THESIS ABSTRACT

Background: The 90–90–90 targets were launched by the Joint United Nations Programme on HIV/AIDS (UNAIDS) and partners with the aim to diagnose 90% of all HIV-positive persons, provide antiretroviral therapy (ART) for 90% of those diagnosed, and achieve viral suppression for 90% of those treated by 2020. In Zimbabwe, a population-based survey in 2016 reported that 74.2% of people living with HIV (PLHIV) aged 15–64 years knew their HIV status. Among the PLHIV who knew their status, 86.8% self-reported current use of Antiretroviral treatment (ART), with 86.5% of those who self-reported being virally suppressed. For these 90–90–90 targets to be met, prevalence and incidence rate estimates are crucial in understanding the current status of the HIV epidemic and determining whether the trends are improving to achieve the 2030 target. Ultimately, this will contribute to the achievement of Sustainable Development Goals 3 (SDG 3) and the broader goal of promoting sustainable development and eradicating poverty worldwide by 2030.

Using data from household surveys, this thesis provides a unique statistical approach for estimating the incidence and prevalence of the Human Immunodeficiency Virus (HIV). To properly assess the efficacy of focused public health interventions and to appropriately forecast the HIV-related burden placed on healthcare systems, a comprehensive assessment of HIV incidence is essential. Targeting certain age groups with a high risk of infection is necessary to increase the effectiveness of public health interventions. To jointly estimate age-and-time-dependent HIV incidence and diagnosis rates, the methodological focus of this thesis was on developing a comprehensive statistical framework for age-dependent HIV incidence estimates. Additionally, the risk of HIV infection was also evaluated using interval censoring methods and machine learning. Finally, geospatial modelling techniques were also utilised to determine the spatial patterns of HIV incidence at district levels and identify hot spots for HIV risk to guide policy.

The main aim of this thesis was to estimate and predict HIV risk using statistical and machine learning methods.

Study objectives: The study objectives of this thesis were:

1. To determine the effect of several drivers/factors of HIV infection on survival time over a decade in Zimbabwe, using current status data.
2. To determine common risk factors of HIV positivity in Zimbabwe and the prediction capability of machine learning models.
3. To estimate HIV incidence using the catalytic and Farrington models and to test the validity of these estimates at the national and sub-national levels.
4. To estimate the age- and time-dependent prevalence and HIV Force-of-infection (FOI) using current status data by comparing parametric, semi-parametric and non-parametric models; and determining which models best fit the data.
5. To investigate the HIV incidence hotspots in Zimbabwe by using geographically-weighted regression.

Methods: We performed secondary data analysis on cross-sectional data collected from the Zimbabwe Demographic Health Survey (ZDHS) from 2005 to 2015. Datasets from three Zimbabwe Demographic Health Survey HIV test results and adult interviews were merged, and records without an HIV test result were excluded from the analysis. The outcome variable was HIV status.

Survey and cluster-adjusted logistic regression were used to determine variables for use in survival analysis with HIV status as the outcome variable. Covariates found significant in the logistic regression were used in survival analysis to determine the factors associated with HIV infection over the ten years. The data for the survival analysis was modelled assuming age at survey imputation (Model 1) and interval-censoring (Model 2). To determine the risk of HIV

infection using machine learning methods, the prediction model was fit by adopting 80% of the data for learning/training and 20% for testing/prediction. Resampling was done using the stratified 5-fold cross-validation procedure repeatedly. The best algorithm was the one with the highest F1 score, which was then used to identify individuals with a higher likelihood of HIV infection. Considering that the proportion of those HIV negative and positive was imbalance with a ratio of 4.2:1, we applied resampling methods to handle the class imbalance. We performed the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes. We evaluated two alternative methods for predicting HIV incidence in Zimbabwe between 2005 and 2015. We estimated HIV incidence from seroprevalence data using the catalytic and Farrington-2-parameter models. These models were validated at the micro and macro levels using community-based cohort incidence and empirical estimates from UNAIDS EPP/SPECTRUM, respectively. To ascertain the age-time effects of HIV risk, we estimated the age- and time-dependent HIV FOI using current status data. Five generalised additive models were explored, ranging from linear, semi-parametric, non-parametric and non-proportional hazards additive models. The Akaike Information Criteria was used to select the best model. The best model was then used to estimate the age- and time-dependent HIV prevalence and force-of-infection. The OLS model was fitted for each survey year to determine the global relationship between HIV incidence and the significant covariates. The Moran's I spatial autocorrelation method was used to assess the spatial independence of residuals. The Getis-Ord Gi* statistic was used for Hotspot Analysis, which identifies statistically significant hot and cold spots using a set of weighted features. Interpolation maps of HIV incidence were created using Empirical Bayesian Kriging to produce smooth surfaces of HIV incidence for visualisation and data generation at the district level. The Multiscale Geographically Weighted Regression method was used to see if the relationship between HIV incidence and covariates

varied by district. The software used in the thesis analysis included R software, STATA, Python, ArcGIS and WinBugs.

Results: Model goodness of fit test based on the Cox-Snell residuals against the cumulative hazard indicated that the model with interval censoring was the best. On the contrary, the Akaike Information Criterion (AIC) indicated that the normal survival model was the best. Factors associated with a high risk of HIV infection were being female, the number of sexual partners, and having had an STI in the past year prior to the survey. The machine learning model indicated that the XGBoost model had better performance compared to the other 5 models for both the original data and SMOTE processed data. Identical variables for both sexes throughout the three survey years for predicting HIV status were: total lifetime number of sex partners, cohabitation duration (grouped), number of household members, age of household head, times away from home in last 12 months, beating justified and religion. The two most influential variable for both males and females were total lifetime number of sex partners and cohabitation duration (grouped).

According to these findings, the catalytic model estimated a higher HIV incidence rate than the Farrington model. Compared to cohort estimates, the estimates were within the observed 95% confidence interval, with 88% and 75% agreement for the catalytic and Farrington models, respectively. The limits of agreement observed in the Bland-Altman plot were narrow for all plots, indicating that our model estimates were comparable to cohort estimates. Compared to UNAIDS estimates, the catalytic model predicted a progressive increase in HIV incidence for males throughout all survey years. Without a doubt, HIV incidence declined with each subsequent survey year for all models. Based on birth year cohort-specific prevalence, the female HIV prevalence peaks at approximately 29 years of age and then declines. Between 15 and 30 years, males have a lower cohort-specific prevalence than females. Male cohort-specific prevalence decreases marginally between ages 33 and 39, then peaks at age 40. In all age

categories, the cohort-specific FOI is greater in females than males. Moreover, the cohort-specific HIV FOI peaked at age 22 for females and age 40 for males. A 18-year age gap between the male and female HIV FOI peaks was observed.

Throughout the decade covered by this study, the Tsholotsho district remained a 99 % confidence hotspot. The impact of STI, condom use and being married on HIV incidence has been strong in the Eastern parts of Zimbabwe with Mashonaland Central, Mashonaland East and Manicaland provinces. From our findings from the Multiscale Geographically Weighted Regression (MGWR), we observed that Matabeleland North's HIV incidence rates are driven by wealth index, multiple sex partners, STI and females with older partners.

Conclusions: The difference between the results from the Cox-Snell residuals graphical method and the model estimates and AIC value may be due to inadequate methods to test the goodness-of-fit of interval-censored data. We concluded that Model 2 with interval-censoring gave better estimates due to its consistency with the published results from the literature. Even though we consider the interval-censoring model as the superior model with regard to our specific data, the method had its own set of limitations. Programmes targeted at HIV testing could use the machine learning approach to identify high-risk individuals. In addition to other risk reduction techniques, machine learning may aid in identifying those who might require Pre-exposure prophylaxis. Based on our results, older men and younger women resembled patterns of higher HIV prevalence and force-of-infection than younger men and older women. This could be an indication of age-disparate sexual relationships. Therefore, HIV prevention programmes should be targeted more at younger females and older males. Lastly, to improve programmatic and policy decisions in the national HIV response, we recommend the triangulation of multiple methods for incidence estimation and interpretation of results. Multiple estimating approaches should be considered to reduce uncertainty in the estimations from various models. The study spread the message that various factors differ from district to

district and over time. The study's findings could be useful to policymakers in terms of resource allocation in the context of public health programs. The findings of this study also highlight the importance of focusing on districts like Tsholotsho, which have consistently had a high HIV burden over time.

The main strength of this study is dependent on the quality of the data obtained from the surveys. These data were derived from population-based surveys, which provide more reliable and robust data. Another strength of this study was that we did not restrict our analysis to one method; however, we had the opportunity to determine the risk and incidence of HIV by exploring different methodologies. However, the limited number of variables accessible to us for this study constituted one of its drawbacks. We could not determine the impact of variables including viral load, health care spending, HIV- risk groups, and other HIV-related interventions. Additionally, there were missing values in the data, which required making assumptions about their unpredictability and utilising imputation methods that are inherently flawed. Last but not least, a number of the variables were self-reported and, as a result, were vulnerable to recall bias and social desirability bias.

***Keywords*: Catalytic model, HIV incidence, Interval Censoring, Machine learning, Multiscale Geographically Weighted Regression, Spatial analysis, Zimbabwe.**