Journal of Data Analysis and Information Processing, 2023, 11, 11-36
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Modelling Key Population Attrition in the HIV and AIDS Programme in Kenya Using Random Survival Forests with Synthetic Minority Oversampling Technique-Nominal Continuous

Evan Kahacho[1*], Charity Wamwea[1], Bonface Malenje[1], Gordon Aomo[2]

[1]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[2]Monitoring and Evaluation, Kenya Red Cross Society-Global Fund Unit, Nairobi, Kenya
Email: *kahacho.evan@gmail.com

How to cite this paper: Kahacho, E., Wamwea, C., Malenje, B. and Aomo, G. (2023) Modelling Key Population Attrition in the HIV and AIDS Programme in Kenya Using Random Survival Forests with Synthetic Minority Oversampling Technique-Nominal Continuous. *Journal of Data Analysis and Information Processing*, **11**, 11-36.
https://doi.org/10.4236/jdaip.2023.111002

Received: October 10, 2022
Accepted: January 28, 2023
Published: January 31, 2023

Abstract

HIV and AIDS has continued to be a major public health concern, and hence one of the epidemics that the world resolved to end by 2030 as highlighted in sustainable development goals (SDGs). A colossal amount of effort has been taken to reduce new HIV infections, but there are still a significant number of new infections reported. HIV prevalence is more skewed towards the key population who include female sex workers (FSW), men who have sex with men (MSM), and people who inject drugs (PWID). The study design was retrospective and focused on key population enrolled in a comprehensive HIV and AIDS programme by the Kenya Red Cross Society from July 2019 to June 2021. Individuals who were either lost to follow up, defaulted (dropped out, transferred out, or relocated) or died were classified as attrition; while those who were active and alive by the end of the study were classified as retention. The study used density analysis to determine the spatial differences of key population attrition in the 19 targeted counties, and used Kilifi county as an example to map attrition cases in smaller administrative areas (sub-county level). The study used synthetic minority oversampling technique-nominal continuous (SMOTE-NC) to balance the datasets since the cases of attrition were much less than retention. The random survival forests model was then fitted to the balanced dataset. The model correctly identified attrition cases using the predicted ensemble mortality and their survival time using the estimated Kaplan-Meier survival function. The predictive performance of the model was strong and way better than random chance with concordance indices greater than 0.75.

Keywords

Random Survival Forests, Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC), Key Population, Female Sex Workers (FSW), Men Who Have Sex with Men (MSM), People Who Inject Drugs (PWID)

1. Introduction

HIV and AIDS has been a major public health concern since the first case was reported in the United States in 1981. Despite tremendous strides taken globally to reduce new HIV infections, the number of new cases is still remarkable with an estimate of 1.7 million people who acquired HIV globally in 2019 [1] with approximately 2.4% (41,408) cases from Kenya [2]. As the target date for delivery on the sustainable development goals (SDGs) looms, there is need for collective action to reimagine the relationship between data and interventions to end the HIV and AIDS epidemic by 2030.

It should be noted that HIV prevalence is not uniform among the entire population. The majority of new infections reported globally in 2019 were among key population and their sexual partners, contributing to 62% of total infections [1]. The situation in Kenya is a reflection of the global perspective where the epidemic is more skewed towards the key population with prevalence among female sex workers (FSW), men who have sex with men (MSM), and people who inject drugs (PWID) being 29.3%, 18.2% and 18.3% respectively, as compared to 4.9% prevalence rate among the general population [3] [4].

Key population refers to persons at an elevated risk of getting infected with HIV, partly due to prejudice and social marginalisation. They include female sex workers, people who inject drugs, men who have sex with men, and transgender people [5]. Kenya Red Cross Society was the main non-state beneficiary of the global fund HIV grant implemented from January 2018 to June 2021, with the goal of decreasing the number of new HIV infections by 75%, minimising HIV and AIDS related mortality by 50%, and reducing stigma and discrimination associated with HIV to less than 25%, while mainly focusing on FSWs, MSM and PWID in 19 targeted counties in Kenya. In spite of such a comprehensive programme, a considerable proportion of key population either defaulted (dropped, transferred out or relocated), were lost to follow up, or died. This study classifies these individuals as attrition, whereas those who remained active and alive at the end of the program were categorised as retention. Most studies have mainly focused on attrition from HIV antiretroviral care [6] [7] [8], with only a few studies focusing on key population [9] [10] [11], and they mainly used cox regression and Kaplan Meir survival curves in their analysis.

The study uses density analysis to examine the spatial differences of key population attrition in the 19 targeted counties, and using Kilifi county as an exam-

ple to map attrition in smaller administrative areas (sub-county level). Density analysis distributes known quantities of a phenomena over a given landscape depending on the amount observed at each area and the spatial connection between the measured quantities' locations. It may be used to gain important insight into the natural and social occurrences using features such as population by observing areas where more points are concentrated [12]. For instance, Goldenberg *et al.*, [13] mapped communities in Vancouver, Canada to identify areas where sex workers reported either being criminalised or experienced violence. Thereafter, the study uses SMOTE-NC to balance the datasets since the cases of attrition account for a very small proportion of the dataset and hence represents the minority class, while the retention cases comprise the majority class. If the majority to minority class ratio of the response variable of a dataset is outstandingly large, then such a dataset is said to be imbalanced [14]. High class imbalance has been proven to suppress the effectiveness and reliability of machine learning models.

Several techniques have been proposed to improve classification when dealing with imbalanced datasets. They are broadly categorised into three: data level or resampling techniques, algorithm level techniques and cost-sensitive techniques [15]. The most common data balancing techniques are undersampling the majority class randomly and also oversampling the class with minority cases under the data level techniques. However, undersampling could lead to the risk of loss of information and small disjuncts problem, while oversampling is prone to overfitting [16]. A hybrid method, Synthetic Minority Oversampling Technique (SMOTE), that simultaneously downsizes the class with majority cases and oversamples minority class by creating synthetic examples was proposed by Chawla *et al.* [17]. Instead of focusing on the data space, SMOTE places emphasis on the feature space, and hence creates the synthetic examples by artificially interpolating pre-existing minority class examples within a defined neighbourhood. However, SMOTE only focuses on continuous variables. Chawla *et al.* [17] extended SMOTE to handle categorical variables, and proposed Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC) where the median of standard deviations of continuous features is introduced to impose a penalty when an instance and its closest neighbours have different labels for a categorical feature. This study adopts SMOTE-NC to handle the binary class imbalance problem.

Finally, the balanced dataset is used to fit the random survival forests algorithm to find out the survival of key population in the HIV and AIDS programme. The cox proportional hazards (Cox-PH) model has long been the go-to method for survival analysis. However, researchers such as Hothorn and Lausen [18], Ishwaran *et al.* [19], Ramezankhani *et al.* [20], and Spooner *et al.* [21] have highlighted shortcomings of Cox-PH model such as the proportional hazard assumption; inability to predict risk of future events because the baseline hazard function is unknown; and it also doesn't scale up well to high dimensions' data-

sets since it was initially designed for small data sets. In consequence, more robust survival models have been put forward such as survival trees and random survival forest. Ishwaran *et al.* [19] proposed random survival forests as an extension of Breiman [17] random forests to accommodate right-censored survival data. Random survival forests is an ensemble supervised machine learning algorithm that operates by building several survival trees grown from bootstrap samples. The survival trees are grown progressively to full size by recursively splitting the daughter nodes while maximising survival difference between the nodes, and then terminates when no new daughters can be formed because of the stopping criterion that each node must contain at least one unique death. Random survival forests has been applied in various domains ranging from financial sector [22] [23] to health sector [15] [19] [24]. Researchers have attributed random survival forests success and adoption in various contexts because it is highly data-adaptive, assumption free, detects interactions among features, and it's also known for automatically and coherently handling the proportionality assumption. Feature selection is an important step when building machine learning models in order to optimise their predictive or classification performances, and this study uses the variable importance measure to determine the most important features for the random survival forest models to predict ensemble mortality and survival function of a key population individual.

The organization structure of this paper is as follows: Section 1 highlights the background and assumption of the study; Section 2 discusses the methodology that is chosen to achieve the objective of the study; the major findings and analysis are reported in Section 3; and in Section 4 conclusions and recommendations for further study are discussed.

Assumption of the Study

The study assumed that no attrition case had been enrolled in the programme in a different location than the one they were reported as either defaulted (dropped, transferred out, or relocated) or lost to follow up.

2. Methodology

2.1. Research Design

This study adopts a retrospective design. The study uses July 2019 to June 2021 secondary data of key population spread across the 19 counties in Kenya grouped into two cohorts, 2019/2020 and 2020/2021. The data were provided by Kenya Red Cross Society under strict privacy measures due to its sensitive nature. The study focuses on modelling the key population attrition in the HIV and AIDS programme in Kenya. The study begins by using density analysis to map attrition cases in the 19 targeted counties and Kilifi county as an example to find out the spatial differences, and also to determine the hotspots of attrition. The study then uses synthetic minority oversampling technique-nominal continuous (SMOTE-NC) to balance the datasets since the cases of attrition are the minority

and retention cases are the majority. Random survival forests then learns from the balanced dataset to predict the ensemble mortality value. This study also utilises the nonparametric measure of variable importance provided by the random survival forests to determine the most significant features associated with attrition.

2.2. Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC)

This study employs the synthetic minority oversampling technique-nominal continuous as suggested by Chawla *et al.* [14] since it balances a dataset by handling continuous and categorical features simultaneously. SMOTE-NC is chosen over other data balancing techniques because it is simple, computationally efficient, and has exceptional performance as shown by different researchers such as Rahmayanti *et al.* [25] and Islahulhaq and Ratih [26] in their studies. In addition, it's also worth mentioning that SMOTE-NC works with any classification approach since it is independent of the classifier. However, SMOTE-NC is intended for datasets with both categorical and numerical features; it's not well suited for datasets with categorical features only [14]. Islahulhaq and Ratih [26] notes that a dataset can be categorised as imbalanced if the cases of minority class are less than 35% of cases from the majority class, that is, if the imbalance ratio is less than 0.35.

Let $\mathcal{A}$ denote a full dataset with dimension $(n \times p)$, where n denotes the size of the dataset while p denotes the number of variables considered per observation. SMOTE-NC algorithm has two hyperparameters: a number, $k \in \mathbb{N}$ showing the count of the nearest neighbours; and a number, $N \in \mathbb{N}$, that controls the extent of oversampling the minority class. Let $M \subseteq \mathcal{A}$ denote the set of the observations associated with the minority class. The procedure to create new synthetic data points using SMOTE-NC is as follows: first, a minority class instance is randomly selected. Let $S_i \in M$ where $S_i = \left(x_{i1}, x_{i2}, \cdots, x_{it}, x_{i1}^{(c)}, x_{i2}^{(c)}, \cdots, x_{i(p-t-1)}^{(c)}\right)$ denote the sample (feature vector) of a minority instance i for which we are interested in computing its nearest neighbour. Let $S_i^* \in M$ where $S_i^* = \left(x_{i1}^*, x_{i2}^*, \cdots, x_{it}^*, x_{i1}^{*(c)}, x_{i2}^{*(c)}, \cdots, x_{i(p-t-1)}^{*(c)}\right)$ be one of the sample considered as a neighbour of S_i. $\left(x_{i1}, x_{i2}, \cdots, x_{it}\right)$ and $\left(x_{i1}^*, x_{i2}^*, \cdots, x_{it}^*\right)$ denote the continuous features while $\left(x_{i1}^{(c)} x_{i2}^{(c)}, \cdots, x_{i(p-t-1)}^{(c)}\right)$ and $\left(x_{i1}^{*(c)} x_{i2}^{*(c)}, \cdots, x_{i(p-t-1)}^{*(c)}\right)$ denote the categorical features of the samples. Assume there are z categorical features that differ between S_i and S_i^*, then the Euclidean distance between S_i and S_i^* is computed as:

$$\left\|S_i^* - S_i\right\| = \sqrt{\left(x_{i1}^* - x_{i1}\right)^2 + \cdots + \left(x_{it}^* - x_{it}\right)^2 + \text{penalty}} \tag{1}$$

where, $\text{penalty} = z \times \text{Median}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{is} - \overline{x}_s\right)^2}\right\}, s = 1, 2, \cdots, t$

In words, the penalty is the median of standard deviations of all the continuous features for the minority class. We utilise the median to penalise the dif-

ference in categorical features by a quantity that is associated with the usual difference in continuous feature values.

Thereafter, N elements are randomly selected from the k neighbours to generate new synthetic examples via interpolation, where $0 < N < k$. Note, k is usually set to 5 (default). SMOTE-NC focuses on the "feature space" instead of "data space". Taking S_i^{**} as one of the nearest neighbours of S_i, then a new synthetic sample is generated as per Equation (2)

$$S_i^{new} = S_i + \omega\left(S_i^{**} - S_i\right), \omega \sim U(0,1) \tag{2}$$

The weight, ω, is a randomly drawn number between 0 and 1. By implementing, the synthetic data point, S_i^{new}, may be thought of as a randomly sampled point on the line segment between two minority instances in the feature space. The categorical feature is assigned the label that is most frequent in the majority of the k-nearest neighbors. As a result, the learner's misclassification cost of the minority class is minimised and the initial bias is reversed in disfavour of the majority class.

2.3. Random Survival Forests

This study uses random survival forests algorithm as proposed by Ishwaran *et al.* [19]. Consider a right-censored training survival dataset $\mathcal{D} = \left\{\left(\mathcal{T}_i, \mathcal{X}_i, \xi_i\right), i = 1,2,\cdots,n\right\}$, where $\mathcal{X}_i$, $\mathcal{T}_i$ and ξ_i denote the feature vector, observed survival time, and censoring indicator for individual *i*, respectively. $\mathcal{T}_i = \min\left(\mathcal{T}_i^0, \mathcal{C}_i^0\right)$, where $\mathcal{T}_i^0$ and $\mathcal{C}_i^0$ denote the true survival and censoring time, respectively. ξ_i is defined as

$$\xi_i = I_{\left\{\mathcal{T}_i^0 < \mathcal{C}_i^0\right\}} = \begin{cases} 1, & \text{Attrition has occurred} \\ 0, & \text{Censored observation}\left(\text{Retention}\right) \end{cases} \tag{3}$$

From $\mathcal{D}$, sample uniformly and independently m datasets with replacement, $\mathcal{D}_1, \cdots, \mathcal{D}_m$. The bootstrap samples should have the same size as the original dataset, that is, $\left|\mathcal{D}_j\right| = \left|\mathcal{D}\right|$. For each bootstrap sample, $\mathcal{D}_j$, train a survival tree to full depth. It is important to note that when growing random survival forests, it is essential to consider the outcome, and given right censored survival data, this involves the survival time, $\mathcal{T}_i$, and censoring status, ξ_i, of the individuals, which are also considered when determining the splitting criterion in each node [27]. Tree methods are characterised by node splitting rules which maximise homogeneity or purity of nodes and stopping rule that decides the optimal size of the tree [28]. In the case of random survival forests, trees are grown until no additional daughter nodes can be created because of the stopping criterion that the terminal node must have at least $d_0 > 0$ unique deaths (at least 1 death). Survival trees are used as base learners in random survival forests, which improves on the ensemble approach by introducing randomness to the learning process. Random bootstrap sampling in tandem with random feature selection introduces the randomness that ensures that the trees are de-correlated. Random feature selection is implemented as follows: before each split, randomly subsam-

ple k features without replacement out of the p input features, such that $k \ll p$, and the best split on these k is used to split the node. Typically, k is set as the square root of the number of all input features, that is, $k = \sqrt{p}$. Splitting a node on best features increases the similarity of observations within the resultant nodes while considering the response variable. In addition, due to random feature selection, not all features entering the tree-growing procedure will be in the final model, but instead, only the features that give the best split at every step in computation while simultaneously meeting the criteria set to ensure the tree's performance is optimal [28]. It is worth mentioning that all terminal nodes in random survival forests are given equal weights [27].

The node impurity of a survival tree is measured by survival difference which informs how effective (pure) a data split is. The splitting rule generates partition in the features space by pushing different cases apart. This study only considers log-rank splitting rule.

2.3.1. Log-Rank Splitting Rule

Let the right-censored training survival dataset at the root node be given as $\mathcal{D} = \{(\mathcal{T}_i, \mathcal{X}_i, \xi_i), i = 1, 2, \cdots, n\}$, and the goal is to split the node on a feature, say x_i. Denote, ζ, as the threshold that splits the root node into right, $\mathcal{R} = \{x_i > \zeta\}$, and left, $\mathcal{L} = \{x_i \le \zeta\}$, daughter nodes. Denote the unique death (attrition) times as $\tau_1 < \tau_2 < \cdots < \tau_m$. The individuals "at risk" of attrition at time τ_j are $\mathcal{Y}_j^{(\mathcal{R})} = \#\{\mathcal{T}_i \ge \tau_j, x_i > \zeta\}$ in the right daughter node, and $\mathcal{Y}_j^{(\mathcal{L})} = \#\{\mathcal{T}_i \ge \tau_j, x_i \le \zeta\}$ in the left daughter node, implying that total individuals "at risk" are $\mathcal{Y}_j = \mathcal{Y}_j^{(\mathcal{R})} + \mathcal{Y}_j^{(\mathcal{L})}$. Also, let $d_j^{(\mathcal{R})}$ and $d_j^{(\mathcal{L})}$ denote the number of deaths in the right and left daughter nodes respectively at τ_j, and thus the total number of deaths in both nodes is: $d_j = d_j^{(\mathcal{R})} + d_j^{(\mathcal{L})}$. The log-rank split-statistic value is given by:

$$\mathcal{L}(x_i, \zeta) = \frac{\sum_{j=1}^{m}\left[d_j^{(\mathcal{L})} - \mathcal{Y}_j^{(\mathcal{L})}\left(\frac{d_j}{\mathcal{Y}_j}\right)\right]}{\sqrt{\sum_{j=1}^{m}\frac{\mathcal{Y}_j^{(\mathcal{L})}}{\mathcal{Y}_j}\left(1-\frac{\mathcal{Y}_j^{(\mathcal{L})}}{\mathcal{Y}_j}\right)\left(\frac{\mathcal{Y}_j - d_j}{\mathcal{Y}_j - 1}\right)d_j}} \tag{4}$$

A better split happens when $|\mathcal{L}(x_i, \zeta)|$ is large since then, it maximises the survival difference between the right and left daughter nodes.

2.3.2. Survival Tree Terminal Node Statistics

1) In-Bag Survival Tree Estimators

This study begins by splitting the datasets into training (in-bag) and testing (out-of-bag) data. Denote the terminal node of the survival tree by $\mathfrak{h}$, and $\tau_{1,\mathfrak{h}} < \tau_{2,\mathfrak{h}} < \cdots < \tau_{m(\mathfrak{h}),\mathfrak{h}}$ be the distinct death times in $\mathfrak{h}$. In addition, let the number of deaths and individuals at risk at time $\tau_{j,\mathfrak{h}}$ be denoted by $d_{j,\mathfrak{h}}$ and $\mathcal{Y}_{j,\mathfrak{h}}$, respectively. The cumulative hazard function (CHF) at terminal node $\mathfrak{h}$ of a single survival tree is estimated using bootstrapped Nelson-Aalen estimator, and it is given by:

$$\hat{\mathcal{H}}_{\mathfrak{h}}(\tau)=\sum_{\tau_j,\mathfrak{h}\le\tau}\frac{d_{j,\mathfrak{h}}}{\mathcal{Y}_{j,\mathfrak{h}}} \tag{5}$$

On the other hand, the bootstrapped Kaplan-Meier estimators is given by:

$$\hat{\mathcal{S}}_{\mathfrak{h}}(\tau)=\prod_{\tau_j,\mathfrak{h}\le\tau}\left(1-\frac{d_{j,\mathfrak{h}}}{\mathcal{Y}_{j,\mathfrak{h}}}\right) \tag{6}$$

All individuals in the terminal node $\mathfrak{h}$ are assigned the same estimates for CHF and survival since they are assumed to have similar (homogeneous) survival behaviour. For a given feature vector, say $\mathcal{X}_i$, CHF and survival estimators are determined by dropping $\mathcal{X}_i$ down the survival tree and it will fall into a unique terminal node, say $\mathfrak{h}$, and thus, $\mathcal{X}_i$'s CHF and survival estimators will be identical to the Nelson-Aalen, Equation (5), and Kaplan-Meier, Equation (6), estimators for the terminal node where $\mathcal{X}_i$ falls into. That is, $\hat{\mathcal{H}}^{IB}(\tau\,|\,\mathcal{X}_i)=\hat{\mathcal{H}}_{\mathfrak{h}}(\tau)$ and $\hat{\mathcal{S}}^{IB}(\tau\,|\,\mathcal{X}_i)=\hat{\mathcal{S}}_{\mathfrak{h}}(\tau)$, which denote the in-bag (IB) data estimators.

2) Out-of-Bag Survival Tree Estimators

Let a survival tree grown from cases where *i* is OOB be defined as $\mathcal{O}_i$. Let the OOB indicator function be denoted by:

$$I_i=\begin{cases}1, & i\in\mathcal{O}_i\\ 0, & \text{Otherwise}\end{cases} \tag{7}$$

Dropping a feature vector, say $\mathcal{X}_i$, for an individual in the OOB data ($i\in\mathcal{O}_i$), down a survival tree and falls into terminal node, say $\mathfrak{h}$, the Nelson-Aalen estimator is given by $\hat{\mathcal{H}}^{OOB}(\tau\,|\,\mathcal{X}_i)=I_{i\in\mathcal{O}_i}\left[\hat{\mathcal{H}}_{\mathfrak{h}}(\tau)\right]$ and $\hat{\mathcal{S}}^{OOB}(\tau\,|\,\mathcal{X}_i)=I_{i\in\mathcal{O}_i}\left[\hat{\mathcal{S}}_{\mathfrak{h}}(\tau)\right]$, respectively. $\hat{\mathcal{H}}_{\mathfrak{h}}(\tau)$ is estimated as per Equation (5) and $\hat{\mathcal{S}}_{\mathfrak{h}}(\tau)$ as per Equation (6).

2.3.3. Ensemble Statistics

1) In-Bag Ensemble Estimators

Let $\mathcal{T}$ denote the total number of survival trees in random survival forest, and let, $\mathcal{X}_i$'s in-bag Nelson-Aalen and Kaplan-Meier estimators for a survival tree, grown from $\mathcal{D}_j^{\text{th}}$ bootstrap sample, be noted by $\hat{\mathcal{H}}^{IB}_{\mathcal{D}_j}(\tau\,|\,\mathcal{X}_i)$ and $\hat{\mathcal{S}}^{IB}_{\mathcal{D}_j}(\tau\,|\,\mathcal{X}_i)$, then the Nelson-Aalen CHF estimator for the ensemble is given by:

$$\bar{\mathcal{H}}^{IB}(\tau\,|\,\mathcal{X}_i)=\frac{1}{\mathcal{T}}\sum_{j=1}^{\mathcal{T}}\hat{\mathcal{H}}^{IB}_{\mathcal{D}_j}(\tau\,|\,\mathcal{X}_i) \tag{8}$$

The Kaplan-Meier survival estimator for the ensemble is given by:

$$\bar{\mathcal{S}}^{IB}(\tau\,|\,\mathcal{X}_i)=\frac{1}{\mathcal{T}}\sum_{j=1}^{\mathcal{T}}\hat{\mathcal{S}}^{IB}_{\mathcal{D}_j}(\tau\,|\,\mathcal{X}_i) \tag{9}$$

The in-bag estimators are used for prediction.

2) Out-of-Bag Ensemble Estimators

Recall, a survival tree grown from cases where *i* is OOB is denoted as $\mathcal{O}_i$, and the size (number) of all those trees is denoted as $|\mathcal{O}_i|$. The OOB Nelson-Aalen CHF estimator for individual $i\in\mathcal{O}_i$ for the ensemble is given by:

$$\bar{\mathcal{H}}^{OOB}(\tau\,|\,\mathcal{X}_i)=\frac{1}{|\mathcal{O}_i|}\sum_{\mathcal{D}_j\in\mathcal{O}_i}\hat{\mathcal{H}}^{IB}_{\mathcal{D}_j}(\tau\,|\,\mathcal{X}_i) \tag{10}$$

The OOB Kaplan-Meier survival estimator for individual $i \in \mathcal{O}_i$ for the ensemble is given by:

$$\bar{S}^{OOB}\left(\tau \mid \mathcal{X}_i\right) = \frac{1}{\left|\mathcal{O}_i\right|} \sum_{\mathcal{D}_j \in \mathcal{O}_i} \hat{S}^{IB}_{\mathcal{D}_j}\left(\tau \mid \mathcal{X}_i\right) \tag{11}$$

2.3.4. Mortality Value

Random survival forest uses this as the predicted value. It can also be referred to as the risk score. It is constructed from the Nelson-Aalen estimators given by Equation (8) and Equation (10) for the In-Bag and Out-of-Bag data, respectively. The mortality value is the estimated risk for each individual which is simply the average number of events for one particular leaf node. Denoting $\tau_1 < \tau_2 < \cdots < \tau_m$ as the distinct event times for the training data, the In-Bag ensemble mortality value for an individual with a feature vector, say $\mathcal{X}_i$, is given as:

$$\bar{\mathcal{M}}^{IB}\left(\mathcal{X}_i\right) = \sum_{j=1}^{m} \bar{H}^{IB}\left(\tau_j \mid \mathcal{X}_i\right) \tag{12}$$

On the other hand, the Out-of-Bag ensemble mortality value is given by:

$$\bar{\mathcal{M}}^{OOB}\left(\mathcal{X}_i\right) = \sum_{j=1}^{m} \bar{H}^{OOB}\left(\tau_j \mid \mathcal{X}_i\right) \tag{13}$$

Note, considering two individuals, say *i* and *j*, *i* is said to have a worst outcome than *j* if $\bar{\mathcal{M}}_i^{OOB} > \bar{\mathcal{M}}_j^{OOB}$. OOB ensemble mortality value is also used in computation of Harrell's Concordance Index (C-Index).

2.3.5. Performance Measures

This study uses Harrell's concordance index to evaluate the performance of the probabilistic risk predictions of the random survival forests model using the out-of-bag (training) data.

1) Harrell's Concordance Index

Concordance index estimates the proportion of pairings in which the observation with the greater survival time has a higher survival probability (smaller risk score) than the model predicts. The index is independent of the model's evaluation time and incorporates individual censoring, thus making it robust [21]. The c-index is constructed by first creating all feasible pairings of observations, say (i, j) from the entire dataset. Denote the survival time for *i* by $\mathcal{T}_i$ and for *j* by $\mathcal{T}_j$ and the censoring status for *i* by ξ_i and for *j* by ξ_j. Moreover, let the OOB ensemble mortality for *i* and *j* be denoted by $\bar{\mathcal{M}}_i^{OOB}$ and $\bar{\mathcal{M}}_j^{OOB}$. For a pair where observation associated with shorter survival time is censored, or if $\mathcal{T}_i = \mathcal{T}_j$ unless $\left(\xi_i = 1, \xi_j = 0\right)$, $\left(\xi_i = 0, \xi_j = 1\right)$, or $\left(\xi_i = 1, \xi_j = 1\right)$, then such a pair is omitted. Let the other remaining pairs after omission be called "permissible" pairs.

Next, for the permissible pairs: count 1 for each pair where the observation associated with shorter survival time had the worst outcome and $\mathcal{T}_i \neq \mathcal{T}_j$, or where for a pair $\mathcal{T}_i = \mathcal{T}_j$ and $\bar{\mathcal{M}}_i^{OOB} = \bar{\mathcal{M}}_j^{OOB}$. On the other hand, count 0.5 for each pair in which $\mathcal{T}_i \neq \mathcal{T}_j$ and $\bar{\mathcal{M}}_i^{OOB} = \bar{\mathcal{M}}_j^{OOB}$, or $\mathcal{T}_i = \mathcal{T}_j$ and $\bar{\mathcal{M}}_i^{OOB} \neq \bar{\mathcal{M}}_j^{OOB}$.

Let concordance be the resultant count across all permissible pairs, and the index be denoted by $\mathcal{C}$. The c-index is thus given by:

$$\mathcal{C} = \frac{\text{concordance}}{\text{permissible}} \tag{14}$$

This statistic, given in Equation (14), is an estimate of the probability that, in a randomly selected pair of cases, the sequence of events that occur is successfully predicted, that is, the probability of classifying a pair of cases correctly. Hence, it is possible to estimate the prediction error using it which is given as $\mathcal{E} = 1 - \mathcal{C}$, $0 \le \mathcal{E} \le 1$. A perfect prediction will result to $\mathcal{E} = 0$ and random guessing would result to $\mathcal{E} = 0.5$.

2.3.6. Feature Selection

Random survival forests uses variable importance (VIMP) measure, which is fully non-parametric, as proposed by (Breiman, 2002) to determine variables which are most important for model to best predict risk score. VIMP utilises OOB data, which makes the algorithm less computationally expensive compared to using cross validation for feature selection. Variable importance measure is computed by subtracting the out-of-bag error before and after a feature vector, say $\mathcal{X}_i$, is randomly permuted. Basically, it computes the cost of misclassification. Thus, the significance for $\mathcal{X}_i$ is determined by the amount by which the new error surpasses the tree's initial OOB error. The larger the value, the more the predictive ability for $\mathcal{X}_i$, whereas a zero or negative value indicates noise variables which should be filtered. Taking the average of every survival tree gives $\mathcal{X}_i$'s permutation importance. Mathematically, recall that the training or learning data was given by $\mathcal{D} = \{(\mathcal{T}_i, \mathcal{X}_i, \xi_i), i = 1, 2, \cdots, n\}$, where $\mathcal{T}_i$ is the response variable (survival time) while $\mathcal{X}_i$ is the *p*-dimensional feature vector. Here, let's ignore the censoring indicator, ξ_i, and denote our training data as $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{T}_i), i = 1, 2, \cdots, n\}$. The aim is to estimate a function, $h(x)$ of the response given that $\mathcal{X}_i = x$.

1) Survival Tree Variable Importance

Denote $\mathcal{D}^*(\theta_m)$ and $\mathcal{D}^{**}(\theta_m)$ as the m^{th} bootstrap sample and its corresponding OOB data, respectively. Let $\mathcal{X} = (\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \cdots, \mathcal{X}^{(p)})$, where $\mathcal{X}^{(j)}$ denotes the j^{th} variable coordinate, and its permuted value be given by $\hat{\mathcal{X}}^{(j)}$. Note, $\hat{\mathcal{X}}^{(j)} = (\mathcal{X}^{(1)}, \cdots, \mathcal{X}^{(j-1)}, \hat{\mathcal{X}}^{(j)}, \mathcal{X}^{(j+1)}, \cdots, \mathcal{X}^{(p)})$ after substitution of the j^{th} coordinate of $\mathcal{X}$. The variable importance measure (denoted as: $I(\mathcal{X}^{(j)}, \theta_m, \mathcal{D})$) of the m^{th} survival tree, which is given by the difference between the predicted error of the original $\mathcal{X}$ and the permuted value $\hat{\mathcal{X}}^{(j)}$ is given by:

$$I(\mathcal{X}^{(j)}, \theta_m, \mathcal{D}) = \frac{1}{\mathcal{N}(\theta_m)} \sum_{i \in \mathcal{D}^{**}(\theta_m)} \left[l\left(\mathcal{T}_i, h\left(\hat{X}_i^{(j)}, \theta_m, \mathcal{D}\right)\right) - l\left(\mathcal{T}_i, h\left(\hat{X}_i, \theta_m, \mathcal{D}\right)\right) \right] \tag{15}$$

2) Ensemble Variable Importance

The ensemble variable importance is given by averaging the survival tree VIMP given by Equation (15), that is,

$$I\left(\mathcal{X}^{(j)},\theta_1,\cdots,\theta_m,\mathcal{D}\right)=\frac{1}{M}\sum_{m=1}^{M}I\left(\mathcal{X}^{(j)},\theta_m,\mathcal{D}\right) \tag{16}$$

It is also possible to define the $100(1-\alpha)$ confidence region for the true variable importance, and is given by $\psi_{1-\alpha}=\hat{\theta}_n^{(j)}\pm Z_{\frac{\alpha}{2}}\sqrt{\hat{v}_n^{(j)}}$, where $\hat{v}_n^{(j)}$ is variance of the predicted error.

3. Data Analysis

3.1. Geospatial Mapping of Attrition Cases

The study began by investigating the spatial differences of attrition in the 19 targeted counties, and Kilifi County as an example to map attrition in smaller administrative areas for all key population groups, that is, female sex workers (FSW), men who have sex with men (MSM) and people who inject drugs (PWID).

3.1.1. At County Level

As shown in Figure 1, for the 2019/2020 cohort, the majority of attrition cases for female sex workers were recorded in West Pokot, Kwale, Trans Nzoia, Nairobi, and Nyeri counties with an attrition rate of 18.91%, 1.94%, 1.34%, 1.20% and 0.77%, respectively; whilst counties with the lowest attrition cases were Vihiga, Nyamira, Laikipia, Kisii, and Kericho with no attrition cases reported. On the other hand, for the 2020/2021 cohort, the majority of attrition cases for female sex workers were recorded in Kilifi, West Pokot, Nyeri, Laikipia, and Kwale counties with an attrition rate of 81.85%, 30.23%, 12.97%, 11.24%, and 10.48%,

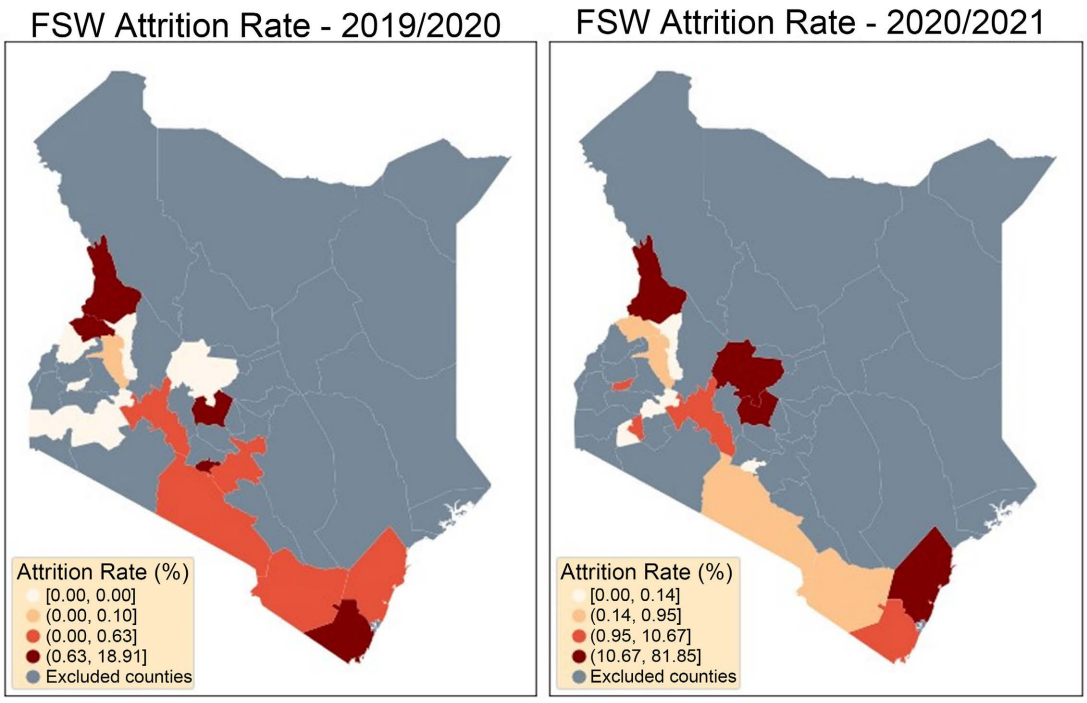

Figure 1. Female sex workers attrition rate.

respectively; while Kericho, Nairobi, Kisii, Elgeyo Marakwet and Kajiado counties had the lowest attrition cases with an attrition rate of 0%, 0.07%, 0.07%, 0. 10%, and 0.16%, respectively.

As shown in Figure 2, for the 2019/2020 cohort, the majority of attrition cases for men who have sex with men were recorded in Kiambu, Mombasa, Kisumu, Taita Taveta, and Kilifi counties with an attrition rate of 8.67%, 6.04%, 4.45%, 1.20%, and 0.53%, respectively; whilst counties with the lowest attrition cases were Vihiga, Uasin Gishu, Trans Nzoia, Nyeri, and Nyamira counties with no attrition cases recorded. For the 2020/2021 cohort, the majority of attrition cases for men who have sex with men were recorded in Kilifi, Laikipia, Nyamira, Mombasa, and Taita Taveta counties with an attrition rate of 6.71%, 2.38%, 2.0%, 1.63%, and 1.27%, respectively; while counties with the lowest attrition cases were Vihiga, Uasin Gishu, Trans Nzoia and Machakos, all with an attrition rate of 0%, and Nairobi county with an attrition rate of 0.08%.

The HIV and AIDS programme for people who inject drugs focused on 4 counties for both 2019/2020 and 2020/2021 cohorts. As shown in Figure 3, for 2019/2020 counties, Nairobi county had the highest number of attrition cases with an attrition rate of 9.13% followed by Kilifi and Mombasa counties with attrition rates of 2.52% and 0.64%, respectively, and Kwale county had the lowest attrition rate of 0.27%. For the 2020/2021 cohort, Nairobi county still reported the highest attrition cases with an attrition rate of 3.18% followed by Kilifi and Mombasa counties with attrition rates of 1.63% and 1.48%, respectively, and Kwale county had the least number of attrition cases with an attrition rate of 0.10%.

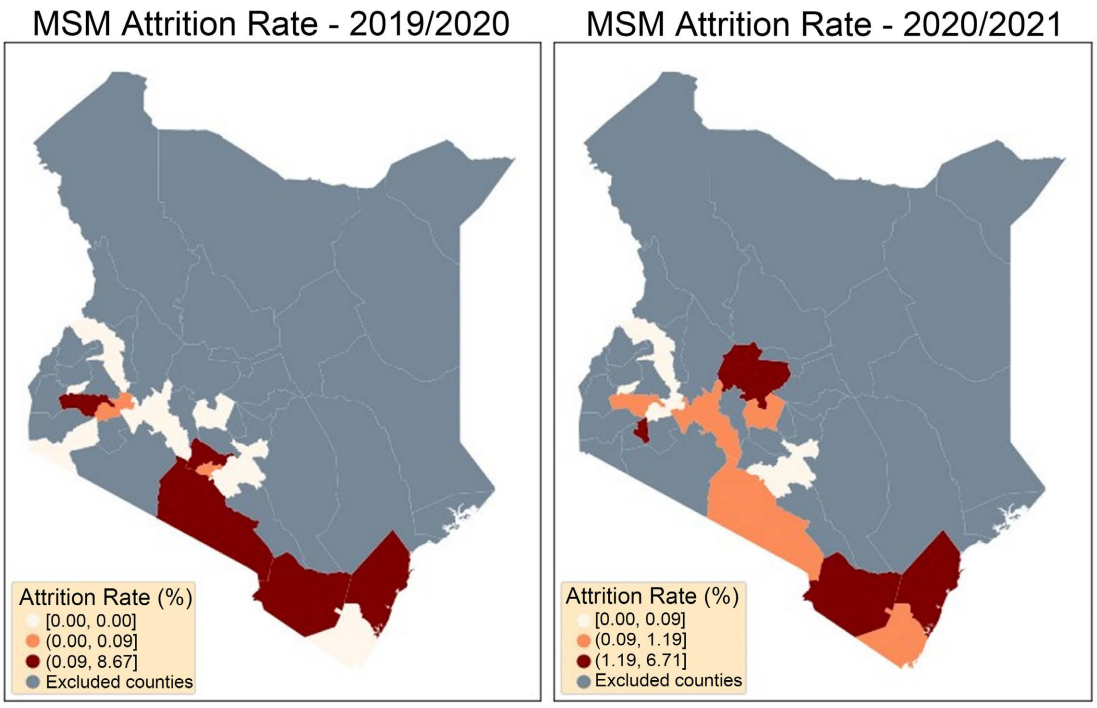

Figure 2. Men who have sex with men attrition rate.

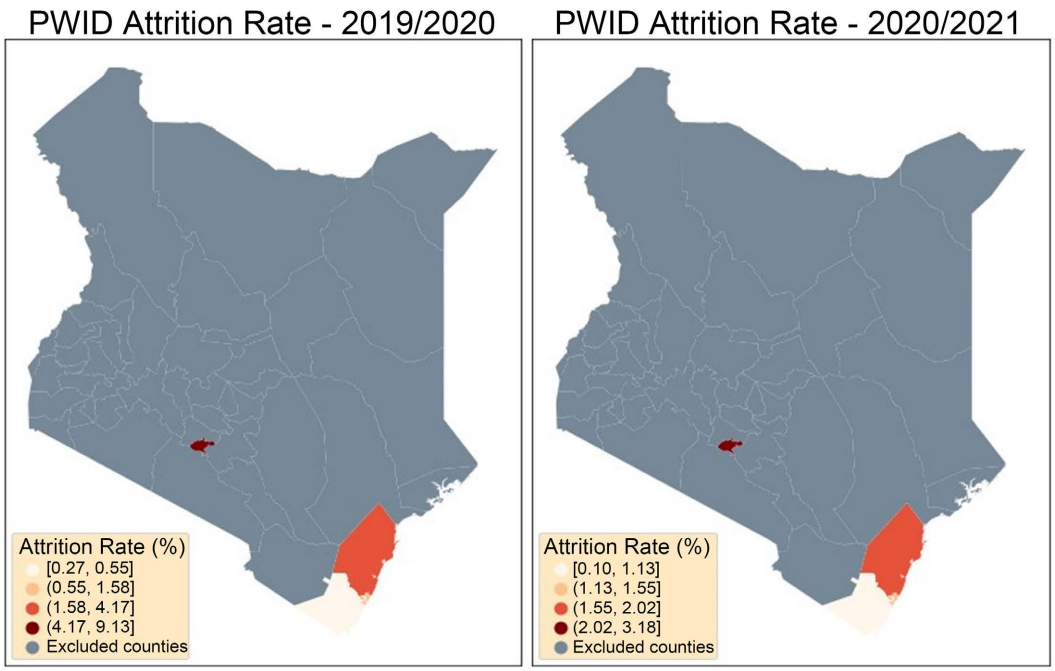

Figure 3. People who inject drugs attrition rate.

3.1.2. At Sub-County Level

This study focused on Kilifi county to find out the spatial differences of attrition cases in smaller administrative areas (sub-counties) for all the three key population groups. The choice of Kilifi county was informed by the high attrition rates reported for the key population.

As shown in **Figure 4**, for both the 2019/2020 and 2020/2021 cohort, the majority of attrition cases for female sex workers in Kilifi were reported in the Southern part of the county. For the 2019/2020 cohort the attrition rate was 0.44% and 0.39% in Rabai and Kaloleni, respectively; while for the 2020/2021 cohort, the attrition rate increased to 96.73% and 78.03% in Rabai and Kaloleni, respectively.

As shown in **Figure 5**, the majority of attrition cases for men who have sex with men in Kilifi was on the Northern part of the county with Malindi leading with an attrition rate of 0.69% and Magarini with an attrition rate of 0.56% for the 2019/2020 cohort. For the 2020/2021 cohort, attrition cases were only reported in Malindi with an attrition rate of 9.26%

As shown in **Figure 6**, Kilifi North and Kilifi South had the highest attrition rate of 7.86% and 6.49%, respectively, for people who inject drugs; while Malindi had the least attrition rate of 0.59% for the 2019/2020 cohort. For the 2020/2021 cohort, Kilifi South and Kilifi North still had the highest attrition rates of 4.82% and 4.65%, respectively, while Malindi and Magarini had an attrition rate of 0.26% and 0.17%, respectively.

3.2. Data Balancing

The cases of attrition compared to retention were much lower making the datasets

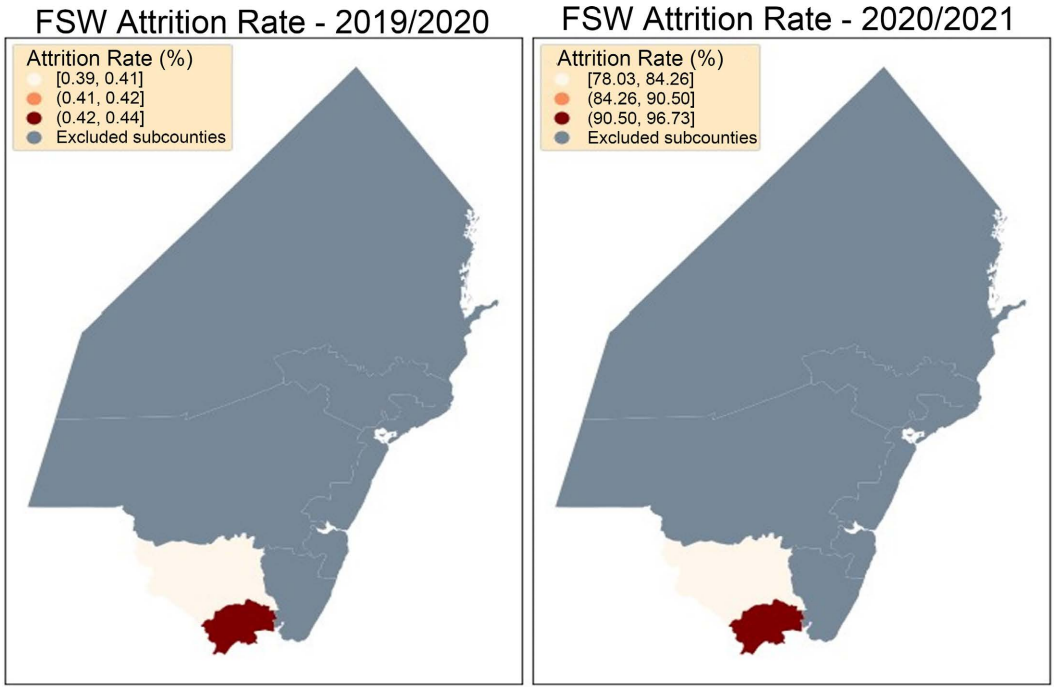

Figure 4. Kilifi female sex workers attrition rate.

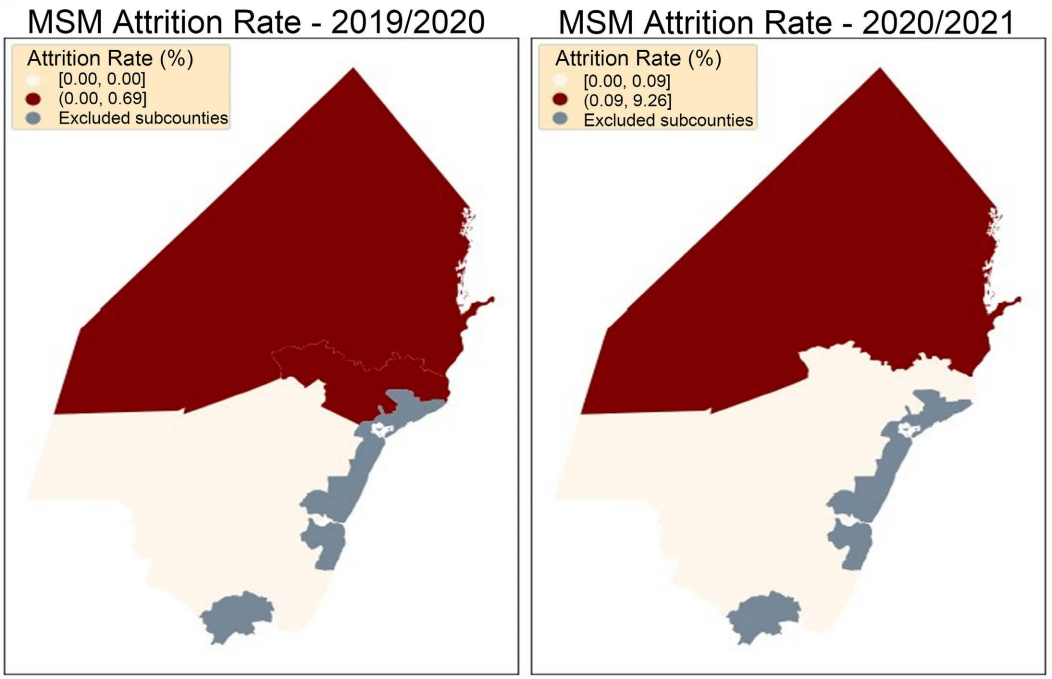

Figure 5. Kilifi Men who have sex with men attrition rate.

highly imbalanced which, if not taken care of, could lead to poor performance of the machine learning algorithm—random survival forests. This study first computed the imbalance ratios of the datasets as:

$$\text{Imbalance Ratio}\left(\text{IR}\right)=\frac{\text{Minority Cases}\left(\text{Attrition}\right)}{\text{Majority Cases}\left(\text{Retention}\right)}$$

The imbalance ratios for the datasets used in this study are shown in Table 1.

From the results in Table 1 and the plots in Figure 7 and Figure 8, it is clear that the datasets are highly imbalanced with all imbalance ratios less than 0.35 - the threshold of a dataset to be classified as imbalanced as per Islahulhaq and

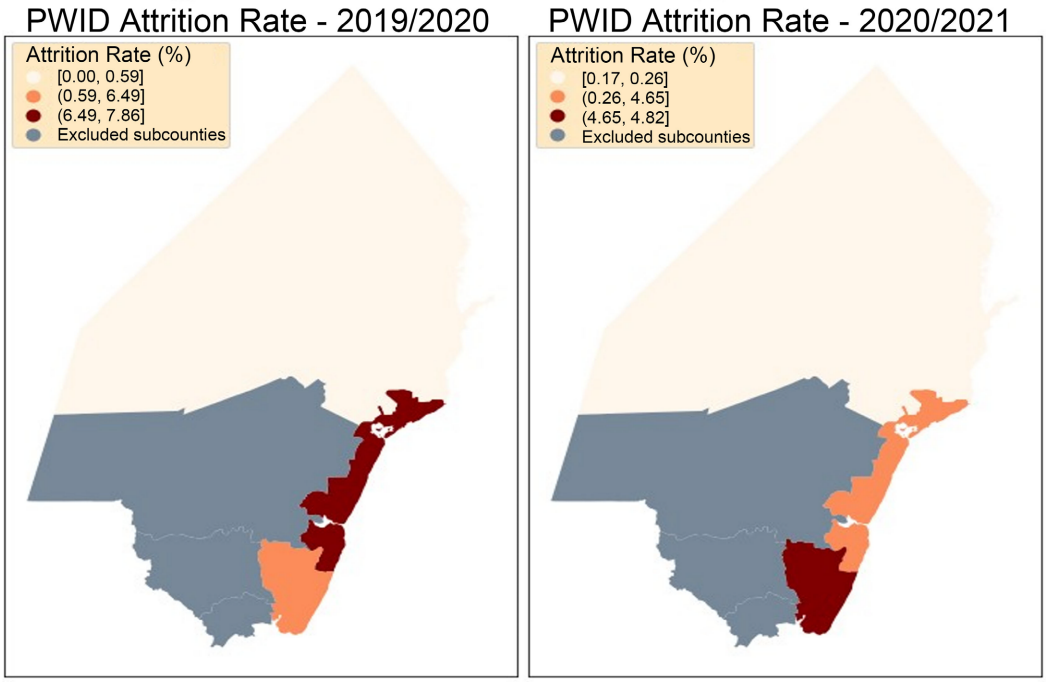

Figure 6. Kilifi people who inject drug attrition rate.

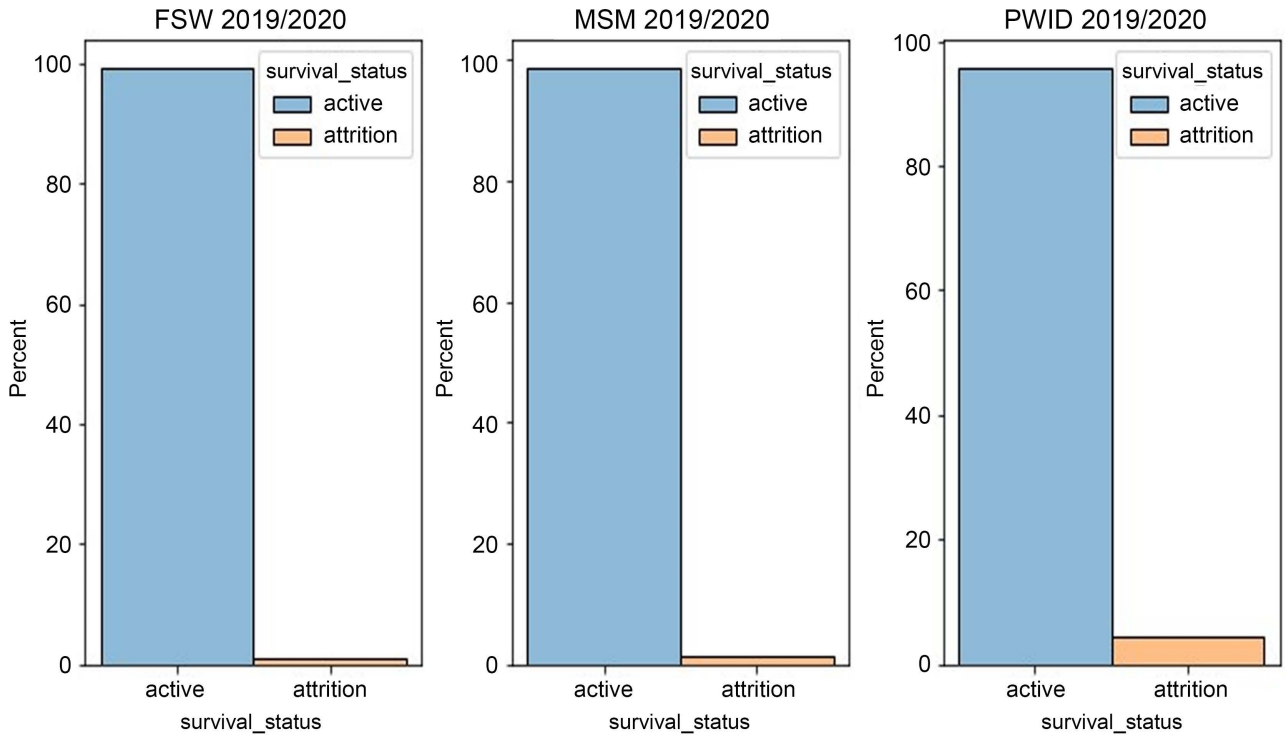

Figure 7. Key population 2019/2020 cohort attrition incidence.

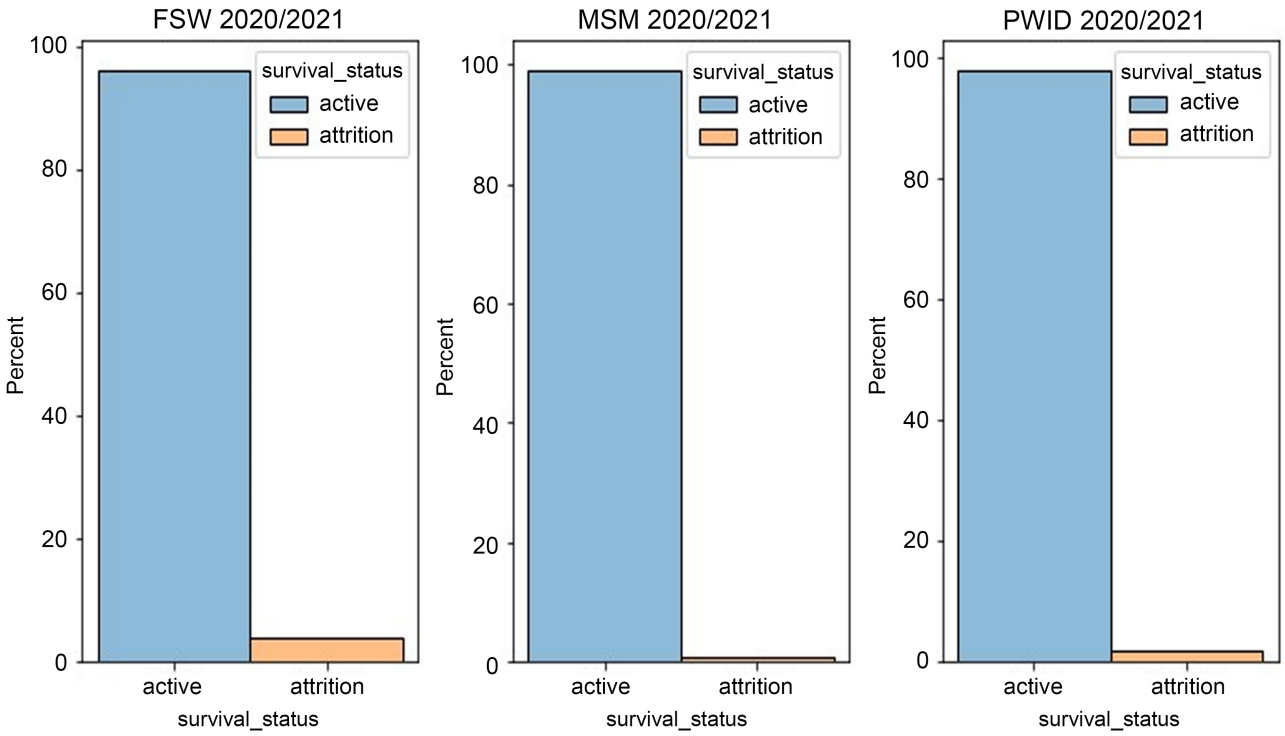

Figure 8. Key population 2020/2021 cohort attrition incidence.

Table 1. Imbalance ratios of the original datasets.

Key Population	2019/2020	2020/2021
FSW	0.0082	0.0401
MSM	0.0141	0.0081
PWID	0.0470	0.0192

Ratih [26]. Synthetic minority oversampling technique-nominal continuous (SMOTE-NC) is thus employed to balance the datasets by generating synthetic data points for the minority class (attrition). The study determines the optimal imbalance ratio in the range [0.35, 1] that yields the least predicted error by considering a sample size of 1000 from each key population dataset. The study uses the default number of nearest neighbours as per Chawla *et al.* [14], that is $k = 5$, to generate the synthetic data points. The optimal imbalance ratios of the datasets after applying SMOTE-NC are as shown in **Table 2** and the plots in **Figures 9-11**.

The study then used the datasets balanced using SMOTE-NC to fit the random survival forest model.

3.3. Ensemble Mortality

This study generated 1000 trees for the random survival forest algorithm, meaning that a 1000 bootstrap samples were used to grow the survival trees. In addition, the log-rank splitting rule was used when growing the survival trees. The

Table 2. Optimal imbalance ratios after applying SMOTE-NC.

Key Population	2019/2020	2020/2021
FSW	0.40	0.40
MSM	0.35	0.40
PWID	0.50	0.50

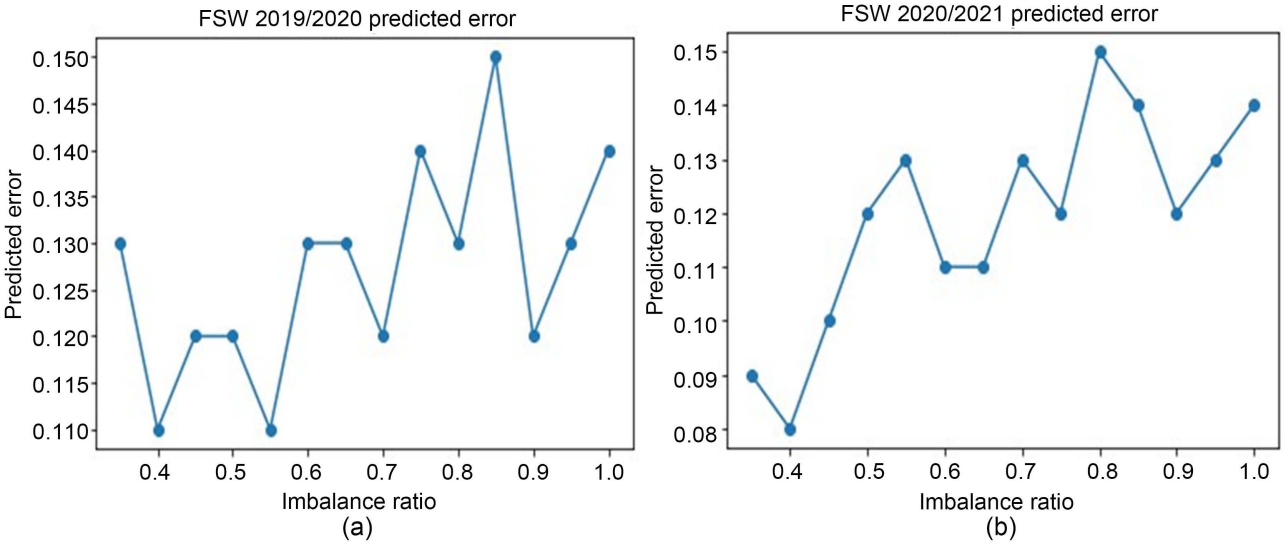

Figure 9. FSW 2019/2020 (a) and FSW 2020/2021 (b) optimal imbalance ratio.

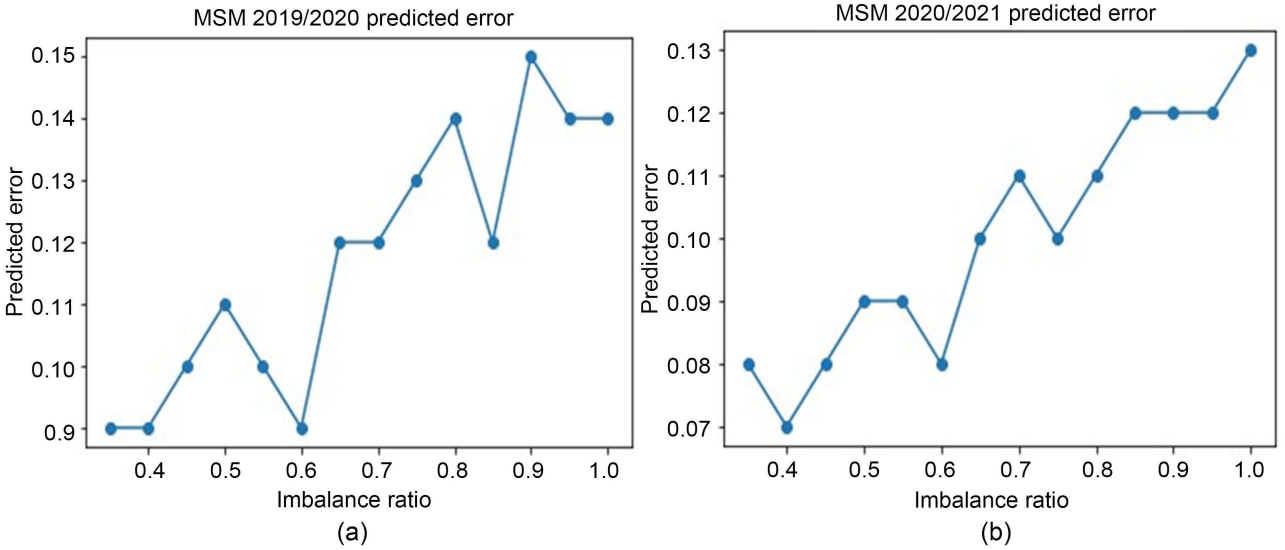

Figure 10. MSM 2019/2020 (a) and MSM 2020/2021 (b) optimal imbalance ratio.

minimum number of samples that was required to split an internal node was 10, while the minimum number of samples that was required to be at a leaf node was 15. The number of features that were considered when looking for the best split was the squareroot of the total number of input features per key population dataset. After training the random survival forests model using the in-bag data,

the study determined the ensemble mortality (risk scores) of 6 randomly chosen individuals from the out-of-bag data for each key population for both the 2019/2020 and 2020/2021 cohorts as shown in Tables 3-8. The study went further and plotted the estimated Kaplan-Meier survival function of the 6 randomly chosen individuals as shown below in Figures 12-14. The actual survival time is

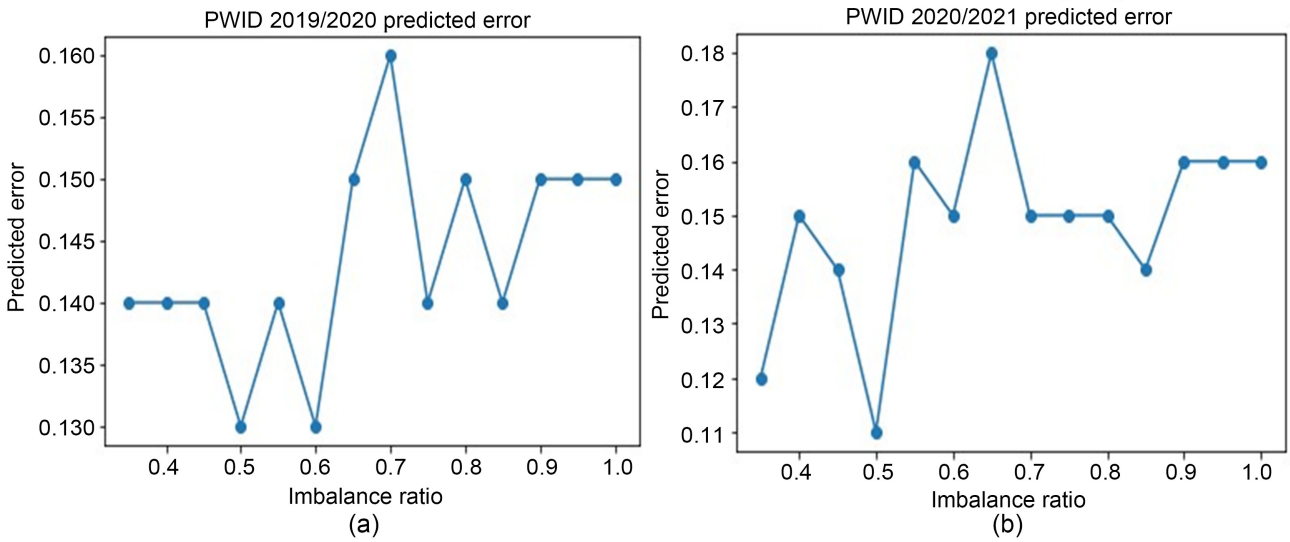

Figure 11. PWID 2019/2020 (a) and PWID 2020/2021 (b) optimal imbalance ratio.

Table 3. Ensemble mortality of 6 randomly sampled FSW from 2019/2020 cohort.

Individuals	Actual survival time	Actual survival status	Predicted ensemble mortality
0	1	active	4.4731
1	12	active	0.1761
2	6	active	0.4195
3	10	attrition	6.4758
4	11	attrition	5.5668
5	2	attrition	7.6227

Table 4. Ensemble mortality of 6 randomly sampled FSW from 2020/2021 cohort.

Individuals	Actual survival time	Actual survival status	Predicted ensemble mortality
0	12	active	0.1330
1	10	active	0.1091
2	12	active	0.4250
3	6	attrition	5.9688
4	6	attrition	7.5175
5	10	attrition	6.1278

Table 5. Ensemble mortality of 6 randomly sampled MSM from 2019/2020 cohort.

Individuals	Actual survival time	Actual survival status	Predicted ensemble mortality
0	1	active	0.2912
1	12	active	0.0114
2	2	active	0.3004
3	3	attrition	8.8722
4	5	attrition	10.7828
5	3	attrition	8.1108

Table 6. Ensemble mortality of 6 randomly sampled MSM from 2020/2021 cohort.

Individuals	Actual survival time	Actual survival status	Predicted ensemble mortality
0	12	active	6.6532
1	4	active	0.0930
2	12	active	0.1811
3	2	attrition	11.7834
4	5	attrition	11.2670
5	2	attrition	13.8415

Table 7. Ensemble mortality of 6 randomly sampled PWID from 2019/2020 cohort.

Individuals	Actual survival time	Actual survival status	Predicted ensemble mortality
0	6	active	0.6887
1	6	active	0.7952
2	12	active	3.9289
3	4	attrition	8.3554
4	3	attrition	9.4311
5	4	attrition	6.8323

Table 8. Ensemble mortality of 6 randomly sampled PWID from 2020/2021 cohort.

Individuals	Actual survival time	Actual survival status	Predicted ensemble mortality
0	4	active	1.7865
1	6	active	3.1029
2	12	active	2.5699
3	2	attrition	6.2759
4	9	attrition	4.2218
5	3	attrition	8.5793

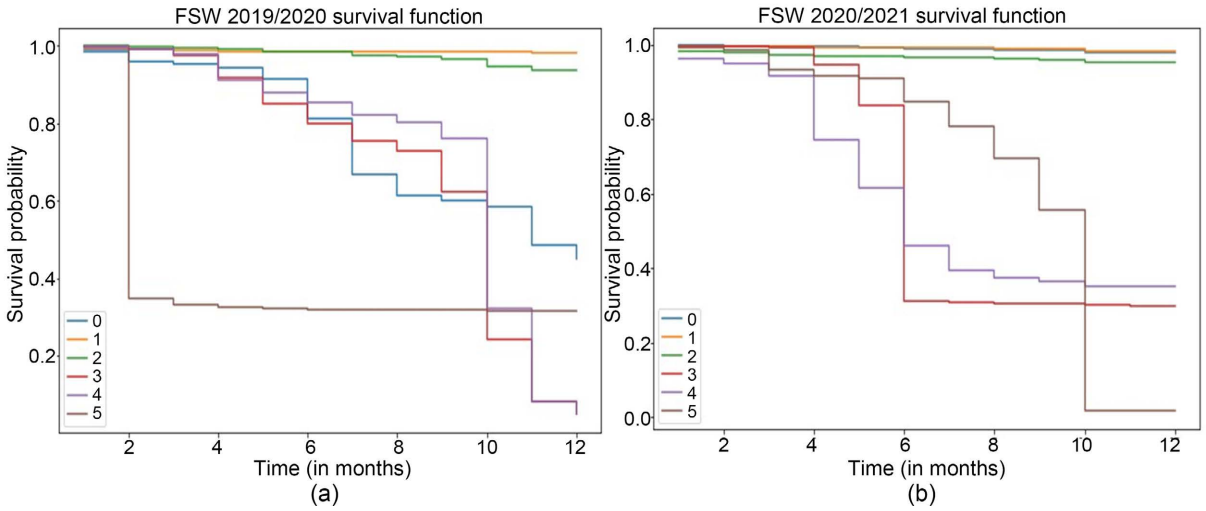

Figure 12. FSW 2019/2020 (a) and FSW 2020/2021 (b) survival function.

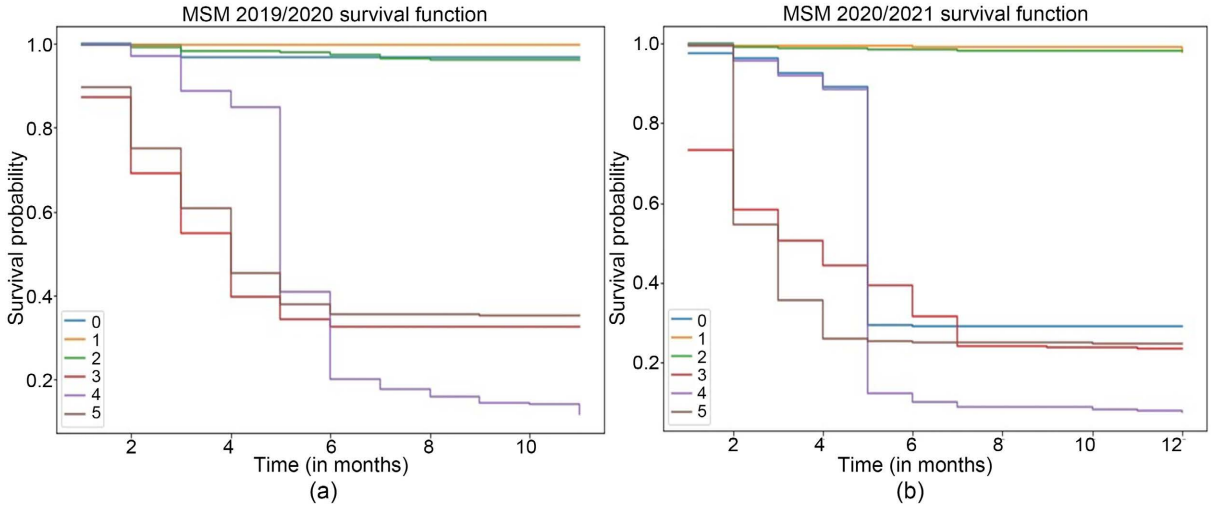

Figure 13. MSM 2019/2020 (a) and MSM 2020/2021 (b) survival function.

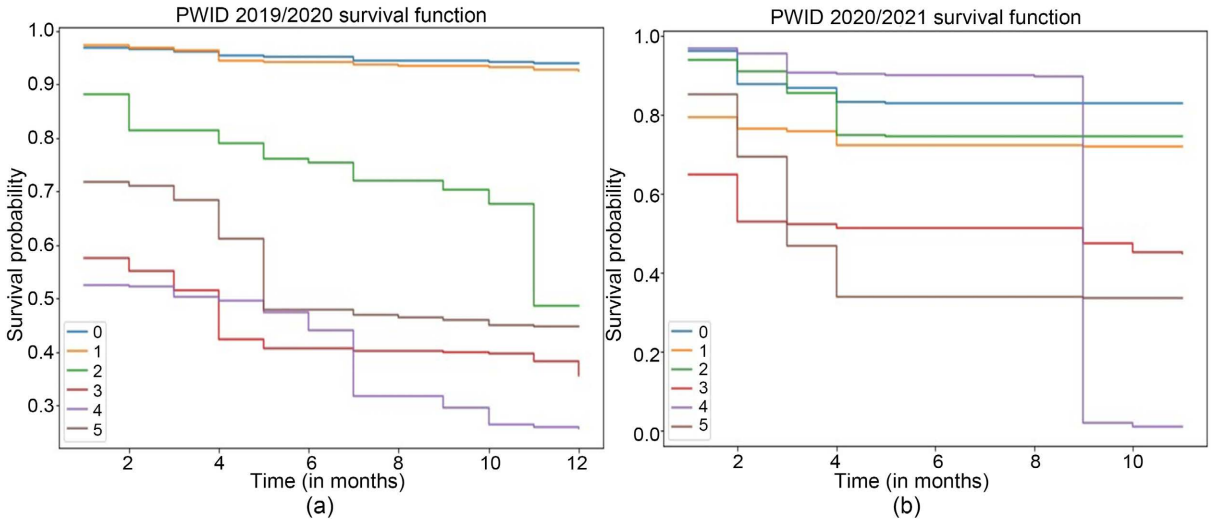

Figure 14. PWID 2019/2020 (a) and PWID 2020/2021 (b) survival function.

compared to when there is a big step in survival function. A big step in the survival function means that there is a huge risk of the event of interest happening, implying that if a big step occurs at time 2, then at that particular point it's when an individual was likely to experience the event (attrition). On the other hand, the actual survival status is compared to the ensemble mortality. Individuals with high ensemble mortality are at higher risk of attrition compared to those whose ensemble mortality is low. The study compares the actual survival status label to the values of predicted ensemble mortality to find out whether the model predicted them correctly.

From Table 4, it can be seen that individuals 3, 4 and 5 were labelled as attrition from the actual dataset, and the model also predicted that they had the highest predicted ensemble mortality. This shows that the model correctly predicted the attrition cases. In addition, from Figure 12(b), it can be seen that the big step of individual 3 and 4 was at time 6, while from the actual data, their survival time was 6 months; while for individual 5, the big step was at time 10, while from the actual data the survival time was 10 months. The same interpretation goes for the other key population groups for both the 2019/2020 and 2020/2021 cohort. However, since our random survival forests model is not a perfect model as seen from concordance indices in Table 9, there were some instances where the model predicted the attrition cases and survival time wrongly. However, it predicted the majority correctly.

3.4. Predictive Performance of the Random Survival Forests Model

This study used the Harrell's concordance index to evaluate the performance of the random survival forests model for each dataset used for both the 2019/2020 and 2020/2021 cohorts. The prediction error which is given as (1 − concordance index) was also computed. The greater the concordance index, the better the performance of the model since it means lower prediction error.

The results in Table 9 show that the predictive performance of the random survival forests model was great. It is also clear that the model's performance was way much better than random guessing. In addition, the concordance indices obtained for each key population dataset are comparable to those obtained in survival analysis as noted by Ptak-Chmielewska and Matuszyk [23].

3.5. Feature Importance

The plots below, Figures 15-17, show the top six most important features for the

Table 9. Concordance index and prediction error

Key	2019/2020		2020/2021	
population	c-index	prediction error	c-index	prediction error
FSW	0.8157	0.1843	0.8811	0.1189
MSM	0.8704	0.1296	0.8721	0.1279
PWID	0.8224	0.1776	0.7517	0.2483

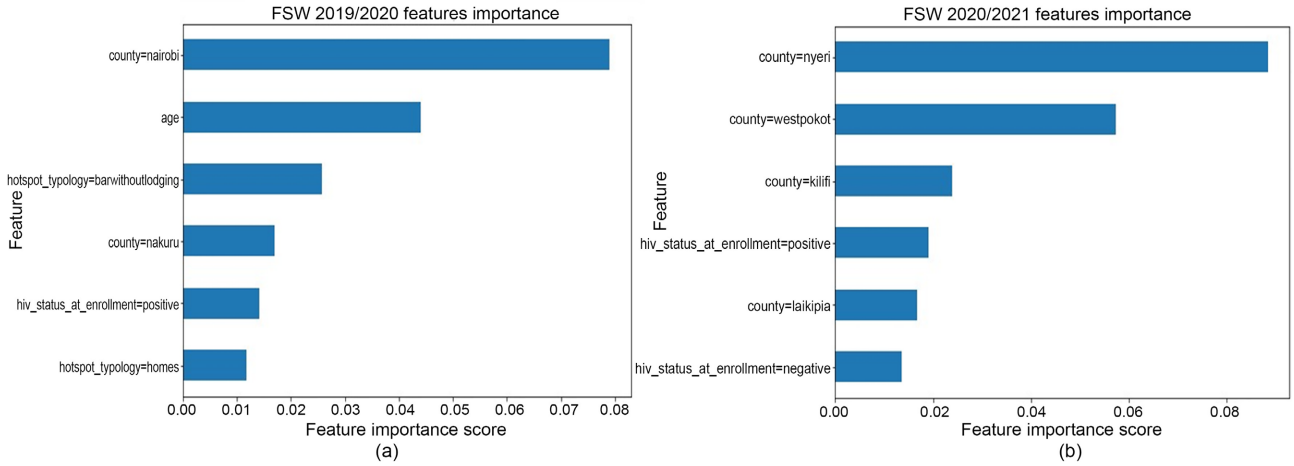

Figure 15. FSW 2019/2020 (a) and FSW 2020/2021 (b) feature importance.

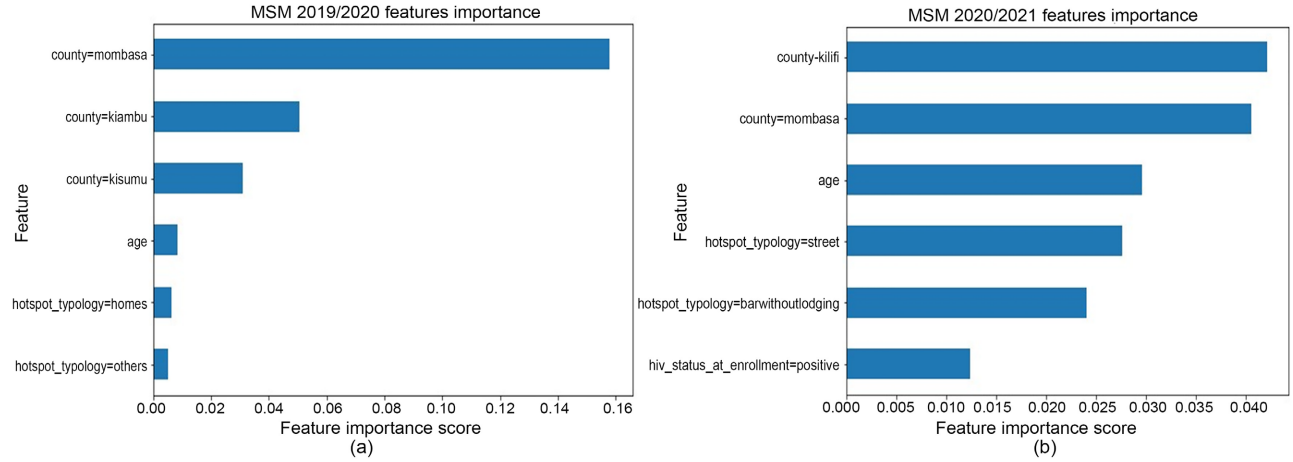

Figure 16. MSM 2019/2020 (a) and MSM 2020/2021 (b) feature importance.

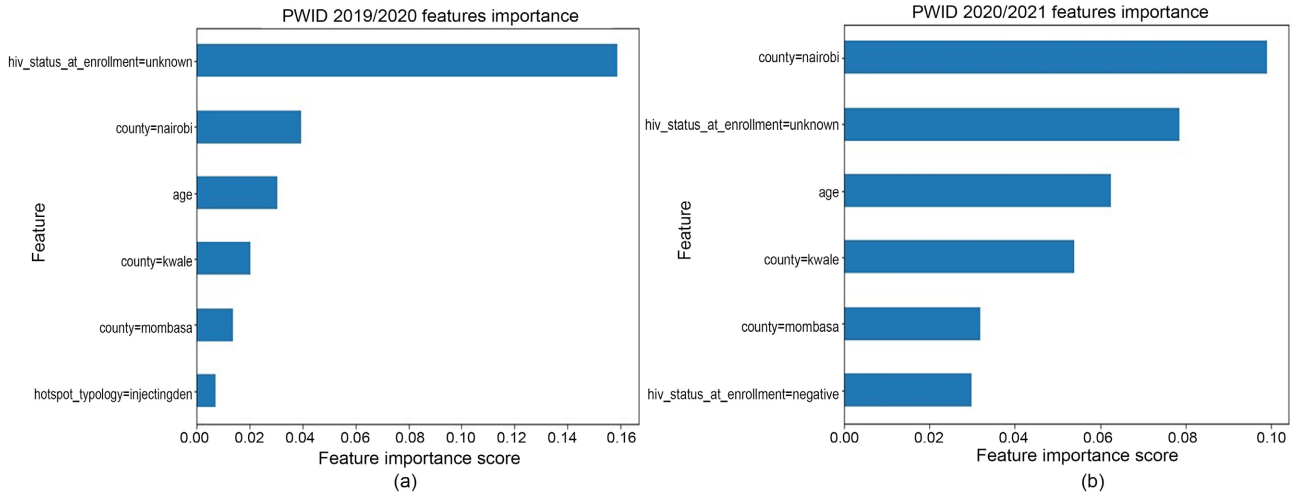

Figure 17. PWID 2019/2020 (a) and PWID 2020/2021 (b) feature importance.

random survival forests model to predict the risk score of attrition (outcome event) for a key population individual using the datasets from the 2019/2020 and

2020/2021 cohorts.

4. Conclusions and Recommendations

According to HIV and AIDS statistics, a lot more needs to be done if the HIV epidemic is to be eradicated in Kenya and throughout the world by 2030. This study adopted machine learning approaches to inform key population attrition since early identification of potential cases and intervention would ensure quality of life within HIV care continuum. The performance of the random survival forests model was great, and thus the model could be adopted by Kenya Red Cross Society and other organisations in the world having similar programmes to inform policy and strategy decision for men who have sex with men, female sex workers, people who inject drugs, or even the general population where cases of attrition have been reported in Antiretroviral Therapy (ART) to ensure that every individual stays in the programme until the end to reap the most benefits. For Kenya Red Cross Society, the model could first be deployed in areas where high cases of attrition were recorded as per the density analysis results for both the 2019/2020 and 2020/2021 cohorts in their subsequent HIV and AIDS programmes.

The findings of this study could also contribute further towards research on modelling key population attrition in HIV and AIDS programme. Possible areas of further research include assessing the performance of the random survival forests under various data balancing techniques. SMOTE-NC has been shown to be prone to over-generalisation in cases where the distribution of the minority samples is highly sparse thus increasing the likelihood of synthetic samples being generated in the majority samples feature space [16]. Therefore, variants of SMOTE such as bordeline SMOTE, support vector machine—SMOTE, penalty-based SMOTE, relocating safe-level SMOTE and other hybrid SMOTE algorithms could be used for a similar study. When building the random survival forests model, different splitting rules such as brier score gradient and log-rank score could be used to find out whether they lead to better performance. Finally, other machine learning algorithms could be used, and their performance compared to the random survival forests model. Some of the machine learning models that could be considered for a similar study includes gradient boosted models and survival support vector machine. Some possible future research topics are hyperparameter tuning for random survival forests and modelling Antiretroviral Therapy (ART) attrition using gradient boosted models with support vector machine SMOTE.

Acknowledgements

This study was approved, supported and funded by Kenya Red Cross Society-Global Fund Unit. The analysis was based on data collected between January 2018 to June 2021 through the New Funding Model II by Kenya Red Cross Society-Global Fund Unit.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] UNAIDS (2020) Unaids Data 2020. https://www.unaids.org/sites/default/files/media_asset/2020_aids-data-book_en.pdf

[2] NASCOP (2020) HIV and Aids Progress Report 2020. https://www.unaids.org/sites/default/files/media_asset/Annual_Progress_Report_HIV_Prevention.pdf

[3] KENPHIA (2018) Kenphia, 2018 Preliminary Report. https://www.health.go.ke/wp-content/uploads/2020/02/KENPHIA-2018-PREL-REP-2020-HR3-final.pdf

[4] NACC (2018) Karpr-Report 2018. https://nacc.or.ke/wp-content/uploads/2018/11/KARPR-Report_2018.pdf

[5] USAID (2021) Key Populations: Achieving Equitable Access to End Aids—U.S. Agency for International Development. https://www.usaid.gov/global-health/health-areas/hiv-and-aids/technical-areas/key-populations

[6] Hassan, A.S., Mwaringa, S.M., Ndirangu, K.K., Sanders, E.J., de Wit, T.F.R. and Berkley, J.A. (2015) Incidence and Predictors of Attrition from Antiretroviral Care among Adults in a Rural HIV Clinic in Coastal Kenya: A Retrospective Cohort Study. *BMC Public Health*, **15**, Article No. 478. https://doi.org/10.1186/s12889-015-1814-2

[7] Makurumidze, R., Mutasa-Apollo, T., Decroo, T., Choto, R.C., Takarinda, K.C., Dzangare, J., Lynen, L., Van Damme, W., Hakim, J., Magure, T., *et al.* (2020) Retention and Predictors of Attrition among Patients Who Started Antiretroviral Therapy in Zimbabwe's National Antiretroviral Therapy Programme between 2012 and 2015. *PLOS ONE*, **15**, e0222309. https://doi.org/10.1371/journal.pone.0222309

[8] Nacarapa, E., Verdu, M.E., Nacarapa, J., Macuacua, A., Chongo, B., Osorio, D., Munyangaju, I., Mugabe, D., Paredes, R., Chamarro, A., *et al.* (2021) Predictors of Attrition among Adults in a Rural HIV Clinic in Southern Mozambique: 18-Year Retrospective Study. *Scientific Reports*, **11**, Article No. 17897. https://doi.org/10.1038/s41598-021-97466-2

[9] Graham, S.M., Mugo, P., Gichuru, E., Thiong'o, A., Macharia, M., Okuku, H.S., van der Elst, E., Price, M.A., Muraguri, N. and Sanders, E.J. (2013) Adherence to Antiretroviral Therapy and Clinical Outcomes among Young Adults Reporting High-Risk Sexual Behavior, Including Men Who Have Sex with Men, in Coastal Kenya. *AIDS and Behavior*, **17**, 1255-1265. https://doi.org/10.1007/s10461-013-0445-9

[10] Madkins, K., Greene, G.J., Hall, E., Jimenez, R., Parsons, J.T., Sullivan, P.S. and Mustanski, B. (2018) Attrition and HIV Risk Behaviors: A Comparison of Young Men Who Have Sex with Men Recruited from Online and Offline Venues for an Online HIV Prevention Program. *Archives of Sexual Behavior*, **47**, 2135-2148. https://doi.org/10.1007/s10508-018-1253-0

[11] Zhang, D., Li, C., Meng, S., Qi, J., Fu, X. and Sun, J. (2014) Attrition of MSM with HIV/Aids along the Continuum of Care from Screening to CD4 Testing in China. *AIDS Care*, **26**, 1118-1121. https://doi.org/10.1080/09540121.2014.902420

[12] Altaweel, M. (2017) Density Mapping with Gis-Gis Lounge. https://www.gislounge.com/density-mapping/

[13] Goldenberg, S.M., Deering, K., Amram, O., Guillemi, S., Nguyen, P., Montaner, J. and Shannon, K. (2017) Community Mapping of Sex Work Criminalization and Violence: Impacts on HIV Treatment Interruptions among Marginalized Women Living with HIV in Vancouver, Canada. *International Journal of STD & AIDS*, **28**, 1001-1009. https://doi.org/10.1177/0956462416685683

[14] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) Smote: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[15] Wanjiru, W.H. (2021) Improved Balanced Random Survival Forest for the Analysis of Right Censored Data: Application in Determining under Five Child Mortality. Ph.D. Thesis, Moi University, Melbourne.

[16] Nekooeimehr, I. and Lai-Yuen, S.K. (2016) Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) for Imbalanced Datasets. *Expert Systems with Applications*, **46**, 405-416. https://doi.org/10.1016/j.eswa.2015.10.031

[17] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[18] Hothorn, T. and Lausen, B. (2003) On the Exact Distribution of Maximally Selected Rank Statistics. *Computational Statistics & Data Analysis*, **43**, 121-137. https://doi.org/10.1016/S0167-9473(02)00225-6

[19] Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S. (2008) Random Survival Forests. *The Annals of Applied Statistics*, **2**, 841-860. https://doi.org/10.1214/08-AOAS169

[20] Ramezankhani, A., Tohidi, M., Azizi, F. and Hadaegh, F. (2017) Application of Survival Tree Analysis for Exploration of Potential Interactions between Predictors of Incident Chronic Kidney Disease: A 15-Year Follow-Up Study. *Journal of Translational Medicine*, **15**, Article No. 240. https://doi.org/10.1186/s12967-017-1346-x

[21] Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N.A., Trollor, J. and Brodaty, H. (2020) A Comparison of Machine Learning Methods for Survival Analysis of High-Dimensional Clinical Data for Dementia Prediction. *Scientific Reports*, **10**, Article No. 20410. https://doi.org/10.1038/s41598-020-77220-w

[22] Mageto, D.K., Mwalili, S.M. and Waititu, A.G. (2015) Modelling of Credit Risk: Random Forests versus Cox Proportional Hazard Regression. *American Journal of Theoretical and Applied Statistics*, **4**, 247-253. https://doi.org/10.11648/j.ajtas.20150404.13

[23] Ptak-Chmielewska, A. and Matuszyk, A. (2020) Application of the Random Survival Forests Method in the Bankruptcy Prediction for Small and Medium Enterprises. *Argumenta Oeconomica*, **44**, 127-142. https://doi.org/10.15611/aoe.2020.1.06

[24] Hamid, O., Tapak, M., Poorolajal, J., Amini, P. and Tapak, L. (2017) Application of Random Survival Forest for Competing Risks in Prediction of Cumulative Incidence Function for Progression to Aids. *Epidemiology, Biostatistics, and Public Health*, **14**, e12663-1.

[25] Rahmayanti, I.A., Sediono, S., Saifudin, T. and Ana, E. (2021) Applying Smote-NC on Cart Algorithm to Handle Imbalanced Data in Customer Churn Prediction: A Case Study of Telecommunications Industry. *Syntax Literate Jurnal Ilmiah Indonesia*, **6**, 1321-1337.

[26] Islahulhaq, W.W. and Ratih, I.D. (2021) Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Over-Sampling Technique-Nominal Continuous (Smote-NC). *International Journal of Advances in Soft Computing and its Applications*, **13**,115-128. https://doi.org/10.15849/IJASCA.211128.09

[27] Mogensen, U.B., Ishwaran, H. and Gerds, T.A. (2012) Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, **50**, 1-23. https://doi.org/10.18637/jss.v050.i11

[28] Zhou, Y. and McArdle, J.J. (2015) Rationale and Applications of Survival Tree and Survival Ensemble Methods. *Psychometrika*, **80**, 811-833. https://doi.org/10.1007/s11336-014-9413-1