3.3.1, the CPH regression model is often used for the analysis of survival data. A modification of this model is suggested to cater for survival data that is multilevel in nature. Consequently, the conventional CPH regression model is enhanced through the incorporation of random effect terms to account for within-cluster homogeneity in outcomes (Goldstein, 2011). The inclusion of random effects into the CPH model shares many similarities with traditional methods for the analysis of multilevel data with continuous, binary, or count outcomes (Section 3.4.1).

The term frailty model is used to denote a survival regression model (typically either a CPH regression model or a PPH survival model) that incorporates random effects (Goldstein, 2011). According to Crowther et al. (2014), A *frailty model* refers to a survival model with only a random intercept, whereas a *mixed effects model* refers to a model that can have multiple random effects. Consequently, a frailty model is a special case of the mixed effects survival models. Frailty models originally only incorporated subject-specific random effects to account for unmeasured subject characteristics that influenced the hazard of the occurrence of the outcome. These models have now been extended to incorporate cluster-specific random effects to account for within-cluster homogeneity in outcomes (Rabe-Hesketh and Skrondal, 2008). These models have been described as shared frailty models, because the same random effect is shared by all subjects within the same cluster (Goldstein, 2011; Rabe-Hesketh and Skrondal, 2008).

When random effects are incorporated in the CPH regression model, these random effects denote increased or decreased hazard for distinct classes (*e.g.* clusters such as hospitals, workplaces, or schools) (Rabe-Hesketh and Skrondal, 2008). The discussion to follow assumes the multilevel survival data have two levels. The term *nested frailty model* is used to refer to survival models with random effects in which there are two or more levels of clustering (Rondeau et al., 2012).

According to Rabe-Hesketh and Skrondal (2008), the mathematical formulation of a CPH regression model with mixed effects is described by

$$h_{ij}(t) = h_0(t) \exp\left(X_i\beta + \alpha_j\right) \tag{17}$$

where α_j denotes the random effect associated with the j^{th} cluster. Rabe-Hesketh and Skrondal (2012) use the term *shared frailty* to denote the exponential of the random effect, $\exp(\alpha_j)$. The random effect is considered to be the random intercept that modifies the linear predictor, whereas the shared frailty term has a multiplicative effect on the baseline hazard function: $h_{ij}(t) = h_0(t) \exp(\alpha_j) \exp(X_i\beta)$.

CPH regression models with mixed effects are characterised by the distribution of the shared frailty terms (Wienke, 2010). The most commonly used distributions include the gamma distribution, and the log-normal distribution (where the frailty terms have a log-normal distribution, and the random effects have a normal distribution). In the case of a gamma frailty model:

- The cluster-specific random effects are distributed as the logarithms of independent, identically distributed gamma random variables, having variance θ.
- The within-cluster correlation of subjects is $\frac{\theta}{\theta+2}$.

CPH regression models with mixed effects resemble the previously described HGLMs (Section 3.4.1). In both cases, the cluster-specific random effect terms have a relative effect on the baseline hazard function (Goldstein, 2011). Consequently, the relative effect of a given covariate pattern on the baseline hazard function varies across clusters. Because of this similarity between Cox shared frailty models and HLM/HGLMs, these models are an attractive approach to fitting survival models to multilevel data (Rabe-Hesketh and Skrondal, 2008). As with HLM/HGLMs, CPH models with mixed effects are able to utilise data with more than one level of clustering.

According to Rondeau et al. (2012), limitations of the CPH regression model with mixed effects are two-fold. Firstly, Cox shared frailty models require the assumption that each subject is the member of only one level two unit, thus one cannot account for more complex multilevel structures (*e.g.* multi-membership multilevel data), in which some subjects are clustered within more than one level two unit (Therneau et al., 2000). Secondly, while Cox models with mixed effects can be extended to account for multilevel data with more than two levels, such extensions have not been incorporated into many statistical software packages.

3.4.2.2 Piecewise exponential survival models with mixed effects

As discussed in Section 3.3.2, in PPH survival models the analyst is required to make specific assumptions about the form of the hazard function. The PWE model is considered an adaption to the PPH model and thus requires specifying the distribution of the hazard (in the case of PWE, the exponential distribution) (Allison, 2010).

The PWE is a survival model in which the time scale is divided into intervals, and the hazard function assumed to be constant in each interval (Allison, 2010). The mathematical formulation for this model is described below:

- Define a set of K intervals, defined by $K+1$ cut points: $T_0, T_1, \ldots, T_k$ (where $T_0 = 0, T_k = \infty$)

In interval k, for $k = 1, \ldots, k$ and given by $[T_{k-1}, T_k)$, the hazard function for a given subject is assumed to be constant and is related to the baseline hazard function by the function

$$h(t) = \lambda_k \exp(\beta X) \tag{18}$$

where λ_k is the baseline hazard function in the k^{th} interval.

According to Allison (2010), there is a benefit in having an approximately equal number of events occur in each interval. Consequently, priority should be placed on

the former over ensuring intervals are of the same length. Alternatively, intervals can be selected using subject matter knowledge whereby it is considered reasonable to believe that the hazard is constant within each interval.

If the hazard function is constant as a function of time, then the exponential survival model and the Poisson regression model can be used interchangeably (Laird and Olivier, 1981). Consequently, the PWE model is equivalent to a Poisson regression model (Rondeau et al., 2012). Given:

- survival data consisting of a (possibly censored) observed survival time t_i for the i^{th} subject, and
- an event indicator d_i denoting whether the event was observed to occur for the i^{th} subject ($d_i = 1$ denoting the event occurred, 0 otherwise)

analogous measures for each duration interval can be defined (Rodríguez et al., 2008). Therefore:

- t_{ij} denotes the survival time for the i^{th} subject in the j^{th} interval,
- d_{ij} is an event indicator that takes the value 1 if the i^{th} subject experienced the event in interval j, and 0 otherwise.

A PWE model can, therefore, be fit by treating the event indicators as if they were Poisson observations with means $\mu_{ij} = \lambda_{ij} t_{ij}$, where λ_{ij} is the hazard for the i^{th} subject in the j^{th} interval. In doing so, an offset variable denoting the logarithm of the time-at-risk during each of the intervals would need to be incorporated into the model formulation (Crowther et al., 2014).

The discovery that a PWE survival model can be fit using a generalised linear model (*i.e.* a Poisson regression model) has important consequences for fitting multilevel survival models. Firstly, cluster-specific random intercepts can be incorporated into model formulation, and thus the model can account for within-cluster homogeneity in outcomes (Allison, 2010; Goldstein, 2011). Secondly, whilst the use of Cox models with random effects allows the baseline hazard function to vary across clusters, the use of a random coefficients Poisson regression model allows the effect of a given covariate to vary across clusters. Consequently, random coefficients are more easily incorporated using this approach than with the CPH model with mixed effects (Rodríguez et al., 2008). Lastly, by using the PWE model and incorporating random effects, existing statistical procedures available for multilevel Poisson regression models can be leveraged.

3.4.2.3 Discrete time survival models with mixed effects

Discrete time survival models are used when survival time is measured in discrete

values (*e.g.* days until death, months until relapse, or years until an attrition event). In such instances, the outcome variable of interest is the probability of an event occurring (*i.e.* event, or non-event) in time interval t, given survival up until time interval t. These models thus utilise a discrete version of the hazard function in model formulation. Even when survival time is approximately continuous, the discrete time survival model can be used by dividing survival time into a finite number of discrete intervals (Rabe-Hesketh and Skrondal, 2008).

The PWE survival model (described in Section 3.4.2.2) divides the time scale into a sequence of intervals, under the assumption that the hazard function is constant within each interval. In fitting the PWE survival model, each subject's duration of exposure, during the interval, is taken into account (as an offset variable) (Allison, 2010; Rodríguez et al., 2008). The process is slightly different in discrete time survival models; intervals are generated by noting whether or not an event occurred within a specified time frame. This method disregards each subject's duration of exposure within the given interval (Barber et al., 2000).

Therefore, a HGLM binomial regression model[5] for binary outcomes can be used to model the probability of the occurrence of an event within each intervals. Possible link functions for the generalised linear model are the logit link function, the probit link function and complementary log–log link function (Rodríguez et al., 2008). An advantage to the complementary log–log link function is that the resultant regression coefficients are identical to those of an underlying proportional hazards regression model (Allison, 2010; Rabe-Hesketh and Skrondal, 2008). Consequently, the estimated coefficients can be interpreted as having a relative effect on the hazard of the occurrence of the event.

Discrete time survival models can easily incorporate the multilevel structure of the data (Barber et al., 2000), and in such cases, are known as multilevel discrete-time event (MDTE) models. As with the PWE mixed effects survival model, random coefficients can be readily incorporated by including random coefficients in the HGLM that is being fit. An advantage to discrete time survival models compared with the PWE survival model is that the assumption that the hazard function is constant within each interval does not need to be made (Rabe-Hesketh and Skrondal, 2008). Moreover, since an HGLM (a binomial model with either a logit link function or a complementary log–log link function) is being fit, existing statistical methods and software can be leveraged.

[5]Binomial regression models: Binomial regression is a general term of a regression model in which the dependent variable has binomial distribution. Binomial regression belongs to HGLM class, which uses a link function to connect linear predictor variables to the expectation of response variables (Barber et al., 2000).

3.4.2.4 Summary of statistical models for multilevel survival analysis

Section 3.4.2 describe three families of regression models for the analysis of multilevel survival data. Firstly, CPH regression models with mixed effects which incorporate cluster-specific random effects that modify the baseline hazard function. Secondly, PWE survival models which partition the duration of follow-up into mutually exclusive intervals and fit a model that assumes that the hazard function is constant within each interval. This is equivalent to a Poisson regression model that incorporates the duration of exposure within each interval. By incorporating cluster-specific random effects, generalised linear mixed models can be used to analyse these data. Lastly, after partitioning the duration of follow-up into mutually exclusive intervals, one can use discrete time survival models that use a complementary log–log generalised linear model to model the occurrence of the outcome of interest within each interval. Random effects can be incorporated to account for within-cluster homogeneity in outcomes.

3.5 Machine learning approaches for predicting attrition rates

Based on the review of literature, three ML models are of interest for this investigation, namely NNs, GLMM trees, and the novel method of TBME models (Section 2.2.2). Although some of these methods have not been applied to multilevel survival data with two or more levels, or to data with a significant number of categorical predictor variables, research proposed the possibility that these methods can potentially cater for such data. Consequently, the theoretical basis behind these methods are explored in Sections 3.5.1 - 3.5.3.

3.5.1 Artificial neural networks for multilevel survival data

Artificial neural networks (ANNs), usually simply called neural networks (NN), are computing systems inspired by the biological neural networks that constitute animal brains. An NN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain (Hagan et al., 1997).

According to Hagan et al. (1997), neural networks consist of many layers of interconnected neuron units, starting with an input layer to match the feature space, followed by multiple layers of non-linearity (hidden layers), and ending with a linear regression or classification layer to match the output space.

MLPs are one of the most widely used and best understood neural computing

paradigms, whereby learning is based on backpropogation of error [6]. Consequently, for each cycle of learning, cases are first passed forward through the network, the error is calculated thereafter and passed backwards to adjust the weights in the model, with the aim of providing better predictions. This process is carried out multiple times (Hagan et al., 1997).

LVQs are defined as a competitive learning algorithm for solving classification problems (Kvamme and Borgan, 2019). Learning in LVQs differs from MLPs as it is not based on backpropogation, but based on a learning algorithm similar to that used in k-means clustering, which takes place in the hidden layer of the network architecture (Kohonen and Kohonen, 1995). In LVQs data are passed from the input layer to neurons in the hidden layer which serve as prototypes. As data flows through the network, cases with similar profiles are assigned to a specific neuron in the hidden layer. The neurons representing each prototype are then clustered on the basis of similarities among them, and the resulting clusters are then assigned to an outcome variable class (Kohonen and Kohonen, 1995). The clusters of prototypes identified during learning represent patterns extracted from the input data, which are then used to classify previously unseen data observations.

Unlike the neural networks of the past, modern deep learning provides training stability, generalization, and scalability with big data. Since it performs quite well in a number of diverse problems, deep learning is quickly becoming the algorithm of choice for the highest predictive accuracy. Although NNs are statistical in nature (Hagan et al., 1997), they differ from statistical methodologies in two important ways. Firstly, NNs are especially well suited to capturing non-linear relationship amongst predictor variables. In other words, they are able to extract and identify nonlinearities in data and use these nonlinear relationships in prediction and classification. Secondly, NNs are non-parametric, and are thus parameterized by the number of neurons in a given NN architecture (Hagan et al., 1997).

The basic framework of MLP and LVQ neural networks can be used to accomplish deep learning tasks. Deep learning architectures are models of hierarchical feature extraction, typically involving multiple levels of non-linearity. Although the review of literature (Section 2.2.2.1) indicated that NNs have not been applied to multilevel survival data with a large number of predictor variables, according to Hagan et al. (1997), deep learning models are considered capable of learning useful representations of raw data and have exhibited high performance on complex data such as images, speech, and text. Due to limitations posed by the scope of this research, only MLP NNs are explored for attrition prediction in this investigation.

[6] Backpropogation of error: process through which differences between predicted and observed values are gradually minimized by calculating the observed error, and then adjusting weights in the model to make the error slightly smaller (Hagan et al., 1997).

According to Hagan et al. (1997), in order to build an MLP NN, the following hyper-parameters need to be set:

- `activation function`: defines how the weighted sum of the input is transformed into an output from a node or nodes in a layer of the network. The choice of activation function has a large impact on the capability and performance of the neural network, and different activation functions may be used in different parts of the model.
- `learning rate`: the rate at which the NN updates the weights during each update.
- `epochs`: number of iterations.
- `epsilon`: adaptive learning rate time smoothing factor.
- `hidden dropout ratios`: percentage of hidden layer units to randomly drop during training to prevent overfitting.
- `hidden`: number of hidden layers and neurons within each hidden layer.
- `input dropout ratio`: percentage of input layer units to randomly drop during training to prevent overfitting.
- `l1 (Lasso regularisation)`: a quadratic regulariser to control model complexity and avoid over-fitting by decreasing the number of features in a large and complex dataset.
- `l2 (Ridge regularisation)`: a quadratic regulariser to control model complexity and avoid over-fitting by dispersing the error terms in all the weights.

In most cases, the optimal combination of hyper-parameters is not immediately known. For this reason, hyper-parameter tuning must be undertaken. As is the case with conventional analytical methods, generalisability of the model to unseen data is important in neural computing (Hagan et al., 1997). Consequently, in order to effectively test the performance of the final model, after hyper-parameter tuning, the test dataset should not be used during hyper-parameter tuning. Tuning NNs thus requires performing a grid search across varying combinations of hyper-parameters by training each hyper-parameter combination on a training dataset and testing performance on a validation set (Kvamme and Borgan, 2019).

Creating a grid search with large unique combinations of hyper-parameters is computationally expensive, especially if the dataset used for training and validation contains a large number possible factor predictor variables with many levels (*i.e.* high number of neurons in the input layer) (Hagan et al., 1997). This is so, as

the model will need to be trained and validated for every unique combination of hyper-parameters.

It is important to note, that the choice of loss function is dependent on the type of problem. For a 2-class classification problem such as the attrition event prediction problem, the distribution of the response variable needs to be specified as *Multinomial*. This distribution is associated with the cross-entropy (or log-loss) loss function (Hagan et al., 1997). For this reason, model performance is assessed by looking at the model's log-loss obtained after training the model, and testing it on a validation set.

There are three non-linear activation functions that can be used during model training for classification MLPs, namely:

- `Tanh`
- `Rectified Linear`, and
- `Maxout`.

The `Tanh` function is a rescaled and shifted logistic function; its symmetry around `0` allows the training algorithm to converge faster (Kvamme and Borgan, 2019). For this reason, `Tanh` is most often used in MLP NNs for binary classification.

The `h2o` package in R follows the model of MLP NNs for predictive modeling (Candel et al., 2016). It will be the package used to develop the NNs for this investigation.

3.5.2 GLMM trees for event prediction

In contrast to traditional GLMMs, recursive partitioning or decision-tree methods describe the association between predictor and outcome variables by a binary decision tree and not as a mathematical formula (Section 3.4.2). Decision-tree methods do not require assumptions as to the distribution of residuals and are capable of handling a large number of potential predictor variables. Compared to other machine learning algorithms, they are also considered the easiest to interpret due to their tree-like structure (Fokkema et al., 2021).

As explored in the review of literature, a recent decision-tree method (GLMM Trees) has become known in research as a prediction algorithm capable of handling longitudinal and multilevel data (Section 2.2.2.2). The GLMM tree algorithm is a special case of the model-based recursive-partition algorithm of Zeileis et al. (2008). GLMM trees fit a recursive-partition based on a generalised linear mixed effect model, whereby the nodes in the GLMM tree are comprised of nested-group specific GLMMs which contain an intercept term and the possible effects of one or

more predictor variables, as well as random effects. Decision-tree methods that allow for the analysis of correlated data structures, such as GLMM trees, have only recently been developed and is thus still considered a rather novel area in applied research (Fokkema et al., 2021).

Moreover, the GLMM tree algorithm is an extension of the unbiased [7] recursive-partitioning (DT) framework. Several earlier developed recursive-partition frameworks suffer from variable selection bias. GLMM tree algorithms mitigate this bias by separating variable and cut-point selection. In other words, in every node the splitting variable is selected first (based on the association between predictor and response variables, quantified using test statistics). At each step, the variable obtaining the lowest p-value is selected as the variable for splitting. Thereafter, the cut-point (*i.e.* splitting value) is selected by optimizing the sum of the loss function in the two resulting nodes (Fokkema et al., 2021). Moreover, utilizing statistical tests for variable splitting ensures that a stopping rule is built into the algorithm. For example, when none of the potential predictor variables in the current node have a p-value lower than the pre-defined α level, splitting is stopped (Brandmaier et al., 2013).

There are a number of algorithms and software packages that allow for recursive-partitioning of GLMM type models, such as `SEM` trees, `longRpart`, and `longRpart2` (Brandmaier et al., 2013). These methods, however, only allow for partitioning data on variables measured at the highest level (Brandmaier et al., 2013), whereas GLMM trees allow for partitioning of variables at any level in the data hierarchy. GLMM trees are thus considered most appropriate when the dataset is highly multilevel with predictor variables of interest at lower levels in the hierarchy (Fokkema et al., 2021). GLMM trees can be modelled in the `R` modelling suite using the `glmertree` package (Fokkema and Zeileis, 2022).

The main functions in package `glmertree` are `lmertree()` for continuous outcome variables, and `glmertree()` for binary or count outcome variables. Both functions require specification of a formula and data argument (Fokkema et al., 2021). For both main functions, the formula argument specifies the model formula, which is composed of a left-hand side (specifying the response variable) and right-hand side (comprising three parts: the predictors for the node-specific model (comprising fixed effects only, with coefficients that are allowed to differ over subgroups), the global model (comprising random and/or fixed effects, for which coefficients are estimated globally, using all observations) and the potential partitioning variables) (Fokkema et al., 2021)

[7] unbiased: methods that do not present with a variable selection bias, in which variables with a large number of categories have a higher probability or likelihood to be selected for partitioning - even if they are no more informative than their competitors Zeileis et al. (2008).

3.5.3 Extreme gradient boosted trees with mixed effects for event prediction

According to the review of literature (Section 2.2.2.3), XGB trees have proven to be rather successful for attrition prediction, however, limited research has been undertaken to apply these methods to multilevel or longitudinal survival data or similar problems where variables within the dataset are highly correlated. Recently, a library, `GPBoost`, to combine tree-boosting algorithms with mixed-effects models has been developed, but its application in literature is sparse, having only been successfully applied to small datasets with few predictor variables, and 2 or fewer levels (Jain and Nayyar, 2018; Liu et al., 2022).

According to Liu et al. (2022), `GPBoost` is a highly efficient package for fitting mixed effects models to data as it utilises tree-boosting to model fixed effects. Tree-boosting in recursive-partitioning refers to the creation of an ensemble of decision-trees for improving the accuracy of a single decision-tree classifier or regressor (Fokkema et al., 2021). In tree-boosting, each of the trees in the collection is dependent on its prior trees. Consequently, as the algorithm proceeds, it learns from the residual of the preceding trees. This allows the predictive model to flexibly approximate the association present in the dataset in a smooth manner. As a result, ensemble methods are known to obtain higher predictive accuracy than normal recursive-partitioning methods (Liu et al., 2022). On the other hand, these methods result in predictive models that consist of a larger number of trees that cannot be visually grasped. Consequently, the increase in predictive accuracy by using these ensemble methods comes at the cost of increased complexity (Sokhansanj and Rosen, 2022).

The `GPBoost` algorithm, in simple terms, is defined as a boosting algorithm that iteratively learns the covariance parameters (*i.e.* hyperparameters) using natural gradient descent or Nesterov accelerated gradient descent, and adds a decision tree to the ensemble using gradient boosting (Sokhansanj and Rosen, 2022). The mathematical formulation for the `GPBoost` algorithm is defined in Formula 19 below.

$$y = F(X) + Zb + xi \tag{19}$$

where, y is the outcome variable of interest for the `GPBoost` algorithm, X the predictor variables, F the non-linear predictor function, Zb the grouped random effects, and xi the independent error term. Consequently, training the `GPBoost` algorithm refers to learning the hyperparameters of the random effects and the predictor function $F(X)$ using an ensemble of decision trees. Further elaboration of the mathematical foundation of TBME models can be found in Sigrist (2020).

3.6 Data format for model development

All of the modelling techniques described in this chapter require that the dataset contain:

- An outcome event of interest: either modelled as a hazard probability (relative probability of an event occurring), or as a binary outcome (event, or non-event)
- All numeric predictor variables to be standardized [8]
- All categorical variables to be encoded: factor encoding

In the case that the data is survival data, the dataset needs to be in the format of a person-period dataset with additional time-indicator variables (discussed in Section 4.1.3).

3.7 Chapter Summary

In this chapter, focus was placed on gaining an understanding of the theory behind the statistical and ML modelling techniques applicable to attrition prediction problems. The chapter explored the theory behind survival analysis (Section 3.1) and univariate, multivariate, and multilevel-multivariate survival analysis and statistical modelling techniques (Sections 3.2-3.4). The chapter further explored hazard and event prediction using multilevel-multivariate statistical modelling, as well as supervised machine learning algorithms (Sections 3.43.5). The choice of modelling methods to be applied in this investigation will be impacted by the dataset available for use in this study, specifically if the dataset contains multilevel survival data.

[8] Data standardization: Including variables in model formulation, that are measured at different scales or by different magnitudes, can heavily contribute to model bias. Standardizing these variables, to have a common reference, helps minimize this risk (Sigrist, 2020).

4 Exploratory data analysis

Based on the thorough review of methodology, discussed in Section 3, there are a multitude of different statistical and ML methods that have been applied to attrition-like prediction problems. However, the choice of model most appropriate for the problem is influenced by the datasets available for use in the investigation, the structure and volume of the data, as well as possible historic and time-series patterns prevalent in the dataset.

Consequently, this chapter includes an exploration of the datasets available for use, as well as the preliminary data analysis and preprocessing required and performed to transform the datasets into suitable formats for use. The chapter further explores the historical patterns present in the cleaned and preprocessed data by means of a thorough historical exploratory data analysis. All data exploration and analysis is conducted using the `R` programming language (Chambers, 2008).

4.1 Data

Two of the key objectives of this investigation, as described in Section 1.3, was to obtain and explore the raw data provided by CHAI, and perform data preprocessing and cleaning to transform data into a format suitable for model development.

Preliminary exploratory data analysis (PEDA) was, therefore, conducted to understand the datasets available for use in this investigation (and the complexities thereof), so as to inform the process of preparing, preprocessing, wrangling, and cleaning the datasets into a form suitable for the development of the modelling methods explored in Section 3. PEDA and data transformation is explored in Sections 4.1.1-4.1.3.

4.1.1 Preliminary exploratory data analysis and preprocessing

All original datasets used in this investigation were provided by CHAI, who are in collaboration with the EC Department of Health (Section 1.2). Several datasets contain HR specific data from 2010 to 2021, and were thus appended to form one overall EC Health HR dataset. The second dataset used in this investigation contains the latest information on healthcare facilities in the EC province. These datasets are to be further referred to as `persal_ec_2010_2021` and `health_care_facility_list` datasets, respectively.

During the process of PEDA, it was identified that majority of the variables in the `persal_ec_2010_2021` dataset contain human-entered data. This resulted in a significant amount of data duplication and data integrity issues within this dataset. Five main data integrity issues arose during PEDA, namely:

- Inaccurate healthcare facility names: the human-entered nature of the data resulted in multiple spelling mistakes arising in the column containing facility name information, making it difficult to determine a unique and accurate set of health facilities in the EC province. The geographical data corresponding with these facility names (*i.e.* district and regional data) were also incorrectly inputted in some cases, resulting in discrepancies in the location of the facility within the province.
- Historical changes to facility data: the dataset being analysed contains data spanning over a ten-year period. In such time frame specific facilities and facility information (*i.e.* facility type, region, district) may have been updated. In such instances, the old facility names are still present in the `persal_ec_2010_2021` dataset, creating a false and inflated sense of unique facilities, facility types, regions, and districts in the dataset. This would skew any further analysis on this data.
- Inaccurate cadres and associated job titles: the `persal_ec_2010_2021` dataset contained three variables, `Occupational Group`, `Occupational Classification`, and `Job Title`, that contained information regarding an individuals job title or position and cadre. These variables, however, contained many spelling mistakes and were often used interchangeably, making it difficult to accurately determine unique sets of job titles and cadres, as well as accurately distinguish these unique sets from each other.
- Healthcare workers working at more than one facility simultaneously: This anomaly only occurs in 0.0087% of the unique individuals in the dataset. After discussions with CHAIs subject matter experts (SME), it was agreed to accept this anomaly due to the incredibly low percentage of these outliers.
- Healthcare workers age incorrectly incrementing: This anomaly only occurs in 0.0394% of the unique individuals in the dataset. After discussions with CHAIs subject matter experts (SME), it was agreed to accept this anomaly due to the incredibly low percentage of these outliers.

The aim of this investigation, as set out in the problem statement (Section 1.2), is to explore and utilise the EC Health HR data (*i.e.* `persal_ec_2010_2021`) to help predict attrition rates within and across cadres, health facilities, and districts in the EC province. Seeing as a majority of the main data integrity issues determined in PEDA impact the main variables of concern for this investigation (cadre, facility name, and geographical data), it was imperative to preprocess the `persal_ec_2010_2021` into a usable and reliable state before a full exploratory data analysis could be conducted. Three phases of preprocessing were undertaken, namely:

- Heathcare facility name cleaning,
- Merging of the `persal_ec_2010_2021` and `health_care_facility_list` datasets on facility name to create one overall `merged_persal_ec_2010_2021` dataset with enhanced geographical variables, and
- Cleaning and clustering of health workers into acceptable cadres and job titles/positions.

Cleaning and preprocessing of data relied heavily on regex manipulation[9], computer-assisted translation (approximate string matching), collaboration with SMEs and data source specialists, and manual matching.

Regex manipulation was used to clean the facility names in the `persal_ec_2010_2021` dataset, creating a unique set of healthcare facilities. Due to the inaccuracy of the geographical data corresponding to these cleaned facility names in the `persal_ec_2010_2021` dataset, they could not be assumed sufficient for further analysis. It was therefore necessary to merge the `persal_ec_2010_2021` and `health_care_facility_list` datasets together, to obtain accurate geographical variables. Both dataset have the variable, `facility_name`, in common creating the possibility for merging. However, the contents of these columns contained slightly different spelling and order of text. Consequently, a computer-assisted translation method known as Fuzzy Matching (FM) was first applied to quantify the dissimilarity between the text strings in the two columns (Cayrol et al., 1982). Since the data being matched is human-entered, the Jaro-Winkler distance metric was used (Cayrol et al., 1982). Facility names unable to be fuzzy matched (*i.e.* string distance value greater than 0) were then either i) manually matched using the `facility_name` variable in the `health_care_facility_list` as the benchmark, or in the case of outdated facility names ii) cross referenced to latest geographic information and updated manually. The final fuzzy matched column was then used to merge the datasets together into one overall `merged_persal_ec_2010_2021` dataset. Lastly, dealing with inaccurate cadres and associated job titles was more complex than cleaning the facility names. This was as a result of the interchangeable use of the three variables, `Occupational Group`, `Occupational Classification`, and `Job Title`, for specifying a health workers' cadre and or job title. All three variables were regex manipulated, but the interwoven nature of these variables made it difficult to identify correct and separate groupings of job titles and cadres using this method alone. Consequently, based on inspection of the regex manipulated columns and careful discussion with CHAI's SMEs, the job titles were manually

[9]Regex manipulation: is the process of applying a specific search pattern to text strings with the aim of extracting, manipulating, or validating text data. The strings used to define these search patterns are called regular expressions, or regex strings (Thompson, 1968).

grouped into overarching job titles which were, then, manually grouped into cadres of relevance.

This preprocessed dataset underwent a final cleaning which involved removing duplicate records, standardizing column naming formats to a single naming convention, and selecting columns of relevance for further analysis (based on discussions with CHAI and SMEs).

4.1.2 Preprocessed data

The cleaned and preprocessed dataset, `preprocessed_persal_ec_2010_2021`, thus contains healthcare worker information for the district of the Eastern Cape from 2010 to 2021 with respect to 12 variables. The details of these variables are described in Table 1.

Variable	Data Type	Description	Example
Persal Number	Factor with 91218 levels	The unique, province specific, identification number for a healthcare worker in South Africa's public healthcare system	8 digit number
Gender	Factor with 2 levels	Gender of healthcare worker	Male, Female
Age	Numeric	Age of healthcare worker, in years	56
Race	Factor with 4 levels	Race of healthcare worker	African, Asian, Coloured, White
Job Title	Factor with 716 levels	Position or job title held by healthcare worker	Medical Specialist, Physiotherapist, General Manager
Cadre	Factor with 38 levels	Healthcare profession grouping	Medical Services, Nursing
Notch Value	Numeric	Annual salary of healthcare worker, in Rands	R100000
Reporting Year	Factor with 12 levels	Indicating the year of the reporting period, commencing on the 1st April and ending on the 31 March	2010
Facility Name	Factor with 849 levels	EC healthcare facility name	Fort England Hospital
Facility Type	Factor with 9 levels	Type of healthcare facility	Clinic
Rural/Urban	Factor with 2 levels	Geographical grouping of healthcare facility	Rural or Urban
District	Factor with 9 levels	The district in which a healthcare facility resides	Sarah Baartman DM

Table 1: Detailed description of the variables in the cleaned and preprocessed, `preprocessed_persal_ec_2010_2021` dataset.

The `preprocessed_persal_ec_2010_2021` dataset contains 571859 observations reflecting the change in the EC health workforce from 2010 to 2021.

In order to determine attrition rates for a given year, a variable indicating an attrition event is required. According to the review of literature, discussed in Section 2.2, there is a lack of internationally comparable definitions and guidelines for measuring attrition within healthcare HR. Consequently, methods and measures for computing healthcare workforce attrition is determined by the data available for use in determining attrition rates, and the outcomes of importance or relevance to the decision

makers. Based on discussions with CHAI, two types of attrition events are of interest to stakeholders, namely:

- an EC healthcare worker exiting the EC public heath system entirely, or
- an EC healthcare worker moving from one facility to another facility in the EC province.

Since the `preprocessed_persal_ec_2010_2021` dataset utilised in this investigation contains annual HR data, the annual attrition rate is to be used for attrition computations.

Consequently, an **attrition event**, for the purpose of this investigation, is defined as either an exit from the EC public health system, or a move from one facility to another facility in the EC province, at the end of a specific reporting year. The `preprocessed_persal_ec_2010_2021` dataset does not explicitly contain such an event variable. According to SMEs and data source specialists at CHAI, an attrition event can be considered to have occurred if an individuals persal number disappeared entirely from the dataset between two intervals (*i.e.* from one reporting year to the next), or if their persal number appeared at different facilities between two intervals. Consequently, an `event` outcome variable was constructed to indicate whether or not an attrition event had occurred for a specific healthcare worker (*i.e.* persal number), at the end of a specific reporting year, conditional on the healthcare worker being employed at a facility in the EC province at the beginning of the reporting year.

This dataset thus contains information that tracks the attrition event history of EC healthcare workers over a 12 year period. A sample of this data is graphically represented by means of Figure 3.

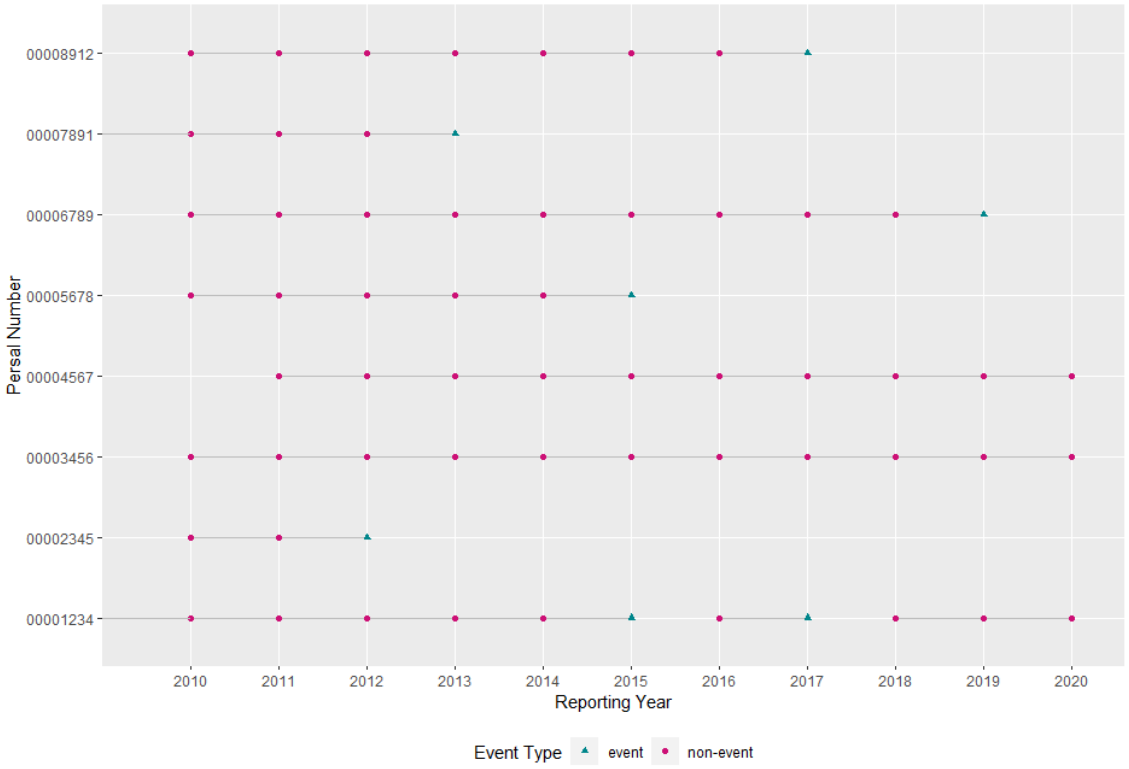

Figure 3: Schematic representation of the attrition event history of eight EC healthcare workers over the study period.

The time of origin for this investigation is defined as the beginning of the 2010 reporting period. According to Figure 3, at the time of origin, three scenarios are possible, namely:

- a healthcare worker began working at their facility at the time of origin (*e.g.* healthcare workers `00003456` and `00006789`),
- a healthcare worker began working at their facility prior to the time of origin, are still employed at that facility at the time of origin and are, therefore, considered left-censored as described in Section 3.1.1 (*e.g.* healthcare workers `00007891` and `00001234`), and
- a healthcare worker is not yet employed in the EC public healthcare system at the time of origin, but enters the system later in the study period (*e.g.* healthcare worker `00004567`).

The `preprocessed_persal_ec_2010_2021` dataset only contains workforce level data up until the beginning of the 2021 reporting year (*i.e.* end of the 2020 reporting year). Seeing as the dataset does not contain adequate data to determine realistic attrition levels for the 2021 `reporting year`, data pertaining to the 2021 `reporting year` is excluded from use in further exploratory data analysis and modelling conducted in this investigation. Consequently, the study termination time for this investigation is defined as the end of the 2020 reporting period. Catering for right censored data is thus not required for this investigation since the full survival times for individuals, at the termination of the study, are known.

Moreover, re-entry of a given healthcare worker into the EC health system is possible. This is visualised in Figure 3, whereby healthcare worker `00001234` enters the health system on three separate occasions (prior to study start, 2016, and 2018) and experiences two different attrition events (in years 2015 and 2017, respectively) over the course of the study period. A partial listing of the data [10] for healthcare worker `00001234`, schematically represented in Figure 3, is provided in Table 2.

Persal Number	Facility Name	Age	Reporting Year	Employment Start Date	Event
00001234	Nompumelelo Clinic	25	2010	2007	0
00001234	Nompumelelo Clinic	26	2011	2007	0
00001234	Nompumelelo Clinic	27	2012	2007	0
00001234	Nompumelelo Clinic	28	2013	2007	0
00001234	Nompumelelo Clinic	29	2014	2007	0
00001234	**Nompumelelo Clinic**	**30**	**2015**	**2007**	**1**
00001234	Nqabara Clinic	31	2016	2016	0
00001234	**Nqabara Clinic**	**32**	**2017**	**2016**	**1**
00001234	Nompumelelo Clinic	33	2018	2018	0
00001234	Nompumelelo Clinic	34	2019	2018	0
00001234	Nompumelelo Clinic	35	2020	2018	0

Table 2: Partial data listing for Persal Number `00001234` over the study period.

The outcome variable of interest, the attrition `event`, is thus not only impacted by whether or not an event occurred, but also the reporting year in which the event occurred. The outcome variable can, therefore, be defined as a *time-to-an-event* variable, where the time variable, `reporting year`, is discretely measured. According to a methodological review, discussed in Section 3.1, the characteristics of this dataset match the characteristics of survival data.

According to the review of literature, discussed in Section 2.2.1, multilevel data structures are often present in healthcare data. This usually exists in situations where patients are nested in a hospital, which is nested in a district. In this context,

[10] Partial data listing: snippet of a dataset that contains some, but not all, variables for ease of display.

a hierarchical data structure of five levels seems to be present, whereby "repeated" measures are nested within a healthcare worker, who is nested in a cadre, which is nested in a facility, which is nested in a district. As per research findings in Section 3.4, subjects who are nested within the same higher level unit within a data hierarchy are likely to have outcomes that are correlated with one another, thus violating the assumption of independent observations. Consequently, the possibility of this within-cluster homogeneity must be considered.

In order to make use of the modelling methods described in the literature and methodological reviews (Sections 2 and 3), the dataset is required to be manipulated into a person-period format. Manipulation of the `preprocessed_persal_ec_2010_2021` dataset into this format is described in Section 4.1.3.

4.1.3 Person-period dataset

A person-period data set contains a record of data for each individual with information on each predictor at each time period the data was recorded. The `preprocessed_persal_ec_2010_2021` dataset was already in the format of a person-period dataset (as visualised in Table 2), but required some additional time indicator variables, survival predictor variables, and data manipulation to some existing covariates.

The newly created person-period dataset, `person_period_persal_ec_2010_2021`, contains the following information, on the i^{th} individual with j records:

- Time indicators:
  - `enter`: indicating the baseline time, in years, since the individual was first recorded (in the context of the study's time of origin) to be working at a specific facility.
  - `exit`: indicating the interval, in years, the record was taking place in, relative to the individuals baseline `enter` time of origin. For example, if an individual has an `enter` value of 0, they are in the first interval (*i.e.* `exit` = 1).
  - `duration`: the duration, in years, that an individual had been working at a specific facility before they experienced the event of interest.
- Predictors: the covariates for individual i at time period tj, described in 4.1.2.
- Event indicator:
  - `event`: indicating if the attrition event of interest had occurred for the i^{th} individual in time period tj (boolean: 1 if attrition event occured, 0 otherwise).

The transformation of the `preprocessed_persal_ec_2010_2021` dataset into the `person_period_persal_ec_2010_2021` dataset is visualised in Figure 4, and the process thereof described below.

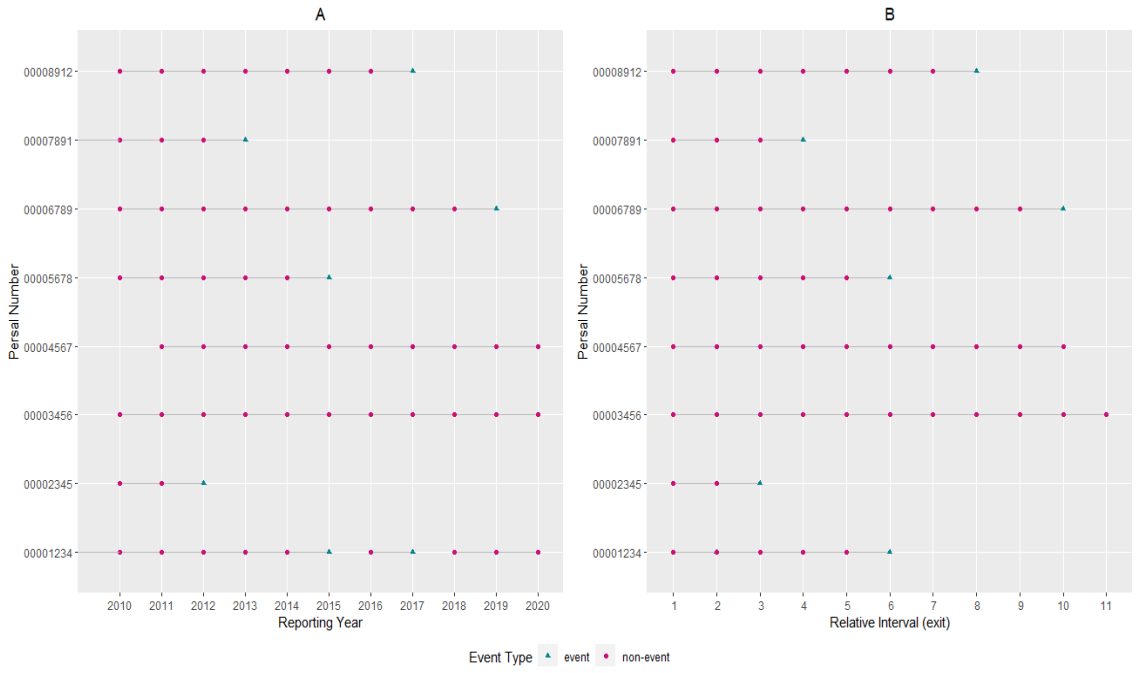

Figure 4: Schematic representation of the transformation of the preprocessed persal dataset (A) into a person-period format (B).

According to the methodological review, discussed in Section 3.1, study subjects in survival data must be comparable at their time of origin. In other words, every individual in the study is to be followed from a baseline date until the end event or termination of study. All first records of an individual working at a specific facility within the study period were set to a starting baseline `enter` time of 0 years, with a relative interval time-indicator of 1 (*i.e.* `exit` = 1). For example, if an individual was in Facility X in the EC health system in the first year of the study period (*i.e.* 2010), the value of `enter` and `exit` for that record would be 0 and 1, regardless of whether or not that individual had been working at that facility prior to 2010 (the study's time of origin). This transformation scenario is graphically depicted for healthcare workers `00008912` and `00007891` in the comparison of Figure 4A and Figure 4B. Similarly, if a healthcare worker only entered the EC health system after the study start (*e.g.* healthcare worker `00004567`, visualised in Figure 4A) the value of enter and `exit` for that record would be set to 0 and 1, respectively (as displayed in Figure 4B).

In order to cater for left-censored data, an additional time-indicator variable, `years_in_system_before_study_start`, was created. This variable indicates the

number of years an individual was working in the EC system prior to the first year of the study (time of origin, 2010), subject to the fact that the facility that the individual worked in prior to the the study's time of origin equaled the facility to which they were still employed at in the first year of the study.

Moreover, and as discussed in Section 4.1.2, re-entry of a given healthcare worker into the EC health system is possible (as depicted by healthcare worker `00001234` in Figure 4A). For the purpose of this investigation, re-entry events are assumed to be independent and this assumption has been accepted by CHAI seeing as understanding the impacts of repeated entries on attrition is not an objective of this investigation. Consequently, no additional variable, to model these *repeated events*, is required. Consequently, repeated events are treated as independent observations. For clarity of understanding, a partial listing of the transformed data (person-period format) for healthcare worker `00001234`, schematically represented in Figure 4, is provided in Table 3.

Persal Number	Facility Name	Age	Reporting Year	Years In System Before Study Start	Enter	Exit	Duration	Event
00001234	Nompumelelo Clinic	25	2010	3	0	1	4	0
00001234	Nompumelelo Clinic	26	2011	3	1	2	5	0
00001234	Nompumelelo Clinic	27	2012	3	2	3	6	0
00001234	Nompumelelo Clinic	28	2013	3	3	4	7	0
00001234	Nompumelelo Clinic	29	2014	3	4	5	8	0
00001234	**Nompumelelo Clinic**	**30**	**2015**	**3**	**5**	**6**	**9**	**1**
00001234	Nqabara Clinic	31	2016	0	0	1	1	0
00001234	**Nqabara Clinic**	**32**	**2017**	**0**	**1**	**2**	**2**	**1**
00001234	Nompumelelo Clinic	33	2018	0	0	1	1	0
00001234	Nompumelelo Clinic	34	2019	0	1	2	2	0
00001234	Nompumelelo Clinic	35	2020	0	2	3	3	0

Table 3: Partial listing of the transformed data (person-period format) for healthcare worker `00001234`, over the study period.

A partial data listing, in person-period format, for the remaining 7 healthcare workers schematically represented in Figure 4, is documented in Section 7.1 of the Appendix.

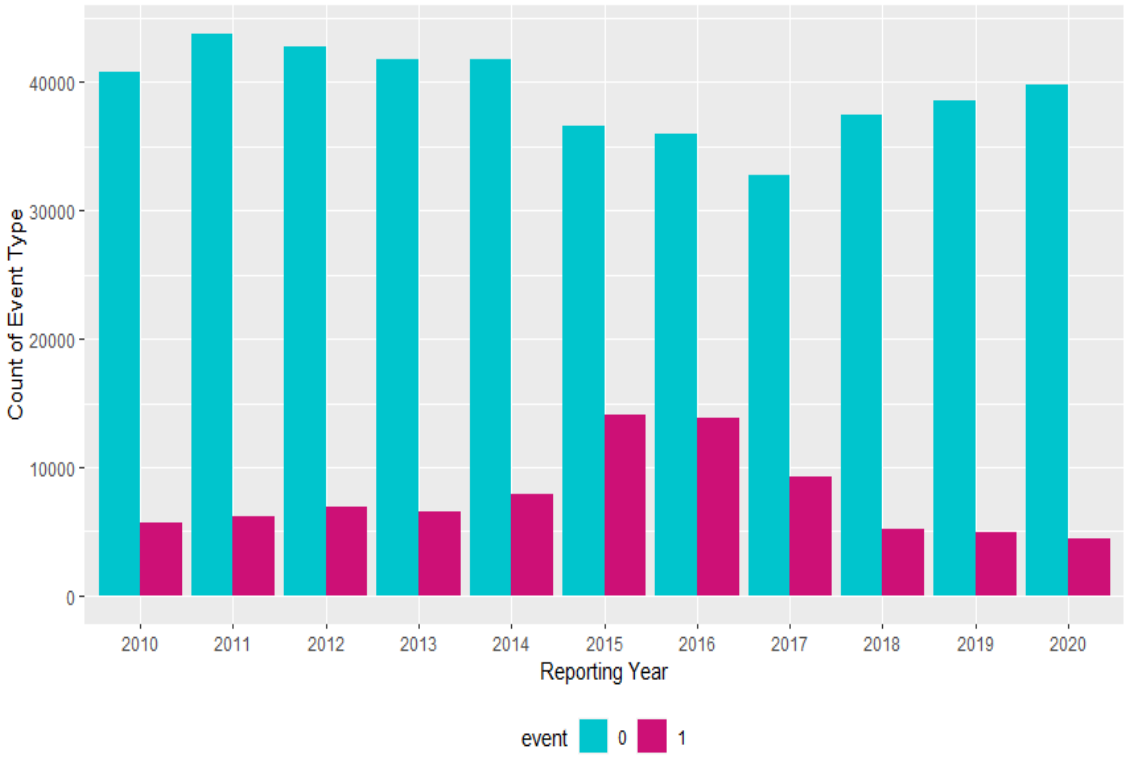

Figure 5: Graphical representation of the presence of imbalanced classes in the outcome variable.

Additionally, according to Figure 5, the number of attrition events versus non-events experienced in each year of the study differs substantially. This indicates that the `person_period_persal_ec_2010_2021` dataset has imbalanced classes. According to Cateni et al. (2014), developing ML and statistical models on significantly imbalanced datasets often results in models that are unable to generalise well to new and unseen data. However, the degree of the effect of imbalanced classes on model performance depends on the problem and the type of error considered acceptable for the problem (*e.g.* Type I error preferred over Type II error) (Menon et al., 2013). According to Cateni et al. (2014), resampling techniques, to cater for imbalanced classes, should be applied to datasets when the ratio of the classes (*i.e.* *minority* : *majority*) constitutes less than 0.1 in a given reporting year. According to the results, tabulated in Table 21 (documented in Section 7.2 of the Appendix), it is evident that the class ratio for the outcome variable, `event`, is never lower than 0.1. Consequently, no resampling techniques are applied to cater for imbalanced classes in this investigation.

The `person_period_persal_ec_2010_2021` dataset, is made up of four time indicator variables, thirteen predictor variables, and one discrete event outcome variable. The variables `district`, `facility_name`, `persal_number`, `gender`, `race`, `reporting_year`, `rural_urban`, `facility_type`, `cadre`, and `job_title` are all factor variables with 9, 849, 91218, 2, 4, 2, 9, 38, and 716 levels, respectively. The variables `age`, `notch_value`, `years_in_system_before_study_start`, `duration`, `enter`, and `exit` are all numeric.

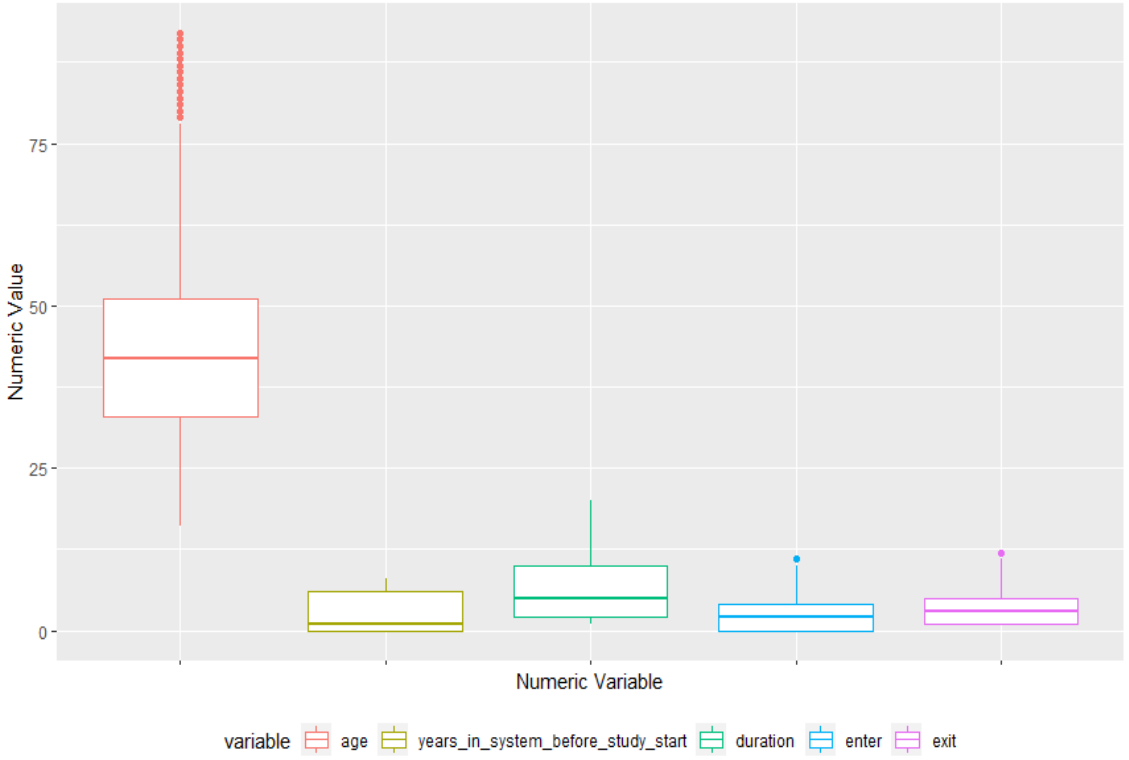

Figure 6: Graphical representation of the 5-number summaries for the numeric vaiables present in the dataset.

Based on the 5-number summaries for these numeric variables, presented in Figure 6, it is noted that the variable values are measured on different scales. Consequently, the numeric variables must be standardized to each have a mean of zero, and the standardized dataset used for all model development in this investigation. After the extensive PEDA and data preprocessing performed, discussed in Section 4.1.1, no missing values exist in the `person_period_persal_ec_2010_2021` dataset.

4.2 Exploratory data analysis

One of the key objectives of this investigation, as described in Section 1.3, was to gain an understanding of historical attrition rates within and across the EC public health system, with the primary aim of gaining insight into possible trends or areas of concern with respect to workforce attrition. Exploratory data analysis (EDA) was, therefore, conducted on the cleaned and preprocessed `person_period_persal_ec_2010_2021` dataset with the intention to answer the pertinent questions raised by stakeholders regarding historical workforce attrition in the EC healthcare system. The findings of this EDA are discussed in the section to follow.

4.2.1 Historical attrition analysis

As identified in a thorough review of literature (Section 2.2), there is a general lack of consistency regarding the definition of heathcare workforce attrition in applied research. Consequently, there is a lack of consistency regarding the computation of attrition and attrition rate. Despite this inconsistency, however, the annual attrition rate is considered to be the most common and the only comparable measure (Section 2.2). Based on findings in PEDA and data preprocessing (Sections 4.1.1-4.1.2), in conjunction with this review of literature, attrition is to be measured on an annual basis. Consequently, Formula 1 will be used to compute healthcare workforce attrition rates in this investigation. This required the computation of the number of employees at the beginning and end of a reporting year, for a given nested grouping, respectively.

EDA was first conducted at a provincial level. The number of employees at the start and end of a reporting year and the annual provincial attrition rates for the years of study (2010 to 2020) are displayed in Table 4.

Year	Number of Employees at Year Start	Number of Employees at Year End	Number of Attrition Events	Annual Provincial Attrition Rate (%)
2010	46407	40718	5689	12.90
2011	49982	43781	6201	14.06
2012	49616	42690	6926	15.70
2013	48283	41796	6487	14.70
2014	49679	41742	7937	18.00
2015	50553	36529	14024	31.79
2016	49760	35927	13833	31.36
2017	42006	32699	9307	21.10
2018	42610	37403	5207	11.80
2019	43396	38502	4894	11.09
2020	44111	39723	4388	9.95

Table 4: Annual provincial level attrition rates for the years of study.

The annual EC provincial level attrition rates, displayed in Table 4, show no specific increasing or decreasing trend in attrition rates between the beginning of 2010 and end of 2020. The highest annual provincial level attrition experienced during the study was in years 2015 - 2017. In such years, the attrition rates were as much as three times as high as as the later and earlier years in the study. The average provincial level attrition rate over the 10 years is, therefore, 18.79%. According to the EC DOH, an annual attrition rate of 5% is assumed during budget formulation for all provinces in SA, including the EC (Section 1.1). Based on these results tabulated in Table 4, the actual annual attrition rates experienced during the study period, year on year and on average, far exceed this expected and planned for 5% annual attrition rate. This suggests that the EC DOH have been incorrectly catering for attrition in their annual budgets for at least the last 10 years.

Moreover, it was important to stakeholders to understand how much of this annual attrition was attributed to individuals leaving the system due to retirement. There is no explicit variable in the dataset that indicates whether or not an attrition event was due to retirement or not. According to SMEs, the legal retirement age in South Africa is 65, but healthcare workers tend to retire between the ages of 60 and 65. Consequently, an individual is assumed to have retired if the individual was over the age of 60 when they experienced their last attrition event and did not re-enter the system, during the time of observation (study period). Based on this definition, it was identified that 0.00746% of the annual attrition rate in each year of the study was attributed to attrition events caused by retirement. This suggests that the proportion of individuals retiring every year, is consistent across all years in the study. Stakeholders can, therefore, assume that 0.00746% of any years annual provincial level attrition rate is attributed to retirement. This low proportion is indicative of the fact that retirement has a small influence on annual attrition rates, when computed at a provincial level. This suggests that other factors must be influencing attrition within and across the EC public heath system.

Consequently, it was important to analyse whether or not annual attrition rates have been historically impacted by specific grouping variables. Historic annual attrition rates, with respect to `district`, `facility_type`, `rural_urban`, `facility_name`, and `cadre`, are computed. Due to the number of levels in each grouping variable, it is not feasible to include the detailed attrition results in the report. Consequently, the detailed year by year attrition rate results for each grouping are provided to CHAI as supplementary documentation. Therefore, only key findings will be discussed in this section.

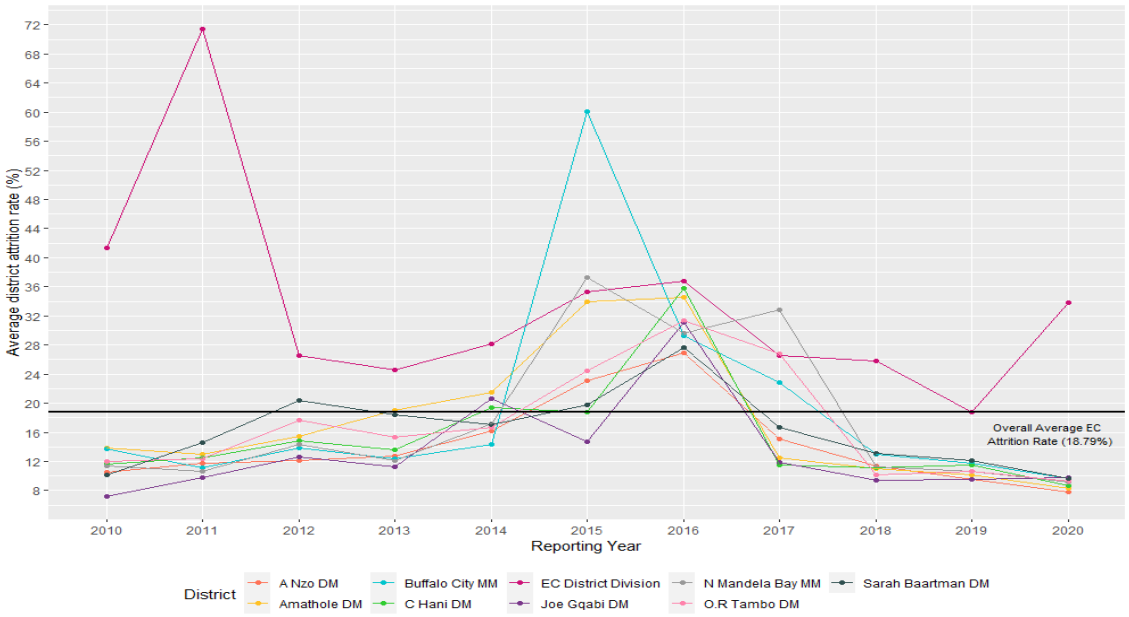

Figure 7: Graphical representation of the historical annual district level attrition rates for the years of study.

The first grouping analysed, with respect to historic attrition rates, is the `district` level grouping. According to Figure 7, one district (the `EC District Division`) seems to continuously experience attrition rates that exceed the overall average EC attrition rate of 18.79%. According to Table 22 (documented in Section 7.3 of the Appendix), this district level experienced an average annual district attrition rate of 33.56% and a coefficient of variation[11] (CV) of 0.78, which is double almost all of the other 8 districts' average annual attrition rates and CV's. According to Figure 7, the spread in annual attrition rates for this district is also alarming as the maximum annual attrition rate experienced in the 10 year period is nearly 4 times the minimum rate experienced. This minimum annual attrition rate of 18.76%, experienced in the 2019 reporting year, is also nearly double the planned for 5% annual attrition rate. It is important to note, however, that the `EC District Division` behaves slightly differently to the other district levels as it is made up of individuals that work for the EC DOH but are not assigned to any specific district post. This may be the reason for the significantly higher average annual attrition rate. Due to the definition of this

[11]Coefficient of Variation (CV): CV relates the standard deviation of the estimate to the value of this estimate. The lower the value of the coefficient of variation, the more precise the estimate. A CV less than 1 is considered low variance, whilst a CV greater than 1 is considered high variance Singer et al. (2003).

district, and based on discussions with CHAI SMEs, it is deemed sufficient to treat this district as an anomaly to be further investigated with modelling techniques.

According to Figure 7, the other 8 districts seem to follow a similar historical attrition pattern, except for `Buffalo City MM`, who experienced a spike in annual attrition in 2015. However, upon comparison of these 8 districts' average attrition rates, documented in Table 22 (Section 7.3, Appendix), very little variation in average attrition rates exists between these 8 district levels in this grouping. This implies that the district level attrition rates for these 8 districts are roughly constant around the overall mean EC attrition rate year on year. Consequently, `district` does not seem to influence the probability of an attrition event occurring. In other words, the probability that a healthcare worker will experience an attrition event in a given year does not, historically, seem to be influenced by the district in which they work. Despite this, however, the average attrition rate for each district in this grouping is higher than the expected and planned for 5% annual attrition rate.

Moreover, the second grouping variable of interest is `rural_urban` (*i.e.* geographic location). The change in annual geographic location level attrition rates, for the years of study, are graphically depicted in Figure 8.

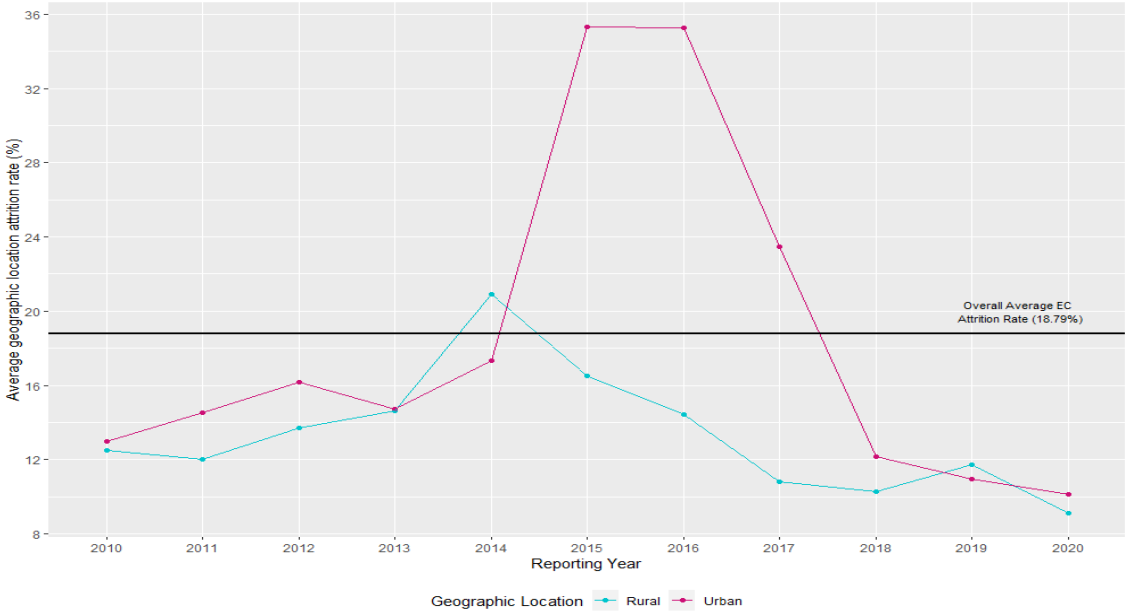

Figure 8: Graphical representation of the historical annual geographic location level attrition rates for the years of study.

According to Figure 8, in conjunction with results in Table 23 (Section 7.3 of the Appendix), facilities in `urban` areas seem to experience higher annual attrition rates on average (18.46%) than facilities in `rural` areas (13.33%). However, it is important to note that the average attrition rate for facilities in `urban` areas, seems to be skewed by the annual attrition rates experienced by in these areas in 2015 and 2016 (Figure 8). Taking these outlier years into account, the average CV across these levels (0.37), computed using the results in Table 23, is considered low enough to accept that little variation exists between these grouping levels. Additionally, the overall average annual attrition rate across all rural/urban facilities of 15.89% is only slightly lower than the overall average EC attrition rate of 18.79%. This implies that `rural_urban` does not seem to influence the probability of an attrition event occurring. Despite this, however, the average attrition rate for each geographic location in this grouping is higher than the expected and planned for 5% annual attrition rate.

Moreover, the third grouping variable of interest is `facility_type`. The change in annual facility type level attrition rates, for the years of study, are graphically depicted in Figure 9.

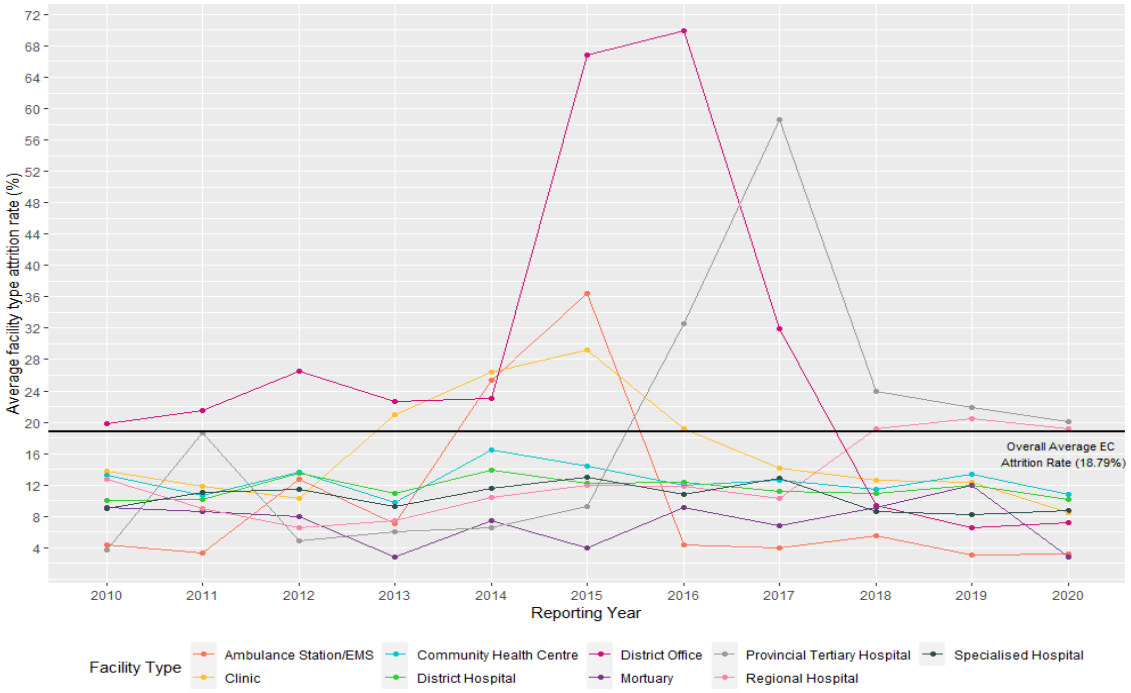

Figure 9: Graphical representation of the historical annual facility type level attrition rates for the years of study.

Using the tabulated results in Table 25 (Section 7.3), the overall average annual attrition rate across all facility types is calculated to be 14.13%, with an average CV of 0.51, respectively. As a result, the overall average attrition rate for this grouping does exceed the expected and planned for 5% annual attrition rate, however, the medium level CV suggests that there is some variation in average attrition rates between the facility type levels.

According to Figure 9 and Table 25 (Section 7.3 in the Appendix), three facility types, `Mortuary`, `Ambulance Station/EMS` and `Specialised Hospital`, do on average exhibit an average annual attrition that resembles the lowest deviation from the budgeted 5% attrition rate. Whilst the CV for `Specialised Hospital's` is low (0.16), indicating little variation around the mean (10.40%), the CV for `Ambulance Station's/EMS` is much higher (1.11), suggesting that the annual attrition rate could deviate significantly from it's mean (9.93%) depending on the year. This within-facility variation for these two facility types is graphically depicted in Figure 9. Moreover, the `Mortuary` facility type obtained the lowest average annual attrition rate of 7.27% (Table 25, Section 7.3), however, in 2019 the annual attrition rate was as high as 12% (Figure 9).

Consequently, the existing attrition rate budget percentage can be considered insufficient when planning for healthcare worker attrition from a facility type level perspective. Additionally, the relatively high average CV across all facility types of 1.95 implies that attrition rates vary across the different facility types, which suggests that `facility_type` may be a factor that influences the probability of an attrition event occurring.

Moreover, the fourth grouping variable of interest is `cadre`. The change in annual cadre level attrition rates, for the years of study, are graphically depicted in Figure 10. Due to the number of levels in this grouping variable, only a sample of the most and least concerning levels are included in this report. The full historical attrition rates for all cadres is provided to CHAI as supplementary documentation.

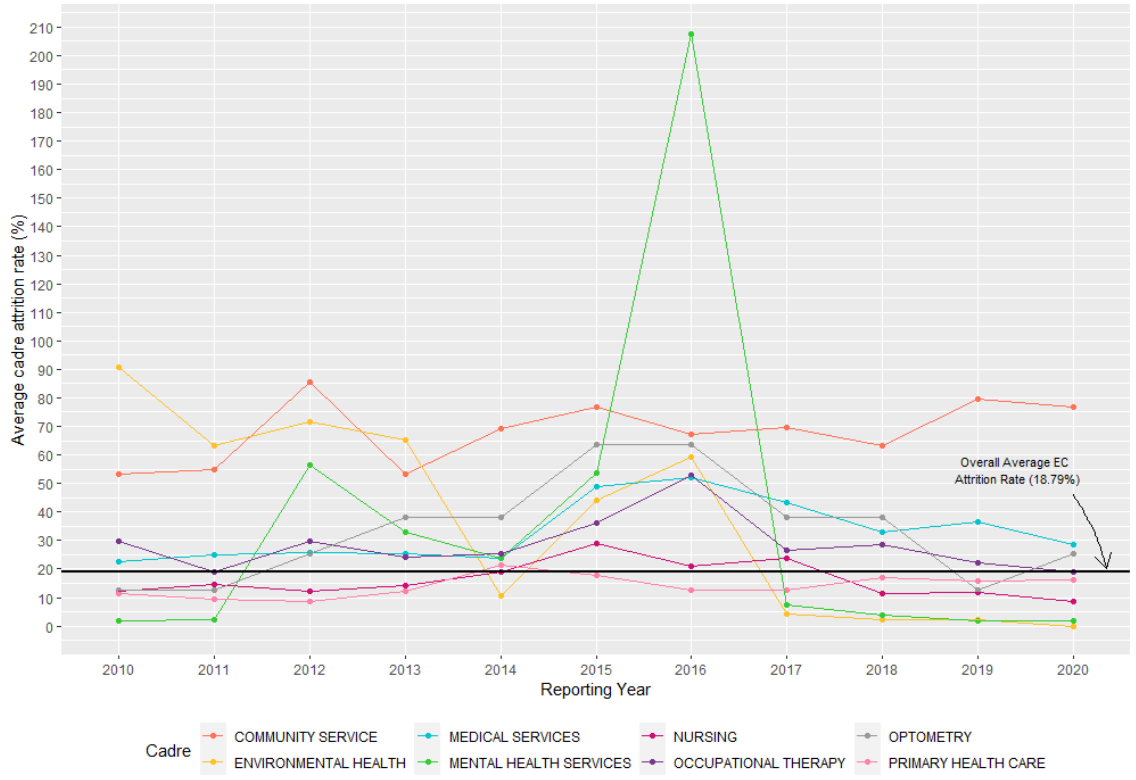

Figure 10: Graphical representation of the historical annual cadre level attrition rates for the years of study.

Using the complete tabulated results of attrition rates for all cadres (provided to CHAI as supplementary documentation), the overall average annual attrition rate across all cadres is calculated to be 18.81%, with an average CV of 0.57, respectively. According to Figure 10, there are zero cadres that experience an average annual attrition rate that is smaller than or roughly equal to 5%. In all cases, the CV experienced in each of these levels varies significantly from one another (between 0.16 and 1.69) . This high level-specific CV, suggests that there is variation in the average attrition rates experienced within and across the 38 cadres. This implies that the probability of an attrition event occurring in a given year may be impacted, or influenced quite significantly, by the cadre in which a healthcare worker is categorized.

In addition to these findings, and according to Figure 10 and Table 26 (Section 7.3 in the Appendix), several cadres seem to be experiencing problematic levels of attrition over the study period, namely, `Community Service`, `Environmental`

Health, Mental Health Services, Medical Services, Occupational Therapy, and Optometry. In all cases the overall average annual attrition rate for these cadres lies above the overall average EC attrition rate of 18.79%. The primary cadres of concern are limited to the cadres that have a high average annual attrition rate, a medium to high CV, and a minimum annual attrition rate that exceeds 5%. Based on this definition, the most problematic cadre is Community Service with an average annual attrition rate, CV, and minimum annual attrition rate (experienced in 2010) of 68.14%, 0.16, and 53.25%, respectively. Seeing as community service contains jobs that are intended to be short-lived, this degree of attrition is to be expected, however, it is currently not accounted for in the budgeted 5% attrition rate. The second most concerning cadre is Medical Services. This cadre is typically made up of job titles described as Medical Specialists, Clinic Specialists, Medical Officers, Medical Advisors, and associated managerial positions. Consequently, an average attrition rate of 33.15%, a CV of 1.69, and a minimum annual attrition rate of 22.48% (experienced in 2010), can be deemed highly concerning from a medical, service delivery perspective. The last two concerning cadres are Occupational Therapy and Optometry with an average attrition rate of 28.48% and 33.48%, a CV of 0.33 and 0.54, and a minimum annual attrition rate of 19.05% (experienced in 2011) and 12.70% (experienced in 2011), respectively.

Moreover, the last grouping variable of interest is facility_name. The change in annual facility level attrition rates, for the years of study, are graphically depicted in Figure 11. Due to the number of levels in this grouping variable, only a sample of the most and least concerning levels are included in this report. The full historical attrition rates for all cadres is provided to CHAI as supplementary documentation.

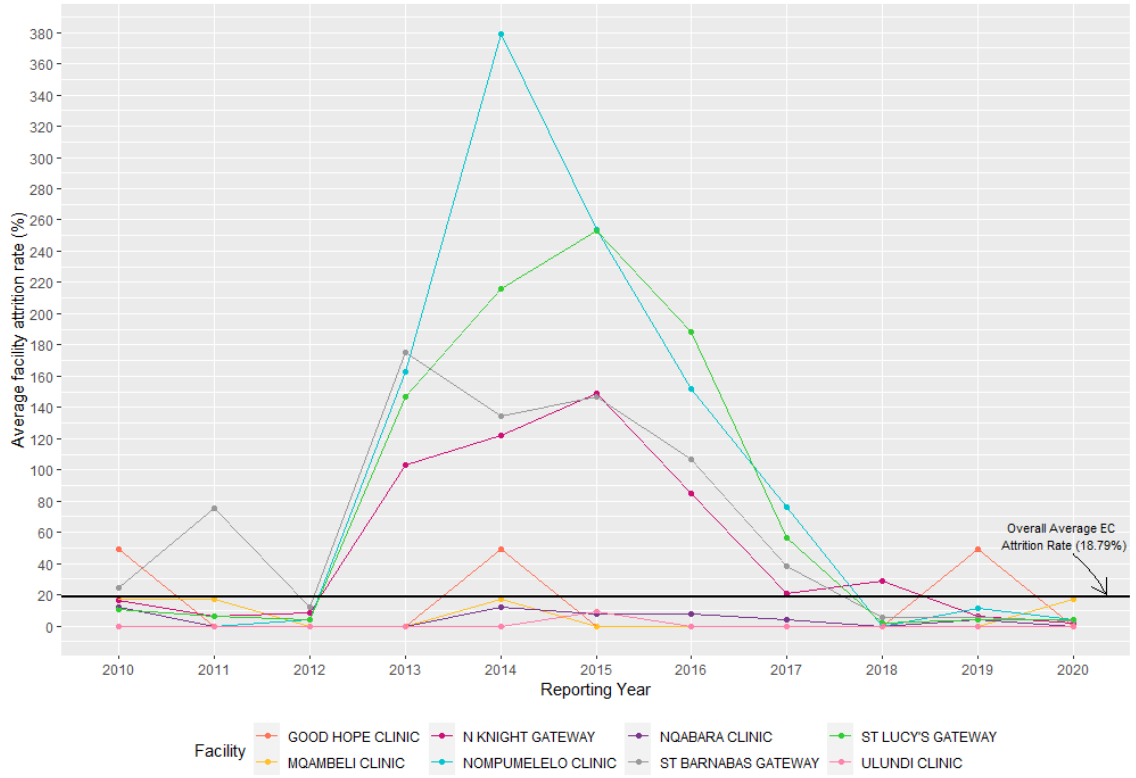

Figure 11: Graphical representation of the historical annual facility level attrition rates for the years of study.

Using the complete tabulated results of attrition rates for all facilities (provided to CHAI as supplementary documentation), the overall average annual attrition rate across all facilities is calculated to be 13.25%, with an average CV of 0.85, respectively. Of all the grouping variables explored thus far, the overall CV across all levels within the `facility_name` grouping, is the greatest. This indicates that there is significant variation in average annual attrition rates across the different health facilities in the EC and is represented in Figure 11. Moreover, whilst the overall average annual attrition rate across all cadres is roughly double the budgeted 5%, this average seems to be significantly skewed to the right, as there are a few facilities (*e.g.* `Nompumelelo Clinic` and `St Lucy's Gateway`) that obtained an average annual attrition rate larger than 80% (Table 27, in Section 7.3 of the Appendix). In such cases, the CV was also significantly higher than the overall average CV for the grouping, with some facilities experiencing a CV larger than 3.30 (*e.g.* `Ulundi Clinic`). The high degree of variability, not only in the average attrition rates between levels, but also the CV within and across these levels implies that the

probability of an attrition event occurring in a given year may be impacted, or influenced quite significantly, by the facility in which a healthcare worker resides.

According to the tabulated results (provided to CHAI as supplementary documentation), 13% of these facilities exhibit an average annual attrition rate that is higher than 20%. Of this, and as displayed in Figure 11 and Table 27, three facilities are of highest concern, namely, `Nompumelelo Clinic`, `St Lucy's Gateway`, and `St Barnabas Gateway`. These facilities obtained the highest average annual attrition rates (94.79%, 81.11%, and 66.15%) and CV's (1.35, 1.22, 0.9%) of all the facilities, respectively. It is important to note, that a few facilities did in fact experience an average annual attrition rate less than 5% and exhibit little to no variation in their annual attrition rate across the study period (*e.g.* `Ulundi Clinic` and `Nqabara Clinic`). In some cases, some facilities experienced an annual attrition rate of 0%, for nearly all years observed. This grouping thus displays the greatest variation in attrition behaviour across the levels when compared to attrition computations for all groupings previously explored.

According to the review of literature, discussed in Section 2.2.1, multilevel data structures are often present in healthcare data. Consequently, it was important to analyse whether or not annual attrition rates have been historically impacted by a multilevel structure in the data. Due to the size of this dataset, and the number of levels within each potential nesting variable, a 2-level hierarchical data analysis was conducted. This analysis was further limited to computing one cadre's (`Medical Services`) average annual attrition rate and CV per facility, with the aim of identifying any nested structure patterns. The change in annual facility level attrition rates, for the `Medical Services` cadre, for the years of study are graphically depicted in Figure 11. Due to the number of levels in each of the grouping variables (facility and cadre), only a sample of the most and least concerning levels are included in this report. The full historical attrition rates for cadres nested in facilities are provided to CHAI as supplementary documentation.

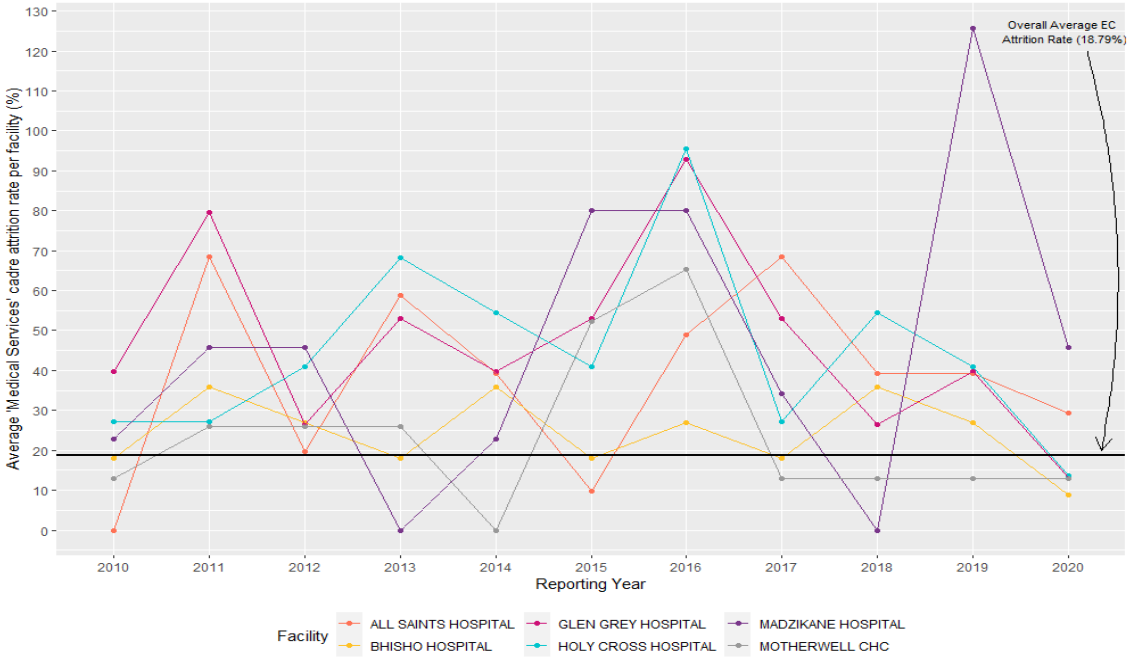

Figure 12: Graphical representation of the `Medical Services` cadres, historical annual facility level attrition rates for the years of study.

Based on the sample results in Figure 12 and Table 28 (Section 7.3 of the Appendix), it is evident that the average annual attrition rate for the `Medical Services` cadre varies significantly across the different facilities. For example, the `Medical Services` cadre in the `Motherwell CHC` experienced an average annual attrition rate of 23.72%, whereas the same cadre in the `Glen Grey Hospital` experienced an average annual attrition rate of 47.01%. Moreover, the CV from the average attrition rate experienced by this cadre also differs significantly per facility, with the CV, for this cadre, for `Motherwell CHC` and `Glen Grey Hospital` being 0.81 and 0.49, respectively. These findings indicate that there is some form of hierarchy in this data and, consequently, this multilevel structure must be considered when analysing historical attrition as well as when building models to predict future attrition.

The historical attrition analysis explored in this section provides useful holistic understanding of historical areas of concern, with regards to healthcare workforce attrition, from a provincial and grouping level perspective. However, since this data is multilevel, these holistic areas of concern may not be areas of concern within all nested levels in the multilevel structure. For example, the cadre `Medical Services`,

was identified in Figure 10 to be one of the most problematic cadres with respect to annual attrition rates, when compared to all other cadres in the dataset over the study period. However, when attrition rates were computed for this cadre, whilst taking into account a 2-level nested structure (cadres within facilities), it became evident that whilst this cadre is problematic in the `Motherwell CHC`, it may not be an area of concern in another facility. Consequently, when building models to predict future attrition events, it cannot be assumed that dataset observations or categorical variable groupings are independent. The event of interest (*i.e.* outcome `event` of prediction), may be heavily influenced by the potential within-cluster homogeneity that is often present in multilevel data.

4.3 Chapter summary

In this chapter, focus was placed on gaining an understanding of the datasets available for use, from a structural and historical perspective, with the aim of identifying suitable modelling methods that can be applied to the data for future attrition prediction, as well as gain insight into areas of concern with respect to workforce attrition.

The chapter, therefore, discussed the process of the preliminary data analysis and data preprocessing (Section 4.1.1) and the required transformation of the dataset into a person-period format (Section 4.1.3). Based on this data analysis, the cleaned, preprocessed, and transformed `person_period_persal_ec_2010_2021` dataset is considered to be structurally appropriate for all modelling techniques described in Chapter 3. The chapter further interrogated the dataset from an historical perspective (Section 4.2.1). Based on this historical EDA, it was identified that the `person_period_persal_ec_2010_2021` dataset can be categorised as multilevel survival data, whereby, variation exists within certain grouping levels (specifically facilities and cadres), as well as between certain grouping levels (*e.g.* cadres nested in facilities, as described in Figure 12). Seeing as the dataset is structurally appropriate for all modelling techniques previously used for attrition prediction and is hierarchical, modelling methods with potential to handle multi-level data should be considered in this investigation for attrition prediction (Sections 3.4-3.5.3).

The historical EDA also identified that across all groupings and levels within these groupings, the average overall attrition rates over the study period exceeded the budgeted and planned for 5% attrition rate. As per Sections 1.1 and 2.1, this may have resulted in excessive costs exacerbated by inadequate staffing levels (such as high overtime). Consequently, there is a financial incentive, backed by data, for developing multilevel statistical and/or ML models capable of accurately predicting future attrition rates within and between multiple levels within the EC province.

5 Model development and results

One of the key aims of this investigation, as described in Section 1.2, was to develop predictive models, appropriate for the dataset used in this investigation, that are capable of effectively predicting future healthcare workforce attrition rates and are able to be utilised to determine factors that positively and negatively influence workforce attrition.

Based on the PEDA (Section 4.1.2) and historical EDA (Section 4.2.1), the `preprocessed_persal_ec_2010_2021` dataset can be categorised as multilevel survival data, containing both categorical and continuous explanatory variables, with a discrete event outcome variable. The dataset is in a longitudinal, person-period format and contains both time-invariant and time-varying explanatory variables.

Two overarching methods for modelling multilevel survival data, with the intent to predict future workforce attrition rates and determine factors influencing workforce attrition, were explored during the review of literature (Section 2). These include statistical modelling as well as machine learning modelling techniques. The following chapter explores the application of these different modelling techniques to this multilevel survival data, and includes an analysis and discussion of the results obtained thereof. Moreover, a comparison of the performance of the models developed using the different techniques is also explored. All models are developed in the `R` programming language (Chambers, 2008).

5.1 Training, validation, and testing sets

The machine learning algorithms identified in literature (Section 2.2.2), typically require to be modelled and evaluated on training, validation, and test datasets. Consequently, in order to compare the performance of the statistical and ML models developed in this investigation, all models must be fitted on training data, tuned on validation data, and thereafter, evaluated using the hold-out (test) dataset.

For the purpose of this investigation, an 80:10:10 training:validation:test split was chosen (Snijders and Bosker, 2011). It is important to note, however, that the partitioning of the data into these datasets was done by the `reporting_year` variable. According to Snijders and Bosker (2011), partitioning multilevel survival data at an observational level (*i.e.* by year) has proven to be effective when the dataset is representative of the whole population. Seeing as the `preprocessed_persal_ec_2010_2021` dataset contains all possible workers in the EC health sector for the study duration, utilizing `reporting_year` for partitioning is deemed sufficient.

5.2 Evaluation metrics

During PEDA (Section 4.1.3), it was identified that the outcome variable of interest for this investigation, `event`, is imbalanced. According to Hossin and Sulaiman (2015), classification accuracy can be highly misleading when the dataset is imbalanced and, therefore, cannot be used in isolation to determine the overall performance of statistical and ML models. Consequently, `Misclassification rate`, `Recall`, `Specificity`, `F1 score`, `ROC AUC` and `P-value` metrics are to be computed. The definitions of these metrics are defined as follows:

- `Classification accuracy`: measures the ratio of correct predictions over the total number of instances evaluated.
- `Misclassification rate`: is the ratio of incorrect predictions over the total number of instances evaluated.
- `Recall`: used to measure the fraction of positive patterns that are correctly classified.
- `Specificity`: used to measure the fraction of negative patterns that are correctly classified.
- `F1 score`: represents the harmonic mean between recall and precision values.
- `ROC AUC`: represents the degree or measure of separability (*i.e.* it indicates how much the model is capable of distinguishing between classes).
- `P-value (p)`: represents is the level of marginal significance within a statistical hypothesis test. For the purpose of this investigation, a `p-value` less than 0.001 is considered statistically significant.

5.3 Statistical models

According to the methodological and literature reviews (Sections 3 and 2.2.1), Multilevel Discrete-Time Event (MDTE) models can, and have been, applied to such data with the intent to predict discrete attrition events and determine factors influencing attrition (or like terms). These models have also proven to be effective at identifying and confirming the multilevel nested (*i.e.* grouping) structure that exists in the data. Research further suggests that these models are very effective when modelling multivariate, multilevel survival data with two nesting levels.

Based on the PEDA (Section 4.1.2) and historical EDA (Sections 4.2.1), a hierarchical data structure of five levels seems to be present in this `preprocessed_persal_ec_2010_2021` dataset, whereby "repeated" measures are nested within a healthcare worker, who is nested in a cadre, which is nested

in a facility, which is nested in a district. Moreover, as identified in Section 4.1.2, Table 1, this dataset is considered complex due to the fact that it contains a significant number of categorical predictor variables that have a large number of levels, respectively.

Although research indicated that these models can be extended to data with more than two nesting levels, to my knowledge no actual application to data with three or more nesting levels has been undertaken. The review of literature was also unable to identify applications of the method to data with a significant portion of its predictor variables being categorical, with these categorical variables having an extremely large number of levels. Despite this, knowledge gained from the methodological and literature review (Section 3 and 2.2.1) suggest that the possibility for application of this modelling technique to such complex and nested data, exists.

Consequently, MDTE models are developed using the `glmer` package in `R` (Bates et al., 2005), and the results discussed in Section 5.3.1.

5.3.1 Multilevel discrete-time event models

The first MDTE model to be fitted is the Baseline Hazard (BH) model. When dealing with multilevel survival data, the BH model should consist of the time-indicator variables as fixed effects and a random varying intercept for a specific nested structure (Hox et al., 2017). Although the `preprocessed_persal_ec_2010_2021` dataset has been previously determined to be multilevel, the actual nested multilevel structure of the data has not yet been established. Consequently, several BH models were trained on the training data, with each model representing a different possible nesting structure combination. Modelling combinations are described in Table 5.

BH Model	Number of Levels	Nested Structure
mlsa_B1	2	"repeated" measures nested within healthcare workers
mlsa_B2	4	"repeated" measures nested within healthcare workers, nested in cadres, nested in facilities
mlsa_B3	3	healthcare workers nested in cadres, nested in facilities
mlsa_B4	5	"repeated" measures nested within healthcare workers, nested in cadres, nested in facilities, nested in districts
mlsa_B5	4	healthcare workers, nested in cadres, nested in facilities, nested in districts
mlsa_B6	3	healthcare workers nested in facilities, nested in districts

Table 5: Baseline Hazard multilevel modelling combinations.

Only three of the BH models described in Table 5 converged, namely `mlsa_B2`, `mlsa_B3`, and `mlsa_B5`. The summary of these BH models are displayed in Table 6.

	(mlsa_B2) **event**			(mlsa_B3) **event**			(mlsa_B5) **event**		
Predictors	*Risk Ratios*	*CI*	*p*	*Risk Ratios*	*CI*	*p*	*Risk Ratios*	*CI*	*p*
(Intercept)	0.14	0.13 – 0.14	**<0.001**	0.14	0.13 – 0.14	**<0.001**	0.14	0.13 – 0.15	**<0.001**
exit	1.34	1.32 – 1.36	**<0.001**	1.30	1.28 – 1.31	**<0.001**	1.30	1.28 – 1.31	**<0.001**
years in system before study start	0.83	0.82 – 0.84	**<0.001**	0.83	0.83 – 0.84	**<0.001**	0.84	0.83 – 0.84	**<0.001**
Random Effects									
σ^2	1.64			1.64			1.64		
τ_{00}	0.05 persal_number:(unit_field:facility_name)			0.64 unit_field:facility_name			0.64 unit_field:(facility_name:district)		
	0.67 unit_field:facility_name			0.08 facility_name			0.08 facility_name:district		
	0.08 facility_name						0.01 district		
ICC	0.33			0.30			0.31		
N	76307 persal_number			38 unit_field			38 unit_field		
	38 unit_field			841 facility_name			841 facility_name		
	841 facility_name						9 district		
Observations	428896			428896			428896		
Marginal R^2 / Conditional R^2	0.027 / 0.344			0.024 / 0.321			0.024 / 0.322		

Table 6: Baseline Hazard model summaries.

The Intra-Class Correlation (ICC) is a correlation coefficient that indicates the proportion of the total variance explained by the grouping/nesting structure in the population. Based on Table 6, it is evident that all three converging BH models display an ICC greater than 30%. Consequently, around 30% of the variation in the outcome variable can be explained by either of the three nested structures modelled in these BH models. Three goodness-of-fit metrics can be used to determine which of the three nested structure combinations results in the best model fit, namely the Likelihood Ratio Test (LRT), the Akaike Information Criterion (AIC), and Misclassification Rate when predicted on the validation dataset. The LRT is useful to test whether the observed difference in model fit is statistically significant (*i.e.* a p-value < 0.001). Different from LRT, AIC deals with the trade-off between goodness-of-fit and model complexity, and as a result, dis-encourages overfitting (Hox et al., 2017). A smaller AIC is preferred. The goodness-of-fit metrics are displayed in Table 7.

BH Model	**AIC**	**Misclassification Rate (%)**	***p***
mlsa_B2	366097	11.76	**<0.001**
mlsa_B3	366159	11.90	
mlsa_B5	366157	11.91	

Table 7: Goodness-of-fit metrics for Baseline Hazard models.

According to the results in Table 7, the `mlsa_B2` BH model obtains the lowest AIC and misclassification rate when predicted on the validation set. Additionally, the LRT resulted in a `p-value` for the `mlsa_B2` model of less than 0.001, indicating that the observed difference in model fit is statistically significant when compared to BH models `mlsa_B3` and `mlsa_B5`, respectively. The `mlsa_B2` BH model will thus be used as the Baseline Hazard model for this investigation.

On further inspection of this model's summary results, displayed in Table 6, it is evident that the variance of the healthcare workforce (`persal_number`) observation level residual errors, symbolized by σ^2, is estimated as 1.64. The variance of the cadre level (nested in facilities) residual errors is the largest group level variance experienced in this nested structure. This suggests that most of the group level variation exists between cadres, when nested in facilities. Both time-indicator variables are highly statistically significant at $p < 0.001$. The Conditional R^2 value obtained for this model is 0.344, indicating that 34.4% of the variance is explained by both the fixed and random factors. This is slightly higher than the ICC, indicating that these time-indicator variables do slightly improve the models ability to explain variation in the outcome event of interest.

Considering that "repeated" measures are nested within healthcare workers, who are nested in cadres, which are nested in facilities, it is important to define the different level-specific variables. This is defined below:

- Observational level (level 1): time-indicator variables (`exit`, `years_in_system_before_study_start`)
- Person level (level 2): `gender`, `race`, `age`, and `notch_value`
- Cadre level (level 3): no level 3 specific variables
- Facility level (level 4): `facility_type` and `rural_urban`

After establishing the BH model for the investigation, level 2 and level 4 predictor variables are to be added to the model formulation in separate iterations, and the performance of these models compared to the BH model using the goodness-of-fit metrics. Adding all level 2 predictor variables to the BH model formulation did result in a model that converged. The summary of this model, `mlsa_L2`, is displayed in Table 8.

	(mlsa_L2) event		
Predictors	*Risk Ratios*	*CI*	*p*
(Intercept)	0.14	0.13 – 0.14	**<0.001**
exit	1.37	1.35 – 1.39	**<0.001**
years in system before study start	0.83	0.82 – 0.84	**<0.001**
gender [Male]	1.08	1.06 – 1.10	**<0.001**
age	1.06	1.05 – 1.07	**<0.001**
race [Asian]	1.14	1.04 – 1.25	**0.005**
race [Coloured]	0.95	0.92 – 0.98	**0.002**
race [White]	1.21	1.17 – 1.26	**<0.001**
notch value	0.71	0.70 – 0.73	**<0.001**
Random Effects			
σ^2	1.64		
$\tau_{00\ persal\_number:(unit\_field:facility\_name)}$	0.04		
$\tau_{00\ unit\_field:facility\_name}$	0.74		
$\tau_{00\ facility\_name}$	0.07		
ICC	0.34		
N_{persal_number}	76307		
N_{unit_field}	38		
$N_{facility\_name}$	841		
Observations	428896		
Marginal R^2 / Conditional R^2	0.035 / 0.366		

Table 8: Summary of the multilevel, discrete-time event model (`mlsa_L2`) with time-indicator and level 2 predictor variables.

Based on the model summary in Table 8, the regression coefficients for all predictor variables are statistically significant. When comparing the summary results of the `mlsa_L2` model to those of the BH model, displayed in Table 6, it is evident that the variance of the healthcare workforce (`persal_number`) observation level residual errors, symbolized by σ^2, are estimated to be the same, at 1.64. The variance of the cadre (nested in a facility) level residual errors in the `mlsa_L2` model is higher than that obtained in the BH model. In both models, this grouping variable is considered to explain the greatest proportion of the total variance explained by the grouping/nesting structure in the population. Although the ICC for the `mlsa_L2` model is only 1% higher than the ICC obtained for the BH model, the Conditional R^2 value obtained for this model is 36.6%, indicating that fixed and random effects of the `mlsa_L2` model are able to explain 2.2% more of the variation in the outcome variable than the BH model. Including the level 2 predictor variables in the model formulation also improved the Marginal R^2 value (variation explained by fixed effects alone), from 2.7% to 3.5%. In order to determine if the `mlsa_L2` model resulted in

a better model fit than the BH model, the three GOF metrics, previously defined, were computed. The GOF metrics are displayed in Table 9.

BH Model	**AIC**	**Misclassification Rate (%)**	***p***
mlsa_L2	364860	11.56	**<0.001**
mlsa_B2	366097	11.76	

Table 9: Comparison of goodness-of-fit metrics for the `mlsa_L2` and `mlsa_B2` models.

According to the results in Table 9, the `mlsa_L2` model obtains a lower AIC, lower misclassification rate, and is considered to be statistically different from the BH model. This model can, therefore, be considered an improvement on the BH model.

In order to determine the impact/influence that these variables have on the outcome variable, it was important to assess them using risk ratios (Hox et al., 2017). According to Table 8, males are 8% more likely to experience an attrition event relative to females. When looking at the race variable (made up of four factors), white individuals appear to be at the highest risk of experiencing an attrition event relative to all of the other race groups, followed by Asian, African, and least at risk being Coloured individuals. Additionally, the risk ratio for the `age` covariate, indicates that an older individual is at higher risk of experiencing an attrition event to younger individuals. The risk ratio for `notch_value` further suggests that the higher the annual salary of an individual in the EC health sector, the lower the risk of experiencing an attrition event.

Moreover, the risk ratio for `exit` suggests that the greater the number of years that an individual has been working, at a specific facility in the system, since the start of the study period, the higher the risk of that individual experiencing an attrition event. Conversely, the risk ratio for `years_in_system_before_study_start` suggests that the individuals who were working at the same facility for a long time before the study period, are less likely to experience an attrition event during the study period. These time-indicator variables (`exit`, `years_in_system_before_study_start`) are both related to the duration that a healthcare worker has been working, in a given facility, in the EC health system. Consequently, the effect of `exit` on the risk of an individual experiencing an attrition event may depend on whether or not that individual had been working at that facility prior to the study start and, if so, may further depend on the duration of this employment.

Based on these findings, further investigation into the potential time-indicator interaction effects is required. In order to identify these effects, these interaction terms are added to the original BH model formulation, without the addition of any level 2 predictor variables. The model summary for the original BH model (`mlsa_B2`)

and the BH model with time-indicator variable interaction effects (`mlsa_B2I`), are displayed in Table 10.

	(mlsa_B2) **event**			(mlsa_B2I) **event**		
Predictors	*Risk Ratios*	*CI*	*p*	*Risk Ratios*	*CI*	*p*
(Intercept)	0.14	0.13 – 0.14	**<0.001**	0.13	0.13 – 0.14	**<0.001**
exit	1.34	1.32 – 1.36	**<0.001**	1.25	1.23 – 1.27	**<0.001**
years in system before study start	0.83	0.82 – 0.84	**<0.001**	0.85	0.84 – 0.85	**<0.001**
exit * years in system before study start				1.22	1.21 – 1.23	**<0.001**
Random Effects						
σ^2	1.64			1.64		
τ_{00}	0.05 persal_number:(unit_field:facility_name)			0.01 persal_number:(unit_field:facility_name)		
	0.67 unit_field:facility_name			0.60 unit_field:facility_name		
	0.08 facility_name			0.08 facility_name		
ICC	0.33			0.29		
N	76307 persal_number			76307 persal_number		
	38 unit_field			38 unit_field		
	841 facility_name			841 facility_name		
Observations	428896			428896		
Marginal R^2 / Conditional R^2	0.027 / 0.344			0.040 / 0.321		

Table 10: Summary of BH models, with and without interaction effects.

According to Table 10, the coefficients for all variables, including the interaction terms, are statistically significant. The ICC and the conditional R^2 for the BH model with interaction is 4% and 2.3% lower than the ICC and conditional R^2 for the original BH model, respectively. However, the marginal R^2 value is higher by 1.3%. This high increase in the marginal R^2 value indicates that by adding these interaction terms, a higher proportion of variance explained by the fixed effects alone, over the overall variance, is achieved. The GOF metrics for these two models are displayed in Table 11.

BH Model	AIC	Misclassification Rate (%)	p
mlsa_B2I	364677	11.72	**<0.001**
mlsa_B2	366097	11.76	

Table 11: Goodness-of-fit metrics for the best BH model with and without interaction.

According to the results in Table 11, the BH model with interaction effects (`mlsa_B2I`) obtains a slightly lower AIC and misclassification rate, and is considered to be statistically different from the original BH model. This model can, therefore, be considered a slight improvement on the BH model. According to research by Hox et al. (2017), it is difficult to determine whether interaction effects are actually present in the data by simply looking at risk ratios or coefficient estimates. Consequently, graphically representing this potential underlying relationship using interaction plots has become industry standard. The interaction plot for the interaction terms modelled in `mlsa_B2I` is displayed in Figure 13.

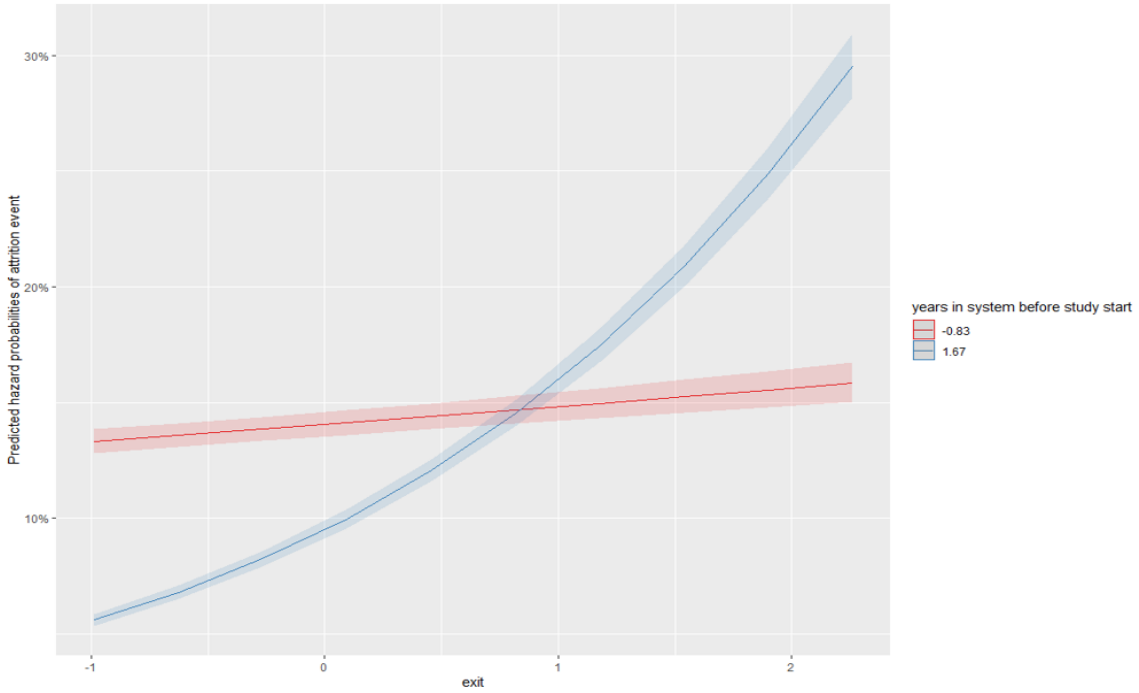

Figure 13: Interaction Plot indicating interaction effects between the time-indicator variables for the standardized dataset.

The intersecting lines in the interaction plot, displayed in Figure 13, indicates that

there is an interaction between the two time-indicator variables (Hox et al., 2017). As the `exit` variable increases, the hazard probability (the probability of experiencing an attrition event) increases regardless of the number of years an individual was employed at a facility in the EC before the study start. However, the hazard probability seems to increase far more drastically as the `exit` variable increases, in cases where an individual was employed in the system for a long time before the commencement of the study. Consequently, the risk ratio obtained for this interaction effect, displayed in Table 11, indicates that the longer an individual is in the system in total, the higher their risk of leaving. This further suggests that an individual who had been in the system for the same number of intervals (*i.e.* `exit`) as another individual, since the study start, would have a higher risk of leaving if they had been in the system prior to the study start. Catering for left-censored observations was, therefore, important in this investigation (Section 3.1.1).

Considering that these interaction effects are significant, it was important to follow the same methodological process, and introduce level 2 predictor variables into the formulation of this BH model (with interaction terms). The summary of this model (`mlsa_L2I`) is displayed in Table 12.

	(mlsa_L2I) **event**		
Predictors	*Risk Ratios*	*CI*	*p*
(Intercept)	0.13	0.12 – 0.13	**<0.001**
exit	1.28	1.26 – 1.30	**<0.001**
years in system before study start	0.84	0.83 – 0.85	**<0.001**
gender [Male]	1.08	1.06 – 1.10	**<0.001**
age	1.07	1.06 – 1.08	**<0.001**
race [Asian]	1.14	1.05 – 1.25	**0.003**
race [Coloured]	0.96	0.93 – 0.99	**0.007**
race [White]	1.21	1.16 – 1.25	**<0.001**
notch value	0.86	0.85 – 0.87	**<0.001**
exit * years in system before study start	1.23	1.22 – 1.24	**<0.001**
Random Effects			
σ^2	1.64		
$\tau_{00\ persal\_number:(unit\_field:facility\_name)}$	0.01		
$\tau_{00\ unit\_field:facility\_name}$	0.69		
$\tau_{00\ facility\_name}$	0.07		
ICC	0.32		
N_{persal_number}	76307		
N_{unit_field}	38		
$N_{facility\_name}$	841		
Observations	428896		
Marginal R^2 / Conditional R^2	0.046 / 0.351		

Table 12: Summary of MDTE model with level 2 predictor variables as well as interaction terms.

According to the summary results for the `mlsa_L2I` model (Table 12), it is evident that the risk ratios for the variables are all roughly the same as previously obtained in the `mlsa_L2` model (Table 8). Consequently, the interpretation of the influencing factors of attrition previously explored still applies. However, the `race` variable is no longer statistically significant for the Asian and Coloured levels, indicating that these levels do not seem to influence the probability of an attrition event occurring. Moreover, interpreting the time-indicator variable risk ratios as individual covariates is not feasible due to the interaction term being statistically significant. The value of the risk ratio obtained for the time-indicator interaction term is the same as what was obtained in the `mlsa_B2I` model. Consequently, the interpretation of this terms effect on hazard holds for this model as well. The conditional R^2 value did decrease by 1.4%, however, the marginal R^2 increased from 3.5% to 4.6%.

In order to determine if the `mlsa_L2I` model resulted in a better model fit than the

`mlsa_L2` or `mlsa_B2I` models, the three GOF metrics were computed. The GOF metrics are displayed in Table 13.

BH Model	AIC	Misclassification Rate (%)	p
mlsa_L2	364860	11.56	**<0.001**
mlsa_B2I	364677	11.72	
mlsa_L2I	363794	11.70	

Table 13: Comparison of goodness-of-fit metrics for the `mlsa_L2`, `mlsa_B2I`, and `mlsa_L2I` models.

Interestingly, the `mlsa_L2` model without interaction, although experiencing a higher AIC than the other two models, obtained the lowest misclassification rate, and has a model fit that is considered to be statistically significant. According to Hox et al. (2017), multilevel models with a high number of nesting levels, as well as categorical variables with many levels, can become increasingly complex very quickly, specifically when having to cater for interaction terms or randomly varying slopes. In such instances, predictive performance tends to decrease for models that do converge, and the addition of any higher level predictor variables in model formulation usually results in models being unable to converge (Hox et al., 2017). This hypothesis was confirmed in this investigation, as all model formulations attempting to add predictor variables higher than level 2, resulted in an inability for the models to converge.

For all multilevel, discrete-time event models developed in this investigation, modelling parameters such as choice of optimizer (`bobyqa`) and link function (`cloglog`) were selected based on findings from the methodological and literature reviews (Sections 3 and 2.2.1). Maximum iterations was also set to 20000 to increase the probability of model convergence during training. Consequently, convergence issues experienced in this investigation can, therefore, be assumed to be attributed to the dataset complexity.

The results of this statistical model investigation for attrition prediction indicates that the `mlsa_L2` model is the best model for attrition prediction, although only by a small amount. Consequently, both the `mlsa_L2` and `mlsa_L2I` models are assessed on their ability to generalise to new, unseen data by using the models to predict attrition using the test dataset. The evaluation statistics for these models are described in Table 14.

Model	Classification Accuracy (%)	Recall (%)	Specificity (%)	F1 Score	ROC AUC
mlsa_L2	89.04	9.57	97.81	27.26	63.58
mlsa_L2I	88.91	7.36	97.92	29.90	66.68

Table 14: Multilevel discrete-time event model evaluation statistics.

According to Table 14, both the `mlsa_L2` and `mlsa_L2I` models obtain high levels of classification accuracy when predicted on unseen test data. During PEDA (Section 4.1.3), it was identified that this data is heavily imbalanced, with respect to the outcome variable of interest. However, the class ratios obtained for each reporting year never subceed the resampling requirement threshold of 0.1 (Section 4.1.3). Despite this, attempts were made to perform stratified sampling on the dataset prior to model development, however, performing this sampling technique on this dataset was not possible due to the highly complex hierarchical data structure present in the dataset. Consequently, this is acknowledged as a potential limitation of the models developed in this investigation. As discussed in Section 5.2, classification accuracy can be highly misleading when the dataset is imbalanced and, therefore, cannot be used in isolation to determine the overall performance of these models. Consequently, recall, specificity, F1 Score and ROC AUC metrics were computed.

Both models obtained high specificity (97.81% and 97.92%), indicating that the models are both able to classify well between correctly and incorrectly labeling non-attrition events. The possibility of a Type-I error (incorrectly labelling an attrition event as a non-event), is therefore very low. Conversely, both models obtain incredibly poor recall results, at 9.57% and 7.36%, respectively. This indicates that these models are highly likely to predict a non-event as an attrition event (*i.e.* the probability of Type-II error is incredibly high). The F1 Scores are also low, and this expected due to the low recall values obtained for both models. The ROC AUC values of 63.58% and 66.68% for models `mlsa_L2` and `mlsa_L2I`, respectively, suggests that there is a relatively high probability that a predicted non-event will in fact be a non-event.

Based on these findings, the `mlsa_L2` model performs better than the `mlsa_L2I` model, albeit by a very small amount. Although the probability of a Type-II error is high, in the context of this problem, it is better to predict a false attrition event, then predict an event to be a non-event when it is in fact an attrition event. This is argued, as the implications for under budgeting attrition rates can lead to sub optimal efficiencies in the system such as poorer staffing capacity levels, which has multiple knock-on effects (*i.e.* increased overtime, increased workload, and a decline in quality of health care provided) (Section 1.1). Additionally, findings from the historical attrition analysis further suggests that high variation in attrition also exists between nested levels across the different years (Section 4.2.1). Incorrectly managing these fluctuations would also exacerbate the sub optimal efficiencies previously defined. Consequently, is it deemed more appropriate, in this context, to accept a higher Type-II error, than a high Type-I error.

The `mlsa_L2` model was thus used to predict the future attrition rates for the 2021 year. These predictions are discussed in Section 5.6.

5.4 Machine learning models

According to Somers (1999) research, discussed in the review of literature (Section 2.2.2.1), accurately predicting employee attrition has been an ongoing topic in applied research and is an area that consistently calls upon researchers to attempt new and unconventional methods to advance the field of study.

Three ML algorithms, discussed in the review of literature (Section 2.2.2), have previously been successfully applied to attrition prediction problems, namely NNs, GLMM trees, and XGB trees. These methods, however, have only been applied to these types of problems in a limited capacity. NNs can, and have been, applied to survival data with the intent to predict discrete events and determine factors influencing the outcome variable of interest, however, they have - to my knowledge - not previously been applied to multilevel survival data. GLMM trees, on the other hand, have been successfully applied to multilevel survival data but have been documented to become incredibly complex when this multilevel data contains many nested levels, and many factor predictor variables comprising of many levels. Although XGB trees have been documented to have the best predictive performance when compared to more traditional statistical models for attrition prediction, the models in these investigations did not take into account the potential multilevel survival nature of the data. According to the review of literature (Section 2.2.2.3), TBME trees are a novel methodology aiming to extend boosted-tree algorithms to cater for this type of data, however, these models have only ever been applied to small, and not very complex, samples of multilevel survival data.

Although there is no applied research detailing the implementation of these methods to data as complex as the dataset used in this investigation, the review of methodology and literature (Sections 3.5 and 2.2.2), suggest that these models, in theory, should be able to cater for such complex data. Moreover, these ML models are known to perform well when the data is non-linear (Section 2.2.2). Due to the fact that interaction terms were identified to be statistically significant during the MDTE model development discussed in Section 5.3.1, non-linearity is present in the dataset creating further motivation to attempt these methods on the `preprocessed_persal_ec_2010_2021` dataset.

Consequently, three ML models - NNs, GLMM trees, and TBME models - are proposed and the results discussed in Sections 5.4.1-5.4.3. These ML models are developed in `R` using the `h2o`, `glmertree`, and `gpboost` packages, respectively.

5.4.1 Multilayer perceptron neural networks

The first ML modelling technique to be attempted in this investigation is the NNs models. In the study conducted by Kvamme and Borgan (2019), discussed in the

review of literature (Section 2.2.2.1), LVQ NNs obtained slightly better performance for attrition prediction than the MLP NN paradigm. However, MLP NNs are considered the most widely used NNs, and for the purpose of this study, will be selected as the modelling technique of choice.

As proposed in the research methodology, discussed in Section 1.4, the first NN to be developed is the baseline model which is to contain the simplest NN architecture structure and serve as a model for which the performance of other NNs can be compared. The baseline NN model architecture consists of an input layer, a hidden layer, and an output layer. As per Section 3.5.1, the input layer must contain 12 neurons to match the feature space, 1 neuron in the hidden layer, and 2 neurons in the output layer to match the discrete outcome variable of interest (*i.e.* attrition event or non-event). Additionally, the activation and loss functions are set to `tanh` and `cross entropy`, respectively. The learning rate is not specified because an adaptive learning rate is chosen instead[12]. This NN is trained on the training dataset and performance tested on the validation set. The evaluation statistics for this basline NN are described in Table 15.

Model	**Classification Accuracy (%)**	**Recall (%)**	**Specificity (%)**	**F1 Score**	**ROC AUC**
NN_B1	44.81	69.08	41.73	19.62	61.90

Table 15: Baseline NN model evaluation statistics when evaluated on the validation set.

According to Table 15, the Baseline NN obtained a low accuracy of 44.81%. As mentioned in Section 5.2, classification accuracy can be highly misleading when the dataset is imbalanced and, therefore, cannot be used in isolation to determine the overall performance of these models. Consequently, recall, specificity, F1 Score and ROC AUC metrics were computed. The specificity obtained by the model when tested on the validation set is also very low (41.73%). This indicates that the model has a high probability of labelling an attrition event as a non-event (*i.e.* a high Type-I error). Conversely, the model is able to obtain a slightly better recall of 69.08% indicating that the model is less likely to predict a non-event as an attrition event (*i.e.* the probability of Type-II error is relatively low). The F1 Score is also rather low, but this is expected due to the low specificity and relatively low recall values obtained. The ROC AUC value of 61.90% suggests that there is a relatively high probability that a predicted non-event will in fact be a non-event. Overall, this model does not seem to be able to generalise well to new, unseen data.

[12]Adaptive learning rate: an optimization of gradient descent methods with the goal of minimizing the objective function of a network by using the gradient of the function and the parameters of the network (Hagan et al., 1997).

As identified in the review of methodology (Section 3.5.1), the optimal combination of hyper-parameters set in the Baseline NN is not immediately known, and for this reason, hyper-parameter tuning must be undertaken. Performing this step can also aid in improving model performance and is a critical step in the iterative process of model development. It is important to note that creating a search space with a large number of unique combinations of hyper-parameters is computationally expensive. For this reason, only a small number of values, per hyper-parameter, are considered in the search space for the second NN model architecture. These include:

- Number of neurons in the hidden layer: varied between 3 and 6
  - There is a lot of conflicting research on the optimal search space for the number of neurons in a hidden layer (Hagan et al., 1997). A large number of neurons in a hidden layer increases model complexity (*i.e.* increased possibility of overfitting) and increases computational time (since more neurons result in an increased number of weights having to be optimized during back propagation, during model training). Conversely, using too few neurons in the hidden layers will result in underfitting (Hagan et al., 1997). For this reason, a range of values have been selected, namely i) a number of neurons around a third that of the input layer, and ii) number of neurons around half to the number of neurons in the input layer
- Number of epochs: varied between 10 and 50
  - Epochs indicate the number times that the learning algorithm will work through the entire training dataset (Section 3.5.1). When the number of epochs used to train a neural network model is more than necessary, the training model learns patterns that are specific to sample data to a great extent. This makes the model incapable to perform well on a new dataset (*i.e.* overfitting occurs) (Hagan et al., 1997). To mitigate overfitting and to increase the generalisation capacity of the neural network, the model should be trained for an optimal number of epochs. In order to determine this optimal epoch value, two different values for epochs are considered in the search space.
- Ridge regression, `l2`: varied between 0.01 and 0.001
  - It is a regularisation technique for neural network models (*i.e.* used to prevent NN's from overfitting) that controls model complexity by dispersing the error terms in all of the weights (Section 3.5.1).

The optimal hyper-parameters identified during tuning, are found to be 6 neurons in the hidden layer, with the number of epochs set to 10 and a ridge regression value of 0.001. The rest of the parameters used for the creation of the Baseline NN remain

the same. The second NN is fitted, with these optimal hyper-parameters specified, and evaluated on the validation set. The models evaluation statistics are displayed in Table 16.

Model	Classification Accuracy (%)	Recall (%)	Specificity (%)	F1 Score	ROC AUC
NN_01	56.97	52.06	63.32	21.03	60.08

Table 16: The optimal hyper-parameter trained NNs evaluation statistics when evaluated on the validation set.

Comparing the baseline NN performance metrics, Table 15, against the model trained on optimal hyper-parameters, Table 16, it is evident that the `NN_01` model does seem to outperform the Baseline NN when assessed on classification accuracy, specificity, and F1 score. The `NN_01` model obtained an accuracy value that is 12.16% higher than the baseline model. Moreover, the specificity obtained for the `NN_01` model is 21.59% higher than what was obtained for the baseline NN. This is a significant improvement and indicates that probability of obtaining a Type-I error (incorrectly labelling an attrition event as a non-event) is lower than the Baseline NN. Conversely, the `NN_01` model obtains lower values for ROC AUC and for recall when compared to the Basleine NN. This lower recall value suggests that this model has a higher probability of experiencing a Type-II error than the baseline NN. As discussed in Section 5.3.1, and for the context of this problem, it is better to predict a false attrition event, than predict an event to be a non-event when it is in fact an attrition event. Consequently, obtaining a lower Type-I error is preferred over a lower Type-II error. For this reason, the `NN_01` model is considered the better model for attrition prediction when compared to the Baseline NN. It is also important to note, that the inclusion of regularisation in the model does seem to improve the overall model performance. Unfortunately, the classification accuracy and specificity is still considered rather low, as both are less than 80% and for this reason, this model may still not be sufficient for attrition prediction in the context of this investigation.

Further hyper-parameter tuning can be applied to potentially improve this performance even more. For example, this could be done by increasing the number of neurons in the hidden layer, adding additional hidden layers, varying activation functions, or using a specified learning rate over an adaptive learning rate (Section 3.5.1). As previously mentioned, however, performing grid searches that contain a large number of unique combinations of hyper-parameters is computationally expensive. The hyper-parameter tuning process performed for model `NN_01` took 6 days and 37 minutes to complete. This grid search only comprised of 12 unique combinations. Consequently, there seems to be a trade-off between sacrificing time and resources to potentially obtain better predictive accuracy. Performing the first grid

search did not result in a large enough increase in overall performance to warrant the significant time required to perform another more comprehensive grid search. For this reason, no additional NNs are developed for the purpose of this investigation, however, this is proposed as future work (Section 6.1).

The results of this NN model investigation for attrition prediction indicates that the `NN_01` model is the best NN model for attrition prediction. Consequently, the `NN_01` model is to be assessed on its ability to generalise to new, unseen data. This is assessed by using the model to predict attrition using the test dataset. The evaluation statistics are described in Table 17.

Model	Classification Accuracy (%)	Recall (%)	Specificity (%)	F1 Score	ROC AUC
NN_01	58.01	52.01	65.02	21.98	60.56

Table 17: The optimal hyper-parameter trained NNs evaluation statistics when evaluated on the test set.

The performance metrics obtained for the `NN_01` model when tested on the set, displayed in Table 17, are virtually the same as the metrics obtained when the model was predicted on the validation set (results displayed in Table 16). The classification accuracy and specificity obtained, however, was about 2% higher when the model was evaluated using the test dataset.

In the cases of both the `NN_B1` and `NN_01` models, the input layer in the NN architecture comprised of all 12 possible predictor variables. Consequently, the effects of all possible nested grouping possibilities (*i.e.* hierarchical data structures) on attrition event prediction are considered in the model formulation. Based on the NNs developed in this investigation, and the model performance obtained thereof, NNs do not seem to be able to cater for the `person_period_persal_ec_2010_2021` datasets multilevel data structure very well. This observation, however, is limited by the number of grid searches undertaken in this investigation.

5.4.2 GLMM trees

The second ML modelling technique to be attempted in this investigation is the GLMM tree models. As defined in the research methodology, discussed in Section 1.4, a baseline GLMM tree is first modelled. This baseline model regresses the discrete event outcome variable (*i.e.* `event`) on the nested groups as random effects and the time-indicator variables as fixed effects. The nested groups used here are the same as what was defined in Section 5.3.1 (*i.e.* four-level nested structure). Consequently, the definitions of the level-specific variables for modelling purposes, also defined in Section 5.3.1, are applicable in this model context.

This baseline GLMM tree is to be fitted on the training dataset and tested on the validation set, and its resulting evaluation metrics interrogated. Thereafter, additional level 2 and level 4 predictor variables should be added to the model formulation, and performance reassessed and compared to the baseline GLMM tree performance. Unfortunately, all GLMM tree models, catering for a four-level nested structure, and attempted in this investigation, failed to converge. Consequently, the development of two and three-level baseline GLMMs were attempted. Unfortunately, these models also failed to converge.

According to the literature review (Section 2.2.2.2), GLMM trees are known to experience convergence issues when applied to highly complex data, specifically:

- data with more than 2 nested levels, and/or
- data with many factor variables comprising of a large number of levels.

The convergence failure of all of the GLMM trees (two, three and four-level trees) attempted in this investigation confirms this complexity limitation of this modelling method. Consequently, the `person_period_persal_ec_2010_2021` dataset contains data that is considered to be too complex for this modelling technique, regardless of the number of levels catered for in the model formulation.

5.4.3 TBME models

The last ML modelling technique to be attempted in this investigation is the TBME models. As proposed in the research methodology, discussed in Section 1.4, the first TBME model to be developed is the baseline model which is to contain a simple ensembled tree structure and serve as a model for which the performance of other TBMEs can be compared. The baseline TBME model parameters are set to the following values:

- Learning rate, α: 0.01
- Maximum depth of the tree: 10
- Minimal number of samples per leaf: 5
- Number of leaves in a tree: set to the default of 31
- Number of boosting iterations: 100
- Objective: Binary (for classification)
- Likelihood: Bernoulli-logit (for classification)

This baseline TBME model is to be fitted on the training dataset and tested on the validation set, and its resulting evaluation metrics interrogated. Thereafter,

and similar to the modelling process followed for NNs, hyper-parameter tuning is required to find the optimal combination of parameters that results in the best possible model performance. The grid search parameter combinations are detailed below:

- Learning rate, α: varied between 0.01 and 0.001
- Maximum depth of the tree: varied across 3, 5, and 10
- Minimal number of samples per leaf: varied across 1, 10, and 100
- Number of boosting iterations: varied between 100 and 150

Unfortunately, all TBME models attempted in this investigation, failed to converge or resulted in excessive computational time[13] forcing an abort. According to the methodological and literature review (Sections 3.5.3 and 2.2.2.3), TBMEs to my knowledge, have never been applied to data of this complex nature. In order to determine the efficacy of this method on HR data such as the `preprocessed_persal_ec_2010_2021` dataset, possible dataset sampling methods could be applied to reduce the dataset size, and thus complexity. However, this is out of scope for this investigation and should be conducted or explored in future work (Section 6.1).

[13]Excessive computational time: in this investigation, models are defined to take an excessive amount of computational time during model training, if the model ran for more than 30 days without reaching a training progress level of 1%.

5.5 Model comparison

In order to achieve the aim of this investigation, defined in Section 1.2, one of the key objectives of the study was to to determine the most effective proposed model for attrition prediction, within the context of this problem statement (Section 1.3). Consequently, it was necessary to compare the generalisability of the best statistical and machine learning prediction models developed (Sections 5.3.1 and 5.4.1), so as to determine the most effective model for predicting attrition rates in the public health sector; and thereafter, use the best model to generate attrition predictions for the 2021 year.

The two models used in this comparison include the `mlsa_L2` multilevel discrete-time event statistical model and the `NN_01` multilayer perceptron neural network ML model. The evaluation statistics for both of these models, when assessed on the test set, are displayed in Table 18.

Model	**Classification Accuracy (%)**	**Recall (%)**	**Specificity (%)**	**F1 Score**	**ROC AUC**
mlsa_L2	89.04	9.57	97.81	27.26	63.58
NN_01	58.01	52.01	65.02	21.92	60.56

Table 18: Comparison of evaluation statistics for the best statistical and ML models developed.

Based on the results displayed in Table 18, the statistical model (`mlsa_L2`) obtains a significantly higher classification accuracy, specificity, F1 score, and ROC AUC when compared to the ML model (`NN_01`). The ML model, however, was able to obtain a far higher recall value, indicating that the probability of obtaining a Type-II error is lower in the ML model than the statistical model.

In an ideal budgeting scenario, one would want to cater for exactly the expected future attrition rates. This would enable capacity plans to effectively cater for required staffing levels. In reality, creating a prediction model capable of minimizing all error is improbable, and for this reason, it is important to determine which type of error has the biggest negative effect on budgeting, and consequently, on staffing. Budgeting for too few actual attrition events, would result in a smaller budget for HR capacity plans. This could, in turn, result in too few staff members being allocated to facilities. Understaffed facilities often result in lower quality of care being provided and increased overtime of healthcare workers, which, according to Castro Lopes et al. (2017), has the potential to further increase turnover rates (Section 1.1). Currently, the compensation for overtime in the EC DOH is 1.5 times the workers normal hourly rate. Consequently, too much overtime can also become increasingly costly, and put additional pressure on the limited budget.

On the other hand, budgeting for more than the expected number of actual attrition events could result in an unnecessary inflation of the budget and inflation of staffing levels, potentially resulting in redundancies and higher operating costs (Castro Lopes et al., 2017). There is, therefore, a trade-off between incurring slightly higher costs and catering for more attrition events than what is actually experienced, or incurring lower costs with the possibility for decline in the quality of healthcare provided and the risk of inflation to these costs due to unknown overtime hours. For the purpose of this investigation, catering for more events than actually experienced is considered most appropriate. Consequently, a higher specificity (lower Type-I error) is preferred. Based on this, the `mlsa_L2` model is considered the best model for attrition prediction in the context of this investigation. It is important to note, however, that this model obtains a very low recall value (9.57%), suggesting that a significant number of non-events will be predicted as events. This may significantly increase the ratio between actual and catered for attrition events, which may over-inflate the estimated staffing levels required and, consequently, inflate the budget. This will need to be taken into consideration when using the model predictions for budgeting.

The `mlsa_L2` model was subsequently used to predict the 2021 attrition events and attrition rates per cadre, facility, and district computed using Formula 1. These predictions are discussed in Section 5.6.

5.6 2021 Attrition predictions

One of the key objectives of this investigation, as described in Section 1.3, was to utilise the best proposed model to predict future attrition rates and use these predictions to identify future areas of concern with regards to healthcare workforce attrition in the EC province. Since the data modelled comprises of a four-level hierarchical structure, the most problematic predicted areas are determined by first identifying the facilities with the highest predicted 2021 attrition rates (*i.e.* highest level in the hierarchy) and, thereafter, analysing these specific facilities' cadre level attrition rates. The predicted results for the full population are provided to CHAI as supplementary documentation.

After analysing the predicted results, four facilities are expected to experience alarming levels of attrition in 2021. These facilities are documented in Table 19.

Facility Name	Predicted Average Annual Attrition Rate (%)
St Barnabas Gateway	106.66
St Lucy's Gateway	90.91
Kwazamukucinga Clinic	85.71
N Knight Gateway	69.23

Table 19: Facilities of high concern with respect to the predicted attrition rates for the 2021 reporting year.

According to Section 4.2.1, `St Barnabas Gateway`, `St Lucy's Gateway`, and `N Knight Gateway` have historically ranked in the top four most concerning facilities with respect to overall average attrition rates computed for the study period. According to Table 19, the average annual attrition rates for these three facilities are predicted to be higher than their historic averages (Table 27, documented in Section 7.3 of the Appendix).

Moreover, it was then important to assess the cadre level attrition rates for these concerning facilities to identify further areas of concern.

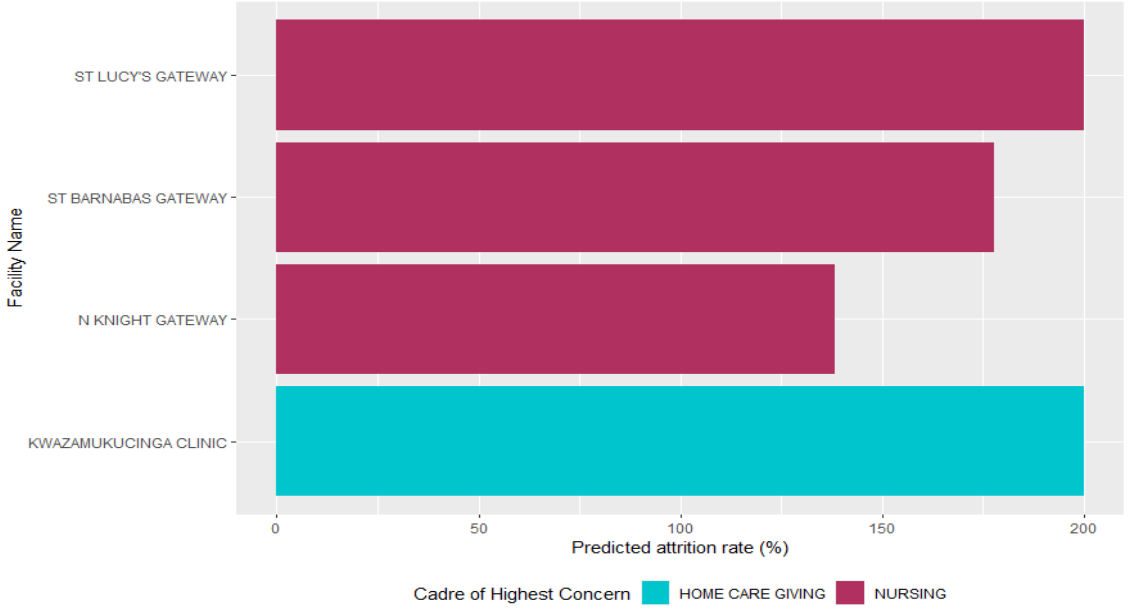

Figure 14: Graphical representation of cadres of highest concern, for the facilities of highest concern.

According to Figure 14, `St Barnabas Gateway`, `St Lucy's Gateway`, and `N Knight`

Gateway are all expected to have the highest attrition within their nursing cadre. This is particularly concerning for St Lucy's Gateway, whereby 200% of the facilities nursing staff are predicted to leave the facility at the end of 2021. Similarly, Kwazamukucinga Clinic is expected to have a 200% staff turnover within their home care giving cadre. According to the historical EDA (Section 4.2.1), over the course of the study period, nursing did experience an overall average attrition rate of 16.10%. It is, therefore, not alarming that three of the four facilities predicted to have high attrition in 2021, are predicted to experience the highest cadre level attrition for the nursing level (Table 26).

The overall annual attrition rate for the EC province, computed from the 2021 predictions, is 4.04%. This is lower than the study periods' historical average of 18.79% (Section 4.2.1), and is also lower than the currently budgeted and planned for 5% annual attrition rate. Despite this overall lower predicted attrition rate, and according to Table 19, some facilities within the province are predicted to experience almost 26 times higher attrition rates, in the 2021 year, than this predicted 4.04% average. This high within and between group level variation is visually confirmed by means of the historical EDA performed on the person_period_persal_ec_2010_2021 dataset (Section 4.2.1), as well as statistically confirmed by means of the ICC value obtained for the model on which these predictions are made (Section 5.3.1).

Consequently, attrition rates need to be determined at a facility level, and reassessed yearly, in order to cater for this high annual variation between and within the hierarchical grouping levels (defined in Section 5.3.1).

5.7 Chapter summary

In this chapter, focus was placed on developing multilevel statistical and ML models capable of handling longitudinal, multilevel survival data with discrete event outcomes, as well as capable of accurately predicting future attrition rates. The chapter discusses the development and results of four modelling techniques applied to the person_period_persal_ec_2010_2021 dataset, namely the application of multilevel discrete-time event models (Section 5.3.1), as well as three novel ML models (Sections 5.4.1-5.4.3). The chapter further discusses the most appropriate model for use, based on a thorough model comparison, explored in Section 5.5. Lastly, the chapter discusses the 2021 attrition predictions based on the best performing model developed in this investigation.

The best performing model was identified to be the multilevel discrete-time event statistical model with four levels, whereby, "repeated" measures are nested within healthcare workers, who are nested in cadres, which are nested in facilities. This model regresses the outcome variable, event, on several fixed effect vari-

ables (namely, `exit`, `years_in_system_before_study_start`, `gender`, `age`, `race`, `notch_value`, `allowance_total`) whilst simultaneously catering for the random effects introduced by the four-level nested structure of the data. According to the summary statistics for this model (Table 8, Section 5.3.1), the `cadre` grouping level explains the greatest proportion of the total variance explained by the grouping/nesting structure in the population. This confirms the findings of the historical EDA (Section 4.2.1), where average annual attrition rates for cadres were identified to, historically, vary significantly across the different facilities. Moreover, the regression coefficients obtained for all of these predictor variables were identified to be statistically significant. This model obtained the best performance when assessed on classification accuracy, recall, specificity, F1 Score, and ROC AUC when analysed holistically. Additionally, it is the only model developed in this investigation that is able to obtain an acceptable degree of Type-I error. The ML models, in the context of this investigation, were identified to be unable to outperform the multilevel discrete-time event model, `mlsa_L2`.

6 Discussion and conclusion

South Africa's public health system budget currently accounts for an annual 5% attrition rate for health facilities in general. This rate does not consider fluctuations in attrition rates between cadres, across facilities, or across districts.

After conducting an historical EDA on the EC DOH cleaned, preprocessed and transformed HR data (Sections 4.2.1), the overall annual provincial level attrition, experienced in the EC province between 2010 to 2020, was 18.79%. Consequently, assuming a 5% annual provincial level attrition rate in budgeting formulations, for the EC health system, is insufficient . Whilst no cyclical, increasing, or decreasing trend in attrition was identified during this analysis, high variation between annual EC provincial level attrition rates was evident, with some years experiencing annual provincial level attrition rates of over 30%. This high variation, however, was not limited to the provincial level only. Upon further investigation, high variation exists within certain grouping levels (specifically facilities and cadres), as well as between certain grouping levels (specifically cadres nested in facilities) for the EC province. This between-group variation indicates that the data is hierarchical in nature (Sections 4.2.1).

The implications of these findings are two-fold. Firstly, the EC DOH have been historically and holistically under budgeting for attrition which, according to literature (Section 2.1), results in poorer staffing capacity levels and may lead to multiple negative knock-on effects (*i.e.* increased overtime, increased workload, and a decline in quality of health care provided). Secondly, the presence of high within and between-group variation in annual attrition rates suggests that fluctuations in annual attrition must be catered for in the annual budgeting process. Neglecting the effects of this variation in the computation of annual attrition could result in an under or over inflated budget. Over budgeting for attrition could result in an inflation of staffing levels which may, in turn, result in redundancies and higher operating costs.

The historical EDA further identified several cadres that consistently experienced high levels of attrition namely, the `Medical Services`, `Nursing`, and `Primary Health Care` cadres (Section 4.2.1). The job titles that fall within these cadres (*i.e.* Medical Specialists, Clinic Specialists, Medical Officers, Medical Advisors, and Nurses) are considered critical to the functioning of any health facility as they are responsible for providing medical care to patients. The historically high attrition levels obtained in these cadres are, therefore, alarming as they suggests that the EC province can expect to consistently see the same or a degrading level of patient care in the years to come. This risk could be mitigated by implementing interventions that aim to improve the attrition levels of these cadres. Since a mul-

tilevel structure is present in the data, whereby cadres are nested in facilities, not all facilities will exhibit the same problematic cadres. Consequently, intervention planning should take place at a facility level. The facilities that were identified to be, historically, most problematic are `Nompumelelo Clinic`, `St Lucy's Gateway`, `St Barnabas Gateway`, and `N Knight Gateway`. Consequently, it is recommended for intervention planning to be prioritized at these facilities.

The findings from the historical EDA (Section 4.2.1), and the potential risks associated with under-budgeting for attrition (Section 18), suggest that there is a financial incentive for the EC DOH to develop models capable of accurately predicting future attrition rates within and between multiple levels within the EC province.

The application of both statistical and machine learning modelling techniques were thus explored in this investigation, however, only one statistical modelling method (MDTE models) and three ML modelling methods (MLP NNs, GLMM trees, and TBME models) were attempted (Chapter 5). This was due to their potential ability to handle and, effectively model, the complex multilevel and longitudinal HR data available in this study.

Unfortunately, all multilevel tree-based ML models (GLMM trees and TMBE models) attempted in this investigation, failed to converge or resulted in excessive computational time forcing an abort (Sections 5.4.2-5.4.3). This confirms the limitations of these models, posed in literature, regarding their potential inability to handle highly complex multilevel data where many factor predictor variables (with many levels) are present (Sections 2.2.2.2-2.2.2.3). The MLP NN modelling technique was the only ML method that successfully converged. Despite this, the best MLP NN developed was only able to obtain a classification accuracy and specificity of 58.01% and 65.02%, respectively, when tested on the test dataset (Section 5.4.1). Based on these findings, and within the limitations of the study scope, it is accepted that these three modelling methods are unable to outperform traditional multilevel statistical methods at this time.

Conversely, several MDTE models were successfully developed in this investigation (Section 5.3.1). These models identified that the data contains a four-level nested structure, whereby, "repeated" measures are nested within healthcare workers, who are nested in cadres, which are nested in facilities. The district level was, therefore, identified to have a negligible effect on the between-group variation in the outcome variable. This further supports the idea that attrition intervention strategies should be targeted at a facility level, and not at a district or provincial level.

All MDTE model formulations that included level 2 and level 4 predictor variables as fixed effects, resulted in the models' inability to converge (Section 5.3.1). Despite this, the `mlsa_L2` multilevel discrete-time event statistical model (with four levels),

was the overall best performing model when tested on unseen data. The model obtained a classification accuracy of 89.04%, a specificity of 97.81% (very low Type-I error) and a recall of 9.57% (high Type-II error). According to Section 18, a low Type-I error is preferred over a low Type-II error (Section 5.3.1). For this reason, this model is considered feasible for use in attrition prediction for the EC DOH.

The `mlsa_L2` model was therefore used to predict future attrition rates as well as determine factors influencing attrition. According to the model output, discussed in Section 5.3.1, males are 8% more likely to experience an attrition event relative to females. When looking at the race-groups, white individuals appear to be at the highest risk of experiencing an attrition event relative to all of the other race groups, followed by Asian, African, and least at risk being Coloured individuals. Additionally, older individual are at higher risk of experiencing an attrition event to younger individuals. Moreover, the higher the annual salary of an individual in the EC health sector, the lower their risk of experiencing an attrition event. The number of years an individual has spent working at a specific facility within the EC, has the largest relative impact on the risk that that individual will experience an attrition event. Additionally, the model summary further indicates that the cadre-facility nested grouping is able to explain the greatest proportion of the variation explained by the hierarchical structure of the data (Section 5.3.1). This insight can be used to further support the process of intervention planning.

6.1 Limitations, recommendations, and future work

While this study has yielded valuable insights into historical attrition as well as the prediction of future attrition in the EC public health system, there are several limitations that must be acknowledged. In this section, these limitations are discussed and opportunities for future research are explored.

Limitation I: *Definition of attrition*

According to the review of literature (Section 2.1), there is a general lack of consistency regarding the definition of healthcare workforce attrition in applied research. Consequently, the definition of attrition used in an investigation is influenced by the outcomes that are of importance or relevance to the decision makers. For this investigation, an attrition event was thus defined as either an exit from the EC public health system, or a move from one facility to another facility in the EC province at the end of a specific reporting year.

This definition heavily impacts the way on which an attrition event is modelled in the dataset which, in turn, influences how attrition is computed and interpreted. For example, the percentage of attrition attributed to retirement may be significantly

downplayed in this investigation due to the incorporation of facility-to-facility attrition events in the definition of attrition.

Consequently, it would be important to perform this same investigation using differing definitions of attrition in order to compare the impact that the choice of definition may have on the key findings of the investigation. This should be conducted as future work.

Limitation II: *Data integrity challenges*

In order to effectively explore and utilise the HR data provided for use in this investigation, a significant amount of cleaning, preprocessing, and transformation was required. Despite best efforts to automate the process of data preparation (using regex manipulation and computer-assisted translation techniques), many variables in the dataset still required manual manipulation. Consequently, the process of data cleaning and preprocessing undertaken in this study may not be easily extended to new datasets, because the manual manipulation required might differ slightly.

In order to utilise the models developed in this investigation for future attrition prediction, future EC HR datasets will need to be acquired and transformed into the same format that the models were trained and validated on. Seeing as this data preparation process is not entirely automated, data scientists or data engineers would need to be provisioned every year in order to perform the manual data manipulation required. This is a time consuming and costly process.

Suggested future work, therefore, includes the following:

- **Proposal I**: *Automate the process of data cleaning, preprocessing, and transformation*
  Due to time limitations, only a few techniques for cleaning and preprocessing messy text data were explored. Other techniques for cleaning messy text data, such as natural language processing (NLP) and machine learning models for text mining, do exists, however, they may not be able to adequately handle the messy text data exhibited in the EC HR datasets. For this reason, it is proposed that this problem be phrased as a new research question, whose results can be used to support the findings and models developed in this investigation.
- **Proposal II**: *Improve the quality of the data at the source*
  A large amount of the data integrity issues identified in the dataset during PEDA (Section 4.1.1), came from columns that stored free-text data manually inputted by users. In some cases, the observations for these columns were so poorly entered (if entered at all) that the variables were considered too poor for use, limiting potential insights that could be gathered from including such column data (*i.e.* `appointment nature`) in the investigation. Some of these

human errors (*i.e.* spelling mistakes, duplicated data, incorrectly assigned job titles) could be prevented or reduced by:

1. Restricting some of these fields to be 'drop-down' list options only (*i.e.* facility name).
2. Ensuring all `persal numbers` only contain 8 numeric characters.
3. Ensuring all punctuation is removed (so no duplicates of the same names exist).
4. Providing data capturers the opportunity to go on data literacy courses to improve their understanding of systems and the impact they have on data quality.
5. Develop an application that manages the storage and use of the EC health system HR data.

Limitation III: *Time, cost, and complexity trade-off in the project scope*

Due to the limited time available to achieve the scope and objectives of this study (Section 1.3), and the estimated time required to complete the training or hyperparameter tuning processes for the ML models, further exploration into the potential of these ML models for complex, multilevel discrete-time event prediction problems was not feasible. Future work that could potentially be implemented to improve the performance of these models include:

- **Proposal I**: *Resampling techniques for dataset reduction*
  According to Section 4.1.3, the cleaned and preprocessed dataset used in this investigation has imbalanced classes. However, since the class ratio for the outcome variable `event` in each reporting year never subceeded 0.1, no resampling techniques were applied. The class ratio, however, was fluctuating around a value of 0.15. For this reason, applying resampling techniques may help reduce the size of the training and validation datasets enough to enable the training/tuning process to conclude faster.
- **Proposal II**: *Reconsider manual grouping of cadres*
  Convergence issues in ML models typically arise when datasets are highly complex, containing multiple factor variables with multiple levels (Section 3.5). One of the data preprocessing steps required during PEDA (Section 4.1.1), was to correctly assign job titles and cadres to healthcare workers. The process of dealing with inaccurate cadres and associated job titles was quite complex due to the interchangeable use of three variables, `Occupational Group`, `Occupational Classification`, and `Job Title` from the original dataset, for specifying a health workers' cadre and or job title. All three variables

were regex manipulated, but the interwoven nature of these variables made it difficult to identify correct and separate groupings of job titles and cadres using this method alone. Consequently, based on inspection of the regex manipulated columns and careful discussion with CHAIs SMEs, the job titles were manually grouped into overarching job titles which were, then, manually grouped into cadres of relevance. This resulted in 38 unique cadres. In future work, it may be worthwhile revisiting this grouping to identify whether or not it could be condensed into a smaller group of unique cadres. This may reduce the complexity of the dataset further, to potentially reduce the time it might take to train and tune the ML models.

- **Proposal II**: *Provision a virtual machine on which training and tuning can be performed*
  A potential reason for the large computational time required for these ML models may be simply due to limitations of the computer used to run the models. Provisioning a virtual machine (VM) on which the `R` code could run, may help improve training/tuning time through better resource allocation, capability for scaling the VM up or down, and the utilization of parallel processing (Smith and Nair, 2005).

  It is important to note, however, that VMs are rather expensive to provision and, in the context of this investigation, may be an unnecessary expense considering that the statistical model runs rather quickly and obtains a satisfactory level of performance. Any future work in this domain would, therefore, have a different research aim, whereby the intention of the research is to advance the academic field of predictive ML models for data that is complex, multilevel and longitudinal.

References

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925.

Allison, P. D. (2010). *Survival analysis using SAS: a practical guide*. Sas Institute.

Austin, P. C., Manca, A., Zwarenstein, M., Juurlink, D. N., and Stanbrook, M. B. (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of clinical epidemiology*, 63(2):142–153.

Barber, J. S., Murphy, S. A., Axinn, W. G., and Maples, J. (2000). 6. discrete-time multilevel hazard analysis. *Sociological methodology*, 30(1):201–235.

Bates, D. et al. (2005). Fitting linear mixed models in r. *R news*, 5(1):27–30.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., and Lindenberger, U. (2013). Structural equation model trees. *Psychological methods*, 18(1):71.

Candel, A., Parmar, V., LeDell, E., and Arora, A. (2016). Deep learning with h2o. *H2O. ai Inc*, pages 1–21.

Castro-Leal, F., Dayton, J., and Demery, L. (2000). Public spending on health care in africa: do the poor benefit? *Bulletin of the World health Organization*, 78(1):66–74.

Castro Lopes, S., Guerra-Arias, M., Buchan, J., Pozo-Martin, F., and Nove, A. (2017). A rapid review of the rate of attrition from the health workforce. *Human resources for health*, 15(1):1–9.

Cateni, S., Colla, V., and Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135:32–41.

Cayrol, M., Farreny, H., and Prade, H. (1982). Fuzzy pattern matching. *Kybernetes*, 11(2):103–116.

Chambers, J. M. (2008). *Software for data analysis: programming with R*, volume 2. Springer.

Christensen, E. (1987). Multivariate survival analysis using cox's regression model. *Hepatology*, 7(6):1346–1358.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. R. and Oakes, D. (2018). *Analysis of survival data*. Chapman and Hall/CRC.

Crowther, M. J., Look, M. P., and Riley, R. D. (2014). Multilevel mixed effects parametric survival models using adaptive gauss–hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics in medicine*, 33(22):3844–3858.

Eysenbach, G. et al. (2005). The law of attrition. *Journal of medical Internet research*, 7(1):e402.

Finch, H., Lapsley, D., and Baker-Boudissa, M. (2009). A survival analysis of student mobility and retention in indiana charter schools. *Education Policy Analysis Archives*, 17:18–18.

Fokkema, M., Edbrooke-Childs, J., and Wolpert, M. (2021). Generalized linear mixed-model (glmm) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy research*, 31(3):329–341.

Fokkema, M. and Zeileis, A. (2022). Fitting generalized linear mixed-effects model trees.

Gensheimer, M. F. and Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257.

Goel, M. K., Khanna, P., and Kishore, J. (2010). Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274.

Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons.

Guillory, C. W. (2008). *A multilevel discrete-time hazard model of retention data in higher education*. Louisiana State University and Agricultural & Mechanical College.

Hagan, M. T., Demuth, H. B., and Beale, M. (1997). *Neural network design*. PWS Publishing Co.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Hox, J. J., Moerbeek, M., and Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Jain, R. and Nayyar, A. (2018). Predicting employee attrition using xgboost machine learning approach. In *2018 international conference on system modeling & advancement in research trends (smart)*, pages 113–120. IEEE.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer.

Kleinbaum, D. G., Klein, M., et al. (2012). *Survival analysis: a self-learning text*, volume 3. Springer.

Kohonen, T. and Kohonen, T. (1995). Learning vector quantization. *Self-organizing maps*, pages 175–189.

Kvamme, H. and Borgan, Ø. (2019). Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*.

Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240.

Lawless, J. (2014). Parametric models in survival analysis. *Wiley StatsRef: Statistics Reference Online*.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*. John Wiley & Sons.

Lin, D. Y. and Wei, L.-J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American statistical Association*, 84(408):1074–1078.

Liu, W., Chen, Z., and Hu, Y. (2022). Xgboost algorithm-based prediction of safety assessment for pipelines. *International Journal of Pressure Vessels and Piping*, 197:104655.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. (2013). On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611. PMLR.

Peto, R., Pike, M., Armitage, P., Breslow, N. E., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., and Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British journal of cancer*, 35(1):1–39.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. STATA press.

Rabe-Hesketh, S. and Skrondal, A. (2012). Understanding variability in multilevel models for categorical responses. In *Proceedings of the AERA Annual Meeting, Vancouver, BC, Canada*, volume 12.

Rodríguez, G., de Leeuw, J., and Meijer, E. (2008). Handbook of multilevel analysis.

Rondeau, V., Marzroui, Y., and Gonzalez, J. R. (2012). frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47:1–28.

Sigrist, F. (2020). Gaussian process boosting.

Singer, J. D., Willett, J. B., Willett, J. B., et al. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Smith, J. E. and Nair, R. (2005). The architecture of virtual machines. *Computer*, 38(5):32–38.

Snijders, T. A. and Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage.

Sokhansanj, B. A. and Rosen, G. L. (2022). Predicting covid-19 disease severity from sars-cov-2 spike protein sequence by mixed effects machine learning. *Computers in Biology and Medicine*, 149:105969.

Somers, M. J. (1999). Application of two neural network paradigms to the study of voluntary employee turnover. *Journal of Applied Psychology*, 84(2):177.

Therneau, T. M., Grambsch, P. M., Therneau, T. M., and Grambsch, P. M. (2000). *The cox model*. Springer.

Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6):419–422.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879.

Wienke, A. (2010). *Frailty models in survival analysis*. Chapman and Hall/CRC.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

7 Appendix

7.1 Partial data listing in person-period format

Persal Number	Facility Name	Age	Reporting Year	Years In System Before Study Start	Enter	Exit	Duration	Event
00008912	Fort Beaufort Hospital	30	2010	0	0	1	1	0
00008912	Fort Beaufort Hospital	31	2011	0	1	2	2	0
00008912	Fort Beaufort Hospital	32	2012	0	2	3	3	0
00008912	Fort Beaufort Hospital	33	2013	0	3	4	4	0
00008912	Fort Beaufort Hospital	34	2014	0	4	5	5	0
00008912	Fort Beaufort Hospital	35	2015	0	5	6	6	0
00008912	Fort Beaufort Hospital	36	2016	0	6	7	7	0
00008912	**Fort Beaufort Hospital**	**37**	**2017**	**0**	**7**	**8**	**8**	**1**
00007891	Dimbaza CFC	48	2010	5	0	1	6	0
00007891	Dimbaza CFC	49	2011	5	1	2	7	0
00007891	Dimbaza CFC	50	2012	5	2	3	8	0
00007891	**Dimbaza CFC**	**51**	**2013**	**5**	**3**	**4**	**9**	**1**
00006789	Frere Hospital	35	2010	0	0	1	1	0
00006789	Frere Hospital	36	2011	0	1	2	2	0
00006789	Frere Hospital	37	2012	0	2	3	3	0
00006789	Frere Hospital	38	2013	0	3	4	4	0
00006789	Frere Hospital	39	2014	0	4	5	5	0
00006789	Frere Hospital	40	2015	0	5	6	6	0
00006789	Frere Hospital	41	2016	0	6	7	7	0
00006789	Frere Hospital	42	2017	0	7	8	8	0
00006789	Frere Hospital	43	2018	0	8	9	9	0
00006789	**Frere Hospital**	**44**	**2019**	**0**	**9**	**10**	**10**	**1**
00005678	St Elizabeth's Hospital	50	2010	0	0	1	1	0
00005678	St Elizabeth's Hospital	51	2011	0	1	2	2	0
00005678	St Elizabeth's Hospital	52	2012	0	2	3	3	0
00005678	St Elizabeth's Hospital	53	2013	0	3	4	4	0
00005678	St Elizabeth's Hospital	54	2014	0	4	5	5	0
00005678	**St Elizabeth's Hospital**	**55**	**2015**	**0**	**5**	**6**	**6**	**1**
00004567	Frontier Hospital	26	2011	0	0	1	1	0
00004567	Frontier Hospital	27	2012	0	1	2	2	0
00004567	Frontier Hospital	28	2013	0	2	3	3	0
00004567	Frontier Hospital	29	2014	0	3	4	4	0
00004567	Frontier Hospital	30	2015	0	4	5	5	0
00004567	Frontier Hospital	31	2016	0	5	6	6	0
00004567	Frontier Hospital	32	2017	0	6	7	7	0
00004567	Frontier Hospital	33	2018	0	7	8	8	0
00004567	Frontier Hospital	34	2019	0	8	9	9	0
00004567	**Frontier Hospital**	**35**	**2020**	**0**	**9**	**10**	**10**	**0**
00003456	St Patrick's Hospital	25	2010	0	0	1	1	0
00003456	St Patrick's Hospital	26	2011	0	1	2	2	0
00003456	St Patrick's Hospital	27	2012	0	2	3	3	0
00003456	St Patrick's Hospital	28	2013	0	3	4	4	0
00003456	St Patrick's Hospital	29	2014	0	4	5	5	0
00003456	St Patrick's Hospital	30	2015	0	5	6	6	0
00003456	St Patrick's Hospital	31	2016	0	6	7	7	0
00003456	St Patrick's Hospital	32	2017	0	7	8	8	0
00003456	St Patrick's Hospital	33	2018	0	8	9	9	0
00003456	St Patrick's Hospital	34	2019	0	9	10	10	0
00003456	**St Patrick's Hospital**	**35**	**2020**	**0**	**10**	**11**	**11**	**0**
00002345	St Elizabeth's Hospital	60	2010	0	0	1	1	0
00002345	St Elizabeth's Hospital	61	2011	0	1	2	2	0
00002345	**St Elizabeth's Hospital**	**62**	**2012**	**0**	**2**	**3**	**3**	**1**

Table 20: Partial listing, in person-period format, for 7 healthcare workers for the study period.

7.2 Preliminary exploratory data analysis tabulated results

Reporting Year	Number of Non-Events	Number of Events	Class Ration
2010	40718	5689	0.14
2011	43781	6201	0.14
2012	42690	6926	0.16
2013	41796	6487	0.16
2014	41742	7937	0.19
2015	36529	14024	0.38
2016	35927	13833	0.39
2017	32699	9307	0.28
2018	37403	5207	014
2019	38502	4894	0.13
2020	39723	4388	0.11

Table 21: Class ratios for the outcome variable for all reporting years.

7.3 Historical exploratory data analysis tabulated results

District	Minimum Attrition Rate (%)	Maximum Attrition Rate (%)	Standard Deviation	Coefficient of Variation	Average Attrition Rate (%)
A Nzo DM	7.82	26.92	5.85	0.41	14.29
Amathole DM	8.24	34.60	9.09	0.52	17.59
Buffalo City MM	9.70	60.08	14.71	0.76	19.26
C Hani DM	8.73	35.78	7.49	0.49	15.40
Joe Gqabi DM	7.24	31.09	6.85	0.51	13.45
N Mandela Bay MM	9.21	37.30	10.25	0.57	17.88
EC District Division	18.77	71.45	14.13	0.43	33.56
O.R Tambo DM	9.29	31.32	7.43	0.44	16.97
Sarah Baartman DM	9.68	27.68	5.25	0.32	16.33

Table 22: A partial 5-number summary of the historical annual district level attrition rates for the years of study.

Geographical Description	Minimum Attrition Rate (%)	Maximum Attrition Rate (%)	Standard Deviation	Coefficient of Variation	Average Attrition Rate (%)
Rural	9.13	20.92	3.31	0.25	13.33
Urban	10.13	35.32	9.07	0.49	18.46

Table 23: A partial 5-number summary of the geographical (rural/urban) level attrition rates for the years of study.

Year	Annual Rural Level Attrition Rate (%)	Annual Urban Level Attrition Rate (%)
2010	12.50	12.99
2011	12.00	14.53
2012	13.71	16.16
2013	14.62	14.726
2014	20.92	17.31
2015	16.51	35.32
2016	14.44	35.26
2017	10.79	23.48
2018	10.27	12.16
2019	11.71	10.95
2020	9.13	10.13

Table 24: Annual attrition rates experienced for the rural and urban geographic location grouping levels for the years of study.

Facility Type	Minimum Attrition Rate (%)	Maximum Attrition Rate (%)	Standard Deviation	Coefficient of Variation	Average Attrition Rate (%)
Ambulance Station/EMS	3.10	36.38	10.98	1.11	9.93
Clinic	8.43	29.20	6.76	0.42	16.27
Community Health Centre	9.79	16.52	1.93	0.15	12.58
District Hospital	9.96	13.90	1.32	0.11	11.56
District Office	6.55	69.95	21.66	0.78	27.76
Mortuary	2.86	12.00	2.91	0.40	7.27
Provincial Tertiary Hospital	3.75	58.62	16.24	0.87	18.74
Regional Hospital	6.51	20.42	4.85	0.38	12.62
Specialised Hospital	8.21	12.94	1.72	0.16	10.40

Table 25: A partial 5-number summary of the facility type level attrition rates for the years of study.

Cadre	Minimum Attrition Rate (%)	Maximum Attrition Rate (%)	Standard Deviation	Coefficient of Variation	Average Attrition Rate (%)
Community Service	53.25	85.66	11.14	0.16	68.13
Mental Health Services	1.81	207.60	60.68	1.69	35.82
Medical Services	22.48	52.08	10.62	0.29	35.82
Occupational Therapy	19.05	52.91	9.51	0.33	28.48
Optometry	12.70	63.49	18.20	0.54	33.48
Environmental Health	0.00	90.69	34.20	0.91	37.58
Nursing	8.46	28.78	6.22	0.38	16.10
Primary Health Care	8.61	21.22	3.83	0.27	14.01

Table 26: A partial 5-number summary of the cadre level attrition rates for the years of study.

Facility	Minimum Attrition Rate (%)	Maximum Attrition Rate (%)	Standard Deviation	Coefficient of Variation	Average Attrition Rate (%)
Nompumelelo Clinic	0.00	379.15	128.43	1.35	94.79
St Lucy's Gateway	2.16	252.75	99.23	1.22	81.12
St Barnabas Gateway	3.50	174.93	64.45	0.9	66.15
N Knight Gateway	2.07	148.84	54.13	1.08	49.80
Good Hope Clinic	0.00	48.98	22.88	1.71	13.36
Mqambeli Clinic	0.00	17.39	8.77	1.38	6.32
Nqabara Clinic	0.00	11.88	4.84	1.14	4.23
Ulundi Clinic	0.00	8.99	2.71	3.30	0.82

Table 27: A partial 5-number summary of the facility level attrition rates for the years of study.

Facility	Minimum Attrition Rate (%)	Maximum Attrition Rate (%)	Standard Deviation	Coefficient of Variation	Average Attrition Rate (%)
Glen Grey Hospital	13.26	92.82	23.24	0.49	47.01
Madzikane Hospital	0.00	125.71	37.56	0.84	44.71
Holy Cross Hospital	13.64	95.45	22.89	0.51	44.63
All Saints Hospital	0.00	68.57	22.53	0.59	38.29
Bisho Hospital	8.96	35.82	9.04	0.37	24.42
Motherwell CHC	0.00	65.22	19.19	0.81	23.72

Table 28: A partial 5-number summary of the `Medical Services` cadres, historical annual facility level attrition rates for the years of study.

7.4 Model formulations in R

Example model formulations for a subset of the models attempted in this investigation are defined and described below.

The Baseline Hazard MDTE model with 2 levels ("repeated" measures are nested in healthcare workers) is formulated in R as follows:

```
mlsa_B0 <- glmer(event ~ exit +
                         years_in_system_before_study_start +
                         (1|persal_number),
                 family=binomial(link='cloglog'),
                 data=train_data,
                 control=glmerControl(optimizer='bobyqa',
                                      optCtrl=list(maxfun=2e5)))
```

The MDTE model, regressed on all time-indicator variables and level 1 predictor variables, and catering for a 4 level nested structure (whereby "repeated" measures are nested in healthcare workers, who are nested in cadres, which are nested in facilities) is formulated in R as follows:

```
mlsa_L2 <- glmer(event ~ exit +
                         years_in_system_before_study_start +
                         age + race + gender + notch_value +
                         (1|facility_name/unit_field/persal_number),
                 family=binomial(link='cloglog'),
                 data=train_data,
                 control=glmerControl(optimizer='bobyqa',
                                      optCtrl=list(maxfun=2e5)))
```

The baseline MLP NN, regressed on all time-indicator variables and level 1 predictor variables, and catering for a 5 level nested structure (whereby "repeated" measures are nested in healthcare workers, who are nested in cadres, which are nested in facilities, which are nested in districts) is formulated in R as follows:

```
NN_B1 <- h2o.deeplearning(x=c(1:11, 13),
                          y=12,
                          training_frame=train_h2o,
                          activation='Tanh',
                          balance_classes=TRUE,
                          hidden=c(1),
                          distribution='multinomial',
                          seed=1,
                          reproducible=TRUE,
                          loss=c('CrossEntropy'),
                          adaptive_rate=TRUE,
                          epochs=10,
                          variable_importances=TRUE,
                          export_weights_and_biases=TRUE)
```

The baseline GLMM tree, catering for a 4 level nested structure (whereby "repeated" measures are nested in healthcare workers, who are nested in cadres, which are nested in facilities) is formulated in R as follows:

```
glmm_01 <- glmertree(event ~ 1 |
                             facility_name/unit_field/persal_number |
                             exit + years_in_system_before_study_start,
                     family=binomial(link='cloglog'),
                     data=train_data)
```

The baseline TBME model, catering for a 4 level nested structure (whereby "repeated" measures are nested in healthcare workers, who are nested in cadres, which

are nested in facilities) is formulated in R as follows:

```
tbme_01 <- gpboost(data=x,
                   label=y$event,
                   gp_model=GPModel(group_data=group_data,
                                    likelihood='bernoulli_logit'),
                   verbose=-1,
                   params=list(objective='binary',
                               learning_rate=0.01,
                               max_depth=10,
                               min_data_in_leaf=5,
                               nrounds=100))
```

where

```
x <- as.matrix(train_data[ ,c('years_in_system_before_study_start',
                              'exit')])

group_data <- train_nn[,c('facility_name','unit_field','cadre')]
```