

**ESTIMATING AND PREDICTING HIV RISK
USING STATISTICAL AND MACHINE LEARNING
METHODS: A CASE STUDY USING THE 2005 TO
2015 ZIMBABWE DEMOGRAPHIC HEALTH
SURVEY DATA**



By

Rutendo Beauty Birri Makota

A Thesis Submitted to the Faculty of Health Sciences,

University of the Witwatersrand

in fulfilment of the requirements for the

Degree of DOCTOR OF PHILOSOPHY

by publication in the Major Area-Subject:

BIostatistics and Epidemiology

DECLARATION

I, Rutendo Beauty Birri Makota, declare this research report to be my own, with the exception of specific references to the work of others. This work has not been submitted for any other degree or award besides the Doctor of Philosophy at The University of Witwatersrand. This research represents my independent work, with the exception of instances where explicit citations credit other sources.

Signature of Candidate:  Date: 10 June 2023

University of the Witwatersrand, Faculty of Health Sciences,

School of Public Health

7 York road, Johannesburg

DEDICATION

This PhD dissertation is dedicated to Raphael, my devoted husband, my pillar of strength, and my confidant! Without you, none of this would have been possible. The assistance, the affection, the commitment, the encouragement! You, my love, have made all of this possible! And to my son Imela, the ray of light that extinguished the darkness, my miracle and my hope. I love you. Finally, I dedicate this work to the Almighty God, because of Him, I am!

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 G_i^* 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.

PREFACE

This dissertation is submitted to the University of Witwatersrand, Faculty of Health Sciences, in fulfilment of the requirements for the degree of Doctor of Philosophy by publication in the major field of Biostatistics and Epidemiology. The doctoral research was conducted at the School of Public Health in the Department of Biostatistics and Epidemiology under the supervision of Professor Eustasius Musenge.

ACKNOWLEDGEMENTS

I would like to give a special thank you to my husband. I can never thank you enough, because it was you alone who knew how the journey had so many downs, but you held my hand and lifted me up. I could have never done it without you. You remained my constant cheerleader in this journey.

I am extremely grateful to Professor Eustasius Musenge for his professional guidance and mentorship throughout my doctoral studies. His efforts are not unnoticed, as he has been by my side since the beginning of this journey, providing support during difficult times. Thank you for your perseverance until the end. You worked tirelessly to ensure that this journey feels lighter and less taxing for me, as it should. I am extremely grateful for your efforts.

I would like to thank the Witwatersrand School of Public Health (SPH). Through the Wits School of Public Health, I received assistance from the PhD administration unit and was constantly supported with writing retreats within and outside the school, in addition to other individualized assistance. To my friend Dr. Maritha Kasambala, for persistently nagging me to complete this PhD without fail. Not forgetting my friends Belinda Makuchete and Dr. Zvifadzo Matsena, who are genuine friends who only want the best for me. I am here because you supported me, so I will always value you.

Special thanks to my family, especially my muzukuru Ngonidzashe Moyo, for being there for me. You effortlessly assumed the "mother" role when I was writing my thesis and had no time to prepare meals or make dinner plans. Thank you, Tapiwa and Tinashe Makota, for always cheering me on.

My contributions in the thesis papers

My contributions as the first author in the thesis papers in chapters 3 to 7 were as follows:

1. Chapter 3: Conceptualisation of the research question, structuring of the paper, conducting the statistical analysis and writing all drafts of the paper, responding to reviewer and co-author comments, and proofreading the final accepted manuscript.

Published paper:

Birri Makota R, Musenge E. Factors Associated With HIV Infection in Zimbabwe Over a Decade From 2005 to 2015: An Interval-Censoring Survival Analysis Approach. *Front Public Health*. 2019 Sep 18;7:262. doi: 10.3389/fpubh.2019.00262. PMID: 31620418; PMCID: PMC6759818.

2. Chapter 4: The conceptualisation of the research question, the organisation of the paper, the statistical analysis, the writing of all drafts of the paper, and the response to the comments of the peer reviewers and co-authors.

Published paper:

Birri Makota RB, Musenge E (2023) Predicting HIV infection in the decade (2005–2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. *PLOS Digital Health* 2(6): e0000260. <https://doi.org/10.1371/journal.pdig.0000260>

3. Chapter 5: The conceptualisation of the research question, the organisation of the paper, the statistical analysis, the writing of all drafts of the paper, and the response to the comments of the peer reviewers and co-authors.

Paper submitted for review:

Birri Makota R, Musenge E. Estimating HIV incidence over a decade in Zimbabwe: A comparison of the catalytic and Farrington models. Under review in *Plos Global Public Health*

4. Chapter 6: Conceptualisation of the research question, structuring of the paper, conducting the statistical analysis and writing all drafts of the paper, responding to reviewer and co-author comments, and proofreading the final accepted manuscript.

Published paper:

Birri Makota R, Musenge E. Estimating age specific prevalence and force of infection in Zimbabwe using combined cross-sectional surveys from 2005 to 2015. *Frontiers in Epidemiology*. 2022;2.

5. Chapter 7: The conceptualisation of the research question, the organisation of the paper, the statistical analysis, the writing of all drafts of the paper, and the response to the comments of the peer reviewers and co-authors.

Paper submitted for review:

Birri Makota R, Musenge E. Spatial Heterogeneity in relationship between district patterns of HIV incidence and covariates in Zimbabwe: A Multiscale Geographically Weighted Regression Analysis. Under review in *Geospatial Health*.

TABLE OF CONTENTS

DECLARATION-----	I
DEDICATION-----	II
THESIS ABSTRACT -----	III
PREFACE-----	X
ACKNOWLEDGEMENTS -----	XI
MY CONTRIBUTIONS IN THE THESIS PAPERS -----	XII
TABLE OF CONTENTS-----	XIV
LIST OF FIGURES-----	XXIII
LIST OF TABLES -----	XXVII
TABLE OF EQUATIONS -----	XXVIII
LIST OF ABBREVIATIONS-----	XXX
CHAPTER 1-----	1
1.1 INTRODUCTION -----	1
1.2 LITERATURE REVIEW-----	6
1.2.1 Statistical Methods for Monitoring the HIV/AIDS Epidemic-----	6

1.2.2 The Use of Cross-Sectional Data to Estimate HIV Incidence and Risk-----	9
1.3 RESEARCH PROBLEM AND JUSTIFICATION-----	12
1.3.1 Thesis Broad and Specific Objectives -----	14
1.4 THESIS FRAMEWORK AND THEMES INTER-LINKAGES OF OBJECTIVES-----	15
1.5 THESIS STRUCTURE-----	17
CHAPTER 2-----	19
2.1 STUDY METHODOLOGY -----	19
2.1.1 Study Site -----	19
2.1.2 Data Sources and Description -----	20
2.2 STATISTICAL ANALYSIS OUTLINE -----	22
2.2.1 Statistical Methods for Objective 1-----	22
2.2.1.1 Survival analysis -----	22
2.2.1.2 Model Building and selection -----	23
2.2.2 Statistical Methods for Objective 2-----	24
2.2.2.1 Data Pre-processing -----	24
2.2.2.2 Model training and validation-----	26
2.2.2.3 Model Testing and algorithm selection -----	28
2.2.3 Feature selection -----	28
2.2.4 Statistical Methods for Objective 3-----	30
2.2.4.1 Catalytic and Farrington model-----	30
2.2.4.2 Model Validation -----	31
2.2.5 Statistical Methods for Objective 4-----	32
2.2.5.1 Generalised Additive Models (GAM) -----	32
2.2.6 Statistical Methods For Objective 5-----	33
2.2.6.1 Spatial Analysis-----	33

2.2.6.2 Geographically Weighted Regression (GWR) and Multiscale Geographically Weighted Regression (MGWR)	33
2.2.7 Ethical Considerations	36
CHAPTER 3	37
THESIS PAPER 1: FACTORS ASSOCIATED WITH HIV INFECTION IN ZIMBABWE OVER A DECADE FROM 2005 TO 2015: AN INTERVAL-CENSORING SURVIVAL ANALYSIS APPROACH.	37
ABSTRACT	38
3.1 INTRODUCTION	39
3.2 MATERIALS AND METHODS	41
3.2.1 Study Design and Area	41
3.2.2 Zimbabwe Demographic Health Survey Data and Sampling	41
3.2.3 Measurement of The Outcome (HIV Status) and Explanatory Variables	42
3.3 ETHICAL CONSIDERATIONS	42
3.4 STATISTICAL METHODOLOGY	43
3.4.1 Application To Zimbabwe Demographic Health Survey	43
3.4.1.1 Data Management	43
3.4.2 Statistical Analysis	44
3.5 RESULTS	45
3.6 DISCUSSION	54
3.7 ACKNOWLEDGEMENTS	57

3.8 AUTHOR CONTRIBUTIONS	58
3.9 DATA AVAILABILITY STATEMENT	58
3.10 CONFLICT OF INTEREST	58
3.11 SUPPLEMENTARY INFORMATION	59
3.11.1 Model Specification	59
3.11.1.1 Model 1	59
3.11.1.2 Model 2	60
CHAPTER 4	62
THESIS PAPER 2: PREDICTING HIV INFECTION IN THE DECADE (2005-2015) PRE-COVID-19 IN ZIMBABWE: A SUPERVISED CLASSIFICATION-BASED MACHINE LEARNING APPROACH.	62
ABSTRACT	63
AUTHOR SUMMARY	64
4.1 INTRODUCTION	65
4.2 METHODOLOGY	67
4.2.1 Data and study design	67
4.2.2 Data Pre-Processing	67
4.2.3 Train, Test and Validation Procedure	69
4.2.4 Feature Selection	69
4.2.5 Data Balancing	70
4.2.6 Machine Learning Models	70

4.2.7 Algorithm Evaluation	71
4.2.8 Statistical Analysis	71
4.3 RESULTS	72
4.3.1 Summary statistics	72
4.3.2 Feature Selection	73
4.3.3 Algorithm comparison	74
4.4 Discussion.....	89
4.5 Conclusion.....	92
4.6 Declarations.....	93
4.6.1 Availability of Data and material.....	93
4.6.2 Competing interests.....	93
4.6.3 Authors' Contributions	93
4.7 SUPPLEMENTARY INFORMATION.....	94
4.7.1 Variables used in the study.....	94
4.7.2 ROC and Precision-Recall curves for combined data - Females	95
4.7.3 ROC and Precision-Recall curves for combined data - Males	96
CHAPTER 5.....	97
 THESIS PAPER 3: ESTIMATING HIV INCIDENCE OVER A DECADE IN ZIMBABWE: A COMPARISON OF THE CATALYTIC AND FARRINGTON MODELS.	 97
ABSTRACT.....	98
5.1 BACKGROUND.....	99
5.2 METHODS.....	101

5.2.1 Survey Data -----	101
5.2.2 Estimation Models-----	102
5.2.2.1 Catalytic Model -----	102
5.2.2.2 Farrington Model-----	103
5.2.3 Validation of the Farrington and Catalytic Models. -----	104
5.2.3.1 EPP/Spectrum Model-----	105
5.2.3.2 Community-Based Cohort Study – Manicaland-----	106
5.3 ETHICAL APPROVAL AND CONSENT TO PARTICIPATE -----	106
5.4 RESULTS -----	107
5.4.1 HIV Incidence for the Two Models -----	108
5.4.2 Models’ Validation Using a Cohort Study-----	110
5.4.3 Models Validation Using UNAIDS EPP/SPECTRUM Estimates-----	113
5.5 DISCUSSION -----	114
5.6 CONCLUSION-----	116
5.7 DECLARATIONS-----	117
5.7.1 Availability of Data and Material-----	117
5.7.2 Competing Interests -----	117
5.8 AUTHORS’ CONTRIBUTIONS -----	117
CHAPTER 6-----	118
THESIS PAPER 4: ESTIMATING AGE SPECIFIC PREVALENCE AND FORCE OF INFECTION IN ZIMBABWE USING COMBINED CROSS-SECTIONAL SURVEYS FROM 2005-2015.-----	118
ABSTRACT-----	119

6.1 INTRODUCTION	120
6.2 METHODS	122
6.2.1 Zimbabwe Survey Cross-Sectional Data	122
6.2.2 Outcome Variable Measurement: HIV Testing Procedure	122
6.2.3 Statistical Model	123
6.2.4 Statistical Analysis	127
6.2.5 Model 1-5 Specification and Goodness-of-Fit	128
6.3 RESULTS	129
6.4 DISCUSSION	134
6.5 AUTHOR CONTRIBUTIONS	136
6.6 CONFLICT OF INTEREST	136
6.7 ACKNOWLEDGMENTS	136
CHAPTER 7	138
THESIS PAPER 5: SPATIAL HETEROGENEITY IN RELATIONSHIP BETWEEN DISTRICT PATTERNS OF HIV INCIDENCE AND COVARIATES IN ZIMBABWE: A MULTISCALE GEOGRAPHICALLY WEIGHTED REGRESSION ANALYSIS.	138
ABSTRACT	139
7.1 INTRODUCTION	141
7.2 MATERIALS AND METHODS	142
7.2.1 Data Source	142
7.2.2 Study Measures	143

7.2.3 Exploratory Data Analysis-----	144
7.2.3.1 Hot Spot Cluster and Outlier Analysis -----	144
7.2.3.2 Exploratory Regression -----	145
7.2.4 Spatial Statistical Models -----	145
7.2.4.1 OLS regression – Global Regression-----	145
7.2.4.2 Geographically Weighted Regression (GWR) model-----	147
7.2.4.3 Multiscale Geographically Weighted Regression (MGWR)-----	148
7.3 RESULTS -----	148
7.3.1 Hot spot analyses – Spatial clustering of HIV incidence -----	149
7.3.2 Ordinary Least Squares Regression Results -----	152
7.3.3 Geographical/Spatial Variation of HIV incidence Results -----	153
7.4 Discussion-----	157
7.5 Study Limitations. -----	161
7.6 Conclusion-----	161
7.7 Supplementary Information-----	163
CHAPTER 8-----	169
8.1 DISCUSSION AND CONCLUSIONS-----	169
8.2 HIV RISK ESTIMATION: DETERMINATION OF HIV POSITIVITY -----	170
8.3 HIV INCIDENCE MODELLING -----	172
8.4 SPATIAL REGRESSION AND HIV HOT SPOT ANALYSIS-----	174
8.5 STRENGTHS AND LIMITATIONS OF THE STUDY-----	175
8.5.1 Variable Recall and Selection Bias -----	175

8.5.2 DHS Data, Representativeness, and Generalisability -----	177
8.5.3 Missing Data Handling-----	177
8.5.4 Analysis and Computing-----	178
8.6 CONCLUSION-----	179
8.7 IMPLICATIONS FOR INTERVENTIONS AND POLICY -----	180
8.8 SCOPE FOR FUTURE RESEARCH -----	181
REFERENCES-----	183
APPENDIX 1: PAPERS UNDER REVIEW-REVIEWER COMMENTS -----	215
APPENDIX 2: PAPERS ANALYSIS SCRIPTS/CODES-----	222
APPENDIX 3: ETHICS CERTIFICATE-----	247
APPENDIX 4: PUBLISHED PAPERS-----	248

LIST OF FIGURES

Figure 1.1: Forty years of the HIV/AIDS response ¹⁶	3
Figure 1.2: UNAIDS 90-90-90 Fast track targets ²	13
Figure 1.3: The schematic presentation of the thesis conceptual framework linking to the thesis objectives.	16
Figure 1.4: Themes and interlinkages of objectives	17
Figure 2.1: Map of Zimbabwe ¹⁰³	19
Figure 2.2: DHS sampling strategy.	21
Figure 2.3: Examining parametric baseline fits for six parametric models (right) and best model (left).	24
Figure 2.4: Dataset Split diagram	27
Figure 3.1: Illustration of different modelling strategies when investigating the factors associated with HIV infection.	45
Figure 3.2: Flowchart of records in the 2005-06, 2010-11 and 2015 ZDHS.	46
Figure 3.3: Goodness-of-fit tests for the Zimbabwe DHS 2005-06 (left column), 2010-11 (middle column) and 2015 (right column). Model 1 (top row) and Model 2 (bottom row)	48
Figure 3.4: Cox-Snell residual against cumulative hazard goodness-of-fit tests for the Zimbabwe DHS 2005-06 (left column), 2010-11 (middle column) and 2015 (right column). Model 1 (top row) and Model 2 (bottom row)	50

Figure 3.5: Dot chart showing relative importance of covariates for the Zimbabwe DHS 2005-06 (left column), 2010-11 (middle column) and 2015 (right column). Model 1 (top row) and Model 2 (bottom row).	51
Figure 4.1: Flowchart of the methods and steps used.....	68
Figure 4.2: ROC curves and Precision-Recall curves of the six algorithms for all survey years for females' data.....	78
Figure 4.3: ROC curves and Precision-Recall curves of the six algorithms for all survey years for males' data.....	79
Figure 4.4: Sequential floating forward selection (SFFS) for males and females...	80
Figure 4.5: Shapley values for 2005-06, 2010-11 and 2015 Females ZDHS data.....	83
Figure 4.6: Shapley values for 2005-06, 2010-11 and 2015 Males ZDHS data.	84
Figure 5.1: Prevalence trends by age group and survey year for females (left) and males (right).	107
Figure 5.2: HIV incidence for males and females based on the catalytic and Farrington models.....	109
Figure 5.3: HIV incidence estimates based on the cohort, catalytic and Farrington models.	110
Figure 5.4: Bland-Altman plots demonstrating agreement between estimated incidence (Farrington and catalytic models) and observed incidence.	112
Figure 5.5: HIV incidence for males and females based on the catalytic, Farrington and EPP/SPECTRUM estimates.	113

Figure 6.1: Observed (dotted curve) and fitted prevalence (five solid curves) as a function of age according to females (top) and males (bottom).	130
Figure 6.2: Observed (circles) and fitted (solid curve with 95% Confidence Interval, dotted curve) HIV prevalence as a function of age according to females and males. Circles represent the estimated prevalence by age. Their size is proportional to the number of individuals.	131
Figure 6.3: Fitted cohort-specific prevalence curves as a function of age according to females and males.....	132
Figure 6.4: Fitted cohort-specific force-of-infection curves as a function of age according to females (left panel) and males (right panel).	133
Figure 7.1: Map of Zimbabwe showing the 60 districts.....	143
Figure 7.2: Spatial Variation of HIV incidence per survey year	149
Figure 7.3: HIV incidence Hot spots (a-c), Clusters (d-f) and Interpolation (g-i) per survey year.	151
Figure 7.4: Coefficient estimates of (a) STI in the past 12 months; (b) No condom use; and (c) Married in describing HIV incidence using Multiscale Geographically Weighted Regression (MGWR) for the ZDHS 2005-06 survey year with ± 1.96 significant areas indicated in colour.	155
Figure 7.5: Coefficient estimates of (a) Richer; (b) Single; and (c) Older partner > 5 years in describing HIV incidence using Multiscale Geographically Weighted Regression (MGWR) for the ZDHS 2010-11 survey year with ± 1.96 significant areas indicated in colour.	156

Figure 7.6: Coefficient estimates of (a) Multiple sex partners; (b) STI in the past 12 months; (c) Richer and (d) Older partner > 5 years in describing HIV incidence using Multiscale Geographically Weighted Regression (MGWR) for the ZDHS 2015 survey year with ± 1.96 significant areas indicated in colour. 157

LIST OF TABLES

Table 3.1: HIV prevalence in Zimbabwe and changes in HIV prevalence (weighted) from the Zimbabwe DHS surveys 2005-06, 2010-11 and 2015.....	47
Table 3.2: Akaike Information Criterion and Bayesian Information Criterion values...	49
Table 3.3: Estimated effects of covariates at baseline on the risk of HIV infection based on different survival models, Zimbabwe Demographic Health Survey (ZDHS) 2005/06, 2010/11 and 2015.	53
Table 4.1: Background characteristics of ZDHS data	73
Table 4.2: Description of original data and SMOTE-processed data.....	75
Table 4.3: Results of classification models in original imbalanced and SMOTE-processed data.....	76
Table 4.4: Odds Ratio of female features selected in XGBoost using SFFS procedure.	85
Table 4.5: Odds Ratio of male features selected in XGBoost using SFFS procedure. ..	88
Table 5.1: Overall HIV incidence for the two models by year of survey.....	108
Table 5.2: Measurements and estimates of incidence using Manicaland Province Data	111
Table 6.1: The Akaike information criterion (AIC), Schwarz's Bayesian criterion (BIC), and effective degrees of freedom (edf) for completely parametric (model 1), semi-parametric (models 2 and 3), and non-parametric (models 4 and 5) models	129
Table 7.1: Ordinary Least Squares Regression (OLS) results.....	153
Table 7.2: Regression Results Comparison.....	154

TABLE OF EQUATIONS

Equation 2.1	22
Equation 2.2	22
Equation 2.3	29
Equation 2.4.....	34
Equation 2.5	34
Equation 2.6	35
Equation 2.7	35
Equation 2.8	35
Equation 5.1.....	103
Equation 5.2	103
Equation 5.3	104
Equation 5.4.....	105
Equation 5.5	105
Equation 6.1	123
Equation 6.2.....	123
Equation 6.3.....	123
Equation 6.4.....	124
Equation 6.5.....	124
Equation 6.6.....	124
Equation 6.7.....	125

Equation 6.8.....	125
Equation 6.9.....	125
Equation 6.10	125
Equation 6.11.....	126
Equation 6.12.....	126
Equation 6.13.....	126
Equation 6.14.....	126
Equation 6.15.....	126
Equation 6.16.....	128
Equation 6.17.....	128
Equation 6.18.....	128
Equation 6.19.....	128
Equation 6.20	128
Equation 7.1.....	144
Equation 7.2.....	146
Equation 7.3.....	147
Equation 7.4.....	148

LIST OF ABBREVIATIONS

AIC - Akaike Information Criterion
AIDS – Acquired Immune Deficiency Syndrome
AIM - AIDS Impact Model
ANC - Antenatal Care
ART - Antiretroviral Therapy
ARV – Antiretroviral
AUC - Area Under Curve
AZT – Zidovudine
CD4 - Cluster Of Differentiation 4
CDC - Centers for Diseases Control
CI – Confidence Interval
COVID - Coronavirus disease
DHS - Demographic and Health Surveys
DRC - Democratic Republic of the Congo
DT - Decision Tree
EA - Enumeration Areas
EBK - Empirical Bayesian Kriging
ELISA - Enzyme-Linked Immunosorbent Assay
EPP - Estimation and Projection Package
ER - Exploratory Analysis
FOI – Force-of-infection
GAM - Generalised Additive Model
GBV - Gender Based Violence
GOF - Goodness-of-Fit
GPS - Global Positioning System
GWR - Geographically Weighted Regression
HAART - Highly Active Antiretroviral Therapy
HIV - Human Immunodeficiency Virus
IPV - Intimate Partner Violence
KNN - k-nearest Neighbour Algorithm
LASSO - Least Absolute Shrinkage and Selection Operator
LightGBM - Light Gradient Boosting Model
MGWR - Multiscale Geographically Weighted Regression
MICE - Multiple Imputations with Chained Equations
MRCZ - Medical Research Council of Zimbabwe
NBSZ - National Blood Service of Zimbabwe
OLS - Ordinary Least Squares
OR - Odds Ratio
PAVA - Pool-Adjacent-Violators Algorithm
PLHIV - People living with HIV
PPV - Positive Predictive Value

PrEP - Pre-exposure prophylaxis
RFC - Random Forest Classifier
ROC - Receiver Operating Characteristics
SFFS - Sequential Forward Floating Selection
SIV - Simian Immunodeficiency Virus
STI – Sexually Transmitted Infection
UMP - Uzumba Maramba Pfungwe
UNAIDS - Joint United Nations Programme on HIV/AIDS
USA - United States of America
VIF - Variance Inflation Factor
WHO - World Health Organisation
ZDHS - Zimbabwe Demographic Health Survey
ZIMPHIA - Zimbabwe Population-based HIV Impact Assessment

CHAPTER 1

1.1 INTRODUCTION

One of the most catastrophic global epidemics of the 20th and 21st centuries is the Acquired Immune Deficiency Syndrome (AIDS) epidemic, which is brought on by Human Immunodeficiency Virus (HIV) viral infection. Since the beginning of the epidemic, 84.2 million [64.0-113.0 million] people have acquired HIV, and approximately 40.1 million [33.6-48.6 million] people have died from the disease ¹. At the end of 2021, there were 38.4 million [33.9-43.8 million] HIV-positive individuals worldwide ¹. According to estimates, 0.7% [0.6-0.8%] of individuals in the globe between the ages of 15 and 49 have HIV; however, the epidemic's toll continues to differ significantly between different nations and areas. The World Health Organisation (WHO) African Region continues to be the most badly affected, with approximately 1 in every 25 individuals (3.4%) now carrying HIV, making up more than two-thirds of all HIV-positive persons globally ².

The HIV virus was found to be a mutation of the Simian Immunodeficiency Virus (SIV), which was found in apes ³. However, the mechanisms by which cross-species transmission happened are unknown. The “hunter idea” holds that the exchange of fluids during ape hunting (through a bite or cut) led to SIV transfer to humans ^{4,5}. Certain fluids, including blood, sperm, pre-embryonic fluid, rectal and vaginal secretions, and breast milk, can transmit HIV. Thus, HIV is generally transmitted through blood transfusion, needle and syringe sharing, sexual intercourse and infected mothers to their babies.

The seroconversion of HIV-infected persons typically occurs within three weeks of infection; however, it can take up to three months ^{6, 7}. Seroconversion lasts only a few weeks and is characterised by a robust immune system response to the virus ⁸, including the development of antibodies and a sharp decrease in cluster of differentiation 4 (CD4) counts (defined as the number of CD4 T-helper lymphocytes in a fixed volume of blood), with possible flu-like symptoms (e.g., high fever, rashes, enlarged lymph nodes). HIV testing generally involves identifying antibodies generated during seroconversion; therefore, the results of a test administered between infection and seroconversion may be inaccurate ⁹. After seroconversion, the CD4 count recovers to an almost normal level; however, this is followed by a steady and prolonged decrease in CD4 count and an increase in viral load ¹⁰. This results in immunodeficiency (the loss of the immune system), AIDS-related disorders (such as Kaposi's sarcoma), and death without therapy ¹¹.

The incubation period is defined as the interval between infection and the onset of AIDS symptoms and, without therapy, lasts from eight to ten years ^{3, 12}. Window period describes the time between infection and seroconversion. The length of the window period can vary depending on the type of test used and individual factors such as the person's immune system and the amount of virus in their body. However, the window period is usually within 3 months. During this period, HIV-infected individuals are highly contagious due to the absence of HIV-specific antibodies; consequently, this period is also known as the period of acute HIV infection.

HIV was first discovered in an adult female from the Democratic Republic of the Congo in 1960. Although phylogenetic research indicates that the HIV virus has been prevalent in central Africa since the early 1900s, the pandemic did not really begin to take off until

the 1960s with the expansion of urbanisation ¹³. Only in 1983, with the identification of HIV on its own in AIDS patients in France and the US, was the connection between Human Immunodeficiency Virus (HIV) infection and the incidence of AIDS discovered ^{14, 15}. The first HIV detection test was authorised in the US and Europe in 1985. Refer to Figure 1.1 for 40 years of the HIV/AIDS response.

40 YEARS OF THE AIDS RESPONSE

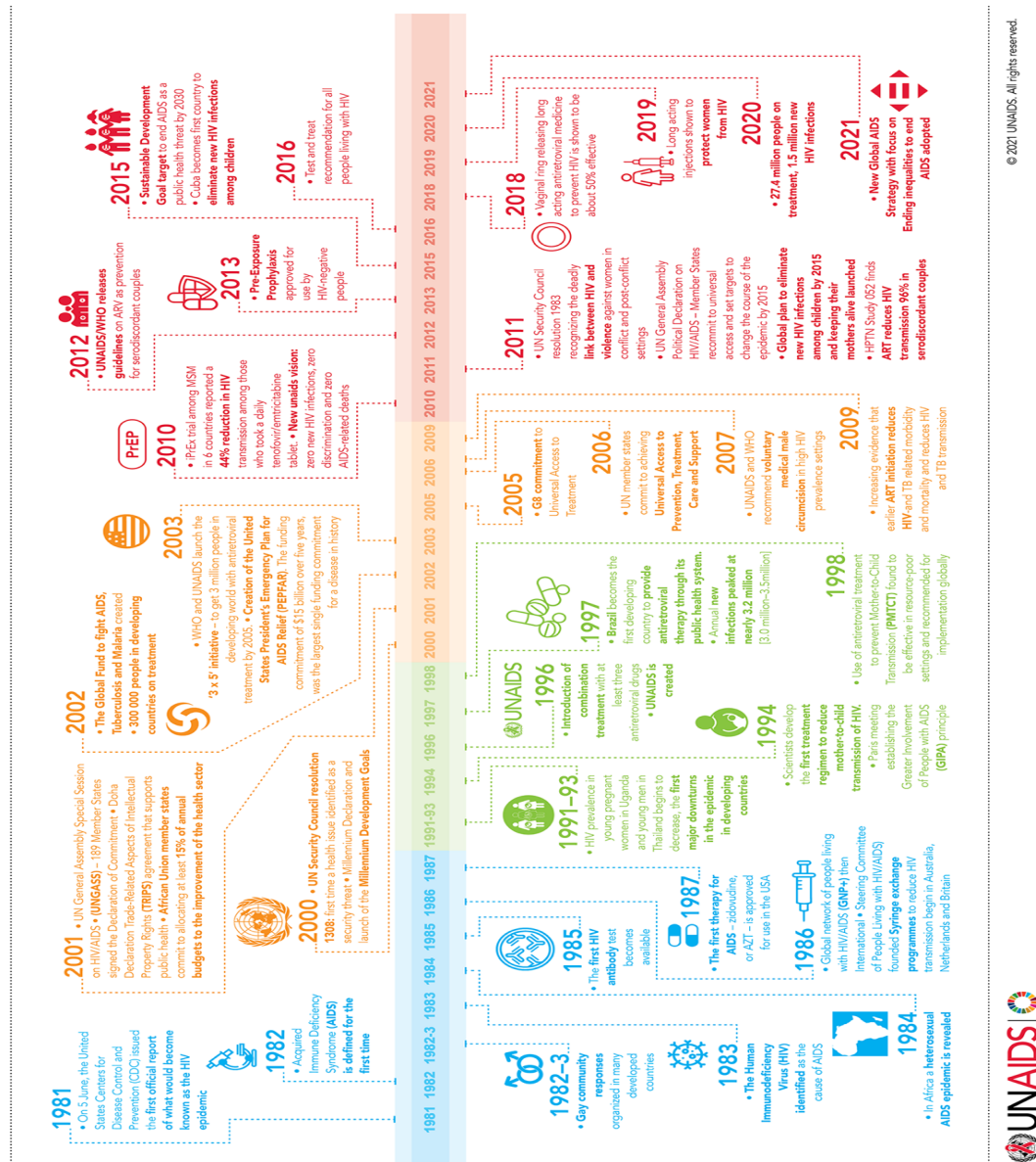


Figure 1.1: Forty years of the HIV/AIDS response ¹⁶.

Zimbabwe reported its first case of AIDS in 1985¹⁷. Since then, many people developed symptoms that were suggestive of HIV infection^{18, 19}. Unfortunately, because of no knowledge of the HIV infection, individuals who presented with suggestive HIV infection symptoms did not react to routine dietary and conventional medical therapies, which suggested an immunodeficiency issue. The first case of HIV infection was formally identified in 1986 in Hurungwe, where a hospital-based surveillance system was implemented to track the progress of the epidemic²⁰. By then, the HIV syndrome was yet to be incorporated into the formal notification system. As a result, the number of cases of AIDS in this region increased dramatically from 19 in 1986 to 290 in 1987, 433 in 1988, and in the first quarter of 1989²⁰. Interestingly, all infections were discovered in the 17-30 age range²¹ as early as 1987, when the prevalence soared to 3.2%.

In response to these cases, the National Blood Service of Zimbabwe (NBSZ) implemented systematic HIV testing of donated blood and blood products¹⁹. As a result, the HIV prevalence among blood donors was estimated to be about 2%. However, ten years later, the HIV prevalence among blood donors increased to 8.8%, with an estimated HIV incidence of 2.1 per 100 person-years, with married individuals aged between 21 to 45 years having the highest prevalence^{22, 23}.

In 1990, a hospital-based cross-sectional survey in Murehwa district hospital, which screened for sexually transmitted infections (STIs) among adults, revealed that 50% of individuals with an STI were HIV positive²⁴. Given the rising statistics, this sparked great concern, evoking questions like: where did the virus originate, and why were there so many cases in a short period? Until 1990, when HIV/AIDS concerns entered the public discourse, there was a great deal of denial of HIV/AIDS by the government of Zimbabwe¹⁷. However, in the same year, routine sentinel HIV surveillance of pregnant women

attending antenatal care clinics (ANC) was initiated, and an HIV prevalence which surpassed 10% was observed ¹⁷.

Despite the vast amount of time and money invested in HIV-related research, no cure for the virus has yet been discovered. In 1987, the first antiretroviral medicine, zidovudine (commonly known as AZT), was introduced in an effort to combat the HIV virus and restrict its deadly effect on the immune system. Unfortunately, despite AZT's therapeutic efficacy, it was linked with several severe adverse effects, with the HIV virus developing resistance after extended medication ²⁵.

An effective HIV therapy was first developed in 1995 ²⁶. This Highly Active Antiretroviral Therapy (HAART), which consists of a fixed-dose combination of multiple antiretroviral drugs, has radically altered the HIV epidemic: the number of AIDS-related deaths has plummeted since the mid-1990s ²⁷, and HIV is now viewed as a chronic, rather than a fatal condition. Although the widespread use of HAART has lowered HIV infection, the epidemic is still yet to be eradicated. A revolutionary Pre-exposure Prophylaxis (PreP) medication was licensed in the United States in 2014: a new antiretroviral medicine (Truvada) for patients who are at high risk of infection but are not HIV positive ²⁸. Clinical study data indicated that PreP is a highly effective method for reducing the risk of infection ²⁹⁻³¹.

Given the history of HIV/AIDS epidemic in Zimbabwe over the years, HIV incidence and risk estimation in Zimbabwe is a critical component of efforts to address the HIV/AIDS epidemic and achieve the Sustainable Development Goals (SDGs), particularly SDG 3, which aims to ensure good health and well-being for all. Accurate estimation of HIV incidence and risk can inform the development of effective prevention,

testing, and treatment strategies, which are key to achieving the UNAIDS 90-90-90 targets and the goals outlined in the Zimbabwe National HIV and AIDS Strategic Plan (2021-2025). By improving our understanding of HIV incidence and risk in Zimbabwe, we can ensure that resources are directed towards the most effective interventions and that progress towards the UNAIDS 90-90-90 targets, and the goals of the Zimbabwe National HIV and AIDS Strategic Plan is maximized. Ultimately, this will contribute to the achievement of SDG 3 and the broader goal of promoting sustainable development and eradicating poverty worldwide by 2030.

1.2 LITERATURE REVIEW

1.2.1 Statistical Methods for Monitoring the HIV/AIDS Epidemic

It is crucial to effectively and reliably measure the dynamics of HIV epidemics to strategically analyse the efficacy of interventions and direct decision-makers toward better HIV/AIDS control. HIV prevalence and incidence are two primary indicators used to analyse and interpret the dynamics of HIV transmission. Using mathematical and statistical modelling techniques ³²⁻³⁴, several academic and government research organisations have estimated the incidence and prevalence of HIV.

Beginning in the 1980s, several modelling studies sought to estimate the incidence of HIV using a range of applicable methodologies. These diverse methodologies have played a crucial role in estimating the incidence of HIV globally. The classical back-calculation method, the CD4+ T cell depletion method, HIV case notification data, cohort study data, serial or cross-sectional prevalence data, and the biomarker approach have been used to estimate HIV incidence over the years.

Back-calculation, one of the most used statistical modelling techniques, is exploited to study the distribution of AIDS incubation times. Using epidemiological monitoring data, the back-calculation approach reconstructs HIV infections through time. The back-calculation approach was initially developed by Brookmeyer et al.³⁵⁻³⁷, who used AIDS incidence data and the maximum likelihood estimation method to estimate discrete HIV incidence. In addition to the incidence of AIDS, the frequency of HIV infections discovered is now available as part of epidemiological surveillance data, which has dramatically aided researchers in expanding the back-calculation method.³⁸⁻⁴⁸ Early investigations relied solely on HIV diagnoses from patients who ultimately developed AIDS³⁸⁻⁴³. Several more methodologies were proposed to encompass all HIV diagnoses, including those of infected people who had not yet developed AIDS.^{38, 44-46}

The second primary HIV incidence modelling strategy involves the use of CD4+ T cell count data. CD4+ T cell count data are now generally accessible when HIV is diagnosed, and Markov process models^{49, 50} have used the CD4+ T cell to estimate HIV incidence. In addition, a structured approach to employ CD4+ T cell counts for diagnosis of HIV infection at the CD4 stage was developed by Birell et al.⁵¹. To estimate HIV incidence, the CD4+ T cell depletion model, which is a combination of the back-calculation method and the use of CD4+ T cell count data, was developed. This methodology first calculates the period between infection and diagnosis (i.e., diagnostic delay) and then estimates the HIV incidence based on the CD4+ T cell depletion⁵²⁻⁵⁴.

Thirdly, in 2017, Xia et al.⁵⁵ developed a straightforward approach using HIV case notification data to estimate HIV incidence. The strategy presupposes that HIV incidence and case testing remain steady every three years. This procedure is straightforward for non-specialists. In addition, only HIV case notification data are required for

implementation. However, this strategy is only practical if incidence and diagnostic rates remain steady over a three-year period.

The fourth method commonly used for estimating HIV incidence is the use of cohort studies where uninfected individuals are followed over time ⁵⁶. Since cohort studies are expensive and time-consuming, it is impossible to track individuals at the national level; hence incidence is estimated among subpopulations ⁵⁷. Although this methodology allows for incidence to be estimated directly, two main sources of error are introduced ⁷. First, individuals who adhere to follow-up visits may not be an accurate representation of the population. Secondly, those individuals who adhere to follow-up visits get counselling repeatedly. As a result, their knowledge of HIV improves over time, affecting their risk of contracting HIV.

The fifth method uses serial or cross-sectional prevalence data to estimate HIV incidence. The Estimation and Projection software Package (EPP) was developed by UNAIDS to obtain HIV prevalence and projections ⁵⁸. Linked to the EPP is another software called SPECTRUM ⁵⁹, which estimates the HIV incidence using the AIDS Impact Model (AIM) model. The software has been updated several times since its initial release in 2004 ⁶⁰⁻⁶³.

Lastly, using the biomarker approach, the “snapshot sampling” is an additional method for determining incidence from a single prevalence survey developed by Quinn and Brookmeyer ⁶⁴. The methodology involves randomly selecting individuals for HIV testing. The HIV incidence is then estimated using the prevalence of the p24 antigen among HIV-negative individuals and the average length of the p24 positive phase. The p24 antigen is only detectable for a limited time before seroconversion and represents recent infections. The biomarker method has been expanded to various indicators of

recent infection across time ⁶⁵⁻⁶⁷, and much effort has been given to estimating their mean window duration ⁶⁸⁻⁷⁰.

1.2.2 The Use of Cross-Sectional Data to Estimate HIV Incidence and Risk

Mathematical modelling of infectious diseases was first documented in 1760 in a publication by Daniel Bernoulli ⁷¹. The concept of estimating the force of infection (which is also known as incidence) as a key parameter in mathematical models was proposed by Hugo Muench, which was documented in his book *Catalytic Models in Epidemiology* ⁷². To estimate the rate a susceptible person becomes infected, Muench proposed using a catalytic model with "current status data." Current status data occurs when an individual is observed at one single point, and the only information obtained is whether the event of interest has occurred ⁷³. Current status data is obtained through a cross-sectional survey. Data on a person's current HIV status, such as that gathered through a demographic health survey, is an example of current status data. For infections that leave permanent markers of immunity in their surviving hosts, the ability to use test results on serological samples from current status data to estimate the force of infection is crucial, which is why Muench's catalytic model is so important.

Many models to estimate the force of infection using current status data have been developed as a result of Muench's work. Muench's model assumed a constant force of infection over time. Griffiths ⁷⁴ then proposed using linear rather than constant FOI. Both Muench and Griffith's methods used these models to model measles incidence. An extension of Griffith's approach encompassing a polynomial FOI with age was proposed by Grenfell and Anderson ⁷⁵. However, these modelling approaches suffered the problem

of monotonicity, which was later addressed by Farrington, who introduced a way to estimate the force of infection using a non-linear model, i.e., an exponentially damped model.

Given the history of modelling serological data obtained from current status data, we observed that most of these modelling strategies mentioned above were used on other infectious diseases other than HIV. Therefore, this thesis aimed to adopt some of these models to estimate the force of infection in Zimbabwe using current status data, in our case, the Zimbabwe Demographic Health Survey data. Besides estimating the incidence or force of infection, we additionally estimated the risk of HIV infection in Zimbabwe. The HIV infection risk was estimated using traditional and machine learning models.

The gold standard for estimating HIV incidence is the periodic testing of previously uninfected people for new infections, as discussed in section 1.2.1, but this procedure is costly and time consuming. In addition, even when an HIV-negative group is tracked throughout time, the exact date of infection is seldom documented ⁷⁶. An interval between the dates of the most recent negative and earliest positive test can be calculated under these conditions. Recently, several strategies have emerged for using interval censoring to infer the risk of HIV infection. Hsu et al. ⁷⁷ proposed a Cox proportional hazard model with time-dependent covariates to identify the nearest neighbours of infection. Chen and Sun ⁷⁸ estimated the regression parameters for interval-censored data. These models have not yet been extended to estimate HIV risk using interval-censored data. Therefore, in this thesis, one of our objectives was to use an interval censoring methodology to estimate the risk of HIV infection using cross-sectional data.

Many studies in literature have used Cox proportional hazards or logistic regression models to establish the HIV risk. However, problems pertaining to the data structure, which are typically non-linear and heterogeneous, resulting in poor performance of these methods, gave rise to machine learning approaches. Compared to the traditional models mentioned above for estimating HIV risk, machine learning algorithms have provided new methods to construct models that perform better since the algorithms can balance the deviation and variance of data. In this thesis we explored the use of machine learning methods to estimate the HIV risk in Zimbabwe.

Machine learning has emerged as a promising tool for predicting HIV risk, by leveraging large datasets of demographic, behavioural, and clinical data to develop accurate risk estimation models. Machine learning algorithms can use a wide range of features, including demographic data such as age, gender, and ethnicity; behavioural data such as sexual behaviours and drug use; and clinical data such as viral loads and CD4 counts, to develop models that can predict an individual's likelihood of HIV acquisition. In some cases, machine learning models can also be used to identify factors that are associated with increased or decreased HIV risk, providing insights into potential risk reduction interventions. There are various machine learning techniques that have been applied to HIV risk estimation, including decision trees, neural networks, logistic regression, and support vector machines. These techniques have been employed to develop risk estimation models in different settings, including community-based HIV prevention programs, HIV testing and counselling services, and HIV treatment centres.

Despite the promising potential of machine learning in HIV risk estimation, challenges remain, including the need for high-quality and comprehensive data, concerns about bias and privacy, and the difficulty of generalizing risk estimation models across different populations and settings. Further research is needed to address these challenges and

develop robust machine learning models that can accurately predict HIV risk in diverse populations. Overall, machine learning has the potential to revolutionize HIV prevention and control efforts by providing healthcare workers with more accurate tools for identifying individuals at risk of HIV acquisition and providing targeted interventions to prevent HIV transmission. In addition, machine learning algorithms mainly focus on making predictions as accurate as possible, while traditional statistical models aim to infer relationships between variables. To date, machine learning approaches have been widely used in the medical field; however, they have not been commonly used in predicting HIV risk, especially in Zimbabwe.

1.3 RESEARCH PROBLEM AND JUSTIFICATION

The Joint United Nations Programme on HIV/AIDS (UNAIDS) and its partners launched the 90–90–90 targets (Figure 1.2) with the goal of diagnosing 90% of all HIV-positive individuals, providing antiretroviral therapy (ART) to 90% of those diagnosed, and achieving viral suppression in 90% of those treated by 2020 ⁷⁹. A 2016 population-based study in Zimbabwe ⁸⁰ revealed that 74.2% of HIV-positive individuals aged 15–64 were aware of their HIV status. In addition, 86.8 percent of HIV-positive individuals who knew their status reported current use of antiretroviral therapy (ART) and 86.5 percent reported viral suppression. For these 90–90–90 goals to be attained, prevalence and incidence rate data are necessary for understanding the current state of the HIV pandemic and determining whether the trend is improving to meet the 2030 target.

Given that we were working with data from 2005-2015, instead of using the 95-95-95 goals which are at present being implemented, the 90-90-90 goals aligned with the data being analysed in this thesis.

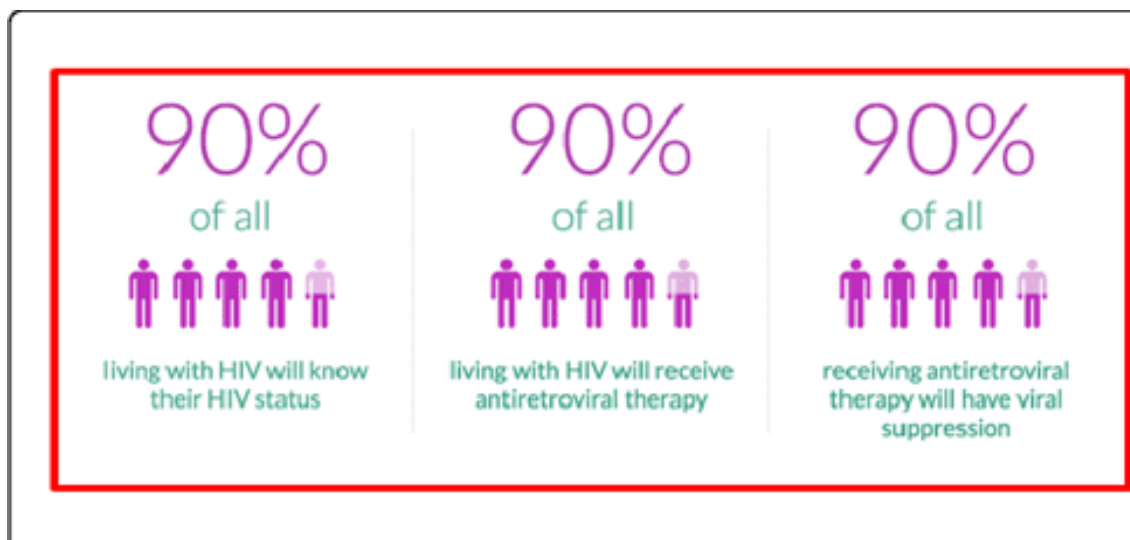


Figure 1.2: UNAIDS 90-90-90 Fast track targets ²

A comprehensive assessment of HIV incidence is required to properly assess the efficacy of targeted public health interventions and forecast the HIV-related burden placed on healthcare systems. Many studies, particularly in Zimbabwe, have estimated HIV incidence ⁸⁰⁻¹⁰², however, none of these studies have used the demographic health survey to estimate HIV incidence. Furthermore, these studies require complex longitudinal designs, are costly, time-consuming, intermittent missingness or attrition, and most of the results are specific and not general to the entire population. Using data from household surveys, this thesis provides unique statistical approaches to address this gap in estimating the incidence and risk of the Human Immunodeficiency Virus (HIV).

Targeting certain age groups with a high risk of infection is necessary to increase the effectiveness of public health interventions. The methodological focus of this thesis was on developing a comprehensive statistical framework for age-dependent HIV incidence

estimates in order to jointly estimate age-and-time-dependent HIV incidence. 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 province levels and identify hot spots for HIV risk to guide policy.

1.3.1 Thesis Broad and Specific Objectives

The main aim/objective of this thesis was to estimate and predict the HIV risk using statistical and machine learning methods using the 2005 to 2015 Zimbabwe Demographic Health Survey (ZDHS) data. To estimate the HIV incidence and risk, different modelling strategies were used. The specific objectives were:

1. To determine the effect of several drivers/factors of HIV infection on survival time to HIV infection over a decade in Zimbabwe, using current status data.

Paper 1 title: Factors associated with HIV infection in Zimbabwe over a decade from 2005 to 2015. An interval-censoring survival analysis approach.

2. To determine risk factors of HIV positivity in Zimbabwe and the prediction capability of machine learning models.

Paper 2 title: Predicting HIV infection in the decade (2005- 2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach.

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.

Paper 3 title: Estimating HIV incidence over a decade in Zimbabwe: A comparison of the catalytic and Farrington models.

4. To estimate the age- and time-dependent prevalence and HIV FOI using current status data by comparing parametric, semi-parametric and non-parametric models; and determining which models best fit the data.

Paper 4 title: Estimating Age-Time Dependent Prevalence and Force of Infection from Serial Prevalence Data from 2005-2015: A case study of Zimbabwe.

5. To investigate the spatial variation of HIV incidence in Zimbabwe by using Multiscale Geographically-Weighted Regression (MGWR).

Paper 5 title: Sociodemographic determinants of HIV incidence in Zimbabwe: Geospatial modelling using Multiscale Geographically-Weighted Regression (MGWR).

1.4 THESIS FRAMEWORK AND THEMES INTER-LINKAGES OF OBJECTIVES

As a way of interlinking objectives, we formulated a conceptual framework (Figure 1.3) for this thesis. The “green” arrow represents objectives 1 and 5, where underlying and proximate determinants of HIV infection were investigated using survival and spatial analysis. The “blue” arrow represents objective 2, which used machine learning algorithms to determine factors associated with HIV positivity. The probability of individuals moving from susceptible to infected individuals was investigated by

estimating the age-dependent and cohort incidence as represented by the “brown” and “purple” arrows, respectively. In all the objectives, the outcome variable was HIV status.

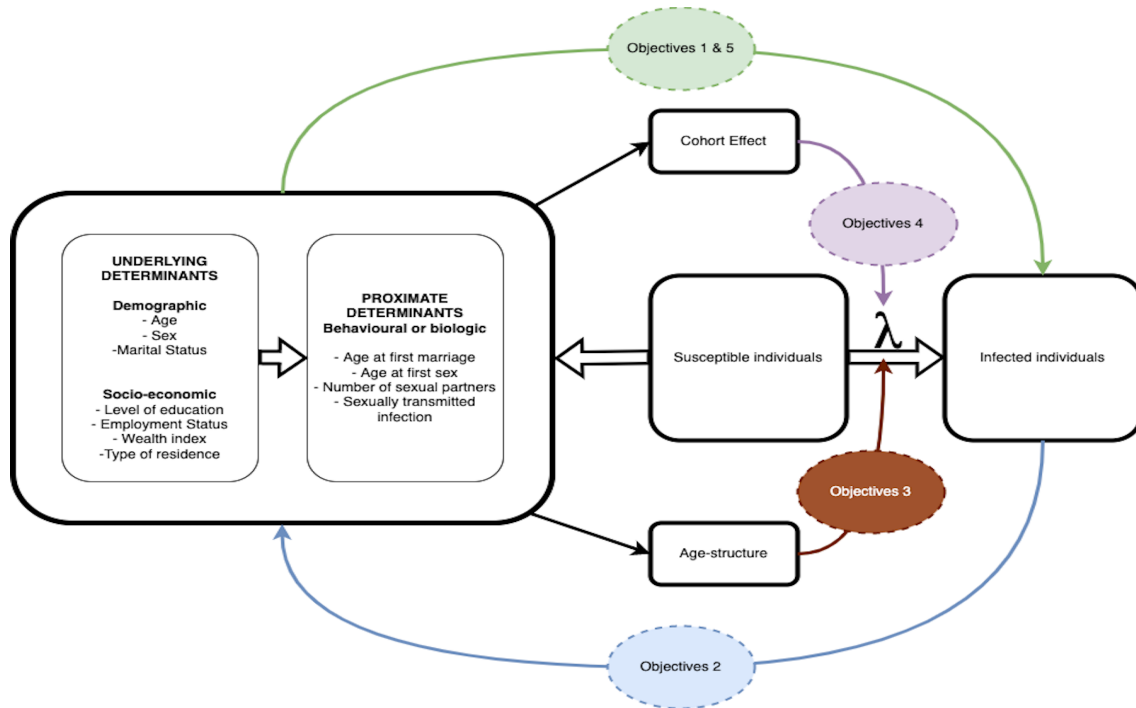


Figure 1.3: The schematic presentation of the thesis conceptual framework linking to the thesis objectives.

Based on the conceptual framework, we identified themes for this thesis linked to the five objectives, summarised in Figure 1.4.

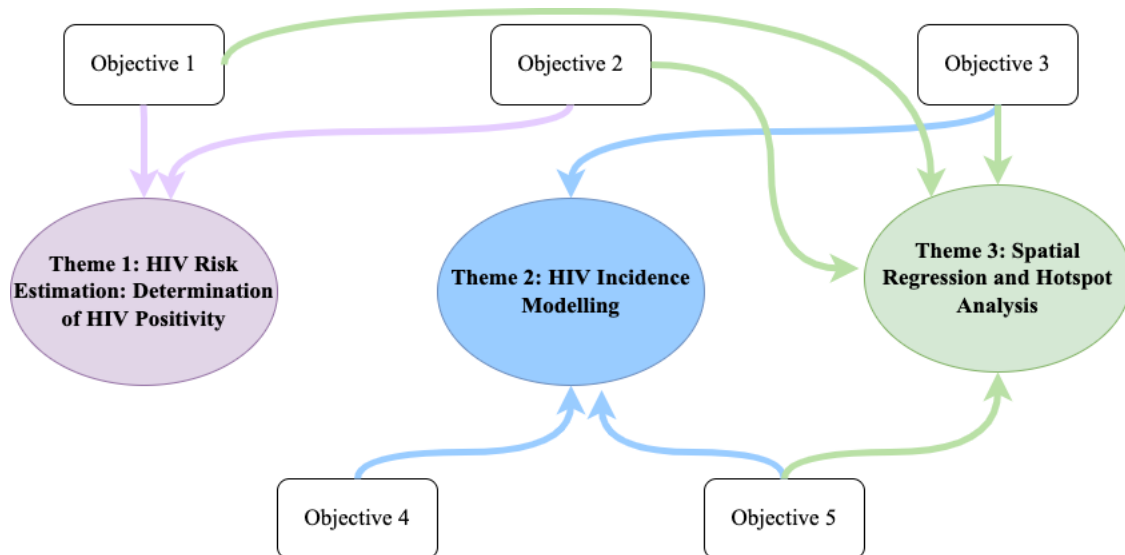


Figure 1.4: Themes and interlinkages of objectives

1.5 THESIS STRUCTURE

Chapter 1 is the background/introduction of the thesis. It describes the brief history of the HIV/AIDS epidemic in Zimbabwe. It also includes the literature review, which gives an overview of the statistical methods for monitoring the epidemic and the use of current status data to estimate HIV incidence and risk. A justification of why this study was done is included, followed by the thesis objectives. Lastly, the chapter consists of a framework of the interlinkages of the objectives and ends with the layout of the entire thesis.

Chapter 2 describes the methodology, including the study setting, data source, statistical methods and ethical approval. Some elements of the methodology are duplicated in subsequent chapters as the research papers comprised in this thesis were written as stand-alone manuscripts.

Chapter 3 (Paper 1) presents the findings of the first objective of this thesis. Using the interval censoring survival approach highlights the factors associated with HIV infection in Zimbabwe.

Chapter 4 (Paper 2) describes the use of machine learning methods to determine the common risk factors of HIV positivity in Zimbabwe and their prediction capability.

Chapter 5 (Paper 3) presents the findings of the third objective. It details HIV incidence estimation using the catalytic and Farrington models. The chapter also describes how the validity of these models was attained using the Estimation and Projection Package (EPP)/SPECTRUM and cohort estimates.

Chapter 6 (Paper 4) describes the methods used to estimate age- and time-prevalence and HIV FOI using current status data. Five models were compared, and the best model was used to estimate the cohort-specific prevalence and incidence.

Chapter 7 (Paper 5) presents the findings of HIV incidence hotspots in Zimbabwe. In addition, this chapter details the geographically-weighted regression methods used to estimate HIV incidence in Zimbabwe spatially.

Chapter 8 summarises the key findings from this PhD work. The chapter also discusses the strength and limitations and finally presents the concluding remarks and recommendations.

CHAPTER 2

2.1 STUDY METHODOLOGY

2.1.1 Study Site

The study site for this thesis was Zimbabwe, a landlocked country with ten provinces: Manicaland, Harare, Bulawayo, Midlands, Masvingo, Mashonaland Central, Mashonaland East, Mashonaland West, Matabeleland North and Matabeleland South (Figure 2.1). Provinces are made up of 59 districts, and districts are subdivided into 1200 wards.

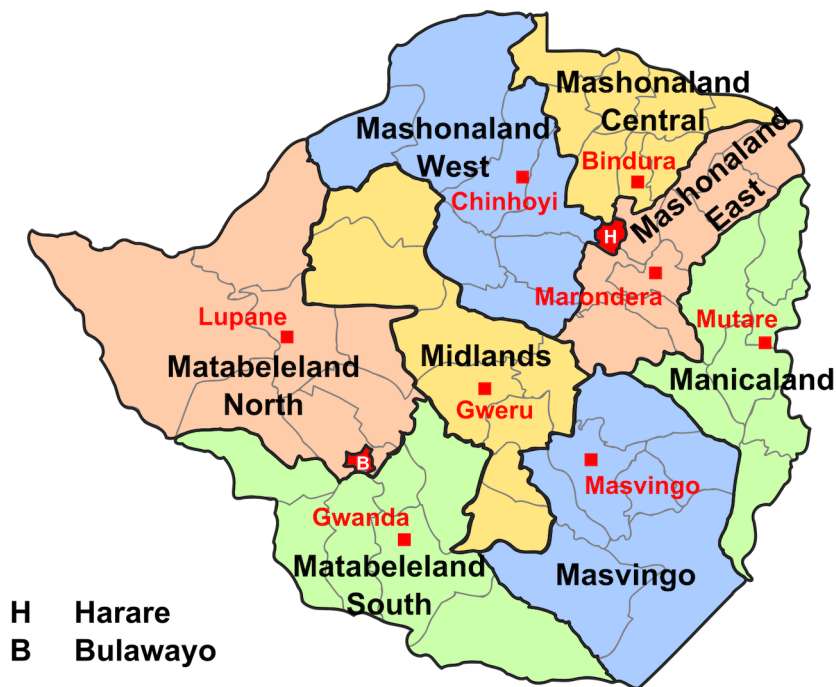


Figure 2.1: Map of Zimbabwe ¹⁰³

2.1.2 Data Sources and Description

Demographic and Health Surveys (DHS) are nationally representative household surveys conducted in about 70 countries ¹⁰⁴⁻¹⁰⁶ since 1984. They give data for various demographic, health, and nutrition monitoring and effect evaluations. The research utilized 2005–2006¹⁰⁴, 2010–2011¹⁰⁶, and 2015 ¹⁰⁵ ZDHS data available from the DHS programme website (<https://dhsprogramme.com>). The ZDHS was designed to collect representative information from all 10 provinces shown in Figure 2.1 for specific indicators such as HIV. Administratively, Zimbabwe is subdivided into provinces, provinces are subdivided into districts and districts are subdivided into wards. The households were selected using a two-stage cluster sampling strategy, Figure 2.2. As a first step, each ward was subdivided into enumeration areas (EAs). The EAs were randomly selected at a rate of 400, 406, and 400, for the 2005–06, 2010–11, and 2015 ZDHS respectively. The sampling frame used to take the EAs was from the 2002, 2002 and 2012 Zimbabwe Population Census for the 2005–06, 2010–11, and 2015 ZDHS, respectively. In the second step, a specified number of households were randomly picked from a complete list of households in the designated EAs. From the ZDHS for the years 2005–06, 2010–11, and 2015, a representative probability sample of 10,800, 10,828, and 11,196 households, respectively, was chosen. This allowed for the assignment of EA-specific weights in the design ¹⁰⁴⁻¹⁰⁶.

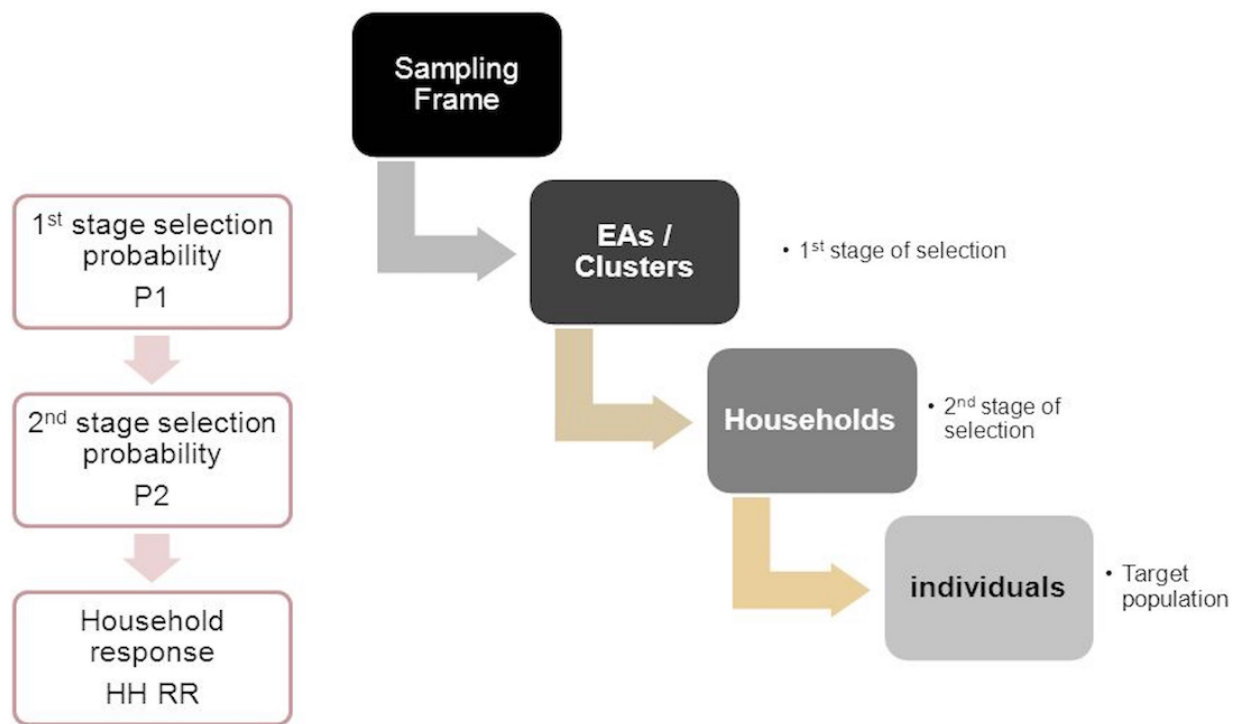


Figure 2.2: DHS sampling strategy.

Having obtained consent from the respondent or parent (i.e. for minors), blood samples were taken in all households from females aged 0–49 and males aged 0–54 for HIV testing in the laboratory. Fingerprick blood spots were collected on filter paper for analysis. A preliminary ELISA test was performed, and then all positive and 5–10% of negative samples were retested with a second ELISA¹⁰⁴⁻¹⁰⁶. If there was discrepancy between the results of the two ELISAs, a third ELISA or a Western blot was performed. This study's data was obtained from the DHS Data Archives¹⁰⁷ and included only participants aged 15–49 years.

2.2 STATISTICAL ANALYSIS OUTLINE

2.2.1 Statistical Methods for Objective 1

2.2.1.1 *Survival analysis*

Survival analysis seeks to determine the link between the survival time and a set of covariates. Numerous strategies have been developed for estimating the survival or hazard function and the impact of covariates in these situations. The survival function $S(t)$ is defined as the probability that a subject survives longer than time t , that is:

$$S(t) = P(T > t) \quad \text{Equation 2.1}$$

Where T denotes the time to event of interest (survival time).

On the other hand, the hazard function is the rate of death/failure at an instant t , given that the individuals survive up to time t :

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \quad \text{Equation 2.2}$$

where $f(t)$ is the probability density function of T .

Survival data are distinguished by the fact that the time to an event is not always accurately observed, and observations are vulnerable to censoring. Censoring is defined as the failure to monitor the time-to-event for all research participants¹⁰⁸. Left censoring describes a situation where the event of interest has already occurred before a person is recruited for the research. Alternatively, if the event of interest is not observed or achieved before the conclusion of the research period, this is referred to as right censoring¹⁰⁹. Of interest is when the event of interest is known to have occurred between two time points, then this becomes interval censoring. When individuals are followed up over time until the desired characteristic is observed, the time to event can be easily estimated¹⁰⁸. In literature, these situations have been analysed using the traditional survival analysis

methodologies. However, considering the data used in this thesis, survival analysis with interval censoring should be considered because each individual is observed at a single time point and the only information recorded is whether or not the event of interest has occurred, which can be defined as current status data. One of the common forms of interval-censored data is current status data, where all individuals in the dataset are either HIV positive (left-censored) or HIV negative (right-censored) ¹¹⁰. Current status data is also referred to as case II interval-censored data ¹⁰⁹. This thesis estimated the risk factors associated with HIV infection by modelling the hazard rate of HIV infection using conventional survival analysis and interval censoring. The time to onset of HIV infection was set as the age when the HIV test was performed on the individual during the survey, i.e., the age at survey date for the standard survival analysis. However, for the interval censoring, the time-to-onset of HIV infection was set between the age of first sex and the age at survey date.

2.2.1.2 Model Building and selection

In this thesis, the binary outcome variable was HIV status. We examined the demographic, socioeconomic, behavioural, and sociocultural aspects linked with HIV infection. Using stepwise regression, the parameters strongly related to HIV infection were determined. Parametric survival analysis was then applied, taking into account the significant factors identified through stepwise regression. We used a parametric model, so we checked the baseline distribution to make sure it was adequate. This was accomplished by contrasting several parametric models with the semi-parametric fit. Since the Weibull model had no systematic deviation from the semi-parametric model, it was found to be the best model, as shown in Figure 2.3.

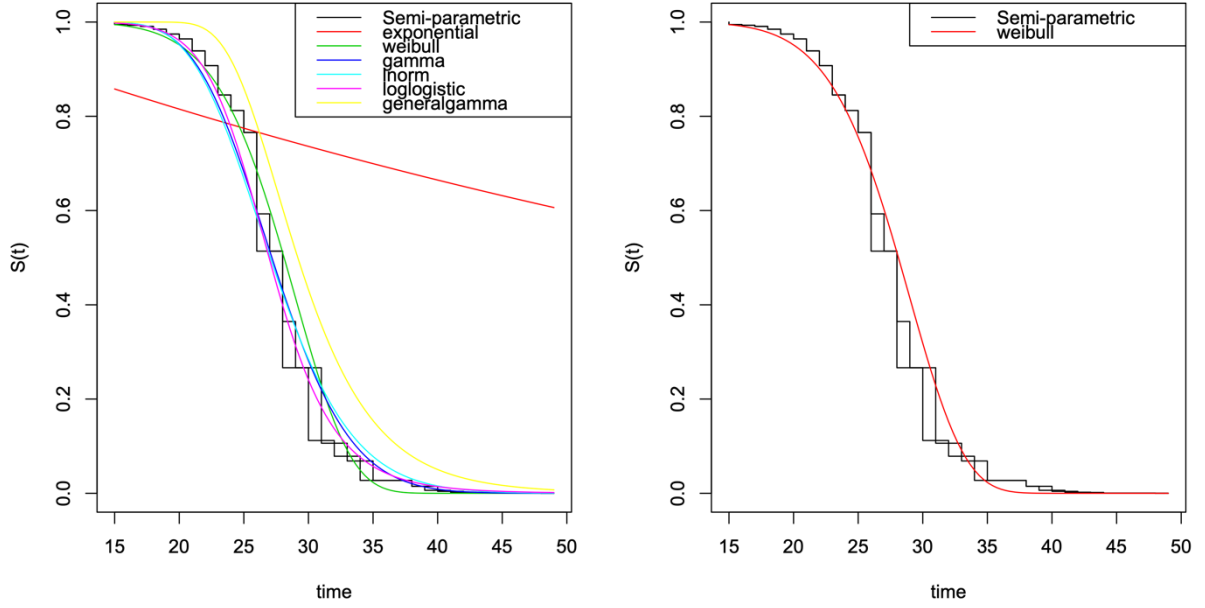


Figure 2.3: Examining parametric baseline fits for six parametric models (right) and best model (left).

Two models were fit using the parametric Weibull survival analysis. Model 1 is the Cox proportional hazards survival analysis, whereas Model 2 is the interval-censored survival analysis. Model fit was evaluated in three ways using the Akaike information criterion, the Cox-Snell residuals, and the parametric baseline fit.

2.2.2 Statistical Methods for Objective 2

2.2.2.1 Data Pre-processing

Data pre-processing consists of transforming, encoding, and filtering data so that an algorithm for machine learning can comprehend and utilise the processed output¹¹¹. In reality, data taken from real-world settings contains noise and missing values. Since machine learning algorithms are not intended to manage missing and noisy data, data pre-

processing seeks to alleviate these issues by treating the data at hand thoroughly. In the initial round of data pre-processing, male and female datasets from three ZDHS were combined individually. The HIV test results were then integrated with the individual adult interviews, using HIV status as the target variable. Those without a positive HIV test result were eliminated from the analysis. Next, data were resampled using sample weights of HIV test results for each survey year to account for non-response, non-coverage, and population total adjustment weights. To make the values comparable, a standard scaler was employed. Since machine algorithms cannot interpret string variables, nominal variables were encoded with a one-hot encoder and ordinal variables were encoded with a label/ordinal encoding. To eliminate redundant data, so the model is less biased, we omitted variables with a correlation coefficient of 80% or higher, non-unique columns, and arbitrary features. All variables with more than 30% missing values were eliminated, while variables with less than 30% missing values were assumed to be missing at random; therefore, Multiple Imputations with Chained Equations (MICE) was used. Lastly, the data were normalised and scaled.

Following the stratification by sex, we left out one of the three ZDHS datasets (rotating left-out survey) to construct three datasets per sex, each consisting of just two surveys. The primary objective of the rotation was to examine the generalizability of the models separately for men and women in subsequent testing. The three freshly formed datasets per sex were then stratified, with 80% of the data serving as training and 20% as testing. Then, standardisation and imputation were performed separately on the train and test datasets to prevent contamination.

2.2.2.2 Model training and validation

The dataset must be divided into three parts for machine learning models to perform well: train set, validation set, and test set. A training set is the subset of a dataset from which an algorithm for machine learning learns the relationships between the features (independent variables) and the target variable (outcome variable). Typically, an algorithm is fed the training dataset to construct a model. Next, the model translates inputs to outputs. Next, a validation dataset is utilised to tune a model's hyperparameters. Hyperparameters are explicitly defined parameters to control the learning process before applying a machine-learning algorithm to a dataset¹¹². Changing the value of hyperparameters modifies the resulting model and is typically used to control how an algorithm learns data relationships. Hyperparameters can be used to set the model's complexity, with complex models learning more complex interactions while simpler models learn more straightforward associations using the same algorithm¹¹². As a result, complex models typically risk the danger of overfitting. The validation dataset is utilised during training to monitor the model's performance with varying hyperparameter values. The validation dataset is utilised to detect overfitting during training. Lastly, a test set is distinct from the training set and is used to evaluate the performance of the trained model. The test set estimates the model's ultimate performance after hyperparameter adjustment. The test dataset is also helpful in comparing the relative performance of various models. Putting the preceding literature on training, test, and validation sets into context, our dataset for this thesis was separated into these three categories. However, instead of having an independent split of the validation set, we opted for stratified k-fold cross-validation. Specifically, we used a stratified 5-fold cross-validation in which we selected 5 different subsets of data as validation sets, and we trained 5 models on the remaining data, as shown in Figure 2.4. Using the validation data set produced from the cross-

validation, 50 sets of hyperparameters were chosen randomly for each algorithm using the validation data set. Then, to evaluate the performance of the models, we averaged the model outputs and picked the set of hyperparameters with the highest mean F1 score throughout the validation dataset.

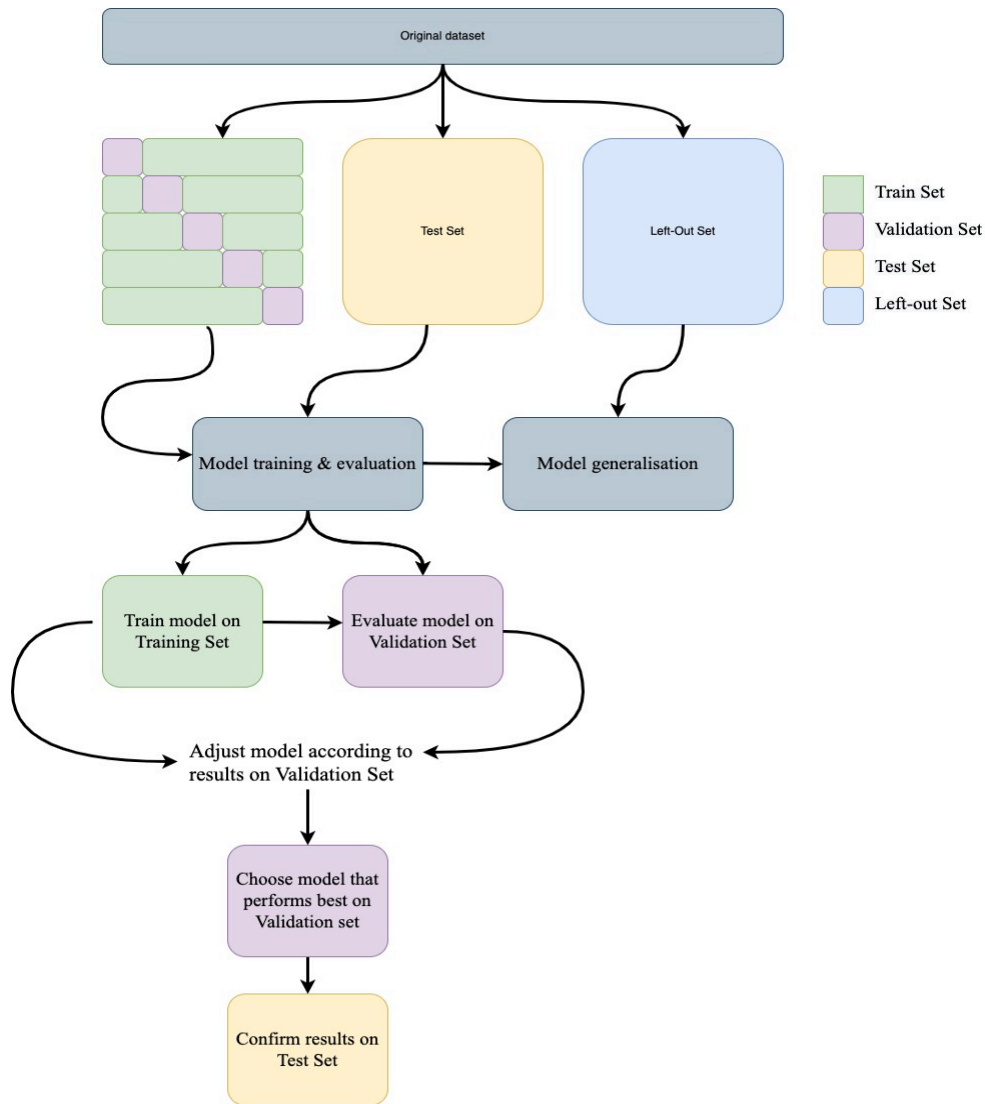


Figure 2.4: Dataset Split diagram

2.2.2.3 Model Testing and algorithm selection

Model testing in machine learning is the process of evaluating the performance of a fully trained model on a testing set ¹¹³. Several statistical metrics can be used to evaluate test findings, including F1 scores, Precision-Recall curves, and receiver operating characteristics (ROC) curves ¹¹⁴. Three of the top models per sex and algorithm were run on the test set from our dataset, and the resulting metric scores were averaged. Six methods were used on the test set: penalised logistic regression (Elastic Net), k-nearest neighbour algorithm (KNN), Random Forest Classifier (RFC), Decision Tree approach (DT), Light Gradient Boosting Model (LightGBM), and XGBoost model. The model with the highest mean F1 score was deemed the best. Each of the seven models was then applied to the left-out survey data set.

2.2.3 Feature selection

Feature selection, also known as variable or attribute selection, is the process of choosing a subset of important features for use in model development¹¹⁵. The primary objective of feature selection in machine learning is to improve a machine learning algorithm's performance, enhance the model's interpretability, and decrease the amount of computing and data-gathering resources required¹¹⁶. Including features irrelevant to the classification problem (outcome variable) will introduce noise into the data set and contribute to overfitting, resulting in a biased classification result with a large variance. Eliminating unnecessary features simplifies the model, which enhances its interpretability. This thesis utilised the Lasso and sequential forward floating selection (SFFS) for feature selection.

Lasso regression is a highly effective automated feature selection method¹¹⁷. If two features are linearly correlated, their simultaneous existence will raise the value of the cost function; thus, Lasso regression will attempt to reduce the coefficient of the less significant feature to zero to choose the most relevant features¹¹⁷. The cost function is explained as follows:

$$\frac{1}{2N_{training}} \sum_{i=1}^{N_{training}} (y_{obs}^i - y_{pred}^i)^2 + \alpha \sum_{j=1}^n |a_j| \quad \text{Equation 2.3}$$

where a_j is the coefficient of the j th feature. The final term is called the $l1$ penalty, and α is a hyperparameter that tunes the intensity of this penalty term. The greater the coefficient of a feature, the greater the cost function's value. Therefore, the objective of Lasso regression is to maximise the cost function by minimising the absolute values of the coefficients¹¹⁸. To minimise the cost function, Lasso regression will automatically choose useful features and reject undesirable or redundant ones. Therefore, we used Lasso regression on a scaled dataset and selected features with coefficients other than zero.

Sequential forward floating selection is a wrapper method that tries to find a subset of the features to optimise an objective function¹¹⁹. First, the space of feasible feature subsets is explored. Then, choose the optimal subset relative to a given objective function. The objective function is a pattern classifier that evaluates the predicted accuracy of feature subsets using statistical resampling or cross-validation¹¹⁹. The capacity to generalise and prevent overfitting is one of the benefits of wrapper approaches, which commonly employ cross-validation metrics of prediction accuracy. In this thesis, the sequential forward floating selection was used to exclude variables based on a given classifier performance criterion for 80% of the training data, and the F1 score was derived for various subsets of

variables. Finally, the optimal subset variable among those for which F1 plateaued was selected.

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. This was achieved by generating synthetic data using the nearest neighbour's algorithm to balance out the classes. Original dataset without implementing the SMOTE procedure was also analysed to compare results.

2.2.4 Statistical Methods for Objective 3

2.2.4.1 *Catalytic and Farrington model*

The first references to the estimation of force-of-infection date back to Muench's early work. Muench recommended using a "catalytic model" to determine the rate at which a susceptible individual acquires an infection¹²⁰. The significance of Muench's catalytic model is the ability to determine the force-of-infection using test results on serological samples⁷². Catalytic epidemic models consider the age distribution at the onset of an infectious disease. The work of Muench has led to the creation of several models for estimating the force-of-infection¹²⁰. Farrington included monotonicity in the calculation of the force-of-infection, but Muench's model did not account for monotonicity with increasing age. Farrington considered a nonlinear model, which was an exponentially damped linear model where the force-of-infection typically rises linearly and then drops off for older ages. Both these models estimate the age-dependent force-of-infection.

Muench's and Farrington's models have been used extensively to estimate the force-of-infection of other infectious diseases such as measles and rubella. Following Muench's catalytic model, we adapted the catalytic model of Mossong et al.¹²¹, which has been used to estimate HIV incidence using seroprevalence data. The catalytic model used in this thesis simultaneously estimated the age-dependent incidence and differential mortality in HIV-infected individuals. On the other hand, the Farrington model assumed non-differential mortality, which indicated that HIV-related mortality did not affect how the observed serological profile was interpreted. In addition, the catalytic model had a lognormal functional form of incidence and assumed homogeneity over time.

2.2.4.2 Model Validation

Model validation is a technique used to ensure that models perform as anticipated. Since age-dependent HIV incidence was predicted using the catalytic and Farrington models, it was necessary to validate the estimates. Validation was established at both the micro and macro levels. First, we used the entire dataset, i.e., country-level data, to estimate the incidence of HIV at the macro level using the catalytic and Farrington models. The estimation findings were then compared to the EPP/SPECTRUM estimates. We used the EPP/SPECTRUM estimates as a comparison since these estimates are the ones used in Zimbabwe, produced by the National AIDS council to inform policy. At the micro level, we used just one province dataset, i.e., Manicaland dataset. From the ZDHS, we only used data from Manicaland province to estimate HIV incidence using the catalytic and Farrington models. These estimates were then compared to the HIV incidence values obtained from a community cohort study carried out in Manicaland HIV/STD Prevention Project¹²². Since cohort studies are deemed the “gold standard” in incidence estimation,

we decided to validate the catalytic and Farrington models against the “gold standard” estimates.

2.2.5 Statistical Methods for Objective 4

2.2.5.1 *Generalised Additive Models (GAM)*

A generalised additive model (GAM) is a generalised linear model with a linear predictor that incorporates a sum of smooth functions of variables¹²³. The interpretability of generalised additive models is simple, their flexible predictor functions can reveal hidden data patterns, and regularisation of predictor functions helps prevent overfitting¹²³. Generalised additive models employ an additive modelling approach in which the influence of the predictive factors is recorded by smooth functions that may be nonlinear depending on the underlying data patterns¹²⁴. When estimating GAMs, the objective is to simultaneously estimate all smoothers and the parametric terms in the model while accounting for the smoothers’ covariance.

This objective focused on estimating age- and time-dependent prevalence and force of infection from serial cross-sectional data. We compared five different generalised additive models. Model 1 adapted the Ades and Nokes¹²⁵ exponential polynomial model. Models 2 and 3 were semi-parametric models modified from Nagarkerke et al.¹²⁶, where a spline smoother was used for the baseline hazard function instead of an isotonic step function. Finally, models 4 and 5 were non-parametric. What was of interest in all five models was modelling the force-of-infection, which followed an exponential family distribution using predictor variables age and year of birth. The Pool-Adjacent-Violators

Algorithm (PAVA) application was used to ensure monotonicity in the age dimension. The Akaike Information Criterion was used to choose the best model.

2.2.6 Statistical Methods For Objective 5

2.2.6.1 *Spatial Analysis*

In this thesis, we explored spatial analysis regression methods to determine the district patterns of HIV incidence over space. Estimation of HIV incidence was carried out using the catalytic method procedure explained in section 2.2.3.1 and Chapter 5. Using the estimated HIV incidence per district, we also explored the different hot spot and cluster areas in Zimbabwe over three different survey points. In order to perform the spatial regression, the response variable was defined as the HIV incidence per district. Some of the explanatory variables were chosen from the results obtained from Chapter 3 as a starting point. Details of how the significant variables were obtained can be found in Chapter 7. To determine the global relationship between the response and independent variables, an Ordinary Least Squares (OLS) Regression was performed. Residuals for the OLS models were checked for spatial autocorrelation using the Moran's I statistic. Independent variables which were significant in the OLS model were then subjected to the Geographically Weighted Regression (GWR) and Multiscale Geographically Weighted Regression (MGWR).

2.2.6.2 *Geographically Weighted Regression (GWR) and Multiscale Geographically Weighted Regression (MGWR)*

One of the most widely used methods to explain the spatial nonstationarity of relationships is the GWR. The GWR model can be expressed as follows:

$$\begin{aligned}
y_i &= \beta_0(u_i v_i) + \beta_1(u_i v_i) x_{i1} + \beta_2(u_i v_i) x_{i2} + \dots + \beta_j(u_i v_i) x_{ij} + \varepsilon_i \\
&= \sum_{j=0}^n \beta_j(u_i v_i) x_{ij} + \varepsilon_i
\end{aligned}
\tag{Equation 2.4}$$

Where y_i are observations of HIV incidence, $u_i v_i$ are the longitude and latitude geographical coordinates for the i th observation, $\beta_j(u_i v_i)$ are the parameters for x_{ij} independent covariates describing the relationship around geographical locations $(u_i v_i)$ and ε_i is the error term.

Traditionally, regression methods produce an average global estimate for the relationship between the response and explanatory variables, however, the GWR allows these relationships to vary over space. The basic idea of the GWR is that, since each location has a neighbour, data from one location can be “borrowed” from nearby locations and weighted according to the distance each neighbour is from the regression line. This weighting scheme is called the Weighted Least Squares approach, where:

$$\hat{\beta}_i = (\mathbb{X}^T \mathbb{W}_i \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W}_i y \tag{Equation 2.5}$$

Where y is the vector of observed values, $\hat{\beta}_i$ is the vector of estimated parameters at regression point i , $\mathbb{X}$ being the design matrix which contains a column of 1s and the independent variables, and $\mathbb{W}$ is a weight matrix conditioned on location i with the form:

$$\mathbb{W}_i = \begin{pmatrix} w_{i1} & 0 & 0 & \dots & 0 \\ 0 & w_{i2} & 0 & \dots & 0 \\ 0 & . & w_{i3} & \dots & 0 \\ . & . & . & \dots & . \\ 0 & 0 & 0 & \dots & w_{in} \end{pmatrix}$$

Therefore, the main strengths of the GWR are its capability to identify spatial heterogeneity and, at the same time, take into account the spatial dependence of the data. One of the assumptions of the GWR is that it assumes that all spatial varying processes in a model operate at the same spatial scale. This may be disadvantageous if we have processes that operate at a local scale and others at a broader scale. However, if this assumption is relaxed, and we assume that processes vary and operate at different scales, then we introduce the Multiscale Geographically Weighted Regression (MGWR) model. The MGWR has the potential to produce a more powerful spatial model than the GWR. A Generalised Additive Model (GAM) can be formulated as:

$$y = f_1(X_1) + f_2(X_2) + \dots + f_k(X_k) + \varepsilon_k \quad \text{Equation 2.6}$$

Where y is the vector of the response variable; $f_1, \dots, f_k$ are smooth functions; $X_1, \dots, X_k$ are the explanatory variables and ε_k is a vector of random error. The MGWR can be formulated using the GAM framework as follows:

$$y = f_{bw1}(X_1) + f_{bw2}(X_2) + \dots + f_{bwk}(X_k) + \varepsilon_k \quad \text{Equation 2.7}$$

where;

$$f_{bwj}(X_j) = \begin{pmatrix} \beta_{1j}x_{1j} \\ \beta_{2j}x_{1j} \\ \vdots \\ \beta_{nj}x_{1j} \end{pmatrix} \begin{pmatrix} u_1v_1 \\ u_2v_2 \\ \vdots \\ u_nv_n \end{pmatrix}$$

therefore;

$$y_i = \beta_{bw0}(u_i v_i) + \beta_{bw1}(u_i v_i)x_{i1} + \beta_{bw2}(u_i v_i)x_{i2} + \dots + \beta_{bwj}(u_i v_i)x_{ij} + \varepsilon_i$$

$$= \sum_{j=0}^n \beta_j(u_i v_i) x_{ij} + \varepsilon_i \quad \text{Equation 2.8}$$

Where y_i are observations of HIV incidence, $u_i v_i$ are the longitude and latitude geographical coordinates for the i th observation, $\beta_j(u_i v_i)$ are the parameters for x_{ij} describing the relationship around geographical locations $(u_i v_i)$ and ε_i is the error term. Hence in this thesis, we explored both the GWR and the MGWR, and the results of the model, which was a better fit, were presented.

2.2.7 Ethical Considerations

The ZDHS HIV testing protocol for all three surveys was reviewed and approved by the ethical review boards Medical Research Council of Zimbabwe (MRCZ) in Harare, Zimbabwe; the ORC Macro Institutional Review Board in Calverton, Maryland, USA; and the Centers for Diseases Control (CDC) in Atlanta, Georgia, USA (8–10). The Human Research Ethics Committee (Medical) at the University of Witwatersrand also granted ethical clearance to this study (No. M151154). The research dataset was obtained on May 16, 2017, after a request was submitted to and accepted by the Measure DHS program. In accordance with its mandate, the DHS Program may release unrestricted survey data files at no cost for approved academic research. User registration was needed before data could be accessed.

CHAPTER 3

THESIS PAPER 1: Factors Associated with HIV Infection in Zimbabwe Over a Decade from 2005 to 2015: an Interval-Censoring Survival Analysis Approach.

Rutendo Birri Makota^{1*}, Eustasius Musenge¹

¹ Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

***Correspondence:**

Rutendo Birri-Makota

rutendobbirri@gmail.com

Published paper:

Birri Makota R, Musenge E. Factors Associated With HIV Infection in Zimbabwe Over a Decade From 2005 to 2015: An Interval-Censoring Survival Analysis Approach. *Front Public Health*. 2019 Sep 18;7:262. doi: 10.3389/fpubh.2019.00262. PMID: 31620418; PMCID: PMC6759818.

Abstract

Objectives: The main objective of this study was to compare results from two approaches for estimating the effect of different factors on the risk of HIV infection and determine the best fitting model.

Study design: We performed secondary data analysis on cross-sectional data which was collected from the Zimbabwe Demographic Health Survey (ZDHS) from 2005 to 2015.

Methods: Survey and cluster adjusted logistic regression was 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 were modelled assuming age at survey imputation (Model 1) and interval-censoring (Model 2).

Results: Model goodness of fit test based on the Cox-Snell residuals against the cumulative hazard indicated that Model 1 was the best model. On the contrary, the Akaike Information Criterion (AIC) indicated that Model 2 was the best model. Factors associated with a high risk of HIV infection were being female, number of sexual partners, and having had an STI in the past year prior to the survey.

Conclusion: The difference between the results from the Cox-Snell residuals graphical method and the model estimates and AIC value maybe due to the lack of adequate 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 literature. Even though we consider the interval-censoring model as the superior model with regards to our specific data, the method had its own set of limitations.

Keywords: Interval-censoring¹, HIV², Survival³, Prevalence⁴, Zimbabwe⁵.

3.1 INTRODUCTION

The 90–90–90 targets was 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 ¹²⁷.

In Zimbabwe, a population-based survey carried out 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, are virally suppressed ¹²⁸. In order for these 90-90-90 targets to be met, prevalence and incidence rates estimates are crucial in understanding the current status of the HIV epidemic and determine whether the trends are improving in a manner which can facilitate to achieve the 2020 target.

The gold standard for estimating the HIV incidence is to test uninfected individuals for new infections periodically, however this method is feasible though costly and time-consuming. In addition, even if an HIV negative cohort is followed over-time, the exact date of infection is rarely observed ¹²⁹. In this scenario, an interval can be determined between the latest negative and the earliest positive test dates. Taking into consideration the issue of cost and time, cohort analysis for estimating the HIV incidence rate in a general epidemic will not produce estimates which are representative of the whole population. Due to these reasons, sentinel surveillance systems have been set up to monitor the spread of the pandemic ¹³⁰. In addition, population-based surveys, in which HIV tests are performed, are carried after every 5-years in Zimbabwe. The advantage of the population-based survey is that data is more representative of the population than a

cohort. On the other hand, the same data does not provide the exact date of the infection but rather provide what is called “current status data”.

Current status data occurs when an individual is observed at one single point, and the only information obtained is whether the event of interest has occurred ¹³¹. An example of current status data includes information collected during a demographic health survey, in which an individual is tested whether they are HIV positive or negative. If an individual were found to be HIV positive, the individual was recorded as left censored at the time the test was done. If an individual were HIV negative, they were recorded to be right censored. Sometimes current status data can be referred to as case interval-censored data, with case II interval-censored data referred to as the general case ¹³². Interval censoring takes into account the range, that is, an interval inside of which one can say the outcome of interest has occurred ¹³³. Given that we would want to determine the factors associated with the hazard of infection using data from these surveys, then survival analysis can be implemented.

In the setting of standard survival analysis, modelling the hazard rate of HIV infection can be achieved by imputing the time-to-onset of disease as the time at diagnostic visit ¹²⁹. Modelling current status data using the mentioned two approaches may overestimate the hazard rate. However, analysing this type of data using interval censoring will be a better approach. Although they are documented advantages of interval censoring compared to the standard cause-specific survival model, the superiority of the interval-censoring remains unclear in estimating the effect of different exposures on the risk of HIV infection. With this argument in mind, the main objective of this study was to

compare results from these different approaches for estimating the effect of different factors on the risk of HIV infection and determine the best fitting model.

3.2 MATERIALS AND METHODS

3.2.1 Study Design and Area

This study used data from Zimbabwe, a landlocked country, bordered by Mozambique on the East, South Africa on the South, Botswana on the West, and Zambia on the North and Northwest. Zimbabwe is sub-divided into 10 Provinces which are: Matabeleland South, Matabeleland North, Mashonaland East, Mashonaland Central, Mashonaland West, Midlands, Masvingo, Manicaland, Harare, and Bulawayo. Each province is subdivided into districts, and each district is made up of wards. The designs for the three 5 yearly surveys were cross-sectional.

3.2.2 Zimbabwe Demographic Health Survey Data and Sampling

Demographic and Health Surveys (DHS) are nationally representative household surveys that have been implemented in approximately 70 countries since 1984¹⁰⁴⁻¹⁰⁶. They provide data for a wide range of monitoring and impact evaluation in the areas of population, health, and nutrition. Data used for the analysis were obtained from 2005-06, 2010-11 and 2015 ZDHS and were retrieved from the DHS programme website (<https://dhsprogram.com>)¹³⁴. A representative probability sample of 10,800, 10,828 and 11,196 households were selected for the 2005-06, 2010-11 and 2015 ZDHS respectively. A two-stage cluster sampling technique was used to select the households. The first stage selected 400, 406 and 400 enumeration areas (EAs) for 2005-06, 2010-11 and 2015 ZDHS respectively. At the second stage, using a complete listing of households in the

selected EAs, a fixed number of households were randomly chosen. This allowed the use of EAs specific weights to be assigned in the design ¹⁰⁴⁻¹⁰⁶.

3.2.3 Measurement of The Outcome (HIV Status) and Explanatory Variables

With consent from the respondent or parent/guardian (for minors), blood samples were collected in all households for HIV testing in the laboratory for females aged 0-49 and males aged 0-54. Blood spots were collected on filter paper from a finger prick and transported to a laboratory for testing. An initial ELISA test was performed, and then retesting of all positive and 5-10 % of the negative tests with a second ELISA was done. If they were discordant results on the two ELISA tests, a new ELISA or a Western Blot was performed. The data used for this study was obtained from the DHS Data Archives¹³⁴ and only included individuals aged 15-49 years. The following explanatory variables were extracted from the dataset: sex, marital status, education level, religion, currently employed, place of residence, STI treatment in the past 12 months and number of sexual partners ¹⁰⁴⁻¹⁰⁶.

3.3 ETHICAL CONSIDERATIONS

The ZDHS HIV testing protocol for all the three surveys was reviewed and approved by the ethical review boards Medical Research Council of Zimbabwe (MRCZ) in Harare, Zimbabwe; the ORC Macro Institutional Review Board in Calverton, Maryland, USA; and the Centres for Diseases Control (CDC) in Atlanta, Georgia, USA ¹⁰⁴⁻¹⁰⁶. This work was granted ethical clearance by the University of Witwatersrand's Human Research Ethics Committee (Medical) (No. M151154). The dataset used in this study was obtained

through an application made to Measure DHS program, which was approved on the 16th of May 2017. The DHS Program is authorised to distribute, at no cost, unrestricted survey data files for legitimate academic research. Registration was required for access to data.

3.4 STATISTICAL METHODOLOGY

3.4.1 Application To Zimbabwe Demographic Health Survey

3.4.1.1 *Data Management*

The socio-demographic datasets for men and women records were appended to provide a single analysis dataset for all the three surveys. The appended dataset was then merged using the unique combination of the individual line number, household line number and cluster (EA) number to the HIV prevalence dataset. All individuals without an HIV test result, never been sexually active, and individuals who did not have an age at first intercourse were excluded from the analysis. In the case of individuals who were HIV positive when the survey was conducted, the age at HIV infection was defined as age at survey date and for individuals who were HIV negative, the age at HIV infection was right-censored at the date of survey in Models 1 (**Figure 3.1**). Accounting for interval-censoring, the age at HIV infection was interval-censored between age at first sexual intercourse and age at date of survey, but right-censored at the age at survey, for individuals who were HIV negative for Model 2. All the models assumed a parametric Weibull distribution for the baseline hazard $\lambda_0(t)$ which allowed estimation of β which is the vector of regression coefficient. Model Specification for all the models, refer to supplementary data.

3.4.2 Statistical Analysis

In this study, the outcome or response variable was the HIV status, a binary variable. The study investigated the socio-cultural, socio-economic, behavioural and demographic factors, which are associated with HIV. Trends in HIV prevalence were assessed using the non-parametric trend test in STATA. A stepwise logistic regression approach was adopted in STATA SE version 15.1 statistical software ¹³⁵ using the command **svy: swaic** ¹³⁶ on some selected explanatory variables highlighted earlier. Factors that were significantly associated with HIV from the stepwise survey logistic regression were then considered for the parametric survival analysis. The most suitable baseline hazard function was investigated using the package **icenReg** ¹³⁷ in R software. The data were modelled assuming age at survey imputation and interval-censoring (**Figure 3.1**). The model goodness-of-fit (GOF) test was assessed using the Akaike Information Criterion (AIC) and two graphical methods which included the Cox-Snell residuals. All analysis were performed in STATA SE version 15.1 and R statistical software.

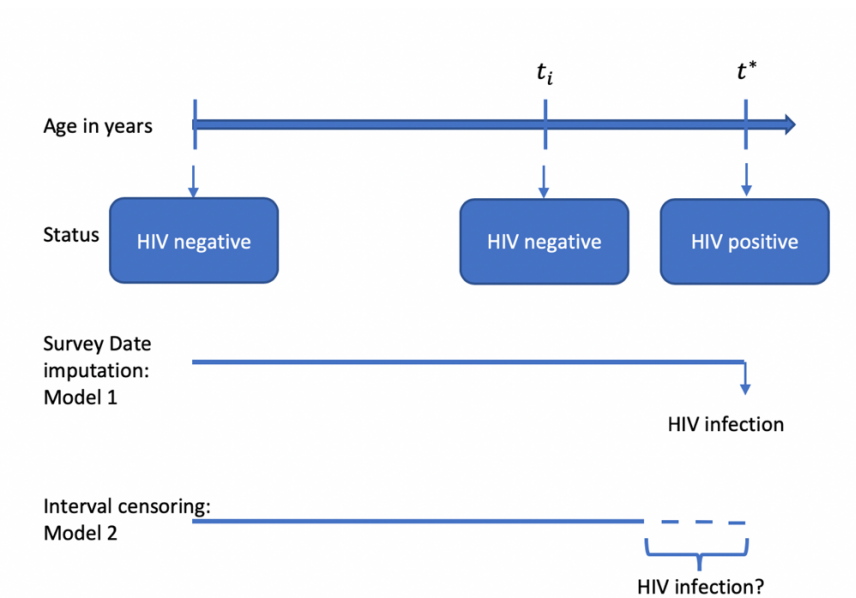


Figure 3.1: Illustration of different modelling strategies when investigating the factors associated with HIV infection.

3.5 RESULTS

The 2005-06 ZDHS database had 10,800 households in which 42,698 records were retrieved. Of the 42,698 records, 16,082 records were for individuals aged 15-49 years with 7,175 (44.6%) males and 8,097 (55.4%) females. The 2010-11 ZDHS database had 10,828 households in which 41,946 records were retrieved. Of the 41,946 records, 16,651 records were for individuals aged 15-49 years with 7,480 (44.9%) males and 9,171 (55.1%) females. The 2015 ZDHS database had 11,196 households in which 43,706 records were retrieved. Of the 43,706 records, 18,351 records were for individuals aged 15-49 years with 8,396 (45.8%) males and 9,955 (54.2%) females. The above information is represented in **Figure 3.2**.

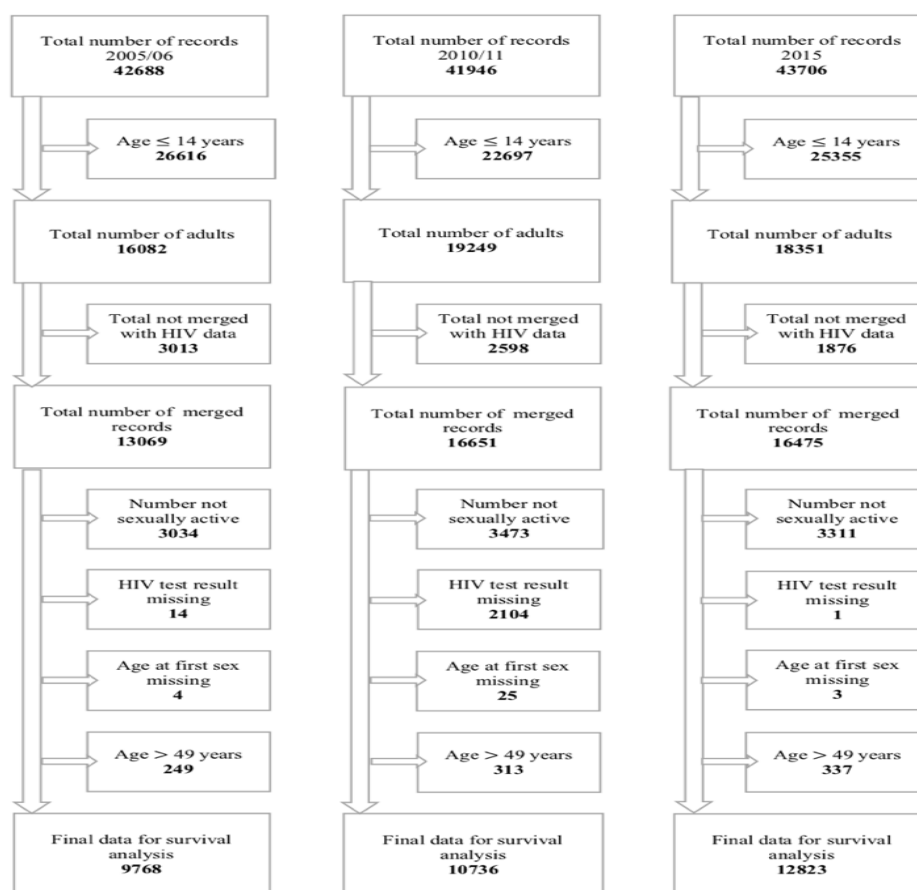


Figure 3.2: Flowchart of records in the 2005-06, 2010-11 and 2015 ZDHS.

Nationally, the non-parametric trend test ($p < 0.001$) showed a significant decline of HIV prevalence from 22.4% to 19.6% to 17.7% for 2005-06, 2010-11 and 2015 ZDHS respectively (see **Table 3.1**). A similar decline trend was observed for gender, marital status, place of residence, education level, current employment status, STI in the past year preceding the survey and number of sexual partners.

The mean survival time for age at HIV infection for Model 1 was 41.1 years for females, 42.8 years for males in 2005/06 ZDHS; 41.9 years for females, 43.4 years for males in 2010/11 ZDHS and 42 years for females, 44 years for males in 2015 ZDHS. The mean

survival time for age at HIV infection for Model 2 was 24.9 years for females, 28.3 years for males in 2005/06 ZDHS; 25.7 years for females, 29.4 years for males in 2010/11 ZDHS and 26.5 years for females, 30.5 years for males in 2015 ZDHS. According to the survival times, Model 2 produced lower times of survival before HIV infection.

Table 3.1: HIV prevalence in Zimbabwe and changes in HIV prevalence (weighted) from the Zimbabwe DHS surveys 2005-06, 2010-11 and 2015.

<i>Survey year</i>	<i>2005-06</i>	<i>2010-11</i>	<i>2015</i>	<i>p-value</i>
<i>Characteristic</i>	<i>Prevalence (CI)</i>	<i>Prevalence (CI)</i>	<i>Prevalence (CI)</i>	
<i>National summary</i>	22.4 (21.6 , 23.2)	19.6 (18.8, 20.3)	17.7 (17.0 , 18.4)	<0.001
<i>Sex: male</i>	18.2 (16.7 , 19.4)	16.2 (15.1 , 17.3)	13.6 (12.7 , 14.5)	<0.001
<i>female</i>	25.1 (24.0 , 26.2)	21.8 (20.8 , 22.8)	20.7 (19.8 , 21.6)	<0.001
<i>Marital Status: never married</i>	10.8 (9.36 , 12.3)	9.66 (8.29 , 11.0)	7.86 (6.78 , 8.95)	0.001
<i>married/cohabiting</i>	20.5 (19.5 , 21.5)	17.8 (16.9 , 18.7)	16.8 (16.0 , 17.6)	<0.001
<i>separated/divorced/widowed</i>	45.6 (43.0 , 48.2)	42.2 (39.6 , 44.8)	38.3 (35.9 , 40.8)	<0.001
<i>Place of residence: urban</i>	23.7 (22.2 , 25.3)	22.0 (20.5 , 23.4)	18.3 (17.3 , 19.3)	<0.001
<i>rural</i>	21.8 (20.8 , 22.8)	18.5 (17.7 , 19.4)	17.3 (16.4 , 18.1)	<0.001
<i>Education level: no education/primary</i>	22.6 (21.3 , 23.9)	21.1 (19.7 , 22.4)	20.7 (19.3 , 22.1)	0.05
<i>secondary</i>	22.6 (21.6 , 23.7)	19.4 (18.5 , 20.3)	17.4 (16.5 , 18.2)	<0.001
<i>higher</i>	16.3 (12.6 , 20.0)	12.6 (9.86 , 15.4)	12.1 (10.3 , 13.9)	0.05
<i>STI treatment in past 12 months: no</i>	21.4 (20.7 , 22.2)	18.9 (18.2 , 19.7)	17.2 (16.5 , 17.8)	<0.001
<i>yes</i>	46.2 (41.2 , 51.2)	39.0 (34.0 , 44.0)	37.2 (32.0 , 42.3)	0.01
<i>No. of sexual partners: 0</i>	31.3 (28.9 , 33.6)	28.2 (25.9 , 30.5)	24.5 (22.2 , 26.9)	<0.001
<i>1</i>	20.8 (19.9 , 21.7)	18.2 (17.4 , 19.1)	17.0 (16.2 , 17.7)	<0.001
<i>2+</i>	19.5 (16.1 , 22.9)	17.7 (14.8 , 20.7)	15.3 (13.2 , 17.4)	0.03
<i>Currently employed: no</i>	22.7 (21.5 , 23.9)	18.9 (17.9 , 20.0)	17.9 (16.9 , 19.0)	<0.001
<i>yes</i>	22.1 (21.0 , 23.2)	20.2 (19.2 , 21.3)	17.5 (16.7 , 18.4)	<0.001
<i>*p-values below 0.05 are considered statistically significant</i>				

The parametric Weibull distribution was used to investigate the factors associated with HIV infection. The Weibull parametric function was suitable to model the baseline hazard distribution, as shown in **Figure 3.3**. Model goodness of fit test indicated that Model 2 was the best model based on the Akaike Information Criterion (AIC) presented in **Table**

3.2. A graphical goodness of fit test was performed, where the semi-parametric model was compared to the parametric model. According to the results in **Figure 3.3**, Model 2 fits the data better than Model 1. Based on **Figure 3.3**, Model 1 overestimates the survival rates between 15-35 years and underestimates the survival rates between 35-49 years.

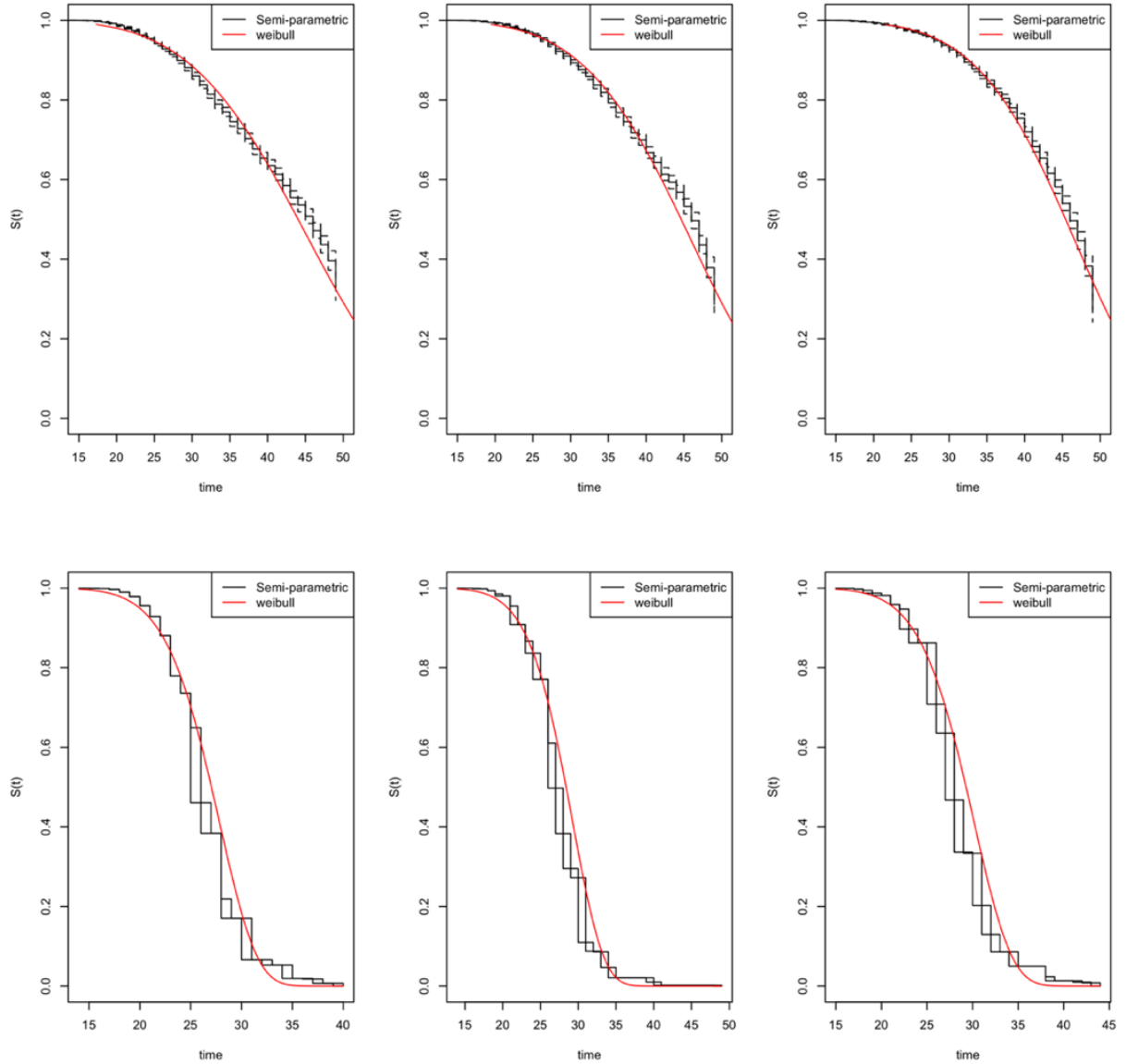


Figure 3.3: Goodness-of-fit tests for the Zimbabwe DHS 2005-06 (left column), 2010-11 (middle column) and 2015 (right column). Model 1 (top row) and Model 2 (bottom row)

Table 3.2: Akaike Information Criterion and Bayesian Information Criterion values.

<i>Year</i>	<i>Model</i>	<i>Obs.</i>	<i>ll(null)</i>	<i>ll(model)</i>	<i>df</i>	<i>AIC</i>	<i>BIC</i>
2005-06	1	9768	-2995.77	-2605.36	14	5238.7	5339.3
	2	9768	-2177.37	-1912.55	14	3853.09	3953.71
2010-11	1	10734	-2767.33	-2516.62	14	5061.23	5163.17
	2	10734	-2080.36	-1777.98	14	3583.95	3685.89
2015	1	12822	-2887.20	-2580.70	14	5189.41	5293.83
	2	12822	-2141.63	-1759.41	14	3546.83	3651.25

However, **Figure 3.4** with the Cox-Snell residuals shows that Model 1 is better than Model 2 as the estimated cumulative hazards are close to the reference line which is formed by the Cox-Snell residuals. Furthermore, dot charts depicting the importance of variables in the two models were plotted (**Figure 3.5**). Place of residence was the least important variable for all the models, while marital status and sex were the most important variables for Model 1 and Model 2 respectively (**Figure 3.5**).

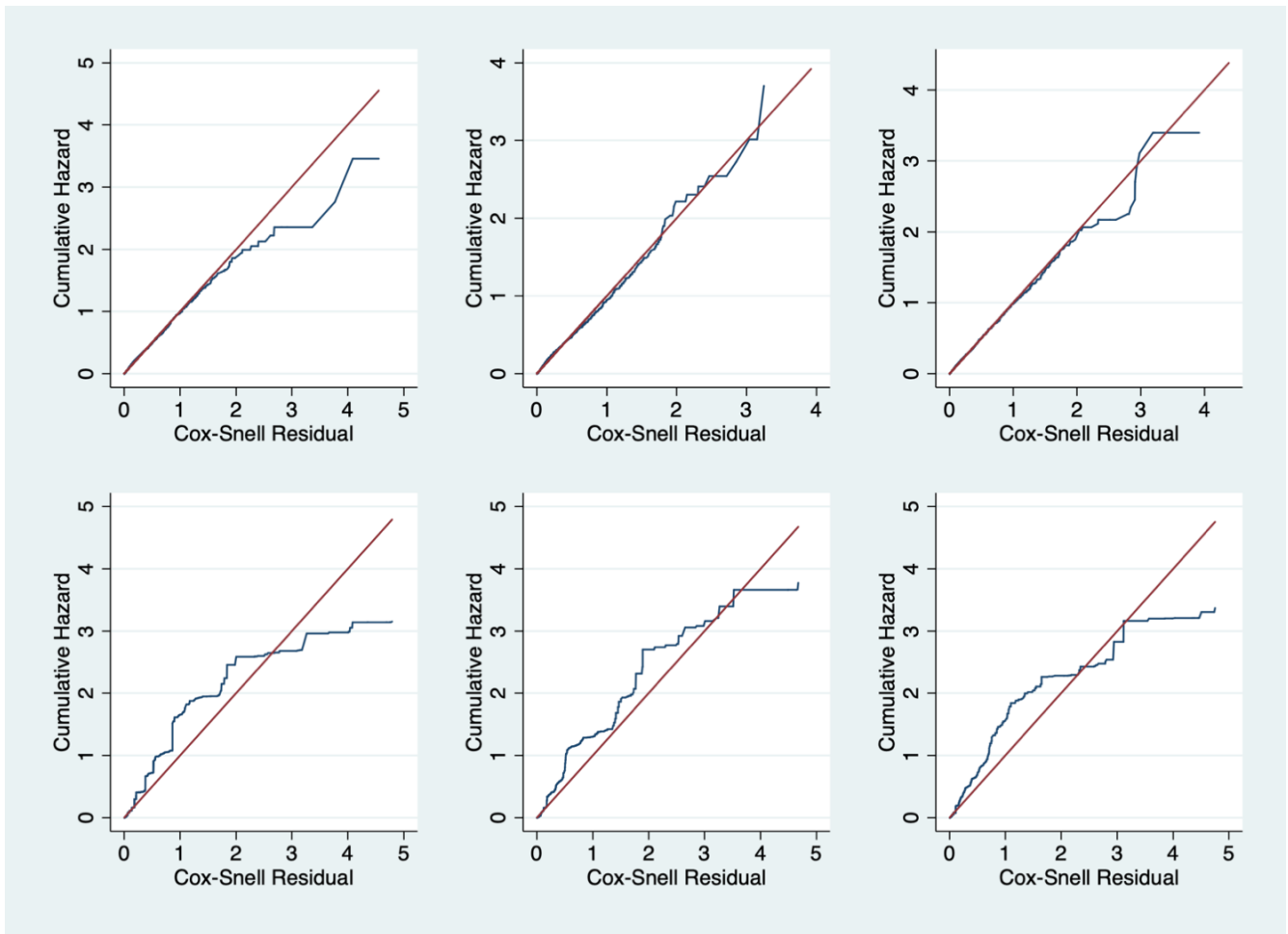


Figure 3.4: Cox-Snell residual against cumulative hazard goodness-of-fit tests for the Zimbabwe DHS 2005-06 (left column), 2010-11 (middle column) and 2015 (right column). Model 1 (top row) and Model 2 (bottom row)

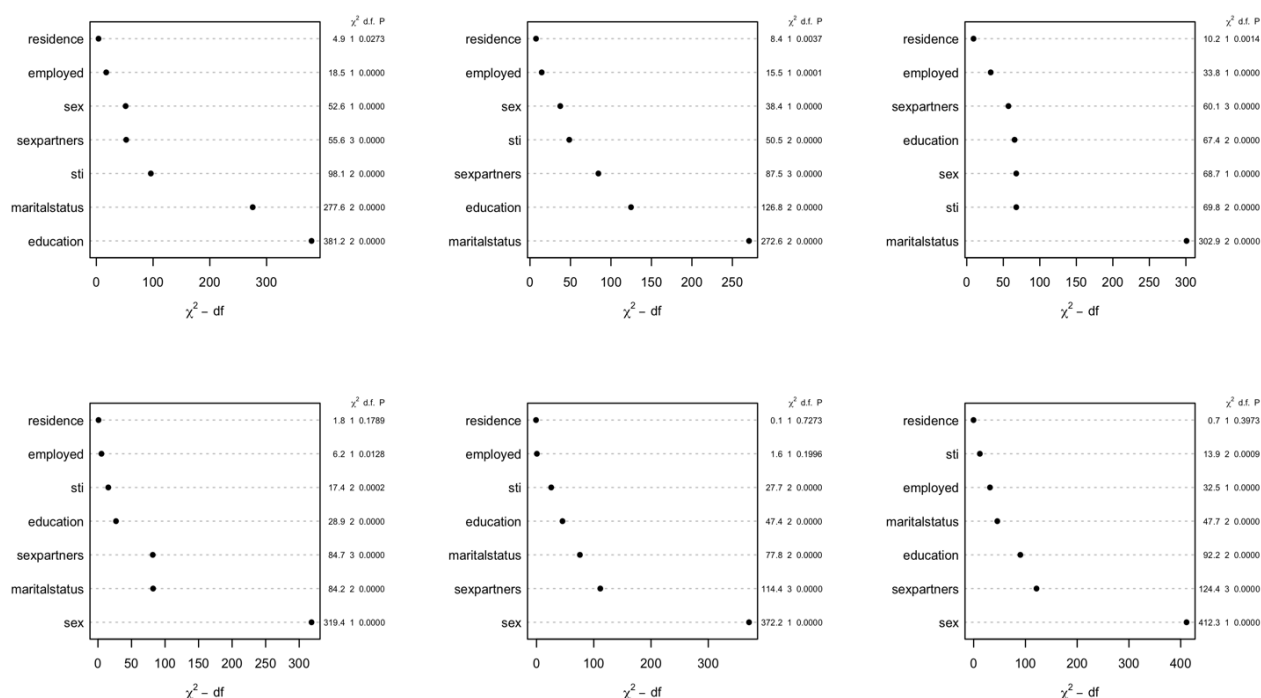


Figure 3.5: Dot chart showing relative importance of covariates for the Zimbabwe DHS 2005-06 (left column), 2010-11 (middle column) and 2015 (right column). Model 1 (top row) and Model 2 (bottom row).

The risk of HIV infection was lower in males than females, as shown in **Table 3.3** for all the models. The risk for HIV infection had a slight decrease from 2005-06 to 2010-11 to 2015 (HR=0.26, 95% CI: 0.23, 0.31), (HR=0.25, 95% CI:0.21,0.29), (HR=0.22, 95% CI:0.19,0.26) respectively with reference to Model 2. Individuals who were married or cohabiting had a lower risk of HIV infection as compared to those who were single. These results were consistent for all three models. However, the risk of HIV infection for all the survey years was lower in individuals who were separated/divorced/widowed in reference to Model 1, but Model 2 results indicated that the risk was higher for that particular group of individuals as compared to those who were single (**Table 3.3**). According to the results

in Model 1, the risk of HIV infection was almost the same for those who had one or more than two sexual partners, as compared to those who did not have any sexual partners in the past 12 months prior to the survey. However, Model 2, the risk was more than four times for those who had more than two sexual partners, as compared to those who did not have any sexual partners in the past 12 months prior to the survey. Of interest, the risk of HIV infection for those who had more than two sexual partners increased over time for Model 2, with Model 1 having a decreasing trend. (**Table 3.3**).

Table 3.3: Estimated effects of covariates at baseline on the risk of HIV infection
based on different survival models, Zimbabwe Demographic Health Survey
(ZDHS) 2005/06, 2010/11 and 2015.

<i>Model</i>	<i>MODEL 1</i>			<i>MODEL 2</i>		
<i>YEAR</i>	2005/06	2010/11	2015	2005/06	2010/11	2015
<i>Variable</i>	HR (CI)	HR (CI)	HR (CI)	HR (CI)	HR (CI)	HR (CI)
<i>Sex</i>						
<i>female</i>	1	1	1	1	1	1
<i>male</i>	0.68 (0.62-0.76)	0.73 (0.66-0.80)	0.66 (0.60-0.73)	0.26 (0.23-0.31)	0.25 (0.21-0.29)	0.22 (0.19-0.26)
<i>Marital Status</i>						
<i>single</i>	1	1	1	1	1	1
<i>married/cohabiting</i>	0.30 (0.25-0.35)	0.32 (0.27-0.38)	0.29 (0.25-0.34)	0.40 (0.33-0.50)	0.45 (0.37-0.55)	0.59 (0.48-0.71)
<i>separated/divorced/widowed</i>	0.64 (0.54-0.76)	0.70 (0.59-0.84)	0.60 (0.50-0.71)	0.69 (0.55-0.87)	0.85 (0.67-1.08)	1.08 (0.84-1.38)
<i>Place of residence</i>						
<i>urban</i>	1	1	1	1	1	1
<i>rural</i>	1.11 (1.01-1.22)	0.87 (0.79-0.95)	0.86 (0.79-0.94)	1.10 (0.96-1.27)	1.03 (0.89-1.19)	1.06 (0.92-1.22)
<i>Education level</i>						
<i>no education/primary</i>	1	1	1	1	1	1
<i>secondary</i>	2.53 (2.29-2.79)	1.62 (1.47-1.79)	1.05 (0.95-1.15)	1.12 (0.97-1.29)	0.91 (0.79-1.07)	0.63 (0.53-0.74)
<i>higher</i>	1.32 (1.01-1.72)	0.75 (0.58-0.97)	0.51 (0.43-0.62)	0.43 (0.30-0.62)	0.30 (0.21-0.43)	0.25 (0.19-0.33)
<i>STI in the last 12 months</i>						
<i>no</i>	1	1	1	1	1	1
<i>yes</i>	2.19 (1.88-2.56)	1.87 (1.57-2.22)	2.15 (1.79-2.58)	1.73 (1.33-2.25)	2.09 (1.58-2.76)	1.80 (1.31-2.46)
<i>No. of sexual partners</i>						
<i>0</i>	1	1	1	1	1	1
<i>1</i>	1.62 (1.42-1.86)	1.84 (1.61-2.12)	1.64 (1.43-1.89)	2.12 (1.71-2.64)	2.32 (1.86-2.89)	2.26 (1.77-2.88)
<i>2+</i>	1.95 (1.54-2.47)	2.19 (1.70-2.67)	2.06 (1.68-2.54)	4.91 (3.39-7.10)	5.58 (3.95-7.89)	6.88 (4.84-9.76)
<i>Currently employed</i>						
<i>no</i>	1	1	1	1	1	1
<i>yes</i>	0.82 (0.75-0.90)	0.83 (0.76-0.91)	0.77 (0.70-0.84)	0.84 (0.74-0.96)	0.92 (0.80-0.99)	0.66 (0.57-0.76)

HR, hazard ratio; CI, confidence interval.

Model 1 was the weibull parametric survival model with survey age imputation as the time and Model 2 was a weibull interval censoring model

3.6 DISCUSSION

Frequently, researchers are interested in using standard survival models in determining the failure times, however interval censoring has become increasingly common. The purpose of this study was to identify risk factors for HIV infection using three different models and determine the best fitting model. To identify the best fitting model, we utilised the Akaike Information Criterion (AIC), where the model with the least AIC value was the best fitting model. We also used graphical methods to ascertain the best fitting model, however graphical goodness-of-fit test for interval-censored data is rare with the available methods still lacking in implementation ¹³⁸⁻¹⁴¹. We managed to check the goodness-of-fit for all the models by overlaying the semi-parametric model with the fitted parametric model of the survival function and also plotting the Cox-Snell residuals against the cumulative hazard.

The model with interval censored data resulted in better estimates of the risk of HIV infection as compared to the standard survival model i.e., Model 2 based on the AIC value and **Figure 3.3**. The superiority of Model 2 was due to the ability to precisely mimic some of the results in literature from previous studies which also used population-based HIV surveys or specific cohorts in Zimbabwe. For example, a study in Zimbabwe conducted between 1999-2001 reported that they observed an increasing trend of HIV incidence among the educated individuals, which was rather unexpected; however, this might have been due to a higher socio-economic status, a factor reported to be associated with HIV infection in the region of Africa ^{142, 143}. However, a systematic review which explored time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa, reported a shift in the HIV epidemic from educated to the

uneducated ¹⁴⁴. According to a study in which they used the 2005/06 Zimbabwe Demographic Health Survey (ZDHS) to determine the relationship between HIV status and the demographic and socio-economic characteristics among adults in Zimbabwe by constructing the risk profile of the average adult, they concluded that there was a significant negative association between HIV infection and education ¹⁴⁵. They further clarified that an extra year of schooling to an average of 8 years (i.e. secondary education and above) was associated with a 0.5 percent point decrease in the probability of HIV infection for Zimbabwe ¹⁴⁵. Another study which used the 2010/11 ZDHS also reported a lower risk of HIV infection of individuals with secondary level education and above ¹⁴⁶. Based on these findings from literature, and the trend observed, the model with interval-censoring was consistent with the reported results, however, the other two models reported a higher risk of HIV infection for individuals with secondary level education, which was contrary to the findings from literature. according to an in-depth analysis of the 2005/06 ZDHS on the risk factors associated with HIV infection, it was reported that individuals who never had any sexually transmitted infection 12 months prior the survey were significantly associated with a 0.437 times lower risk of HIV infection compared with their counterparts who had a sexually transmitted infection during the same period ¹⁴⁷. This result was close to the results obtained for the hazard of HIV infection in the model with interval-censoring. Another example was a study of the analysis of 2005/06 ZDHS, which reported that the likelihood of being HIV infected increased with the number of sexual partners and decreased with the level of faithfulness to a spousal partner. In the same study, it was reported that the odds of being HIV infected were 3 to 4 times greater among those who had two more sexual partner ¹⁴⁸. These results were again consistent with results from the model with interval-censoring rather than the other

two models for the 2005/06 ZDHS. On a general note, a 2013 study noted that having a large number of life partners increased HIV infection in a cohort from Manicaland ¹⁴⁹.

Comparisons of the three models used in this study reveal a consistent match on the factors associated with HIV infection estimated from ZDHS data and the results obtained from previous studies using population-based HIV surveys or specific cohorts in Zimbabwe. For example, in all the three models, and based on all the three surveys, i.e., ZDHS 2005/06, 2010/11 and 2015, females were more at risk of HIV infection than men. Similarly, studies which determined the factors associated with HIV infection using the 2005/06 and 2010/11 ZDHS reported the same findings ¹⁴⁵⁻¹⁴⁷. However, though the surveys were the same, they did not use the same analytical methods to reach the same conclusions. Results suggest that marriage was associated with a lower risk of infection based on all the models. This is further supported by a study which analysed the 2005/06 and 2010/11 ZDHS data ¹⁴⁶. The study reported that marital union was positively associated with the decline of HIV infection for both men and women ¹⁴⁶. Another study to determine the baseline predictions of HIV-1 acquisition among women reported that being unmarried was the strongest risk factor for HIV-1 acquisition ¹⁵⁰. Results from these studies are again consistent with what all the models in our studies. Even though the models reported results similar to what had already been reported in literature, the precision of the model with interval-censoring in explaining some of the covariates is what stood out the most. However, the Cox-Snell residuals clearly showed that Model 1 was the best fitting model. The difference between the Cox-Snell residuals graphical method and the model estimates maybe due to the lack of adequate methods to test the goodness-of-fit of interval-censored data as cited by other authors ¹³⁸⁻¹⁴¹.

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 in reality provides more reliable and robust data. Another strength of this study was due to the fact that we did not restrict our analysis to one method, however, we had the opportunity to determine the best model to fit the hazard of infection by comparing two different scenarios. For instance, if the median survival time for HIV infection was 5 years given the type of data we had, and the intervals were about 3 to 6 months wide, then we would have no reason to complicate the analysis by considering interval censoring. On the other hand, if the intervals were about 1 year or longer, then accounting for uncertainty in the analysis was necessary, which we did when we implemented the interval-censoring approach. Another reason for concluding that interval-censoring gave better estimates was due to its consistency with the published results from literature. Even though we consider the interval-censoring model as the superior model with regards to our specific data, the method had its own set of limitations. These limitations included the wide range of intervals used, which could have underestimated or overestimated the effect of other factors on the risk of HIV infection. Inclusion of competing risks factors in the model would have greatly improved the modelling approach. Further studies can be done on imputation models, which imputes an estimated time of HIV infection based on the data.

3.7 ACKNOWLEDGEMENTS

This work was supported through the Developing Excellence in Leadership Training and Science Africa (DELTA) initiative. The DELTA's Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for

Africa's Development Planning and Coordinating (NEPAD) Agency with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the United Kingdom (UK) government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

3.8 AUTHOR CONTRIBUTIONS

The RB and EM conceived of the presented idea. RB developed the theory and performed the data analysis. RB and EM verified the analytical methods and models. EM supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

3.9 DATA AVAILABILITY STATEMENT

The datasets used for this study can be found on the Demographic Health Survey Program website [<https://dhsprogram.com>].

3.10 CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.11 SUPPLEMENTARY INFORMATION

3.11.1 Model Specification

3.11.1.1 *Model 1*

Weibull regression model can be expressed as a proportional hazards model:

$$\lambda(t|\mathbb{X}) = \lambda_0(t) \exp(X_1\beta_1 + \dots + X_k\beta_k) \quad (\text{S1})$$

where;

$$\lambda_0(t) = \exp(\beta_0)pt^{p-1}$$

or

$$\lambda(t|\mathbb{X}) = pt^{p-1}\exp(\mathbb{X}\beta) \quad (\text{S2})$$

where $\lambda(t|\mathbb{X})$ is the hazard of getting HIV infection at age t for a subject with covariate vector $\mathbb{X}$, $\lambda_0(t)$ is the unspecified baseline hazard and β is the vector of regression coefficient that measured the effect of $\mathbb{X}$ on the hazard of getting HIV infection, p is the Weibull shape parameter and $\pi = \exp(\mathbb{X}\beta)$ is the Weibull scale parameter.

Censored Likelihood Function

Suppose we have n individuals all with observed events t_i (no censoring), then the likelihood function is the product of the PDF:

$$L = \prod_{i=1}^n f(t_i)$$

Given that they are individuals that are still HIV negative at the time the survey is taken, t^* , then the likelihood function is:

$$L = \prod_{t_i \leq t^*}^n f(t_i) \prod_{t_i > t^*}^n S(t_i), \quad (\text{S3})$$

Therefore, the uncensored individuals contribute information to the likelihood function through the event times, and the censored individuals contribute information through the survival function at the endpoint. This can be re-expressed as:

$$\sigma_i \begin{cases} 1 & \text{if } t_i \leq t^* \\ 0 & \text{if } t_i > t^* \end{cases}$$

With the likelihood function as:

$$\prod_{i=1}^n [f(t_i)]^{\sigma_i} [S(t_i)]^{1-\sigma_i} \quad (\text{S4})$$

which shows the bias imposed by omitting right censored individuals.

3.11.1.2 Model 2

We considered $t_1, \dots, t_n$ be the lifetimes for individuals from the 2005/06, 2010/11 and 2015 ZDHS sample of size n where the probability density function (pdf) (for $t_i = t$) is represented by $f(t, \alpha, \beta)$, the cumulative distribution function (cdf) is $F(t, \alpha, \beta)$ and the survival function is $S(t, \alpha, \beta)$ given the two-parameter Weibull distribution we have, respectively,

$$f(t, \alpha, \beta) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right), \quad (\text{S5})$$

$$F(t, \alpha, \beta) = 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right), \quad (\text{S6})$$

$$S(t, \alpha, \beta) = \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right), \quad (\text{S7})$$

Where α represents the scale parameter and β the shape parameter.

Let $[L_i, R_i]$ denote the interval-censored data and let T represent the unknown time of HIV infection, that is, $L_i \leq T_i \leq R_i$, where L_i is the age at first sexual intercourse and R_i the age at survey date. If censoring occurs non-informatively and if the law governing L

and R does not involve any of the parameters of interest, we can base our inferences on the likelihood function $L(L_i, R_i | \alpha, \beta)$ as stated by Oller et al.¹⁵¹, which is given by

$$\begin{aligned} L(L_i, R_i | \alpha, \beta) &= \prod_{i=1}^n [F(R_i, \alpha, \beta) - F(L_i, \alpha, \beta)] \\ &= \prod_{i=1}^n [S(L_i) - S(R_i)] \\ &= \prod_{i=1}^n \text{prob}\{L_i \leq T_i \leq R_i\}. \end{aligned} \quad (\text{S8})$$

Using (S7), we have

$$L(L_i, R_i | \alpha, \beta) = \prod_{i=1}^n \left\{ \exp\left(-\left(\frac{L_i}{\alpha}\right)^\beta\right) - \exp\left(-\left(\frac{R_i}{\alpha}\right)^\beta\right) \right\}. \quad (\text{S9})$$

And taking the natural log of (S9), we have

$$\ell = \sum_{i=1}^n \ln \left\{ \exp\left(-\left(\frac{L_i}{\alpha}\right)^\beta\right) - \exp\left(-\left(\frac{R_i}{\alpha}\right)^\beta\right) \right\}. \quad (\text{S10})$$

For us to find the values of α and β that maximise (S10), we differentiate the equation to obtain partial derivatives with respect to α and β and set the resulting equations to zero.

Therefore,

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n - \frac{\left(\left(\frac{L_i}{\alpha}\right)^\beta \ln\left(\frac{L_i}{\alpha}\right) \exp\left(-\left(\frac{L_i}{\alpha}\right)^\beta\right) \right) - \left(\left(\frac{R_i}{\alpha}\right)^\beta \ln\left(\frac{R_i}{\alpha}\right) \exp\left(-\left(\frac{R_i}{\alpha}\right)^\beta\right) \right)}{\exp\left(-\left(\frac{L_i}{\alpha}\right)^\beta\right) - \exp\left(-\left(\frac{R_i}{\alpha}\right)^\beta\right)} = 0 \quad (\text{S11})$$

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^n - \frac{\left(\left(\frac{L_i}{\alpha}\right)^\beta \left(\frac{\beta}{\alpha}\right) \exp\left(-\left(\frac{L_i}{\alpha}\right)^\beta\right) \right) - \left(\left(\frac{R_i}{\alpha}\right)^\beta \left(\frac{\beta}{\alpha}\right) \exp\left(-\left(\frac{R_i}{\alpha}\right)^\beta\right) \right)}{\exp\left(-\left(\frac{L_i}{\alpha}\right)^\beta\right) - \exp\left(-\left(\frac{R_i}{\alpha}\right)^\beta\right)} = 0 \quad (\text{S12})$$

The maximum likelihood estimates of α and β can then be obtained by using the Newton Raphson method.

CHAPTER 4

THESIS PAPER 2: Predicting HIV infection in the decade (2005-2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach.

Rutendo Beauty Birri Makota^{1*}, Eustasius Musenge¹

¹ Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

* Corresponding author

E-mail: rutendobbirri@gmail.com

Published paper:

Birri Makota RB, Musenge E (2023) Predicting HIV infection in the decade (2005– 2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. PLOS Digital Health 2(6): e0000260. <https://doi.org/10.1371/journal.pdig.0000260>

Abstract

The burden of HIV and related diseases have been areas of great concern pre and post the emergence of COVID-19 in Zimbabwe. Machine learning models have been used to predict the risk of diseases, including HIV accurately. Therefore, this paper aimed to determine common risk factors of HIV positivity in Zimbabwe between the decade 2005 to 2015. The data were from three two staged population five-yearly surveys conducted between 2005 and 2015. The outcome variable was HIV status. 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. Feature selection was done using Lasso regression, and the best combination of selected features was determined using Sequential Forward Floating Selection. We compared six algorithms in both sexes based on the F1 score, which is the harmonic mean of precision and recall. The overall HIV prevalence for the combined dataset was 22.5% and 15.3% for females and males, respectively. The best-performing algorithm to identify individuals with a higher likelihood of HIV infection was XGBoost, with a high F1 score of 91.4% for males and 90.1% for females. 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). In addition to other risk reduction techniques, machine learning may aid in identifying those who might require Pre-exposure prophylaxis, particularly women who experience intimate partner violence. Furthermore, compared to traditional statistical approaches, machine learning uncovered

patterns in predicting HIV infection with comparatively reduced uncertainty and, therefore, crucial for effective decision-making.

Keywords: Demographic Health Survey, HIV infection, HIV risk, Intimate partner violence, Machine Learning, XGBoost

Author Summary

The Joint United Nations Programme (UNAIDS) set up fast-track targets to reach HIV epidemic control by 2030, where it is expected that 95% of people living with HIV know their status, and of those, 95% should be on treatment and of those on treatment 95% should have reached viral suppression. In Zimbabwe, by 2020, it was found that 86.8 percent of adults living with HIV were aware of their status and of those aware of their status, 97.0 percent were on antiretroviral treatment. Furthermore, among those on treatment, 90.3 percent achieved viral load suppression. In order to achieve these targets, modern predictive algorithms using machine learning have the power to enhance HIV prevention and prediction capability. Furthermore, studies have reported that machine learning could accurately predict future HIV infection. This paper, therefore, aims to use these machine learning tools to predict key HIV populations based on socio-behavioural characteristics obtained from survey data.

4.1 INTRODUCTION

In the era before and after COVID-19, HIV ranks amongst the most serious infectious diseases globally. Despite the tremendous advancements in diagnosis and access to antiretroviral therapy (ART), Zimbabwe Population-based HIV Impact Assessment survey (ZIMPHIA 2020) predictions suggest that 1.23 million adults were living with HIV and that the incidence rate is about 0.21% annually ¹⁵². HIV/AIDS as a threat to the public's health is to be eradicated by 2030, according to the Joint United Nations Programme (UNAIDS) ^{153, 154}. The COVID-19 pandemic, however, was already reversing the gains obtained, and it may have a negative impact by increasing the number of AIDS-related fatalities in sub-Saharan Africa ^{155, 156}. Furthermore, the HIV epidemic is unevenly spread throughout several geographic locations, with higher prevalence in some areas and among particular populations. Focused interventions are necessary for this situation to manage the HIV pandemic effectively since they have been found to maximise the prophylactic benefit at the lowest possible cost ¹⁵⁷. Therefore, it is essential to understand where and among which groups new infections develop to inform targeted interventions.

Monitoring and surveillance techniques, such as behavioural risk assessments, interviews, and laboratory test results, should form the basis of a nation's HIV response ¹⁵⁸. Unfortunately, most nations' surveillance systems, including structured government surveillance and ad hoc surveys, are insufficient to accurately monitor epidemic and risk patterns ¹⁵⁸. In addition, there is frequently a reporting lag before data are made public, and data collecting and aggregation techniques are time- and resource-intensive ¹⁵⁹.

Non-requirement of statistical inferences or assumptions is one advantage of machine learning algorithms for developing predictive models. Since machine learning algorithms are data-driven, their greatest benefit is their ability to automatically learn from data that identifies complex nonlinear patterns and exploits complex interactions between risk factors. Machine learning models have been used to predict the future risk of other conditions ¹⁶⁰⁻¹⁶⁵, including HIV. Studies have reported that machine learning could accurately predict future HIV infection ^{113, 153, 166-168}.

Three primary studies have been published that used machine learning methods to predict HIV status in Zimbabwe ^{153, 169, 170}. First, Mutai et al. ¹⁵³ predicted HIV status using Demographic Health Survey data from Sub-Saharan Africa. The results obtained from the study by Mutai et al. ¹⁵³ were not specific to Zimbabwe but sub-Saharan Africa. Second, using data from the Zimbabwe Ministry of Health and Child Care, Chingombe et al. ¹⁶⁹ predicted HIV status among men who had sex with men in Zimbabwe's two major cities, Bulawayo and Harare. The findings of this study were limited to men who had sex with men in Bulawayo and Harare and could not be generalised to other cities in Zimbabwe or the general population. Third, using nationally representative data from a cross-sectional survey from the Zimbabwe Population-Based HIV Impact Assessment (ZIMPHIA15-16), again, Chingombe et. ¹⁷⁰ predicted the HIV status using machine learning techniques. Based on the three primary studies that employed machine learning techniques to estimate the HIV status in Zimbabwe, our study and findings will add to the existing body of knowledge. In contrast to Mutai et al.'s ¹⁵³ research, the findings of our study will be more applicable to Zimbabwe. In addition, unlike the two studies by Chingombe et al. ^{169, 170}, we have access to a vast quantity of data from 2005-2015, i.e., ten years of DHS data. Therefore, the primary aim of this paper was to determine common

risk factors of HIV positivity in Zimbabwe. Secondly with the aid of machine learning algorithms, these risk factors were used to formulate a model that predicts HIV positivity.

4.2 METHODOLOGY

4.2.1 Data and study design

Situated in the southern parts of Africa, Zimbabwe is a landlocked nation with a population of 15.1 million people based on the 2022 census ¹⁷¹. Three nationally representative surveys were held in 2005-06, 2010-11, and 2015 and all named Zimbabwe Demographic Health Survey (ZDHS). The data utilised for analysis in this paper can be obtained from the Demographic Health Survey (DHS) programme website (<https://dhsprogram.com>)¹⁷². Since 1984, around 70 nations have conducted DHS, which are nationally representative household surveys ¹⁰⁴⁻¹⁰⁶. In the fields of population, health, and nutrition, they offer data for various monitoring and impact evaluation. Accordingly, blood samples were taken in all households with the consent of the respondent or parent/guardian (for minors) for HIV testing in the lab for females aged 0-49 and males aged 0-54. This was a retrospective cross-sectional study design that employed secondary data analysis.

4.2.2 Data Pre-Processing

The datasets from the three ZDHS HIV test results and the datasets from adult interviews were merged, and records without an HIV test result were excluded from the analysis. The outcome variable, HIV status, was divided into two categories (0 for HIV negative and 1 for HIV positive). To account for non-response, non-coverage, and population total

adjustment weights, data were resampled utilising sample weights of HIV test results for each survey year. Variables with a correlation coefficient of 80% or higher, non-unique columns, arbitrary features, more than 30% missing values, and uninformative features were all eliminated. The label-code and one-hot encoding methods were used to encode nominal and ordinal variables derived from survey data. As part of step 1, represented in **Fig 4.1**, missing values were imputed using Multiple Imputations with Chained Equations (MICE), and in each of these categories, we assumed that missing was at random. The data were further normalised and scaled.

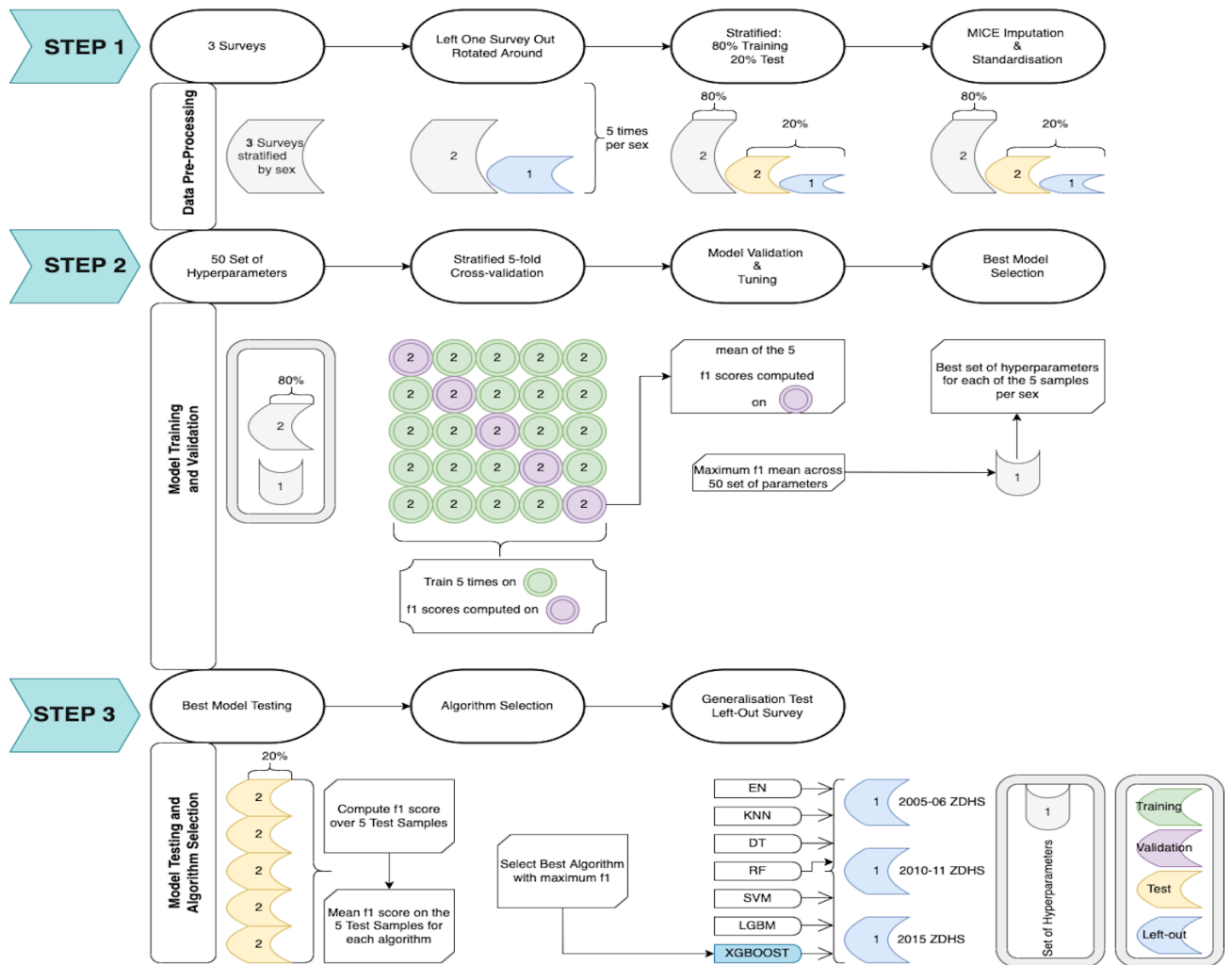


Figure 4.1: Flowchart of the methods and steps used.

4.2.3 Train, Test and Validation Procedure

The data were stratified by sex, and subsequent data training, testing and validation were done following the stratification. Following step 1 from **Fig 4.1**, one survey year was left out, with all the survey years being rotated and left out to produce three distinct datasets for each sex, each with only two surveys. The primary goal of the rotation was to evaluate the generalisation of the models separately for males and females in later testing. Eighty percent of the three newly constructed datasets for each sex were selected for training, and 20% were utilised as test samples. To prevent the test dataset from being contaminated, MICE imputation and data standardisation were carried out independently for the training and test datasets.

4.2.4 Feature Selection

Feature selection was done in two stages. We first used LASSO (Least Absolute Shrinkage and Selection Operator) regression to select features important in determining HIV status using combined dataset from the three surveys. Features selected through LASSO regression were then used to fit the algorithms under investigation. On variables selected through LASSO regression, subsequent analysis utilised the method with the algorithm with the highest F1 score and analysis was done separately for each survey year stratified by sex. We utilised the sequential forward floating selection (SFFS) approach with 80% of the training samples to select the final features to use in best algorithm among the six algorithms. Feature selection using the SFFS was implemented by considering features whose F1 score plateaued from the saturation point. Using SHapley Additive exPlanations (SHAP) ¹¹⁴, the contribution of each feature selected through SFFS to the probability of being HIV-positive was then examined.

4.2.5 Data Balancing

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. This was achieved by generating synthetic data using the nearest neighbour's algorithm to balance out the classes. Original dataset without implementing the SMOTE procedure was also analysed to compare results.

4.2.6 Machine Learning Models

A supervised machine learning binary classification was implemented. The following machine learning algorithms were compared: a penalised logistic regression (Elastic Net), k-nearest neighbour algorithm (KNN), Random Forest Classifier (RFC), Decision Tree approach (DT), Light Gradient Boosting Model (LightGBM) and the XGBoost model. To achieve step 2, training datasets were used, and hyperparameter tuning from a grid of 50 sets was performed using a randomised grid search. Then, using the stratified five-fold cross-validation technique over the validated samples, the average F1 scores for each of these 50 sets was calculated, and the most powerful set of hyperparameters was chosen.

Each of the best three models (for each survey year) by sex and algorithm was then run on the corresponding test dataset described in Step 3, **Fig 4.1**, and the resulting metric scores were averaged. Next, the algorithm with the best average F1 score was selected

from the three test datasets. Finally, each selected model was applied to the survey left-out dataset.

4.2.7 Algorithm Evaluation

Due to the unequal class distribution of the classification variable with a ratio of 4.2:1 in the original dataset, the Precision-Recall curve is recommended over the Receiver Operating Characteristic (ROC) curve ¹⁷³ due to its insensitivity to imbalanced datasets. Other metrics were also considered to evaluate the algorithm's performance: accuracy, F1, precision and recall scores. To determine the ratio of correct classification, accuracy score was used. To determine the prediction of true positive cases, precision was used. To determine the proportion of positive cases that are correctly predicted, recall score was used. Lastly, the F1 score which is the weighted harmonic mean of recall and precision was used to determine the predictive power of the algorithm. An algorithm with the highest F1 score was considered to have the best predictive power. Finally, each evaluation used the same segmentation and repetition of data to ensure a fair comparison of models.

4.2.8 Statistical Analysis

The features that were chosen using the SFFS XGBoost procedure were subjected to a logistic regression analysis. This step was added to confirm the outcomes that the SHAP plots yielded. In addition to the results of the SHAP plots, the logistic regression also provided the magnitude and direction of the risk of contracting HIV.

4.3 RESULTS

4.3.1 Summary statistics

The overall HIV prevalence for the combined dataset, as shown in **Table 4.1**, was 22.5% and 15.3% for females and males, respectively. Those who were not currently employed had a higher prevalence (54.1%) than those who were currently employed (45.9%) for females. However, the opposite relationship is seen in males, where those currently employed had a higher HIV prevalence (74.2%) than those not currently employed (25.8%). Adults had a higher HIV prevalence than youths for both sexes. Of interest was a higher HIV prevalence noticed in those who had the highest level of education as secondary for both sexes.

Table 4.1: Background characteristics of ZDHS data

Characteristic	FEMALES	MALES
	HIV Positive n(%)	HIV Positive n(%)
<i>Overall summary</i>	4457(22.5)	2215(15.3)
<i>Survey year: 2005-06</i>	1604(36.0)	760(34.3)
<i>2010-11</i>	1332(29.9)	733(33.1)
<i>2015</i>	1521(34.1)	722(32.6)
<i>Current age: youths (15-24 years)</i>	730(16.4)	179(8.1)
<i>Adults (25-54 years)</i>	3727(83.6)	2036(91.9)
<i>Marital Status: single</i>	302(6.8)	183(8.3)
<i>cohabiting/married</i>	2595(58.2)	1723(77.8)
<i>Divorced/separated/widowed</i>	1560(35.0)	309(14.0)
<i>Place of Residence: urban</i>	1625(36.5)	877(39.6)
<i>rural</i>	2832(63.5)	1338(60.4)
<i>Education level: no education</i>	123(2.8)	23(1.0)
<i>primary</i>	1510(33.9)	612(27.6)
<i>secondary</i>	2674(60.0)	1431(64.6)
<i>higher</i>	150(3.4)	149(6.7)
<i>STI treatment in past 12 months: no</i>	4127(92.7)	2102(94.9)
<i>yes</i>	325(7.3)	113(5.1)
<i>Number of unions: once</i>	2942(70.9)	1059(52.2)
<i>more than once</i>	1210(29.1)	970(47.8)
<i>Currently employed: no</i>	2410(54.1)	571(25.8)
<i>yes</i>	2047(45.9)	1644(74.2)
<i>Wealth index: poorest</i>	781(17.5)	366(16.5)
<i>poorer</i>	727(16.3)	411(18.6)
<i>middle</i>	929(20.8)	350(15.8)
<i>richer</i>	1198(26.9)	603(27.2)
<i>richest</i>	822(18.4)	485(21.9)

4.3.2 Feature Selection

LASSO regression was performed to determine the possible predictors of HIV infection.

A total of 32 predictors were considered for both males and females (see Supplementary Information for list of predictors). Out of the 32 predictors, type of place of residence, religion, number of household members, relationship to household head, sex of household head, age of household head, times away from home in last 12 months, wealth index,

current contraceptive method, current contraceptive by method type, currently/formerly/never in union, number of unions, cohabitation duration (grouped), time since last sex (in days), recent sexual activity, beating justified, ever heard of AIDS, reduce risk of getting HIV, had any STI in last 12 months, ever been tested for HIV, wife justified asking husband to use condom if he has STI, total lifetime number of sex partners, can ask partner to use condom, were selected as predictors of HIV acquisition for females. While, type of place of residence, highest educational level, religion, number of household members, age of household head, times away from home in last 12 months, current contraceptive method, number of injections in last 12 months, respondent circumcised, currently/formerly/never in union, cohabitation duration (grouped), currently working, beating justified, reduce risk of getting HIV, had any STI in last 12 months, wife justified asking husband to use condom if he has STI, total lifetime number of sex partners, were selected as predictors of HIV acquisition for males.

4.3.3 Algorithm comparison

Model/algorithm's performance comparison was made based on the LASSO selected features. The comparison was done in two-phases, 1) using the original data and 2) using the SMOTE processed data. **Table 4.2** describes the original and SMOTE data samples for the training, testing and validation sets for the three survey years stratified by sex.

Table 4.2: Description of original data and SMOTE-processed data.

		MALES DATA				FEMALES DATA			
Dataset	Survey Year	Majority class	Minority Class	Total Samples	Ratio	Majority class	Minority Class	Total Samples	Ratio
Training	2005-06	7124	1164	8288	6:1	8809	2282	11091	4:1
	2010-11	6684	1186	7870	5:1	8176	2500	10676	3:1
	2015	5739	1194	6933	5:1	7583	2349	9932	3:1
Training-smote	2005-06	7124	7124	14248	1:1	8809	8809	17618	1:1
	2010-11	6684	6684	13368	1:1	8176	8176	16352	1:1
	2015	5739	5739	11478	1:1	7583	7583	15166	1:1
Testing	2005-06	1782	291	2073	6:1	2202	571	2773	4:1
	2010-11	1672	296	1968	5:1	2045	625	2670	3:1
	2015	1435	299	1734	5:1	1897	587	2484	3:1
Testing-smote	2005-06	1782	1782	3564	1:1	2202	2202	4404	1:1
	2010-11	1672	1672	3344	1:1	2045	2045	4090	1:1
	2015	1435	1435	2870	1:1	1897	1897	3794	1:1
Validation	2005-06	3312	760	4072	4:1	4345	1604	5949	3:1
	2010-11	3862	733	4595	5:1	5135	1332	6467	4:1
	2015	5044	722	5766	7:1	5876	1521	7397	4:1
Validation-smote	2005-06	3312	3312	6624	1:1	4345	4345	8690	1:1
	2010-11	3862	3862	7724	1:1	5135	5135	10270	1:1
	2015	5044	5044	10088	1:1	5876	5876	11752	1:1

Six algorithms were trained and tested using year-specific datasets on the original and SMOTE processed data. **Table 4.3** gives results of the performance of the prediction capabilities of the algorithms. The SMOTE processed data performed better in all metrics than the original data. The precision and the F1 score were extremely low for most of the algorithms in the original data. This means that the original data had many false positives from the low precision and low accuracy based on the low F1 score. Overall, the XGBoost was best performing algorithm in both the original data and SMOTE processed data for all the survey year for each sex.

Table 4.3: Results of classification models in original imbalanced and SMOTE-processed data

		SMOTE-processed data				Original unbalanced data			
	MALES								
Survey Year	Models	Accuracy %	Precision %	Recall %	F-Score %	Accuracy %	Precision %	Recall %	F-Score %
2005-06	DT	83	87%	78	83	65	25	77	38
2010-11		82	88	75	81	67	29	80	42
2015		81	83	77	80	68	32	74	44
2005-06	EN	64	59	98	71	63	22	65	33
2010-11		64	59	95	72	60	23	69	34
2015		63	59	80	68	57	25	73	37
2005-06	KNN	86	92	78	84	93	79	65	72
2010-11		86	91	80	85	93	81	69	75
2015		82	88	75	81	90	77	63	77
2005-06	RFC	83	79	90	84	70	29	78	42
2010-11		81	78	87	82	66	28	80	42
2015		83	80	86	83	69	32	71	44
2005-06	LIGHTGBM	94	98	90	93	84	45	77	57
2010-11		94	98	89	94	82	45	75	56
2015		92	98	86	91	81	46	71	56
2005-06	XGBOOST	95	95	96	95	93	75	74	74
2010-11		94	95	94	94	92	71	76	73
2015		94	93	94	94	90	72	71	71
	FEMALES								
2005-06	DT	83	86	79	82	72	40	74	52
2010-11		81	84	76	80	70	42	72	53
2015		81	83	79	81	72	45	70	54
2005-06	EN	63	58	88	70	72	39	62	48
2010-11		61	58	80	67	67	38	60	46
2015		59	55	90	69	67	38	63	47
2005-06	KNN	83	90	74	81	89	79	66	72
2010-11		83	89	75	81	89	80	68	74
2015		81	88	72	80	89	81	68	74
2005-06	RFC	84	84	84	84	75	44	70	54
2010-11		81	81	81	81	73	45	68	54
2015		83	83	83	83	75	47	70	56
2005-06	LIGHTGBM	91	96	86	91	82	55	73	63
2010-11		90	94	84	89	79	54	72	62
2015		90	95	85	89	79	54	72	61
2005-06	XGBOOST	93	93	93	93	90	74	76	75
2010-11		92	92	93	92	88	73	77	75
2015		92	92	92	92	88	75	76	76

ROC curves and Precision-Recall curves of the six algorithms for all survey years per sex were shown in **Figs 4.2 and 4.3**. The results showed that the XGBoost model had better performance compared to the other 5 models for both the original data and SMOTE processed data. The Elastic Net model performed poorly compared to the other models. Additional ROC and Precision-Recall curves for the six algorithms based on combined survey years can be found in Supplementary Information.

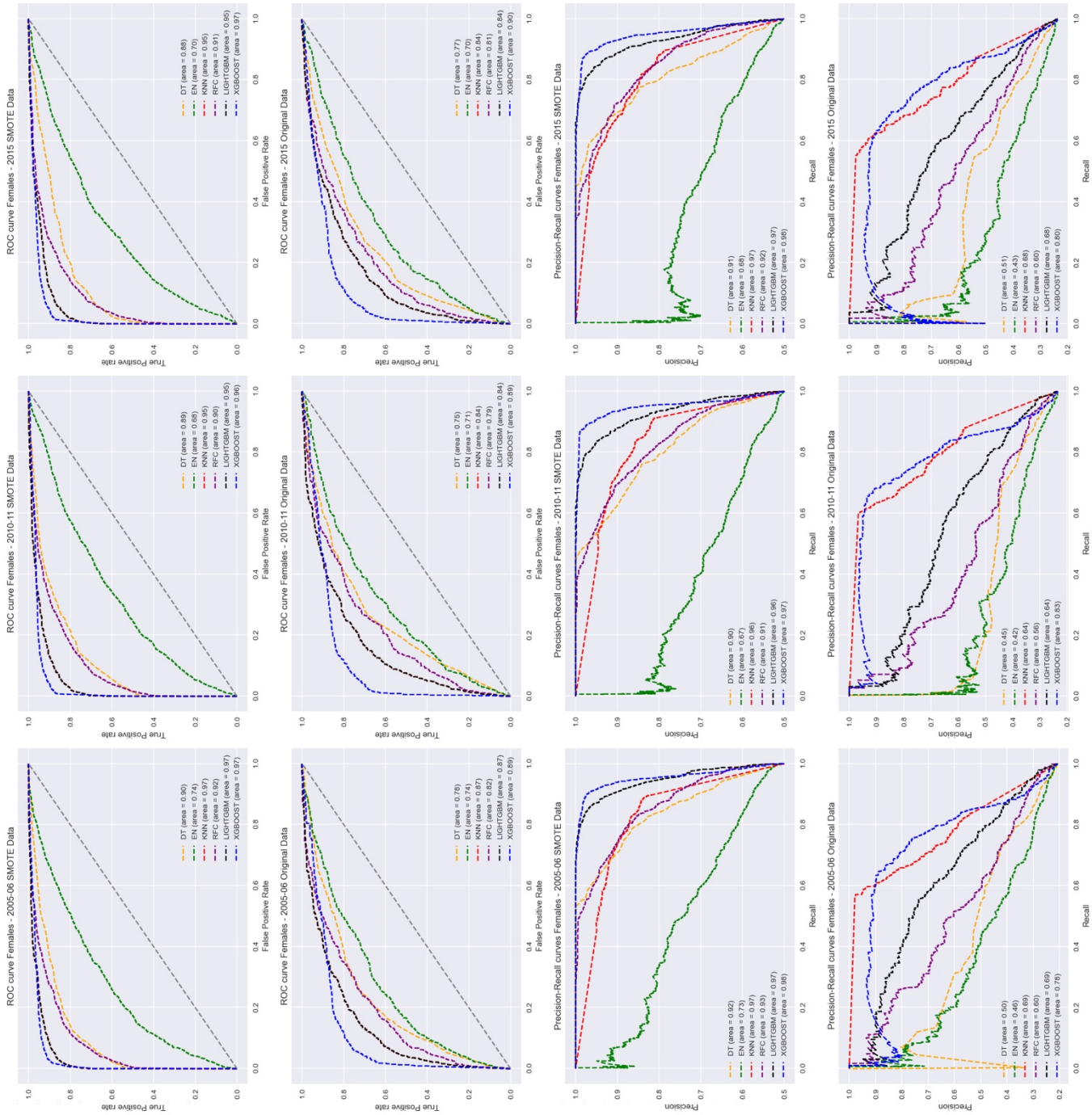


Figure 4.2: ROC curves and Precision-Recall curves of the six algorithms for all survey years for females' data.

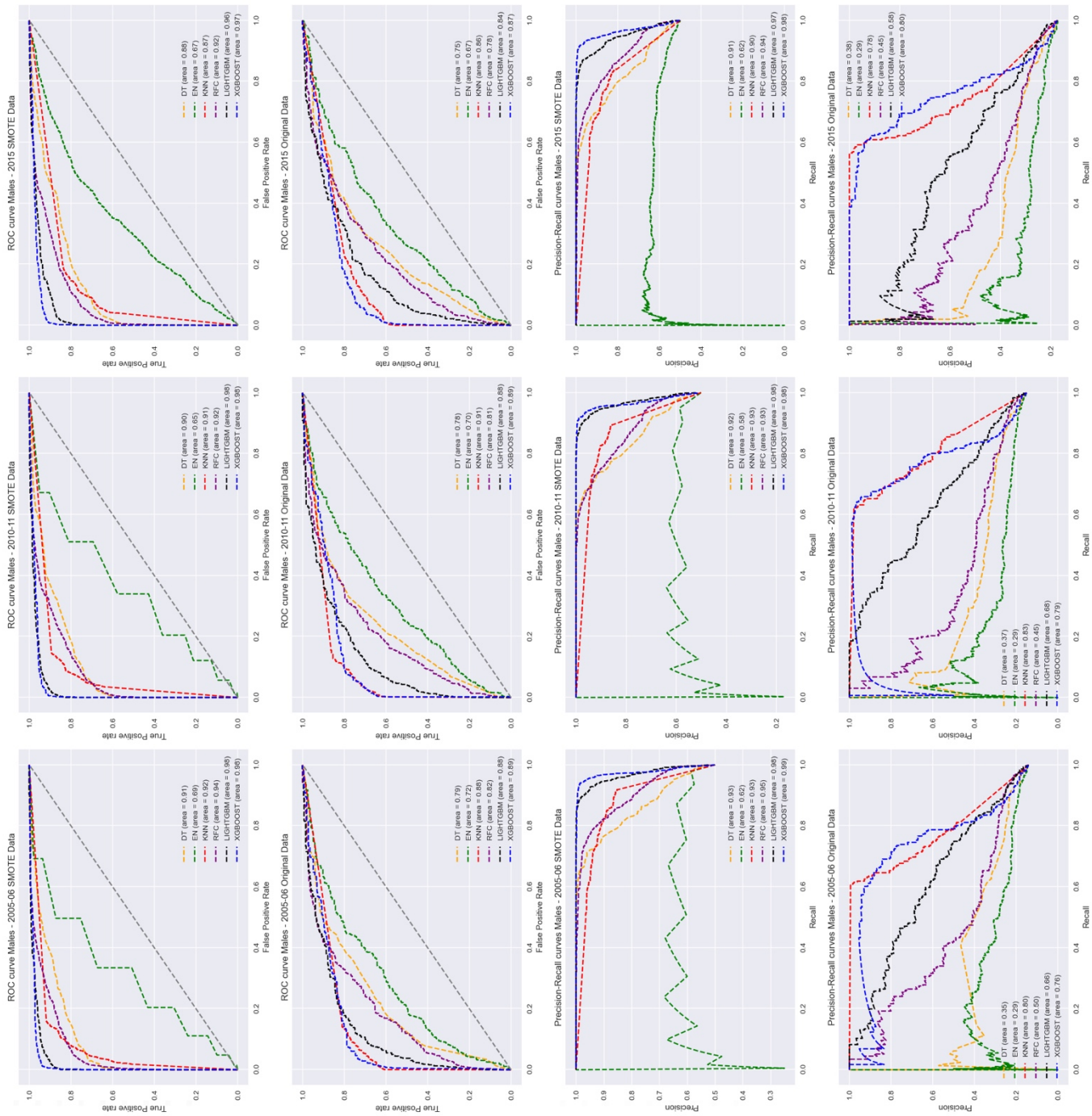


Figure 4.3: ROC curves and Precision-Recall curves of the six algorithms for all survey years for males' data.

Subsequent results were obtained utilising the best algorithm model which was the XGBoost. From the features obtained through LASSO, SFFS procedure was implemented on XGBoost to further determine features associated with HIV positivity.

Fig 4.4 shows the saturation limit obtained by the SFFS procedure for selecting variables based on F1 scoring, with 17, 17 and 14 most influential features selected for females, in the 2005-06, 2010-11 and 2015 ZDHS respectively. Whereas, 13, 13, 11 most influential features were selected for males, in the 2005-06, 2010-11 and 2015 ZDHS respectively.

The F1 score plateaued at 99.9% for both males and females for all survey years.

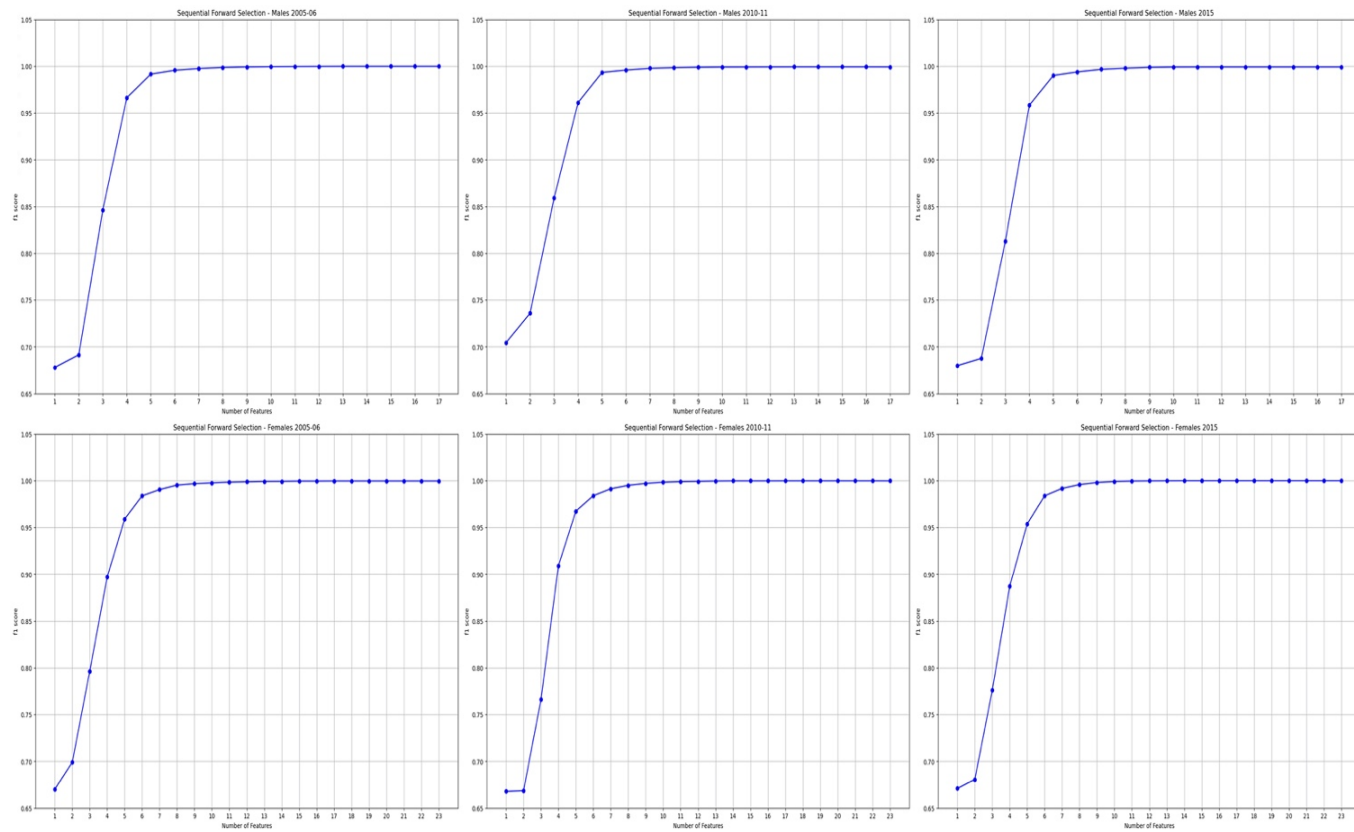


Figure 4.4: Sequential floating forward selection (SFFS) for males and females.

Figs 4.5 and **4.6** displays the subset of features most important to predicting an individual's HIV status following the SFFS procedure. As seen in **Figs 4.5** and **4.6**, the variables are ordered from most important to least important (from highest to lowest Shapley value). According to **Fig 4.5**, features for predicting HIV status which appeared throughout for the three survey years for females were: total lifetime number of sex partners, cohabitation duration (grouped), number of household members, age of household head, time since last sex (in days), times away from home in last 12 months, beating justified, can ask partner to use condom, wealth index, reduce risk of getting HIV, wife justified asking husband to use condom if he has STI and religion.

Features for predicting HIV status which appeared throughout for the three survey years for males were: cohabitation duration (grouped), total lifetime number of sex partners, number of household members, age of household head, times away from home in last 12 months, type of place of residence, beating justified, currently working and religion. 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).

A higher number of total lifetime number of sex partners, more years cohabitating, fewer household members, older age of household head, higher number of days since last sex, fewer times away from home, whether beating was justified or not, living in the rural areas, wearing a condom as current contraceptive method, having a female household head, many unions, can ask partner to use condom, higher wealth index, whether knowing

the risk of getting HIV or not, ever been tested, wife justified asking husband to use condom if he has STI and being affiliated to a religion were factors associated with HIV positivity for females in the 2005-06 ZDHS survey. A higher number of total lifetime number of sex partners, more years cohabitating, fewer household members, higher number of days since last sex, younger age of household head, fewer times away from home, never been tested, weather beating was justified or not, living in the rural areas, not knowing the risk of getting HIV, having a female household head, can ask partner to use condom, wearing a condom as current contraceptive method, relationship to household head, higher wealth index, not affiliated to a religion and wife justified asking husband to use condom if he has STI were factors associated with HIV positivity for females in the 2010-11 ZDHS survey. A higher number of total lifetime number of sex partners, more years cohabitating, fewer household members, younger age of household head, fewer number of days since last sex, fewer times away from home, living in the rural areas, beating was justified, never been tested, can ask partner to use condom, not knowing the risk of getting HIV, higher wealth index and wife justified asking husband to use condom if he has STI were factors associated with HIV positivity for females in the 2015 ZDHS survey.

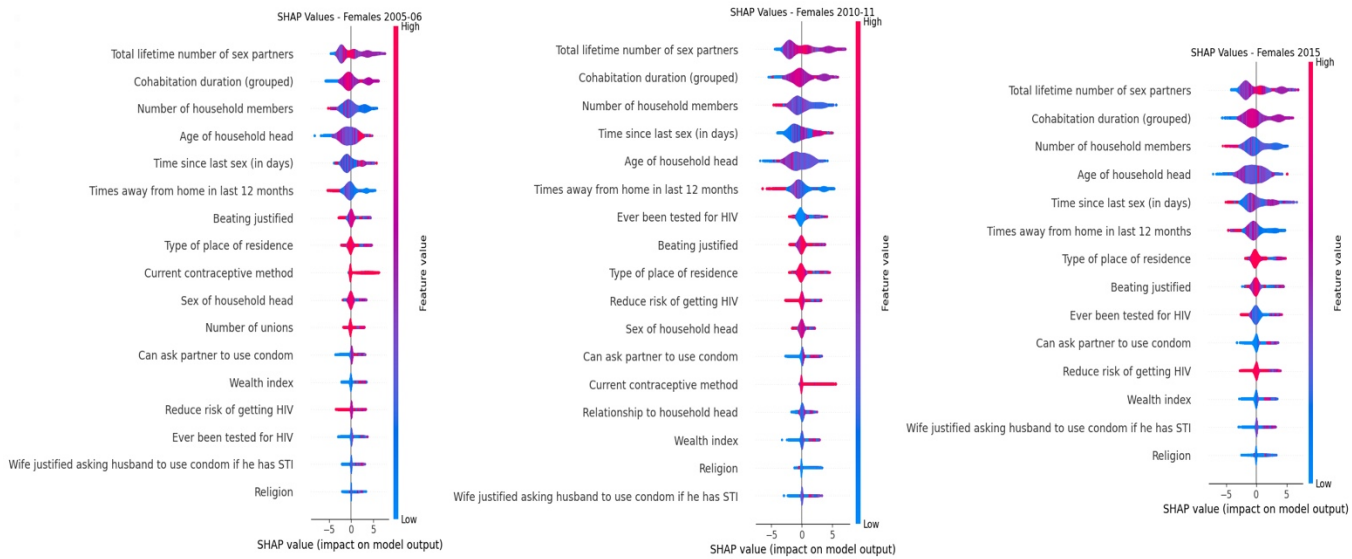


Figure 4.5: Shapley values for 2005-06, 2010-11 and 2015 Females ZDHS data.

Based on **Fig 4.6**, a higher number of years cohabitating, fewer number of total lifetime number of sex partners, fewer household members, younger age of household head, fewer times away from home, not wearing a condom as current contraceptive method, living in the urban areas, beating was justified, currently working, being affiliated to a religion, less number of injections in the last 12 months, respondent not circumcised and wife justified asking husband to use condom if he has STI and were factors associated with HIV positivity for males in the 2005-06 ZDHS survey. A higher number of years cohabitating, fewer number of total lifetime number of sex partners, fewer household members, older age of household head, fewer times away from home, not wearing a condom as current contraceptive method, living in the urban areas, beating was justified, being affiliated to a religion, respondent not circumcised and had an STI in the last 12 months were factors associated with HIV positivity for males in the 2010-11 ZDHS survey. A higher number of years cohabitating, fewer number of total lifetime number of

sex partners, fewer household members, older age of household head, fewer times away from home, living in the urban areas, fewer number of injections in the last 12 months, beating not justified, being affiliated to a religion, currently working and wife justified asking husband to use condom if he has STI were factors associated with HIV positivity for males in the 2015 ZDHS survey.

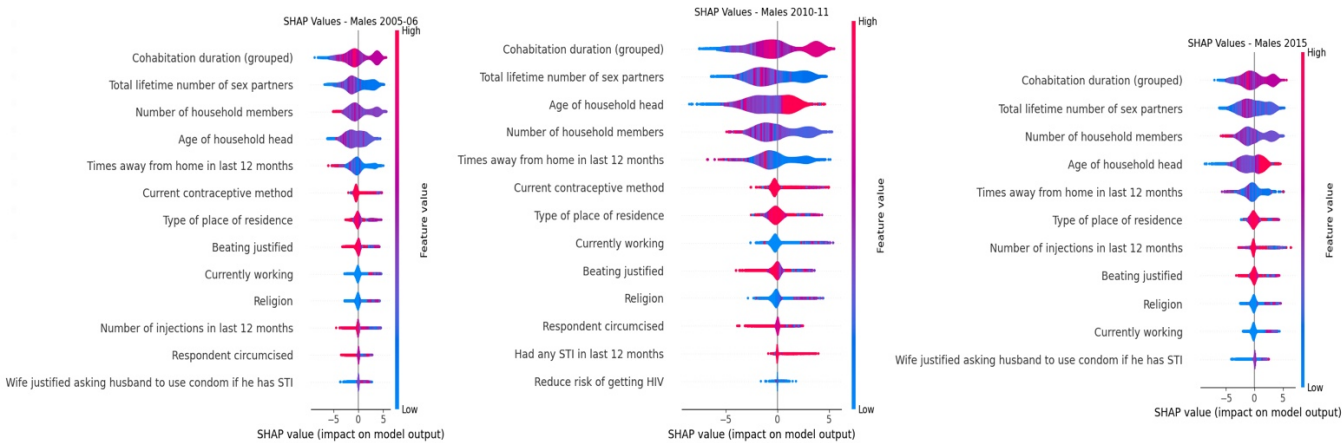


Figure 4.6: Shapley values for 2005-06, 2010-11 and 2015 Males ZDHS data.

Table 4.4 shows the odds ratio of HIV infection performed on the selected features obtained through SFFS for females. The results in **Table 4.4** agree with the results shown in **Fig 4.5**. For instance, each additional increase in the number of sex partners is associated with a 32% (OR: 1.32, $p<0.001$), 11% (OR: 1.11, $p=0.04$) and 41% (OR: 1.11, $p<0.001$) increase in the odds of one being HIV infected for 2005-06, 2010-11 and 2015 ZDHS females data, which is corroborated by **Fig 4.5** were total lifetime number of sex partners is associated with HIV positivity as the number of total lifetime sex partners increases for females in all survey years data. **Table 4.4** also shows a protective effect of HIV infection in individuals who live in rural areas than those who reside in urban areas.

This, again, is corroborated by **Fig 4.5**, where living in urban areas is associated with HIV positivity.

Interestingly, for the 2005-06 and 2010-11 survey data for females, individuals who have been cohabiting for 5-19 years were at higher risk of HIV infection than those who never cohabited. On the other hand, individuals who have more than 19 years of cohabitation have a protective effect of HIV infection than those who have not cohabited. However, this phenomenon changes in the 2015 survey data for females as all individuals who are cohabiting are more at risk of HIV infection than those who are not. HIV positivity was associated with whether one had the knowledge of reducing the risk of HIV transmission or not, according to **Fig 4.5**, 2005-06 survey data for females. However, for the 2010-11 and 2015 surveys, HIV positivity was associated with not having any knowledge of reducing the risk of getting HIV. This is similar to the logistic regression in **Table 4.4** results, where having knowledge of reducing the risk of HIV transmission was a protective effect against HIV infection.

Table 4.4: Odds Ratio of female features selected in XGBoost using SFFS procedure.

Survey	2005-06		2010-11		2015	
Variable	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
Total lifetime number of sex partners	1.32	<0.001	1.109939	0.037	1.410084	<0.001
Cohabitation duration (grouped)						
Never Married	ref		ref		ref	
5-9	1.34	0.011	1.29	0.034	1.47	0.002
10-14	1.39	0.012	1.46	0.004	1.67	<0.001
15-19	1.33	0.032	1.43	0.018	2.22	<0.001
20-24	0.88	0.046	0.90	0.471	1.84	<0.001
25-29	0.55	0.003	0.72	0.011	1.41	0.041
30+	0.49	0.031	0.60	0.048	1.37	0.031
Number of household members	0.96	0.02	0.88	<0.001	0.85	<0.001

Age of household head	1.11	0.012	1.13	<0.001	1.23	<0.001
Time since last sex (in days)	1.00	<0.001	1.00	0.287	1.00	0.001
Intimate Partner Violence (IPV)						
no	ref		ref		ref	
yes	1.58	0.153	1.61	0.0466	1.97	0.0298
Type of place of residence						
urban	ref		ref		ref	
rural	0.84	0.251	0.63	<0.001	0.67	0.013
Current contraceptive method						
condom use	ref		ref			
No condom use	2.02	0.011	5.64	<0.001		
Sex of household head						
male	ref		ref			
female	0.93	0.466	0.99	0.955		
Number of unions						
once	ref					
more than once	2.26	<0.001				
Can ask partner to use condom						
no	ref		ref		ref	
yes	1.04	0.671	1.29	0.007	2.03	<0.001
Wealth index						
poor	ref		ref		ref	
poorer	1.32	0.032	0.87	0.285	0.90	0.441
middle	1.17	0.245	0.94	0.652	1.05	0.685
richer	1.43	0.022	0.83	0.201	0.76	0.092
richest	0.97	0.878	0.51	<0.001	0.55	0.002
Reduce risk of getting HIV						
no	ref		ref		ref	
yes	0.92	0.33	0.86	0.018	0.88	0.025
Ever been tested for HIV						
no	ref		ref		ref	
yes	1.06	0.509	1.11	0.273	1.50	0.045
Wife justified asking husband to use condom if he has STI						
no	ref		ref		ref	
yes	0.87	0.275	1.16	0.22	1.03	0.843
Religion						
No religion						

Christian	0.99	0.974	1.03	0.844	0.76	0.101
Apostolic	0.86	0.313	0.93	0.681	0.79	0.156
Other	0.53	0.025	0.70	0.378	0.55	0.162
Times away from home in last 12 months			0.99	0.916	0.99	0.06
Relationship to household head						
head			ref			
wife			0.73	0.166		
daughter			0.73	0.283		
daughter-in-law			0.41	0.001		
grand-daughter			1.10	0.864		
sister			1.39	0.453		
other relative			1.03	0.904		
not related			1.12	0.893		

Table 4.5 shows the odds ratio of HIV infection performed on the selected features obtained through SFFS for males. Males who are currently working are less likely to be HIV infected compared to those not working with an odds ratio (OR): 0.75 ($p=0.01$) and (OR): 0.71 ($p<0.001$) for the 2005-06 and 2015 survey years, respectively. If there was intimate partner violence (IPV), the risk of HIV infection was two times more than when there was no intimate partner violence for both males and females, as shown in **Tables 4.4** and **4.5**. As the age of household head increases, the risk of infection within the household increases by 11-20% between 2005-2015 for males and females, as shown in **Tables 4.4** and **4.5**. Males cohabiting are more than 3-9 times at risk of HIV infection than those not cohabiting. This is true for all survey years; as shown in **Table 4.5**, the risk increases over the years. The risk of HIV transmission being higher in males is also exhibited in **Fig 4.6**, which shows that cohabitation is the most influential feature for males in all survey years associated with HIV positivity.

Table 4.5: Odds Ratio of male features selected in XGBoost using SFFS procedure.

Survey	2005-06		2010-11		2015	
Variable	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
Cohabitation duration (grouped)						
Never Married	ref					
0-4	3.35	<0.001	4.23	<0.001	4.38	<0.001
5-9	5.87	<0.001	7.05	<0.001	6.53	<0.001
10-14	7.12	<0.001	7.87	<0.001	7.52	<0.001
15-19	6.79	<0.001	9.92	<0.001	8.42	<0.001
20-24	7.74	<0.001	9.48	<0.001	9.52	<0.001
25-29	3.73	<0.001	7.96	<0.001	7.07	<0.001
30+	6.44	<0.001	5.55	<0.001	6.92	<0.001
Total lifetime number of sex partners	1.01	0.002	1.01	<0.001	1.01	<0.001
Number of household members	0.91	<0.001	0.92	<0.001	0.91	<0.001
Age of household head	1.11	<0.001	1.12	<0.001	1.20	<0.001
Times away from home in last 12 months	1.00	0.425	1.00	0.747	0.99	0.234
Current contraceptive method						
condom use	ref		ref			
no condom use	1.24	0.155	2.37	<0.001		
Type of place of residence						
urban	ref		ref		ref	
rural	0.93	0.484	0.90	0.324	0.87	0.119
Intimate Partner Violence (IPV)						
no	ref		ref		ref	
yes	1.61	0.288	1.88	0.0236	1.78	0.01
Currently working						
no	ref		ref		ref	
yes	0.75	0.006	0.96	0.678	0.71	<0.001
Religion						
No religion	ref		ref		ref	
Christian	0.76	0.01	0.81	0.055	0.71	0.002
Apostolic	0.71	0.008	0.81	0.081	0.80	0.048
Other	0.93	0.592	0.93	0.703	1.32	0.126
Number of injections in last 12 months	1.17	<0.001			1.01	0.252
Respondent circumcised						
no	ref		ref			

yes	1.16	0.258	0.94	0.663		
Wife justified asking husband to use condom if he has STI						
no	ref				ref	
yes	1.56	0.003			1.42	0.017
Had any STI in last 12 months						
no			ref			
yes			2.31	<0.001		
Reduce risk of getting HIV						
no			ref			
yes			0.84	0.167		

4.4 Discussion

Although there have been a few studies utilising machine learning techniques to predict HIV in the generalised HIV pandemic, this is, to our knowledge, the first one in Zimbabwe using routinely collected survey data. This study's primary goal was to determine the most prevalent risk factors for HIV infection and the predicted accuracy of machine learning models based on these risk factors.

By using socio-demographic factors obtained from three ZDHS, this study was able to predict the HIV status of individuals. After comparison of different algorithms, XGBoost was the best algorithm to predict HIV status. The algorithm was able to determine the most predictive features/variables of HIV infection common to both sexes. Using SHAP plots, this study additionally evaluated the direction of the relationship between HIV infection and the predictive variables.

The total number of sexual lifetime partners was the main influential feature for females, and cohabitation duration for males in all survey years. Previous studies have shown that the total number of sexual partners is a risk factor for HIV infection ¹⁷⁴⁻¹⁷⁷. This was in agreement with our results which indicated that the higher the total number of sexual

partners an individual has, the more at risk of HIV infection. To add to the same evidence, a study by Armstrong et al.¹⁷⁸ indicated that sexual partner number was an important HIV risk measure. Our findings also showed that cohabitation was a risk factor of HIV positivity. This finding can be corroborated by other studies which reported that most heterosexual rates of HIV transmission take place within cohabiting or married couples¹⁷⁹⁻¹⁸³.

Numerous studies have shown that child-head families are more at risk of HIV infection¹⁸⁴⁻¹⁸⁷. This is contrary to our findings, as our results indicated that households headed by older individuals are more at risk of HIV infection. This may be attributed to the fact that the older the household head is, the less financially stable they are, hence their only source of income for that family might be from their pension fund, which might not be adequate, resulting in the other household members seeking financial assistance from risky behaviours^{188, 189}.

In addition, living in urban areas was found to increase the probability of HIV positivity. This was corroborated by other studies¹⁹⁰⁻¹⁹², which reported that people in urban settings had a greater chance of contracting HIV than people in rural areas. Magure et. al¹⁹³ described three potential reasons why urban areas may have higher HIV prevalence rates compared to rural areas in Zimbabwe. Firstly, urban areas are often characterised by high population densities, which can facilitate the spread of HIV through sexual networks. The increased mobility of urban residents, as well as the higher prevalence of sex work and transactional sex, also contribute to the higher HIV rates in urban areas¹⁹³. Secondly, access to HIV prevention and care services may be more limited in rural areas compared to urban areas. In Zimbabwe, for instance, the study found that rural residents were less likely to have ever tested for HIV or to have ever used condoms, compared to their urban counterparts. This lack of access or utilisation of prevention and care services may

contribute to the higher HIV prevalence rates in rural areas. Thirdly, there may be a lack of knowledge and awareness of HIV prevention and transmission in rural areas compared to urban areas. This knowledge gap may be due to limited access to information and educational resources in rural areas, which may make it harder for people to understand how to reduce their risk of HIV acquisition. Urban areas, on the other hand, tend to have higher levels of education and more access to health information and resources.

Our results indicated a higher risk of HIV infection if there was pro-intimate partner violence. To support this evidence, research done among ever-married and cohabiting women in Zimbabwe by Henderson et.al ¹⁹⁴ studied the relationship between intimate partner violence and HIV status, and they found out that women who had been victims of any kind of intimate partner violence were more likely to be HIV positive. The study also commented that the patriarchal and hypermasculist culture in Zimbabwe contributed to the likelihood of HIV infection. Further research utilising the ZDHS 2005-06 indicated that approximately six out of every ten women who reported experiencing some form of violence in their lifetimes had a much higher risk of being HIV-positive than women who had not suffered any physical or sexual abuse ¹⁹⁵.

To ensure that a substantial percentage of persons tested are HIV positive, a high yield is necessary when targeted HIV case-finding procedures are used to boost testing's cost-effectiveness. The potential for further behavioural-based case-finding techniques to improve or supplement current focused case-finding techniques like index testing is still unknown. For some limited resource conditions and the planned test coverage, it may be necessary to modify the acceptable cut-offs for sensitivity and PPV thresholds.

By better identifying those at high risk of contracting HIV, machine learning algorithms have the potential to enhance the implementation of pre-exposure prophylaxis (PrEP). As

part of an inclusive approach to PrEP, which has gone from development to implementation, programs can utilise these algorithms to spark discussions about PrEP. However, the limited number of variables accessible to us for this study constituted one of its drawbacks. We were unable to determine the impact of variables, including viral load, health care spending, HIV-risk groups, and other HIV-related interventions. The data also contained missing values, which necessitated making assumptions about their unpredictable nature and applying intrinsically flawed imputation techniques. Finally, several variables were self-reported and, as a result, were vulnerable to recall bias and social desirability.

Our technique for predictor identification primarily adapts from Orel's ¹⁹⁶ and Mutai's ¹⁵³ methodology, which also chose the XGBoost algorithm as the best. Contrarily, our findings demonstrate distinct predictors from those discovered in Orel's and Mutai's study. Considering that our study was more specific to Zimbabwe, this might have been the reason why results from Mutai's study were not similar to our findings. Additionally, Orel's population was based on Eastern Africa, which might have different epidemiological factors from the sub-Saharan Africa, therefore making the results statistically incomparable.

4.5 Conclusion

Our findings may help with social-behavioural HIV detection and improve screening procedures in limited resource settings. In addition, the amount of information needed to identify key populations in Zimbabwe will be significantly reduced by features/variables which have been identified in this study through machine learning. Adaptation of HIV screening methods that more effectively target the adult population, those with multiple

partners, those who are frequently away from home, those who reside in urban areas, those who are not currently working and other risk factors associated with HIV positivity are needed. Programmes targeted at HIV testing could incorporate machine learning approaches to adequately and effectively identify high-risk individuals. However, to improve the machine learning approach, further research is required to integrate and implement them in a real-world primary care context. In addition to other risk reduction techniques, machine learning may aid in identifying those who might require PrEP.

4.6 Declarations

4.6.1 Availability of Data and material

The datasets used for this study can be found on the Demographic Health Survey Program website : [<https://dhsprogram.com>].

4.6.2 Competing interests

The authors declare that they have no competing interests.

4.6.3 Authors' Contributions

RBM and EM conceived of the presented idea. RBM developed the theory and performed the data analysis. RBM and EM verified the analytical methods and models. EM supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

4.7 SUPPLEMENTARY INFORMATION

4.7.1 Variables used in the study

Table S1: Male and Female predictors variables.

	Male Features	Female Features
1.	Current age	Current age
2.	Type of place of residence	Type of place of residence
3.	Highest educational level	Highest educational level
4.	Religion	Religion
5.	Number of household members	Number of household members
6.	Relationship to household head	Relationship to household head
7.	Sex of household head	Sex of household head
8.	Age of household head	Age of household head
9.	Times away from home in last 12 months	Times away from home in last 12 months
10.	Wealth index	Wealth index
11.	Current contraceptive method	Ever had a terminated pregnancy
12.	Current contraceptive by method type	Current contraceptive method
13.	Number of injections in last 12 months	Current contraceptive by method type
14.	Respondent circumcised	Number of injections in last 12 months
15.	Currently/formerly/never in union	Currently/formerly/never in union
16.	Number of unions	Number of unions
17.	Age at first cohabitation	Age at first cohabitation
18.	Cohabitation duration (grouped)	Cohabitation duration (grouped)
19.	Time since last sex (in days)	Time since last sex (in days)
20.	Age at first sex (imputed)	Age at first sex (imputed)
21.	Recent sexual activity	Recent sexual activity
22.	Wife justified refusing sex: husband has other women	Currently working
23.	Currently working	Beating justified
24.	Type of earnings	Ever heard of AIDS
25.	Beating justified	Reduce risk of getting HIV
26.	Reduce risk of getting HIV	Had any STI in last 12 months
27.	Had any STI in last 12 months	Number of sex partners, including spouse, in last 12 months
28.	Number of sex partners, including spouse, in last 12 months	Ever been tested for HIV
29.	Ever been tested for HIV	Wife justified asking husband to use condom if he has STI
30.	Paid for sex in last 12 months	Total lifetime number of sex partners
31.	Wife justified asking husband to use condom if he has STI	Can respondent refuse sex
32.	Total lifetime number of sex partners	Can ask partner to use condom

4.7.2 ROC and Precision-Recall curves for combined data - Females

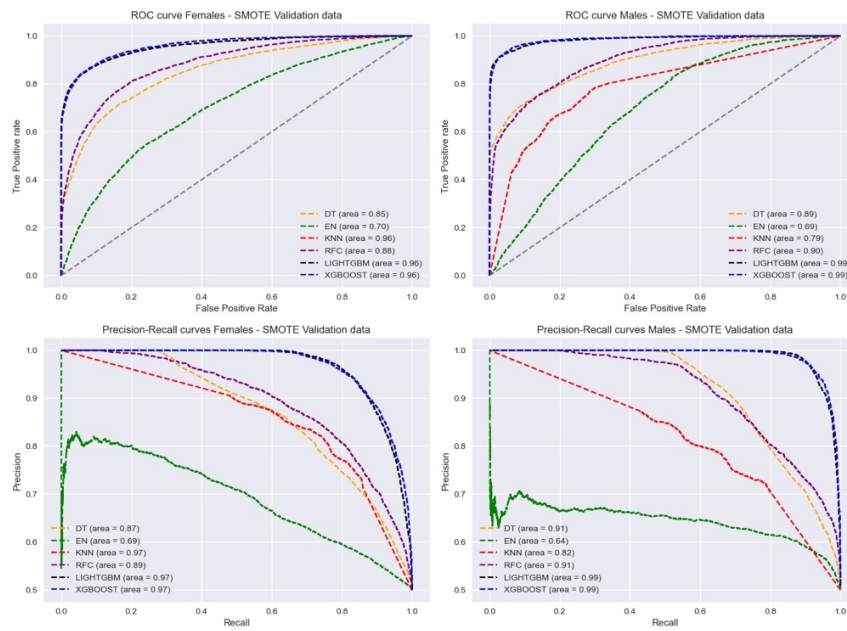


Fig S1: ROC and Precision-Recall Curves for combined SMOTE survey years data – Females.

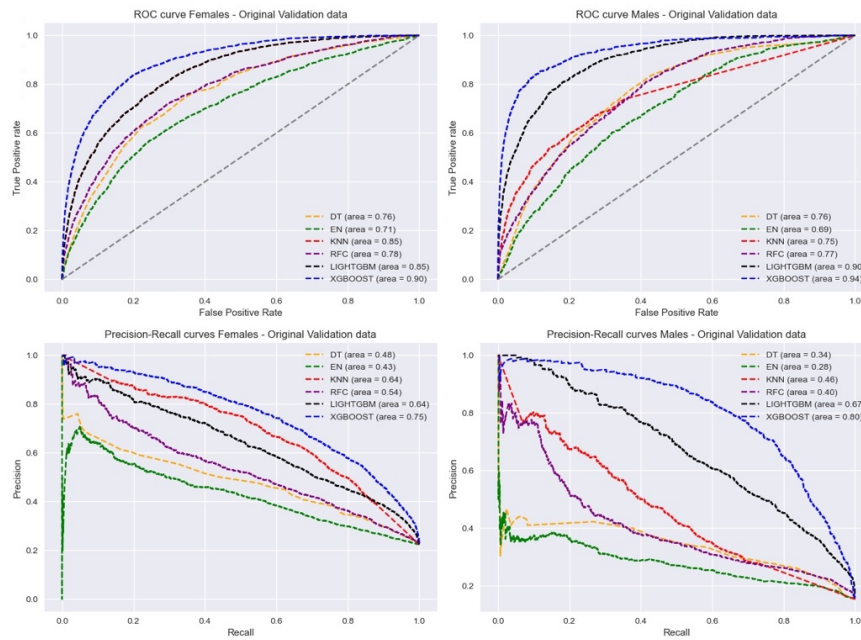


Fig S2: ROC and Precision-Recall Curves for combined ORIGINAL survey years data – Females.

4.7.3 ROC and Precision-Recall curves for combined data - Males

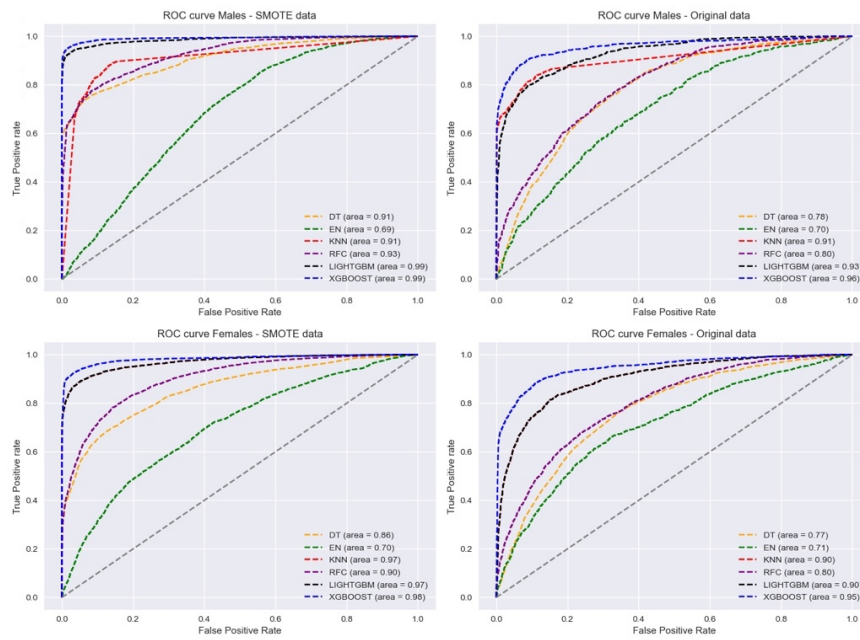


Fig S3: ROC and Precision-Recall Curves for combined SMOTE survey years data – Males.

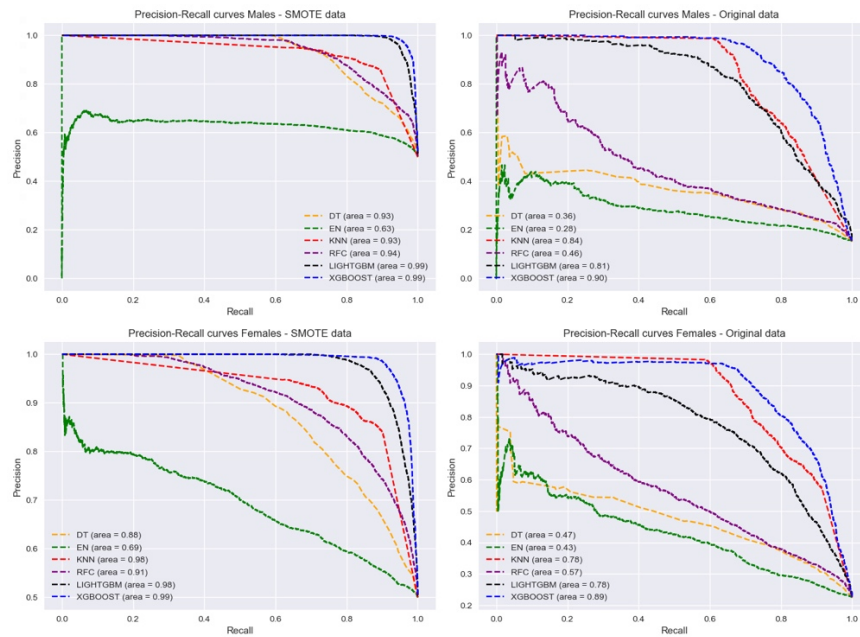


Fig S4: ROC and Precision-Recall Curves for combined ORIGINAL survey years data – Males.

CHAPTER 5

THESIS PAPER 3: Estimating HIV Incidence Over a Decade in Zimbabwe: A Comparison of the Catalytic and Farrington Models.

Rutendo Beauty Birri Makota^{1*}, Eustasius Musenge¹

¹ Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

* Corresponding author

E-mail: rutendobbirri@gmail.com

Abstract

Over the years numerous modelling studies have been proposed to estimate HIV incidence. As a result, the purpose of this study was to evaluate 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. Data were obtained from 2005–06, 2010–11, and 2015 Zimbabwe Demographic Health Survey (ZDHS). These models were validated at the micro and macro-level using community-based cohort incidence and empirical estimates from UNAIDS EPP/SPECTRUM, respectively. The HIV incidence for the catalytic model was 0.032 (CI: 0.028, 0.036), 0.036 (CI: 0.033, 0.039), and 0.028 (CI: 0.026, 0.030), for the years 2005-06, 2010-11, and 2015, respectively. The HIV incidence for the Farrington model was 0.021 (CI: 0.016, 0.026), 0.022 (CI: 0.020, 0.025), and 0.019 (CI: 0.016, 0.022), for the years 2005-06, 2010-11, and 2015, respectively. 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. HIV incidence declined with each subsequent survey year for all models. 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.

5.1 BACKGROUND

Understanding the dynamics of HIV epidemic is essential for determining and evaluating strategies for combating HIV ¹⁹⁷. Since the introduction of antiretroviral therapy decades ago, the number of HIV-positive individuals has continued to rise. HIV has developed into a chronic disease due to antiretroviral treatment, necessitating the establishment of reliable estimates of HIV incidence and prevalence, as it has become a substantial public health problem ^{198, 199}.

Quantifying the dynamics of HIV epidemics is crucial for evaluating the efficacy of interventions and guiding policy to enhance HIV/AIDS control ¹⁹⁷. HIV prevalence and incidence are two primary indicators used to characterise the dynamics of the HIV pandemic. Obtaining accurate estimates of HIV incidence and prevalence has always been a significant concern before the introduction of antiretroviral treatment (ART). Additionally, utilising HIV prevalence as an indicator of the pandemic has proved difficult due to differences in the time interval between infection and mortality due to increased ART availability ²⁰⁰. To get a real-time evaluation of the HIV epidemic, HIV incidence estimation can provide greater insights. Unfortunately, since HIV infection shares similar signs and symptoms with other diseases, individuals who have been recently infected with HIV may remain undiagnosed for significant periods, resulting in failure to detect the HIV incidence ¹⁹⁷.

A significant number of modelling studies have been developed and proposed to predict HIV incidence ^{197, 200-202}. A few mathematical or statistical modelling studies have been developed in Zimbabwe to estimate HIV incidence. Among the few HIV incidence models done in Zimbabwe, Hargrove et al. ⁸² used a simple Susceptible-Infected (SI)

model to estimate the HIV incidence among perinatal women in Zimbabwe between 1990 to 2009. Gregson et al.¹⁰⁰ used mortality data to estimate the HIV incidence in Zimbabwe using back-calculation methods. Ali et al.⁸¹ also estimated HIV incidence from HIV prevalence data obtained from patterns among female sex workers in Zimbabwe. Using population data from a survey done in 2015 under the Zimbabwe Population-Based HIV Impact Assessment Survey (ZIMPHIA 2015–2016), Gonese et al.⁸⁰ compared the HIV incidence estimates obtained through recency testing to UNAIDS Spectrum incidence estimates. Lastly, Hargreaves et al.⁹⁹ estimated HIV incidence among female sex workers (FSW) using HIV prevalence by age and number of years since started selling sex (YSSS) between 2011-2017. These methods have revolutionised how HIV incidence is estimated in various contexts and datasets worldwide²⁰⁰. However, few comparison studies to improve practical estimating settings have been reported^{57, 200, 203, 204}. Therefore, the primary objective of this study was to employ estimation models that have been used for incidence estimation to predict HIV incidence using seroprevalence data.

Serial prevalence surveys have been proposed as a less expensive and invasive alternative to collecting longitudinal data on a cohort of people to determine disease incidence¹²⁵. However, there are two significant disadvantages to this strategy. The first limitation is that any estimate based on serial prevalence data implicitly assumes disease irreversibility due to a lack of exact data on the duration of the condition. Because HIV is an incurable illness, the first limitation has no bearing on our study. The second limitation will be that there will be differential mortality between the infected and uninfected individuals.

Differential selection occurs when the likelihood of being selected varies between infected and uninfected individuals in a cross-sectional sample, resulting in a skewed incidence estimate. For example, when infected people have a higher mortality rate than

uninfected people, differential selection occurs; as a result, diseased individuals are less likely to be included in a cross-sectional sample of the living population, resulting in an underestimate of disease incidence ²⁰⁵. Given the possibility that the rate of mortality or survival after infection may be a factor in predicting incidence, our primary objective was to investigate two models in this study: the catalytic model, which assumes differential mortality, and the Farrington model, which assumes non-differential mortality, to determine if there are differences in the estimates of these models. The catalytic model has been used to estimate the incidence of HIV using seroprevalence data. In contrast, the Farrington model has been used to estimate the incidence of infectious diseases other than HIV. Hence, our secondary objective was to determine if the Farrington model behaved similarly to the catalytic model previously used to estimate HIV incidence. Lastly, we wanted to validate the estimates obtained from these models at micro and macro-level using community-based cohort incidence and empirical estimates from UNAIDS EPP/SPECTRUM, respectively.

The goal of this strategy was to triangulate results such that these two models may be utilised among the existing models that can be used to estimate HIV incidence using seroprevalence data from the Demographic and Health Survey (DHS). Data from three major household surveys in Zimbabwe from 2005 to 2015 was used to apply the models to ascertain the HIV incidence in the country.

5.2 METHODS

5.2.1 Survey Data

Incidence estimation methods in this paper used national household survey data from the Zimbabwe Demographic Health Survey (ZDHS) conducted in 2005-06 ¹⁰⁴, 2010-11 ¹⁰⁶

and 2015¹⁰⁵. The three surveys applied a two-stage random sampling design. The first stage involved selecting enumeration areas (EAs), which allowed specific weights to be assigned in the design. The second stage involved a random sampling of a fixed number of households¹⁰⁴⁻¹⁰⁶. The survey collected data from consenting individuals with age, sex, marital status, and HIV status. Data were collected in all households on females and males aged 0–49 and 0–54 years, respectively. In addition, blood samples were collected from consenting individuals. The finger prick method collected blot spots on filter paper and transported them to a laboratory for testing. Two ELISA tests were performed, with the second being a retest for 5–10% of the negative and all positive tests. A new ELISA or Western Blot test was performed if there were discordant results on the two ELISA tests¹⁰⁴⁻¹⁰⁶.

5.2.2 Estimation Models

The catalytic and Farrington models were used to predict the HIV incidence in Zimbabwe. The HIV incidence was estimated using the ZDHS seroprevalence data in both models. The catalytic model simultaneously accounted for differential mortality, whereas the Farrington model assumed non-differential mortality, which was of interest in this analysis. The models are described in the sections that follow:

5.2.2.1 Catalytic Model

This paper adapted Mossong et al.'s¹²¹ catalytic model to estimate age-dependent HIV incidence. The catalytic model is described fully elsewhere^{121,206}. A lognormal function was used to fit the incidence function (Equation 1), with one survival parameter and three incidence parameters¹²¹ to age-specific HIV seroprevalence data from the 2005-06, 2010-

11 and 2015 ZDHS of individuals aged 15-49 years. In fitting the lognormal function, an offset corresponding to the age of sexual debut was set at 12 years ¹²¹. Homogeneity across time was assumed within the three-time periods considered in this paper, and the chance of survival following HIV infection was assumed to follow a Weibull distribution. The model assumed differential mortality, which was estimated simultaneously with the HIV incidence.

$$\lambda(a) = \begin{cases} 0 & \text{if } a \leq 12 \text{ years} \\ \frac{\theta}{\sqrt{2\pi\sigma}(a-12)} \exp\left(-\frac{(\log(a-12) - \mu)^2}{2\sigma^2}\right) & \text{if } a > 12 \text{ years} \end{cases} \quad \text{Equation 5.1}$$

where: $\lambda(a)$ is the incidence function with parameters mean μ and standard deviation σ , the scale of θ and an offset corresponding to the age of sexual debut fixed at 12 years.

5.2.2.2 Farrington Model

The Farrington model is a non-linear model that has been used to estimate the force-of-infection for many infectious diseases, including measles, mumps, and rubella ²⁰⁷. The Farrington model is fully described elsewhere ²⁰⁸, but briefly, the model assumes that the force-of-infection increases to a peak in a linear trend, then declines exponentially ²⁰⁹. Additionally, the model assumes that the Force-of-infection (FOI) begins at zero at birth and reaches a maximum at an age corresponding to the maximum transmission rate among susceptible and infected people. This results in the Farrington-3-parameter model represented in Equation 5.2:

$$\lambda(a) = (\alpha a - \gamma) \exp(-\beta a) + \gamma \quad \text{Equation 5.2}$$

where; the age-specific Force-of-infection (FOI) is represented by $\lambda(a)$, a is the age in years, and α , β and γ are model constants to be estimated with γ as the long-term residual value of the FOI.

If the residual value is set to zero (i.e. $\gamma=0$), the Force-of-infection declines to zero as age tends to infinity ($a \rightarrow \infty$), resulting in Equation 5.3, which is a Farrington-2-parameter model:

$$\lambda(a) = (\alpha a) \exp(-\beta a) \quad \text{Equation 5.3}$$

The Farrington model assumed non-differential mortality, which indicates that disease-related mortality did not affect how the observed serological profile was interpreted.

The age-specific incidence function, $\lambda(a)$, in Equation 5.3 was estimated using age-specific seroprevalence data from the 2005-06, 2010-11, and 2015 ZDHS of persons aged 15-49 years.

5.2.3 Validation of the Farrington and Catalytic Models.

The Farrington and catalytic models were validated at micro and macro levels. The objective was to determine if our models could estimate HIV incidence at the provincial and national levels. At the micro-level, ZDHS data for Manicaland province was used to estimate the HIV incidence using the Farrington and catalytic models. Estimates obtained from these models were then compared to the observed incidence from a Manicaland community-based cohort. Empirical estimates from the UNAIDS EPP/SPECTRUM were used to validate the Farrington and catalytic model at the national level.

5.2.3.1 EPP/Spectrum Model

UNAIDS created the Estimation Projection Package (EPP)/SPECTRUM ²¹⁰⁻²¹² to estimate HIV prevalence and predictions ²⁰⁰. The HIV prevalence estimated by EPP will then be used to estimate the HIV incidence using another software called SPECTRUM, which uses the AIDS Impact Model (AIM) module. A summary of the SPECTRUM methodology is as follows:

Let the number of HIV infections, the total number of adults in the population and the HIV prevalence of individuals aged a at time t be denoted by $I_{a,t}$, $N_{a,t}$ and $P_{a,t}$, respectively. therefore;

$$I_{a,t} = N_{a,t} \times P_{a,t} \quad \text{Equation 5.4}$$

and;

$$\lambda_{a,t} = I_{a,t} - (I_{a-1,t-1} - d_{a-1,t-1}^A - d_{a-1,t-1}^{NA}) \quad \text{Equation 5.5}$$

where;

$\lambda_{a,t}$ is the number of new HIV infections in the year t minus the number of survived infections from the year $t-1$;

$d_{a-1,t-1}^A$ is the number of deaths caused by AIDS in year $t-1$; and

$d_{a-1,t-1}^{NA}$ is the number of deaths caused by other reasons other than AIDS in year $t-1$.

EPP was used to calibrate HIV prevalence data from household surveys using the maximum likelihood method, fitting a simple epidemiological model to surveillance data from prenatal clinics in Zimbabwe. Curves expressing the uncertainty in the prevalence data were created using Bayesian melding ²¹³. Finally, using assumptions about the average survival of HIV-infected people and the impact of ARV therapy on survival, adult HIV incidence was computed from HIV prevalence through time.

5.2.3.2 Community-Based Cohort Study – Manicaland

HIV incidence rates from community-based cohort studies in Zimbabwe from 2005 to 2015 were abstracted from published papers and conferences. Instead, data from a community-based cohort study in Manicaland HIV/STD Prevention Project in Zimbabwe under the Alpha networks was used ²¹⁴. In addition, the Manicaland cohort HIV incidence data was taken from the supplemental material of research conducted by Hallet et al. ²¹⁵. The Manicaland cohort was a “closed cohort” in which those who joined within a certain time period, were not seen at a follow-up, or were not confirmed to have died were excluded.

5.3 ETHICAL APPROVAL AND CONSENT TO PARTICIPATE

This work was granted ethical clearance by the University of Witwatersrand’s Human Research Ethics Committee (Medical) (No. M151154). The HIV testing protocol for all three ZDHS was reviewed and approved by the Centres for Diseases Control (CDC) in Atlanta, Georgia, USA, the Medical Research Council of Zimbabwe (MRCZ) in Harare, Zimbabwe; and the ORC Macro Institutional Review Board in Calverton, Maryland, USA ¹⁰⁴⁻¹⁰⁶. Application to the Measure DHS program, granted on May 16, 2017, yielded the dataset utilized in this investigation. The DHS Program is permitted to provide free survey data files to qualified academic researchers. Registration was required for access to data¹⁰⁴⁻¹⁰⁶. All methods were carried out in accordance with relevant guidelines and regulations. An informed consent statement was read to each respondent prior to their participation in a DHS interview or biomarker test, giving them the option to accept or

decline. Before a child or adolescent could participate, a parent or legal guardian had to provide their consent ¹⁰⁴⁻¹⁰⁶.

5.4 RESULTS

To inform indirect model-based incidence estimate, the ZDHS prevalence data from 2005–06, 2010–11, and 2015 served as the primary input data. **Fig. 5.1** displays the changes in HIV prevalence by age group, sex, and survey year. Generally, the HIV prevalence was higher in females than in males for both youth and adults. In addition, the HIV prevalence was higher for adults (25–49 years) than for youth (15–24 years) for both sexes. The HIV prevalence also declines with each successive survey year for both sexes.

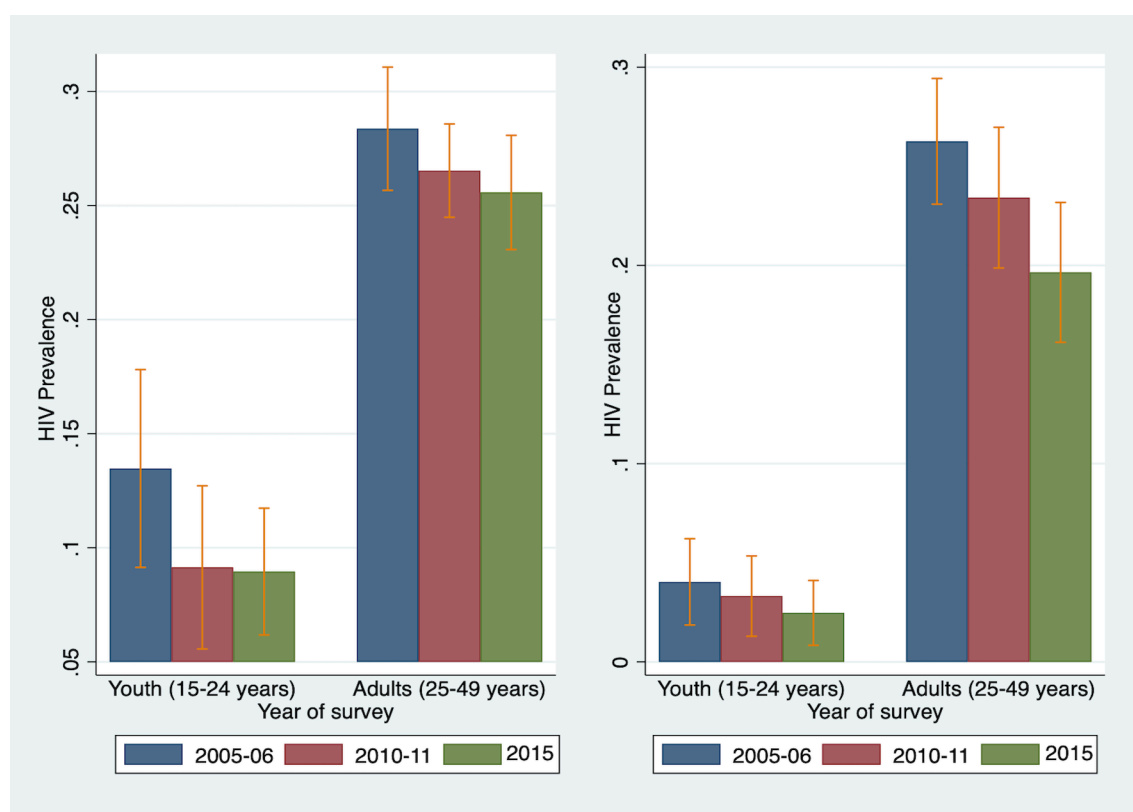


Figure 5.1: Prevalence trends by age group and survey year for females (left) and males (right).

5.4.1 HIV Incidence for the Two Models

Table 5.1 shows the HIV incidence for each modelling method stratified by survey year. The HIV incidence for the catalytic model was 0.032 (CI: 0.028, 0.036), 0.036 (CI: 0.033, 0.038) and 0.028 (CI: 0.025, 0.030) for the year 2005-06, 2010-11, and 2015, respectively. The HIV incidence for the Farrington-2-parameter model was 0.021 (CI: 0.016, 0.026), 0.022 (CI: 0.020, 0.025) and 0.019 (CI: 0.016, 0.022) for the year 2005-06, 2010-11, and 2015, respectively. Based on these results, HIV incidence estimates were higher for the catalytic model than for the Farrington model. **Table 5.1** also shows that HIV incidence increased from 2005-06 to 2010-11 and decreased in 2015 for both the catalytic and the Farrington-2-parameter models.

Table 5.1: Overall HIV incidence for the two models by year of survey

Year	Catalytic	Farrington
	Incidence(CI)	Incidence(CI)
2005-06	0.032(0.028,0.036)	0.021(CI: 0.016, 0.026)
2010-11	0.036(0.033,0.039)	0.022(CI: 0.020, 0.025)
2015	0.028(0.026,0.030)	0.019(CI: 0.016, 0.022)

As seen in **Fig 5.2**, the Farrington model underestimates HIV incidence compared to the catalytic model. Rather than the typical pattern observed in HIV incidence estimates for females, in which the incidence increases to a peak and then declines, HIV incidence estimates for males in 2010–11 and 2015 increased gradually without declining for the

catalytic model. According to the Akaike Information Criteria (AIC), the catalytic model was the best fit for the survey data.

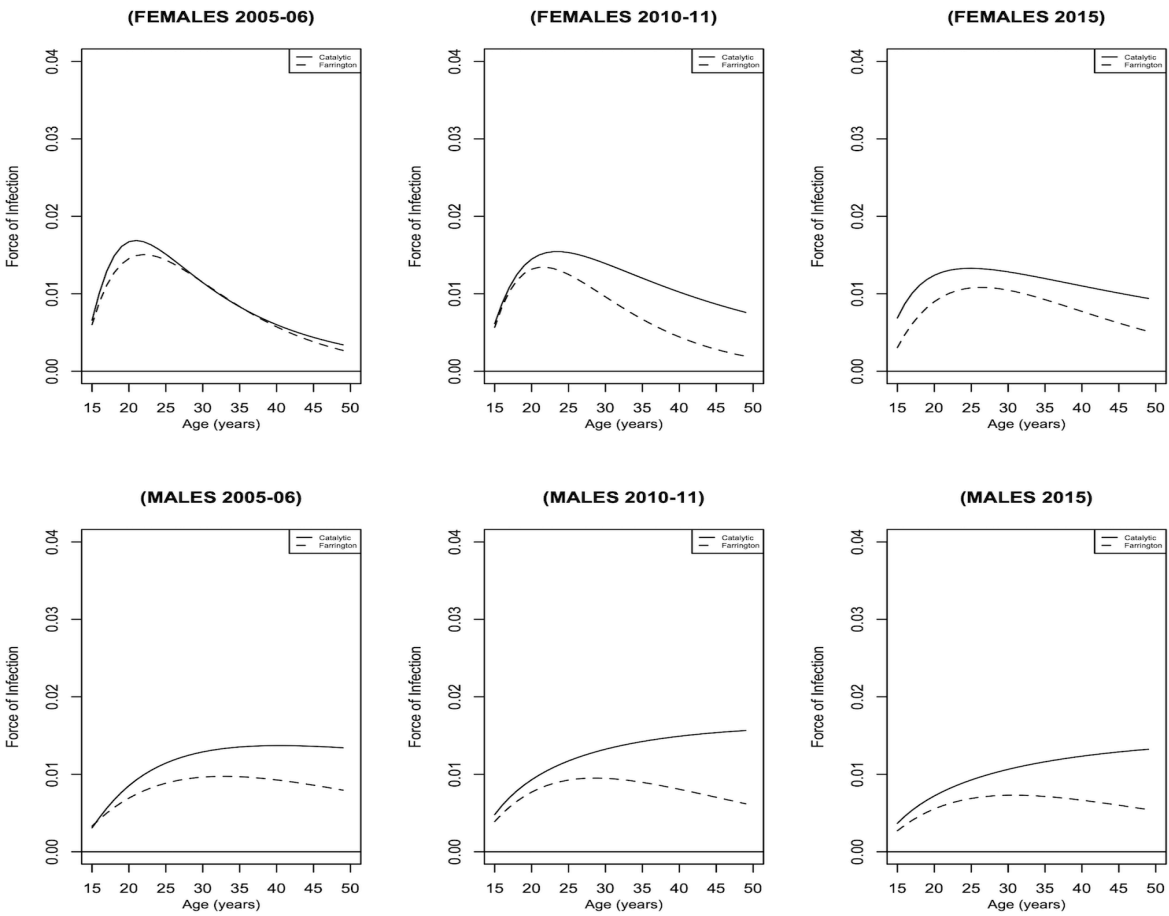


Figure 5.2: HIV incidence for males and females based on the catalytic and Farrington models.

5.4.2 Models' Validation Using a Cohort Study

We obtained Manicaland cohort HIV measured incidence for the years 2005 and 2010. Therefore, comparisons were made based on these two available cohort years, as shown in Fig 5.3.

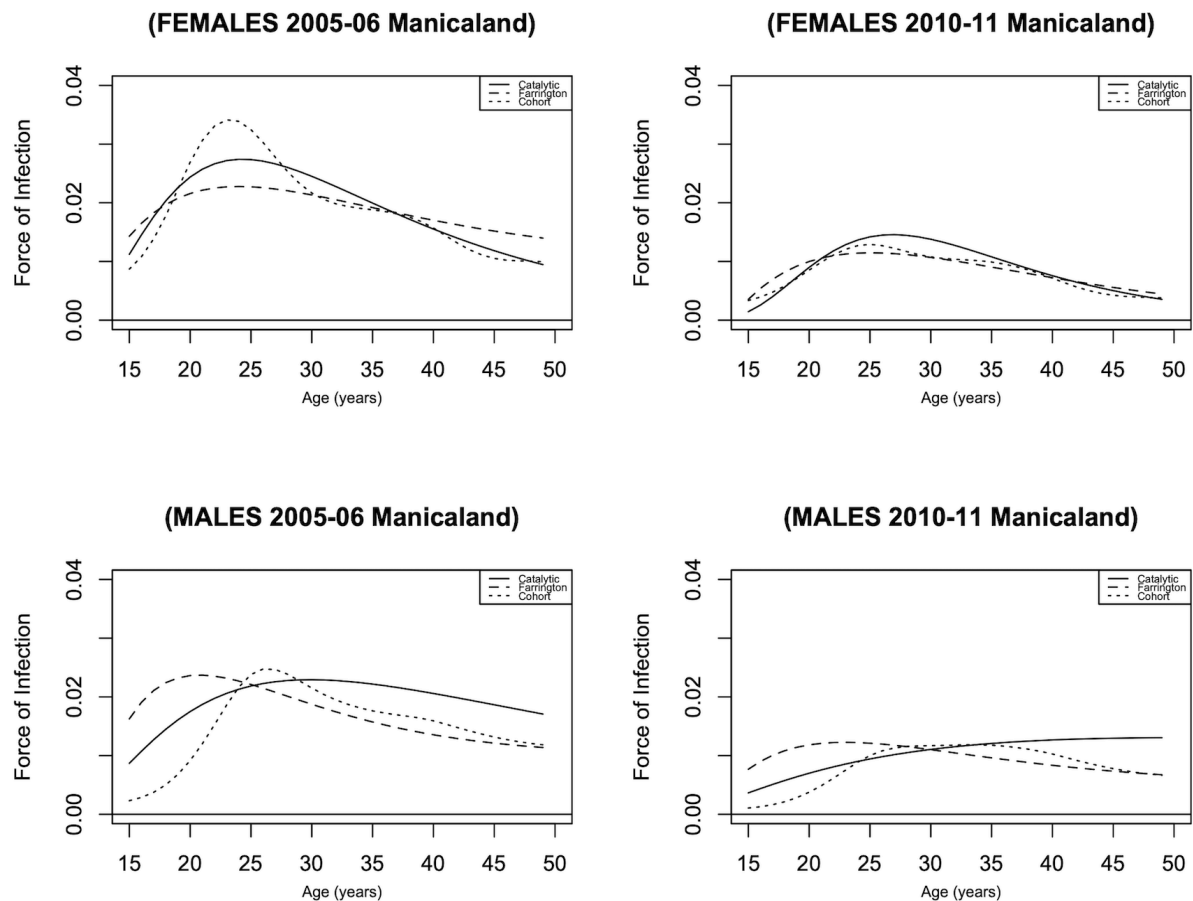


Figure 5.3: HIV incidence estimates based on the cohort, catalytic and Farrington models.

Table 5.2 summarises Manicaland’s estimated and observed incidence in various age groups. Almost all estimations fall within the observed measurements’ 95% confidence range. Both models performed well overall since the pattern and incidence are almost identical to the observed values, as shown in **Table 5.2** and **Fig 5.3**. **Table 5.2** shows that the estimations are within the 95% confidence interval of the observed, with 88% and 75% agreement scores for the catalytic and Farrington models.

Table 5.2: Measurements and estimates of incidence using Manicaland Province

Data

2005-06							
		COHORT		CATALYTIC		FARRINGTON	
Sex	Age Group, Years	Observed Incidence	95% CI ^a	Estimated Incidence	Agreement	Estimated Incidence	Agreement
Females	15-24	0.0232	0.0186, 0.0286	0.0218	*** ^b	0.0201	* ^b
	25-49	0.0179	0.0136, 0.0233	0.0184	*** ^b	0.0184	*** ^b
Males	15-24	0.0097	0.0071, 0.0129	0.0161	—	0.0219	—
	25-49	0.0176	0.0119, 0.0240	0.0209	** ^b	0.0155	** ^b
2010-11							
Females	15-24	0.0079	0.0057, 0.0106	0.0079	*** ^b	0.0088	** ^b
	25-49	0.0083	0.0057, 0.0114	0.0095	** ^b	0.0082	*** ^b
Males	15-24	0.004	0.0025, 0.0061	0.0061	* ^b	0.0109	—
	25-49	0.0101	0.0063, 0.0149	0.0120	** ^b	0.0092	*** ^b

^aCI; confidence interval

^b Agreement between estimate and observed: absolute error values ***<10%; **<20%; *<30%

The agreement between observed (cohort) and estimated (Farrington/catalytic) incidence was assessed using Bland-Altman plots, as illustrated in **Fig 5.4**. The plots give two critical pieces of information: the average of all differences determined using the t-test and referred to as bias and the 95% limit of agreement. The limits of agreement shown in **Fig 5.4** are narrow for all plots, indicating that our model estimates were comparable to cohort estimates. However, a few outlier values were observed in the 2005-06 Farrington females model, the 2010-11 catalytic females model, and the 2005-06 Farrington males model.

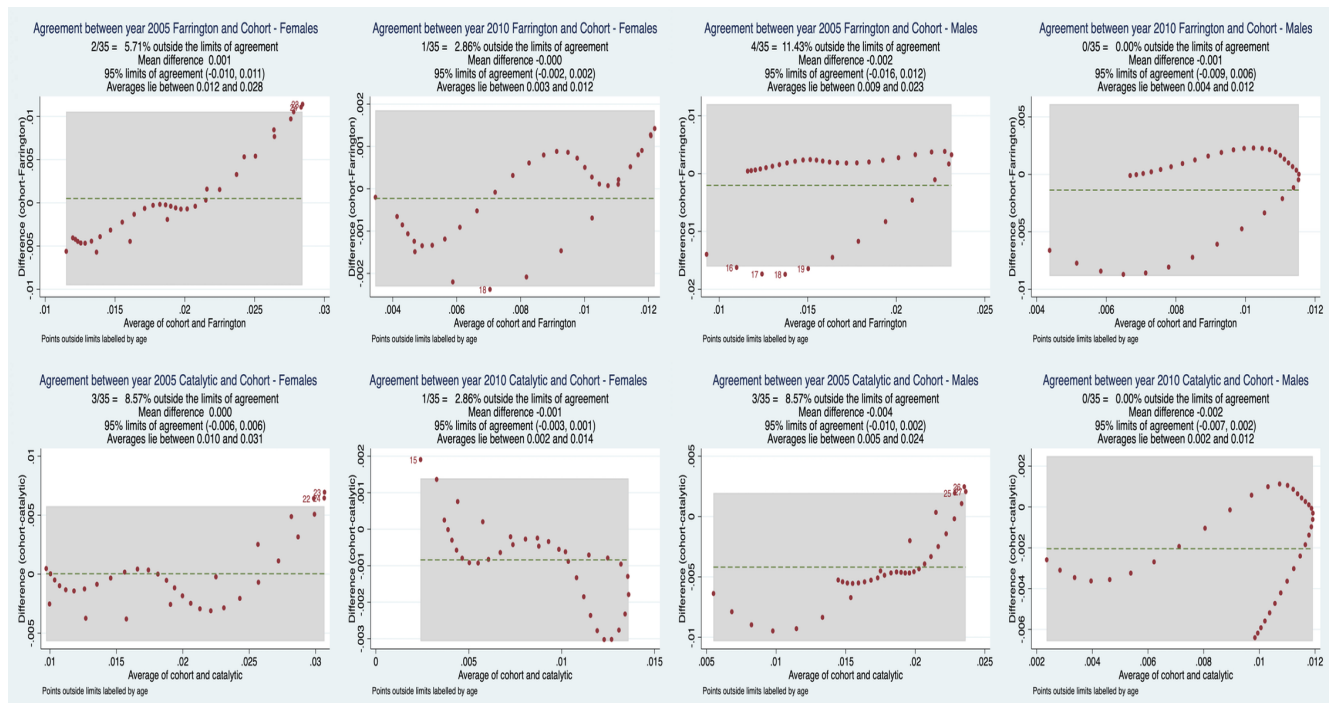


Figure 5.4: Bland-Altman plots demonstrating agreement between estimated incidence (Farrington and catalytic models) and observed incidence.

5.4.3 Models Validation Using UNAIDS EPP/SPECTRUM Estimates

Fig 5.5 illustrates the estimates of HIV incidence using three models: catalytic, Farrington, and EPP/SPECTRUM. From around 15 to 25 years, females had a higher HIV incidence than men; subsequently, the pattern reversed, with males having a higher HIV incidence. It is worth noting how the catalytic model predicts a progressive increase in HIV incidence for males throughout all survey years. HIV incidence declined with each subsequent survey year for all models. According to Fig 5.5, the Farrington model appeared to be more comparable to the UNAIDS EPP/SPECTRUM estimations.

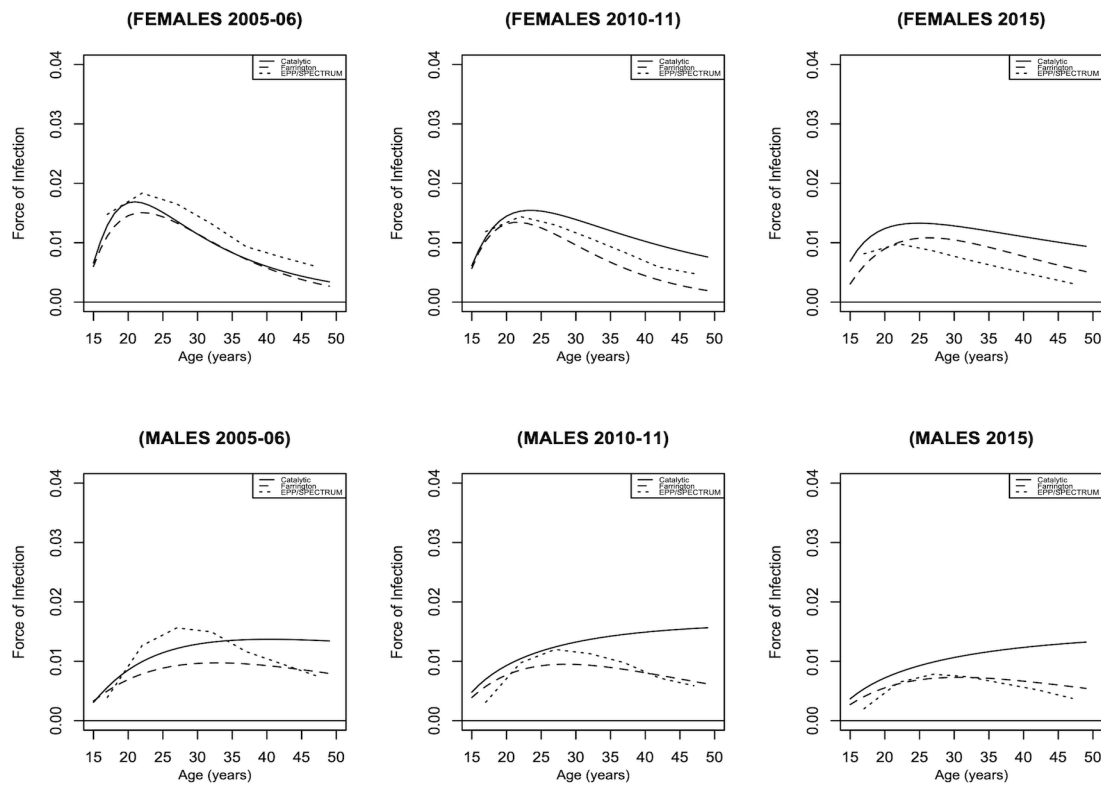


Figure 5.5: HIV incidence for males and females based on the catalytic, Farrington and EPP/SPECTRUM estimates.

5.5 DISCUSSION

Given the high expense of directly estimating HIV incidence using the “gold standard,” which requires long-term follow-up ²¹⁶, indirect approaches based on mathematical or statistical methods have become cost-effective alternatives ²⁰⁰. Therefore, we examined two models in this research that estimated age-specific HIV incidence using age-specific seroprevalence data. The objectives of this study were 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.

The results indicate that HIV incidence estimates from the two models are substantially comparable in estimates and trends; nevertheless, caution should be exercised when interpreting the results due to the models’ different assumptions. The catalytic model best suited the data between the two models employed to predict HIV incidence, evidenced by its low AIC values. The Bland-Altman plots gave an approach to investigate the extent of agreement between the cohort HIV incidence estimates and the two models (catalytic and Farrington). Comparing the catalytic and Farrington model to the cohort observed estimates provides evidence that the two models can be used as part of the models that can be relied on in reporting the subnational HIV estimates on Zimbabwe. Overall, there was reasonable agreement between the two models and the cohort estimates, however, a few discrepancies were seen when some of the data points were not included in the 95% limits of agreement.

Although at the national level comparison, the trends of the two models seemed to be similar to the EPP/SPECTRUM trends, it should be noted that actual HIV estimates fluctuated over time. For example, we noticed that in 2005, the HIV estimates for the two models were underestimated as compared to the EPP/SPECTRUM model however, as

we move with time, the inverse becomes true as the two models now overestimate the HIV incidence as compared to the EPP/SPECTRUM estimates. This may be accounted for by the changes which have been made to the EPP/SPECTRUM model over the years as a way to improve its estimates.

Despite considerable model assumption disparities, two mathematical models produced similar HIV incidence levels and trend estimates. Additionally, the models differed in their assumptions about death, with the catalytic model assuming differential mortality and the Farrington model assuming uniform mortality. As seen in **Fig 5.2**, the Farrington model generates lower HIV incidence estimates, which may be due to the differences in mortality assumptions; however, the differences seem minimal.

Estimating HIV incidence from serial age-specific cross-sectional prevalence data may be subject to several potential biases, including the quality of the age data, the comparability of the various survey populations, and changes over time in factors affecting HIV prevalence, such as socioeconomic status, migration, and fertility²¹⁷. We believe that these possible biases are negligible in our investigation. This is also corroborated by research conducted by Batter et al.²¹⁷, who reported that the effect on mortality in their study was negligible. Other researchers, such as Buycendore et al.²¹⁸ and Miotti et al.²¹⁹, utilized HIV prevalence data from surveys of childbearing women without adjusting for mortality or fertility impacts. Although mortality impacts of HIV infection are difficult to measure, they should ideally be included in HIV monitoring systems.

While estimating HIV incidence across age groups is challenging in the era of ART programs, it is critical for tracking changes in the age-specific pattern of HIV prevalence. Our findings imply that combining various approaches for estimating HIV incidence in

the same population can aid in generating solid estimates of national HIV incidence levels and trends. This has been practiced in many countries, for example, a study by Kim et. al²²⁰ estimated HIV incidence among adults in Kenya and Uganda by comparing multiple methods, and they recommended that triangulation of methods was needed to determine the best supported estimates of incidence to guide programs. In Zimbabwe, Gonese et. al⁸⁰ validated the HIV incidence obtained through recency testing with the EPP/SPECTRUM estimates which are used in Zimbabwe by policymakers. In this study, unlike Gonese et. al⁸⁰ who used recency testing to estimate HIV incidence, we used statistical models to estimate HIV incidence.

To put the results of the two models into context, we note that young women between 2005 and 2015 had the highest risk of HIV incidence. However, our findings reveal a trend toward a decline in the HIV incidence of young women over time. Similarly, men over 30 years are at an increased risk of HIV infection, despite a gradual decline in recent years. Our findings corroborate the issue of age mixing in Zimbabwe^{89, 221} and other sub-Saharan countries²²²⁻²²⁵.

5.6 CONCLUSION

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 youth aged 20 to 30 carry the most significant disease burden in all our models. However, a gradual decline over the years is observed, indicative of a stabilised HIV pandemic as this bathtub pattern is

similar to other demography indicators like fertility. This allows a parallel approach to be feasible as governmental and non-governmental organisations' interventions and campaigns among the youth could carry a combined message on abstinence, monogamous sexual relations and preventative therapy. However, robust public health and societal change behaviour interventions have proven far superior to therapeutics in disease control, especially in the era of COVID-19.

5.7 DECLARATIONS

5.7.1 Availability of Data and Material

The datasets used for this study can be found on the Demographic Health Survey Program website : [<https://dhsprogram.com>].

5.7.2 Competing Interests

The authors declare that they have no competing interests.

5.8 AUTHORS' CONTRIBUTIONS

RBM and EM conceived of the presented idea. RBM developed the theory and performed the data analysis. RBM and EM verified the analytical methods and models. EM supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

CHAPTER 6

THESIS PAPER 4: Estimating Age specific Prevalence and Force of Infection in Zimbabwe using Combined cross-sectional surveys from 2005-2015.

Rutendo Birri Makota^{1*}, Eustasius Musenge¹

¹ Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

*** Correspondence:**

Rutendo Birri Makota

rutendobbirri@gmail.com

Published paper:

Birri Makota R, Musenge E. Estimating age specific prevalence and force of infection in Zimbabwe using combined cross-sectional surveys from 2005 to 2015. Frontiers in Epidemiology. 2022;2.

Abstract

Objective: Age structured sexual mixing patterns have been noted to be associated with HIV prevalence and force of infection. Therefore, this study aimed to estimate the age dependent HIV force of infection using survey cross-sectional data from Zimbabwe.

Methods: We fit generalised additive models namely; linear, semi-parametric, non-parametric and non-proportional hazards models. Using the 2005-06, 2010-11 and 2015 Zimbabwe Demographic Health Surveys data. The Akaike Information Criteria was used to select the best model. The best model was then used to estimate the age dependent HIV prevalence and force-of-infection.

Results: Based on birth year cohort-specific prevalence, the female HIV prevalence reaches the highest peak at around 29 years of age, then declines thereafter. Males have a lower cohort specific prevalence between 15 and 30 years than females. Male cohort-specific prevalence slightly decreases between the ages of 33 and 39, then peaks around the age of 40. The cohort-specific FOI is greater in females than in males throughout all age categories. In addition, the cohort-specific HIV FOI peaked at ages 22 and 40 for females and males, respectively. A 18-year age gap between the male and female HIV FOI peaks was observed.

Conclusion: Our model was appealing because we did not assume that the FOI is stationary over time; however, we used serological survey data to distinguish the FOI's age-and-time effect. The cohort-specific FOI peaked 18 years earlier in females than males, indicative of age-mixing patterns. We recommend interventions that target younger females so as to reduce HIV transmission rates.

Keywords: Force of Infection¹, Prevalence², Demographic Health Survey³, generalised additive models⁴, hazards model⁵.

6.1 INTRODUCTION

There has been a decline in HIV incidence in adults in sub-Saharan (SSA) ²²⁶. The prevalence of HIV has increased in those aged 25 and older despite a decrease in HIV incidence due to increasing life expectancy brought on by the advent of antiretroviral treatment (ART) ²²⁷⁻²³⁰. A cohort study in Manicaland has highlighted that the cohort of individuals living with HIV in Zimbabwe is ageing such that the HIV prevalence is increasing among individuals aged 45 and older and decreasing among younger age groups ²²⁸. Due to the high survival rates of those who are HIV infected earlier in life, a shift in the relative share of infections to older individuals has been observed ²²⁶. Age-structured sexual mixing patterns have been noted to have an association with HIV force-of-infection (FOI) ²²⁴. Force-of-infection (FOI) is defined as the rate at which susceptible individuals acquire an infectious disease, a key parameter that specifies the current state of an epidemic ²³¹. Even though the overall HIV incidence has declined in the era of ART, older cohorts (40-49 years) still account for a growing proportion of HIV incidence ²²⁶. Over time, understanding how the HIV FOI rates have shifted across age groups is needed to realign prevention targets.

Given the information about FOI in a population, interventions can be targeted to those at the most significant risk of infection. Subsequently, information on trends in the FOI over time can be used to evaluate the impact of programmes on the rate of new infections so that resources can be directed to the most effective interventions and plan future health care needs. Additionally, understanding the HIV incidence disparities across ages can help prioritise groups where more effort is needed to lower infection rates.

The most direct method of estimating the FOI is using cohort studies. However, these studies are costly; in most cases, representative samples are not attained ²³². By contrast,

cross-sectional surveys are less expensive, more straightforward, quicker to organise and can be carried out on a mass scale. In addition, serological testing has been more prevalent recently and is now a part of household surveys like the Demographic and Health Survey (DHS), which estimates the prevalence of HIV in the general adult population ^{233, 234}. Given the ease of getting HIV prevalence data from household surveys, mathematical models can help untangle the complicated relationship between HIV FOI and prevalence ^{235, 236}.

One of the modelling approaches to estimating FOI from prevalence data has come from Demography ^{206, 237}. Here, serial cross-sectional population survey estimates of prevalence are used to generate an estimate of the FOI based on the principle that a sample of individuals of age a at time t is under some conditions representative of a group of individuals aged $a - \tau$ at time $t - \tau$, where τ is the interval between these cross-sectional population surveys ²⁰⁷.

Understanding age-specific differences in the infection rates are essential in predicting the magnitude of the FOI, thus, targeting age-groups for different intervention programs. However, temporal trends and estimates of age-specific FOI have to be recovered from age and time-specific prevalence data ²³⁸. In cross-sectional surveys, the information obtained when an individual has been tested is whether that individual has already experienced an infection before the age (a) of testing or not. This is commonly known as “current status” data. If the FOI does not change with time, then a single cross-sectional survey of individuals at different ages is sufficient to estimate age-specific infection rates. While this assumption is reasonable for some infections, this assumption is untenable in the case of HIV infection.

This study aimed to estimate the age- and time-dependent prevalence and HIV FOI using current status data from the Zimbabwe Demographic Health Survey (ZDHS) collected in

2005-06, 2010-11 and 2015. This was achieved by developing parametric, semi-parametric and non-parametric models; and determining which models best fit the data.

6.2 METHODS

6.2.1 Zimbabwe Survey Cross-Sectional Data

Data from 2005–06, 2010–11, and 2015 ZDHS were used in this analysis and random samples of 10,800, 10,828 and 11,196 households were chosen, respectively. The households were selected using a two-stage cluster sampling method. At the first stage, random samples of 400, 406, and 400 enumeration areas (EAs), were selected for the 2005–06, 2010–11, and 2015 ZDHS respectively. At the second stage, households were randomly picked using a comprehensive list of all the households in the specified EAs. This made it possible to apply specific weights to EAs in the design ¹⁰⁴⁻¹⁰⁶.

6.2.2 Outcome Variable Measurement: HIV Testing Procedure

Blood samples from all households were taken with the respondent's consent, or that of the respondent's parent or guardian in the case of minors, for HIV testing in the laboratory on females aged 0–49 years and males aged 0–54 years. Finger-prick blood samples were collected on filter paper and taken to a laboratory for analysis. All HIV-positive and 5–10% of the negative tests were retested with a second ELISA after an initial ELISA (enzyme-linked immunosorbent assay) test. A second ELISA or a Western Blot was conducted if the two ELISA tests yielded inconsistent findings. The DHS Data Archive ¹⁷² provided the data for this study, which only included people between the ages of 15 and 49. We took into account the sampling design by including survey weights in the analysis.

6.2.3 Statistical Model

Over the years, methods have been proposed that estimate the age-specific force-of-infection using seroprevalence data ^{121, 122, 205-207, 220, 233, 237-242}. However, these methods are primarily based on an infection acting on a set of susceptible individuals. Given that $\lambda_A(a)$ is the force-of-infection at age a , if we assume time homogeneity (stationarity) and lifelong immunity, then the proportion susceptible, $p_A(a)$ is:

$$p_A(a) = \exp \left[- \int_0^a \lambda_A(z) dz \right] \quad \text{Equation 6.1}$$

In reality, time homogeneity assumed in equation 1 is not feasible as over the years, the introduction of ART, the evolution of cohorts and mutation of the pathogen has resulted in the change of the force-of-infection over time ²⁰⁷. Therefore, the time-specific force of infection $\lambda_T(t)$ can be estimated instead. By integrating exposure experience between the date of birth, $t - a$, and the survey date t results in the proportion susceptible $p_T(a)$ ²⁰⁷:

$$p_T(a) = \exp \left[- \int_0^a \lambda_T(t - a + z) dz \right] \quad \text{Equation 6.2}$$

Considering that in this analysis, we deal with three-time points, estimating the force of infection $\lambda_T(t)$, will not be ideal ²³⁹. Therefore, while estimating the force-of-infection, especially when working with many time points, it is crucial to model the impacts of age and time effects. The relation between age- and time-specific seroprevalence data and age-time force-of-infection ²⁰⁷, $\lambda(a, t)$, in an individual of age a at time t , results in the proportion susceptible $p(a, t)$ ²⁴³:

$$p(a, t) = \exp \left[- \int_0^a \lambda(z, t - a + z) dz \right] \quad \text{Equation 6.3}$$

This paper proposed five models to model the relation between age- and time-specific seroprevalence data and $\lambda(a, t)$ in Equation 6.3. These models are model 1, representing a parametric model, models 2 and 3, representing semi-parametric models, model 4, representing a non-parametric model, and model 5, representing the non-proportional hazards model. The first is a flexible semi-parametric method suitable for simple exploratory analysis. Then, these methods are applied to age- and time-specific ZDHS HIV seroprevalence data.

The parametric model can capture reasonably complex patterns of age- and time-dependence to determine the degree of curvature required by the data ²⁰⁷. Model 1 constrained the force-of-infection to be positive and is part of the Exponential polynomial (EP) functions which have the following properties ²⁰⁸:

$$\lambda(a, t) = \exp\left(\mu_0 + \sum \mu_i a^i + \sum \theta_j t^j\right), i, j = 1, 2, \dots \quad \text{Equation 6.4}$$

where i represents the time spent by an individual in age band i , and j represents the time spent by an individual in time band j . We can estimate the age-time dependence of the FOI in Equation 6.4 by setting all μ_i and θ_j to zero, thus resulting in Equation 6.5 ²⁰⁷:

$$\lambda(a, t) = \exp(\mu a) \quad \text{Equation 6.5}$$

Equation 6.5 represents model 1 and will be referred to in terms of their time/age parameterisation which is model EP1/EP1 being an exponential polynomial of degree 1 in time and degree 1 in age.

Assuming that the FOI changes exponentially with time represented by the following identity (for two calendar times t_1 and t_2) ²⁰⁸ in Equation 6.6:

$$\lambda(a, t_2) = \exp(\beta(t_2 - t_1))\lambda(a, t_1) \quad \text{Equation 6.6}$$

We applied the methods of Nagelkerke et al. (1999) to estimate age- and time-specific FOI without any assumption about age dependency ²³⁸. To be able to assess the cohort experience, a cohort born at calendar time b and has age a at calendar time $a + b$, then the age-specific FOI for that particular cohort is defined by $\lambda_b(a)$ in Equation 6.7 ²⁰⁸:

$$\lambda_b(a) = \lambda(a, a + b) \quad \text{Equation 6.7}$$

The proportional hazards assumption in Equation 6.6 then translates into Equation 6.8, which leads to models 2 and 3 in this study, which are proportional hazards, semi-parametric models ²⁰⁸.

$$\lambda_b(a) = \exp(\beta(b - b_0))\lambda_{b_0}(a) \quad \text{Equation 6.8}$$

where the hazard at baseline year of birth b_0 , is defined by $\lambda_{b_0}(a)$, which can also be expressed as $\lambda(a, a + b_0)$ ²⁰⁸.

The semi-parametric model in Equation 6.8 can be reformulated as a fully non-parametric proportional hazards model by:

$$\lambda(a, t_2) = \exp(g(t_2 - t_1))\lambda(a, t_1) \quad \text{Equation 6.9}$$

where $g(t)$ is a smooth function with the constraint that $g(0) = 0$.

Substituting Equation 6.9 in the expression given in Equation 6.3 leads to:

$$p_b(a) = p(a, b + a) = \exp\{\exp(g(b - b_0))\log p_{b_0}(a)\} \quad \text{Equation 6.10}$$

where; with $\lambda_{b_0}(a) = \lambda(a, b_0 + a)$,

$$p_{b_0}(a) = \exp\left\{-\int_0^a \lambda_{b_0}(u) du\right\}.$$

Equation 6.10 which is Model 4, may be expressed as a Generalised Additive Model (GAM) for binary data with complementary log-log link function ²⁰⁸ (Equation 6.11), with $\pi_b(a) = 1 - p(a)$ representing the proportion of persons infected at age a or before from the cohort with year of birth b ²⁰⁸:

$$\log \left(-\log(1 - \pi_b(a)) \right) = f_1(a) + f_2(b), \quad \text{Equation 6.11}$$

where

$$f_2(b) = g(b - b_0),$$

and

$$f_1(a) = \log \left(-\log(x_{b_0}(a)) \right).$$

The corresponding FOI for this cohort then equals to ²⁰⁸:

$$\lambda_b(a) \frac{\pi'_b(a)}{1 - \pi_b(a)} = \exp(f_2(b)) \lambda_{b_0}(a), \quad \text{Equation 6.12}$$

with

$$\lambda_{b_0}(a) \frac{\pi'_{b_0}(a)}{1 - \pi_{b_0}(a)} = \frac{d}{da} \exp(f_1(a)) \quad \text{Equation 6.13}$$

The proportional hazards model translates into an additive *cloglog* model for the seroprevalence status ²⁰⁸, as shown by Equation 6.11, represented in this study as Model 4. In addition to the year of birth, this GAM technique may be generalised to include additional subject-specific characteristics, such as gender ²⁰⁸.

A non-proportional hazards model (Model 5) was obtained by modelling the interaction of age a with year of birth b . Taking the GLM model with a two-dimensional smoother ²⁰⁸:

$$\log \left(-\log(1 - \pi_b(a)) \right) = f(a, b), \quad \text{Equation 6.14}$$

some straightforward calculus shows that:

$$\frac{\lambda_b(a)}{\lambda_{b_0}(a)} = \left(\exp(f(a, b) - f(a, 0)) \right) \left(\frac{\partial f(a, b) / \partial a}{\partial f(a, 0) / \partial a} \right) \quad \text{Equation 6.15}$$

The right-hand side of equation (15) is no longer independent of age a . Note that it reduces to Equation 6.12 in case $f(a, b) \equiv f_1(a) + f_2(b)$. The proportional hazards model was

tested by comparing a model with a (parametric or non-parametric) (age $\times$ year of birth) interaction term with the corresponding model with main effects only²⁰⁸.

6.2.4 Statistical Analysis

This study estimated age- and time-dependent prevalence and FOI from serial cross-sectional population survey prevalence data. Our analysis was based on comparing a generalised linear additive model (model 1), generalised semi-parametric additive models (models 2 and 3) and generalised non-parametric additive models (models 4 and 5). All the models were proportional hazards, except for model 5, a non-proportional hazard model. Model 1 adapted Ades and Nokes' ²⁰⁷ exponential polynomial model (EP) of type EP₁/EP₁.

Models 2 and 3 were modified from the semi-parametric model of Nagarkerke et al. ²³⁸ by employing a spline smoother instead of the isotonic step function for the baseline hazard function $\lambda_{b_0}(a)$. The age-specific FOI, $\lambda_b(a)$, for that particular cohort, was estimated in a semi-parametric way using an iterative MM algorithm with a parametric proportionality factor $\exp(\beta(b - b_0))$ and a non-parametric isotonic stepwise estimate for the baseline hazard $\lambda_{b_0}(a)$ as shown in Equation 6.6.

Model 4 was reformulated to a fully non-parametric model using the semi-parametric model of Nagarkerke et al. ²³⁸ as shown in section 6.2.3. Model 5 was obtained by an interaction of age (a) with the year of birth (b), resulting in a non-proportional hazards model. Models 1-5 are fully explained in section 6.2.3.

6.2.5 Model 1-5 Specification and Goodness-of-Fit

We were interested in modelling some response variable y (force-of-infection) which follows an exponential family distribution using predictor variables a which is age and b being the year of birth. Let $\mu = E(y)$, then models 1-5 are:

$$\text{Model 1 (linear additive model): } h(\mu) = \beta_0 + \beta_1 a + \beta_2 b \quad \text{Equation 6.16}$$

$$\text{Model 2 (semi – parametric additive model): } h(\mu) = \beta_0 + \beta_1 b + f_1 a \quad \text{Equation 6.17}$$

$$\text{Model 3 (semi – parametric additive model): } h(\mu) = \beta_0 + \beta_1 a + f_1 b \quad \text{Equation 6.18}$$

$$\text{Model 4 (non – parametric additive model): } h(\mu) = \beta_0 + f_1 a + f_2 b \quad \text{Equation 6.19}$$

$$\text{Model 5 (non – proportional hazards additive model): } h(\mu) = \beta_0 + f_1(a, b) \quad \text{Equation 6.20}$$

where h is a smooth monotonic ‘link’ function of mean μ and f_1 and f_2 are smooth functions of the covariates considered. In order for these models to be distinguishable, the smooth functions were restricted to have a mean of zero using the gam function. Model specifications for models 1-5 and link function were done using the family argument to generalised additive model (GAM) framework²⁰⁵ in the R software package *mgcv*²⁴⁴ with the *gam* function. The Pool-Adjacent-Violators Algorithm (PAVA) application ensured monotonicity in the age dimension. The *mgcv* package’s estimation of GAMs used a penalised likelihood method. Finally, we used the Akaike Information Criteria (AIC) to choose the best model for the model comparison.

$$AIC = \underbrace{-2 \log p(a)}_{\text{goodness of fit}} + \underbrace{2\tau}_{\text{model complexity}}$$

where τ is the effective degrees of freedom (edf). Therefore, if $AIC_{m1} < AIC_{m2}$ choose model 1. The effective degrees of freedom (edf) showed the complexity of the model. The 95% confidence interval of the best model was obtained using the R *confint.gam*

package which calculates the point-wise confidence intervals for the smooth terms of a fitted gam model.

6.3 RESULTS

The dataset was stratified by gender, and all subsequent analysis was done according to the stratification. Generalised non-parametric additive model (model 4) was selected as the best model according to the AIC criteria for females and males (**Table 6.1** and **Figure 6.1**). Furthermore, since model 4 was a proportional hazards model, using analysis-of-deviation, the proportionality assumption was determined to be valid for both the female and male models, with p values of 0.71 and 0.50, respectively.

Table 6.1: The Akaike information criterion (AIC), Schwarz's Bayesian criterion (BIC), and effective degrees of freedom (edf) for completely parametric (model 1), semi-parametric (models 2 and 3), and non-parametric (models 4 and 5) models

	FEMALES					MALES				
Model components	edf	AIC	rank	BIC	rank	edf	AIC	rank	BIC	rank
Model 1: $a + b$	3	22571.69	5	22596.00	5	3	12196.91	5	12220.32	5
Model 2: $f(a) + b$	6.99	21841.76	3	21897.11	2	8.25	11968.76	3	12033.11	1
Model 3: $a + f(b)$	6.61	21945.97	4	21999.54	4	8.03	12003.26	4	12065.93	4
Model 4: $f(a) + f(b)$	9.93	21779.61	1	21860.10	1	12.5	11938.44	1	12035.99	2
Model 5: $f(a,b)$	15.7	21785.40	2	21912.22	3	15.5	11939.81	2	12060.90	3

*a= age, b=year of birth, f=smooth function

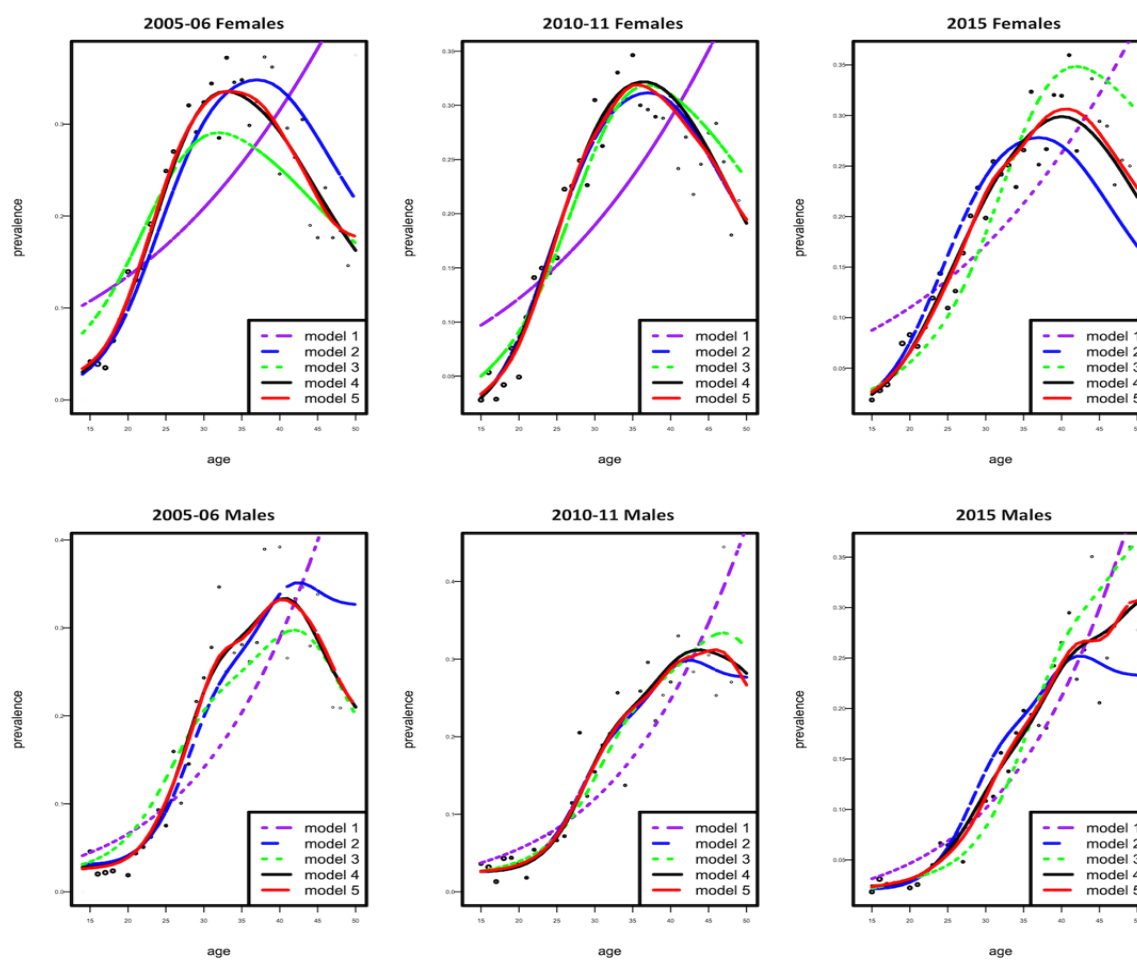


Figure 6.1: Observed (dotted curve) and fitted prevalence (five solid curves) as a function of age according to females (top) and males (bottom).

Figure 6.2 shows the age-specific fitted and observed HIV prevalence estimates for the three calendar periods 2005-06, 2010-11 and 2015, for both females and males. Model 4 showed a good model fit for the observed data. Comparing the fitted HIV prevalence model, we detect a change of high HIV prevalence among 33-year-old females in 2005-06, 36-year-old females in 2010-11, and 40-year-old females in 2015. In 2005-06, the prevalence of HIV was highest among 42-year-old males; however, in 2010-11, the prevalence declined in the same age group. In 2015, the prevalence of HIV steadily rose

with age among males. In general, the prevalence of HIV was greater among females than males over all three time periods.

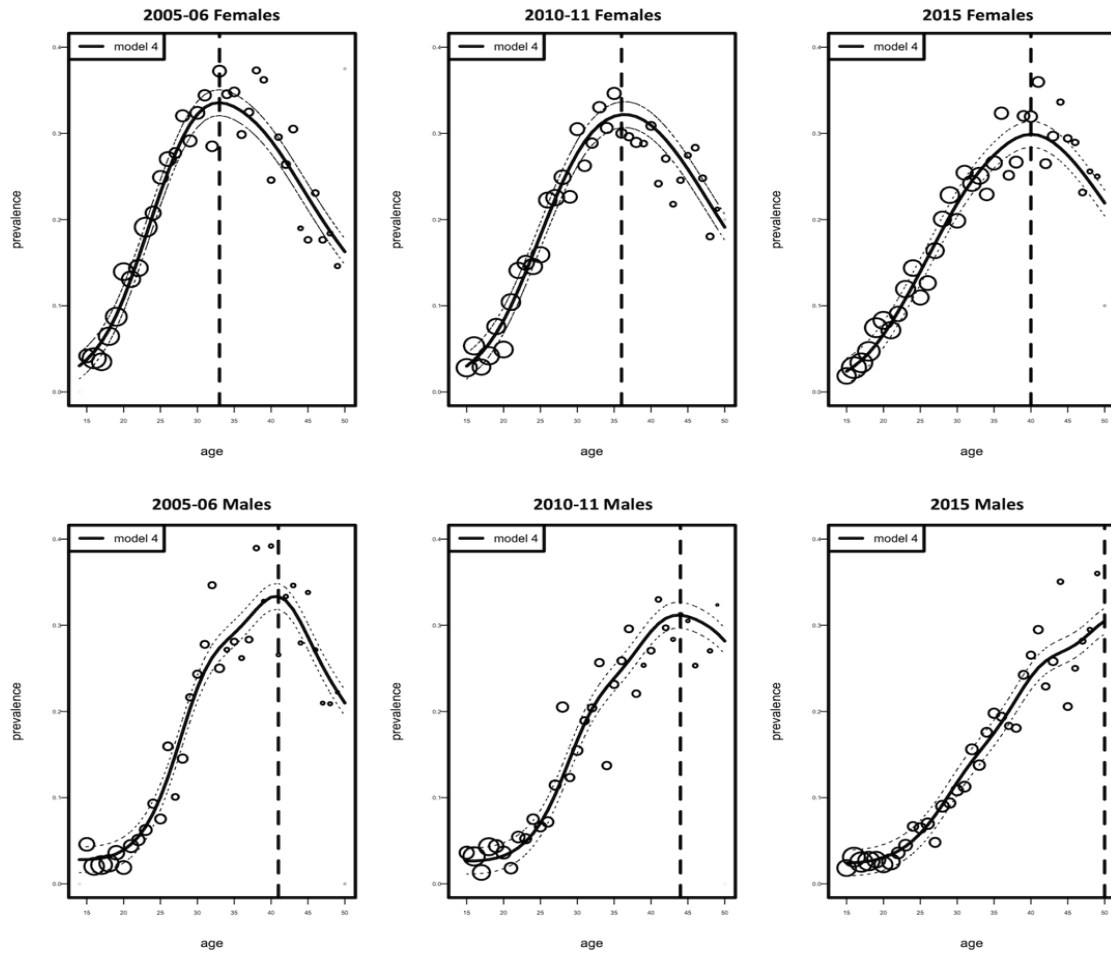


Figure 6.2: Observed (circles) and fitted (solid curve with 95% Confidence Interval, dotted curve) HIV prevalence as a function of age according to females and males. Circles represent the estimated prevalence by age. Their size is proportional to the number of individuals.

Figure 6.3 illustrates the fitted cohort-specific prevalence curves for females and males as a function of age, according to the best model (model 4). Prevalence was estimated for each birth cohort for all survey years combined. The solid red component of each

prevalence curve corresponds to the proportion of the age range (birth cohort) for which data were available, while the dotted portion of each curve was extrapolated using the model. **Figure 6.3** demonstrates that data for both models were available for the age range considered in this study. Based on cohort-specific prevalence, the female HIV prevalence reaches the highest peak at around 29 years of age, then declines up to the age of 49 years (these values can be observed by examining the red lines/curves). Males, have a lower cohort-specific prevalence between 15 and 30 years than females. Male cohort-specific prevalence slightly decreases between the ages of 33 and 39. The highest peak of HIV prevalence for males is around the age of 40.

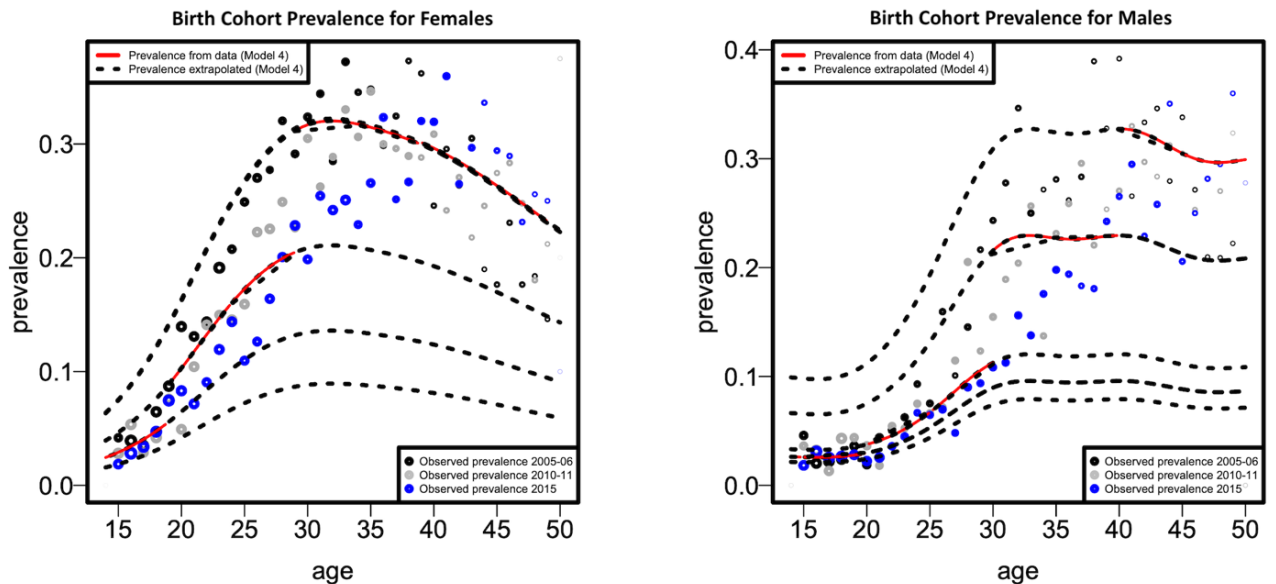


Figure 6.3: Fitted cohort-specific prevalence curves as a function of age according to females and males.

Figure 6.4 depicts the fitted cohort-specific FOI curves as a function of age for females and males, as predicted by Model 4. The FOI was estimated for every birth cohort over the combined survey years. The grey FOI curve displays the baseline FOI $\lambda b_0(a)$ for an

assumed $b_0 = 1955$, 1955 being the minimum time (birth year) in the cohort. The black dotted curves are generated from the proportionality factor $\exp(g(b - b_0))$ in Equation 6.11 resulting in the red curves. **Figure 6.4** indicates that, where data is available, the cohort-specific FOI is greater in females than males throughout all age categories, as indicated by the red curve. The cohort-specific HIV FOI peaked at ages 22 and 40, respectively, for females and males, demonstrating more than 15-year age difference between the HIV FOI peaks of males and females. Nevertheless, based on the extrapolated cohort-specific FOI, males tend to have a greater FOI than females in older age groups.

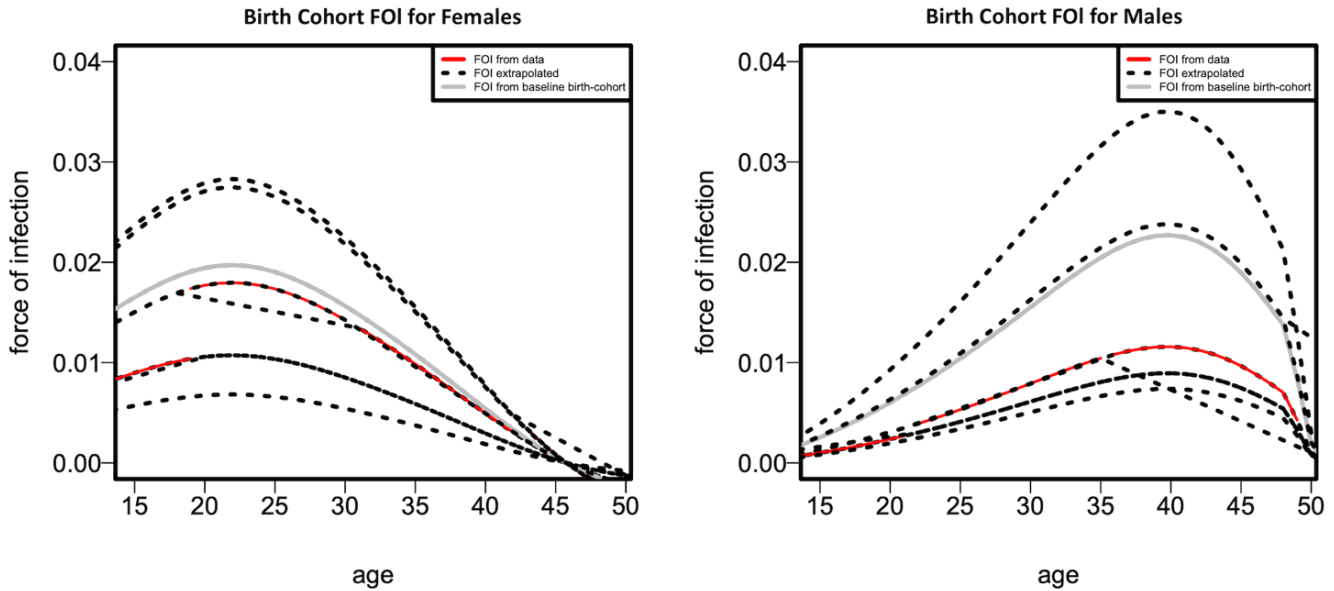


Figure 6.4: Fitted cohort-specific force-of-infection curves as a function of age according to females (left panel) and males (right panel).

6.4 DISCUSSION

This study estimated the age- and time-dependent HIV prevalence and FOI. Our model was appealing because we did not assume that the FOI is stationary over time; however, we used serological survey data to distinguish the FOI's age-and-time effect. It has been reported that the estimation of HIV FOI might lead to over-interpretation of results because of selection bias, partially due to lower consent rates ^{121, 223}. However, recent work has confirmed that non-participation in surveys does not result in significant biases in estimating cross-sectional HIV prevalence ^{121, 227}.

Generally, HIV prevalence decreased for both males and females over the years, with HIV prevalence higher in females than males. Of interest, we observed the peak HIV prevalence for each successive survey year was lower but shifted with an increase in age, which can be explained by either that people are now being infected at older ages or that these are people who were diagnosed with HIV many years after being infected. Unfortunately, data on HIV testing patterns in different age groups are unavailable, and the observed age increase may have resulted from increased testing in cohorts born earlier. In another school of thought, it is believed that due to the scale-up of antiretroviral treatment, there has been an increase in life expectancy; therefore, the prevalence in the population for the cohorts born earlier will increase ²⁴⁵. However, the shift towards increased HIV prevalence in older birth cohorts is typical of populations of people living with HIV (PLHIV) across the world ²⁴⁶. Our model, especially for males, projects similar age profiles (**Figure 6.1**).

A sharp increase in cohort-specific HIV prevalence is observed in our results to occur approximately a decade apart in females (at 29 years) than males (at 40 years). Cohort-specific FOI peaks approximately fifteen years earlier in females (at 22 years) than males

(at 40 years). These findings confirm the issue of age-mixing patterns which can be corroborated by other studies ²²²⁻²²⁴. A study in Zimbabwe found that age-disparate sexual relationships are associated with increased HIV incidence among young women, mainly when partners are ten or more years older ²⁴⁷. In a 2005 national household survey done in South Africa, it was determined that younger females had partners who were at least five years older than them, and it was hypothesised that these connections were linked to an increased HIV risk ²⁴⁸. Gregson et al. offered empirical evidence that age-disparate sexual interactions were responsible for the observed variations in the epidemiological patterns of young women and men in rural Zimbabwe ²⁴⁹. Therefore, age-disparate sexual relationships expose individuals to an increased risk of HIV infection. It has also been noted that age-mixing in sexual relationships is likely to reduce the younger person's ability to negotiate safe sex successfully ²⁵⁰. Increasing female education regarding factors that increase HIV transmission rates might reduce age-disparate relationships, thereby decreasing infection rates in this age group. Pettifor et al. observed that education might lower young women's susceptibility to HIV infection, suggesting that interventions that target structural and partner-level risk factors, such as keeping young women in school, are necessary to lessen this vulnerability ²⁵¹. This notion is supported by a study that found that higher female education was associated with a lower occurrence of age-disparate relationships ²⁵².

A significant limitation of our study was that we did not address the issue of differential selection. This is when disease-related mortality affects the interpretation of observed serological profiles. Instead, we assumed that the excess mortality of individuals testing positive was negligible. For obvious reasons, this is a crude assumption, and other authors have proposed models which address this issue ^{205, 217, 239}. However, our main advantage was combining three cross-sectional surveys from different time points, thereby jointly

modelling age and epidemic effects. Since these surveys are conducted after every five years, the next one was scheduled for 2020 but because of COVID, no recent DHS survey has been conducted for Zimbabwe.

Combined cross-sectional surveys should be used to evaluate control measures and monitor the trends in HIV FOI. The same age groups should also be adopted for each successive survey, assuming that the temporal trend in the FOI can be monitored from the changes in prevalence. Future work would significantly be important to include differential selection and refine model estimates within a Bayesian framework by integrating other data sources.

6.5 AUTHOR CONTRIBUTIONS

RBM and EM conceived of the presented idea. RBM developed the theory and performed the data analysis. RBM and EM verified the analytical methods and models. EM supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

6.6 CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6.7 ACKNOWLEDGMENTS

This work was supported through the Developing Excellence in Leadership Training and Science Africa (DELTA) initiative. The DELTA's Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating

Excellence in Science in Africa (AESAs) and supported by the New Partnership for Africa's Development Planning and Coordinating (NEPAD) Agency with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the United Kingdom (UK) government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

CHAPTER 7

THESIS PAPER 5: Spatial Heterogeneity in Relationship Between District Patterns of HIV Incidence and Covariates in Zimbabwe: A Multiscale Geographically Weighted Regression Analysis.

Rutendo Birri Makota^{1*}, Eustasius Musenge¹

¹ Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

*** Correspondence:**

Rutendo Birri Makota

rutendobbirri@gmail.com

Abstract

Objective: Most studies in Zimbabwe have explored the relationship between HIV prevalence and covariates, mainly at national and provincial levels using global regression models. Therefore, the aim of this paper is to investigate the local/district patterns of HIV incidence in Zimbabwe over the 10-year period between 2005-2015 using the Demographic Health Survey data.

Methods: The outcome variable for our analysis was HIV incidence. The HIV incidence for each district was estimated using the catalytic model. Hot spot Analysis was performed using the Getis-Ord G_i^* statistic, which identifies statistically significant hot and cold spots using a set of weighted features. To produce smooth surfaces of HIV incidence for visualisation and data generation at the district level, interpolation maps of HIV incidence were created through the Empirical Bayesian Kriging. To determine the global relationship between HIV incidence and the significant covariates, the OLS model was fitted for each survey year. Spatial autocorrelation was used to assess the spatial independence of residuals using the Moran's I spatial autocorrelation. The Multiscale Geographically Weighted Regression (MGWR) was used to assess if the relationship between HIV incidence and covariates varied across space.

Results: Tsholotsho district remained a 99% confidence hot spot throughout the decade under observation in this study. 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 MGWR, we observed that Matabeleland North's HIV incidence rates are driven by wealth index, multiple sex partners, STI and females with older partners.

Conclusion: The study disseminated the message that different factors vary from district to district and over time. Findings from this study could be useful to policymakers for the purposes of resource allocation in the context of public health programmes. This study's findings also indicate the need to zoom in on districts like Tsholotsho, which consistently have a high burden of HIV over time.

Keywords: Spatial Analysis¹, HIV Incidence², Demographic Health Survey³, Multiscale Geographically Weighted Regression⁴, Zimbabwe⁵.

7.1 INTRODUCTION

In Zimbabwe, the HIV epidemic has been found to be heterogenous, varying by province, with Matabeleland South having the highest burden ⁸⁰. Given a set of covariates that drive the HIV epidemic in Zimbabwe, this means that because of the heterogeneous nature of the epidemic, the responses to these covariates across the country would be different. Understanding the HIV epidemic and its spatial variation is essential as it allows the implementation of policies targeted explicitly at high risks areas and groups. The use of geospatial analytical methods is useful in achieving this goal. However, among the limited but growing uses of geospatial analysis in Africa, moreover in Zimbabwe, very few studies have analysed the spatiotemporal variation in the HIV epidemic.

Most studies in Zimbabwe have explored the relationship between HIV prevalence and covariates, mainly at national and provincial levels using global regression models. Schaefer et al. explored factors that were likely to influence the equity implications and cost-effectiveness geographically of HIV resource allocation in Manicaland Province ²⁵³. Another study by Oseni et al. investigated the sub-regional changes in HIV prevalence in Zimbabwe ²⁵⁴. Gwitira et al. interestingly tested for the co-existence of the geographical distribution of HIV/AIDS and malaria in Zimbabwe ²⁵⁵. Diego et al. constructed high-resolution maps of intranational estimates of HIV prevalence in Zimbabwe based on spatial variables ²⁵⁶. Lastly, Manyangadze et al. looked at socio-economic factors spatially associated with HIV incidence in Zimbabwe ²⁵⁷. Although significant work has been done in Zimbabwe regarding spatial analysis, besides spatially modelling the HIV epidemic in terms of prevalence and at the province level, not much has been done in modelling the epidemic in terms of incidence and at the district level. This led to the

paper's aim to investigate the local/district patterns of HIV incidence in Zimbabwe over the 10-year period between 2005-2015 using the demographic health survey data.

7.2 MATERIALS AND METHODS

7.2.1 Data Source

The source of data was the Zimbabwe Demographic and Health Survey (ZDHS) conducted in 2005-06¹⁰⁴, 2010-11¹⁰⁶ and 2015¹⁰⁵. Enrolment of individuals in the ZDHS involved a two-stage sampling procedure in selecting households. The first stage involved selecting enumeration areas (EAs) and the second stage involved randomly selecting households from the EAs. Full details of the survey methodology are described elsewhere⁷⁹. The individual records for males and females of each survey year were appended. The appended datasets were then merged to the HIV datasets. Lastly, the datasets of Global Positioning System (GPS) coordinates of all the EAs were merged to the HIV individual records, meaning that all individuals were assigned the location of the centroid of their respective survey EA. After assigning each individual to the respective EA with GPS coordinates, each individual was placed in their respective districts. Zimbabwe is comprised of 60 districts which are shown in **Figure 7.1**.

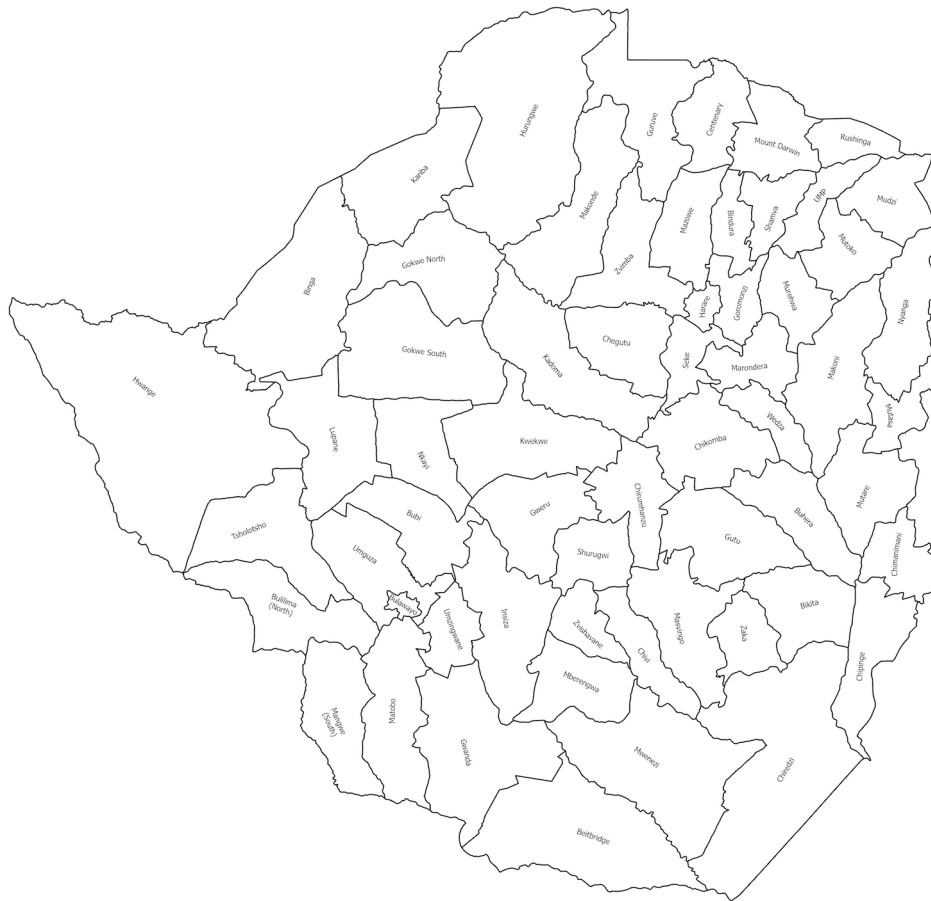


Figure 7.1: Map of Zimbabwe showing the 60 districts.

7.2.2 Study Measures

The outcome variable for our analysis was HIV incidence. The HIV incidence for each district was estimated using the catalytic model. Details of the estimation procedure are explained elsewhere ²⁵⁸. The independent variables included age, sex, marital status, condom use, age of partner, STI in the past year, first sexual debut before the age of 15 years, wealth index, age of household head, multiple sex partners, times away from home, highest level of education, type of earnings and employment status.

7.2.3 Exploratory Data Analysis

7.2.3.1 Hot Spot Cluster and Outlier Analysis

Hot spot Analysis was performed using the Getis-Ord G_i^* statistic, which identifies statistically significant hot and cold spots using a set of weighted features²⁵⁹. Each district was assigned a Z-score, which identified the intensity of the hot spot and cold spot zones for the HIV incidence. The index of the Getis-Ord G_i^* statistic was calculated as:

$$G_i^* = \frac{\sum_j^n w_{ij} y_j - \bar{y} \sum_j^n w_{ij}}{s \sqrt{\frac{n \sum_j^n w_{ij}^2 - (\sum_j^n w_{ij})^2}{n-1}}} \quad \text{Equation 7.1}$$

Where y_j is the HIV incidence for district j , the spatial weights between i and j represented by w_{ij} , n being the total number of districts, $\bar{y}$ is the mean and s the standard deviation.

To produce smooth surfaces of HIV incidence for visualisation and data generation at the district level, interpolation maps of HIV incidence were created through the Empirical Bayesian Kriging. Empirical Bayesian kriging (EBK) through repeated simulations accounts for the error in semi variogram estimation²⁶⁰. Advantages of the EBK include predicting more accurate standard errors than other kriging methods and allowing the accurate prediction of moderately nonstationary data²⁶⁰. Contrary to the Hot spot analysis which identifies groups within an area, the cluster and outlier analysis identifies groupings according to criterion of proximity. Cluster and outlier analysis was performed using the Anselin local Moran's I . Cluster and outlier analysis, Hot spot analysis and interpolation were all performed in ArcGIS Pro version 3.1.

7.2.3.2 Exploratory Regression

To identify independent variables associated with HIV infection, we first conducted an exploratory regression of the 14 variables mentioned above for each survey year separately. The exploratory regression was performed in ArcGIS Pro version 3 using the Exploratory regression tool. The Exploratory regression (ER) tool in ArcGIS is a data mining tool that finds all possible combinations of the explanatory variables that will pass the necessary ordinary least regression (OLS) diagnostics. The advantage of using the ER tool in ArcGIS rather than ordinary stepwise regression is the ability to find models that meet all the requirements and assumptions of the OLS rather than just searching for models with the highest adjusted R^2 .

The significant risk factors obtained from the ER included: having an STI in the past year, being married and single for the 2005-06 ZDHS. For the 2010-11 ZDHS, the significant factors from the ER included: being richer according to the ZDHS wealth index, being single and females with partners five or more years older than them. Lastly, the significant factors from the ER of the 2015 ZDHS included: females with partners five or more years older than them, having an STI in the past year, having multiple sex partners and richer according to the ZDHS wealth index. These significant variables were then run in the ordinary least squares (OLS) model, and residuals were checked for spatial autocorrelation.

7.2.4 Spatial Statistical Models

7.2.4.1 OLS regression – Global Regression

To determine the global relationship between HIV incidence and the significant covariates, the OLS model was fitted for each survey year. The three OLS models fitted

for (1) the 2005-06 survey year, (2) the 2010-11 survey year and (3) the 2015 survey year were as follows:

$$\begin{aligned}
 (1) \text{ HIV} &= \beta_0 + \beta_1 \text{PropSTI}_i + \beta_2 \text{PropMarried}_i + \beta_3 \text{PropRicher}_i + \varepsilon_i \\
 (2) \text{ HIV} &= \beta_0 + \beta_1 \text{PropRicher}_i + \beta_2 \text{PropSingle}_i + \beta_3 \text{OlderPartner}_i + \varepsilon_i \\
 (3) \text{ HIV} &= \beta_0 + \beta_1 \text{PropOldpartner}_i + \beta_2 \text{PropSTI}_i + \beta_3 \text{PropRicher}_i + \\
 &\quad \beta_4 \text{PropSexPartners}_i + \varepsilon_i
 \end{aligned}$$

Where HIV is the estimated incidence, β_0 , the intercept, β_i , regression coefficients, and i represents the 60 districts:

- $\Rightarrow$ *PropMarried*: weighted proportions of individuals married
- $\Rightarrow$ *PropSingle*: weighted proportions of individuals single
- $\Rightarrow$ *PropRicher*: weighted proportions of individuals richer
- $\Rightarrow$ *PropOldpartner*: weighted proportions of individuals with partners five or more years older
- $\Rightarrow$ *PropSTI*: weighted proportions of individuals who had an STI in the past 12 months
- $\Rightarrow$ *PropSexPartners*: weighted proportions of individuals who have multiple sex partners.

Considering that OLS assumes stationarity, meaning that the association between response and predictor variable does not vary spatially. Therefore, to test for stationarity in the spatial pattern, the Moran's I statistics for spatial autocorrelation was used.

$$I = \frac{n \sum_i^n \sum_j^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{(\sum_i^n \sum_j^n w_{ij}) \sum_i (y_i - \bar{y})^2} \quad \text{Equation 7.2}$$

The Moran's I is a measure of spatial autocorrelation in an area, which explains the degree of dependence between values of a variable at different locations. In addition, we tested for multicollinearity using the variance inflation factor (VIF) and local variation using the Koenker (BP) statistic. The OLS was performed in ArcGIS Pro version 3.1.

Using the same set of variables used in the OLS, we performed the Geographically Weighted Regression (GWR) and Multiscale Geographically Weighted Regression (MGWR) to account for nonstationarity between response (HIV incidence) and predictor variables. The GWR and MGWR were performed using the MGWR software which is also integrated in the *mgwr* python package (version 2.2.1).

7.2.4.2 Geographically Weighted Regression (GWR) model

Considering that the covariates' predictability capacity differs from one cluster to the next, these cluster variations can be identified through the use of GWR²⁶¹. The GWR assesses whether the association between HIV incidence and covariates vary across space. The GWR creates an equation for each district while using data from nearby features; hence the coefficient takes different values for each district. The maps created through GWR of the coefficients associated with each explanatory variable provide guidelines for targeted interventions within different districts. The GWR model can be expressed as follows:

$$y_i = \beta_0(u_i v_i) + \beta_1(u_i v_i)x_{i1} + \beta_2(u_i v_i)x_{i2} + \dots + \beta_j(u_i v_i)x_{ij} + \varepsilon_i$$

$$= \sum_{j=0}^n \beta_j(u_i v_i) x_{ij} + \varepsilon_i \quad \text{Equation 7.3}$$

Where y_i are observations of HIV incidence, $u_i v_i$ are the longitude and latitude geographical coordinates for the i th observation, $\beta_j(u_i v_i)$ are the parameters for x_{ij} describing the relationship around geographical locations $(u_i v_i)$ and ε_i is the error term.

7.2.4.3 Multiscale Geographically Weighted Regression (MGWR)

The GWR constrains the local (district) relationships within each model to vary at the same scale. However, the MGWR allows the response variable (HIV incidence) and the different covariates to vary at different spatial scales ²⁶². Meaning that the bandwidths indicating the data-borrowing range can vary across parameter surfaces. Therefore, the equation of the MGWR differs slightly from the GWR by introducing the bandwidth:

$$\begin{aligned} y_i &= \beta_{bw0}(u_i v_i) + \beta_{bw1}(u_i v_i)x_{i1} + \beta_{bw2}(u_i v_i)x_{i2} + \dots + \beta_{bwj}(u_i v_i)x_{ij} + \varepsilon_i \\ &= \sum_{j=0}^n \beta_{bwj}(u_i v_i) x_{ij} + \varepsilon_i \end{aligned} \quad \text{Equation 7.4}$$

Where the bandwidth used for calibration of the j th conditional relationship is represented by bwj in β_{bwj} .

7.3 RESULTS

Figure 7.2 shows the spatial distribution of HIV incidence over the ten years under investigation in this study. Districts with dark red colour had the highest HIV incidence, and light blue colour indicated districts with lower HIV incidence. The highest HIV incidence recorded for one of the districts was 0.029, 0.025 and 0.026 for the years 2005-06, 2010-11 and 2015, respectively. High HIV incidence rates were observed in 11 (18.3%), 8 (13.3%) and 9 (15.0%) districts in the years 2005-06, 2010-11 and 2015, respectively. Most of the districts with high HIV incidence rates in 2005-06 and 2010-11 decreased in 2015.

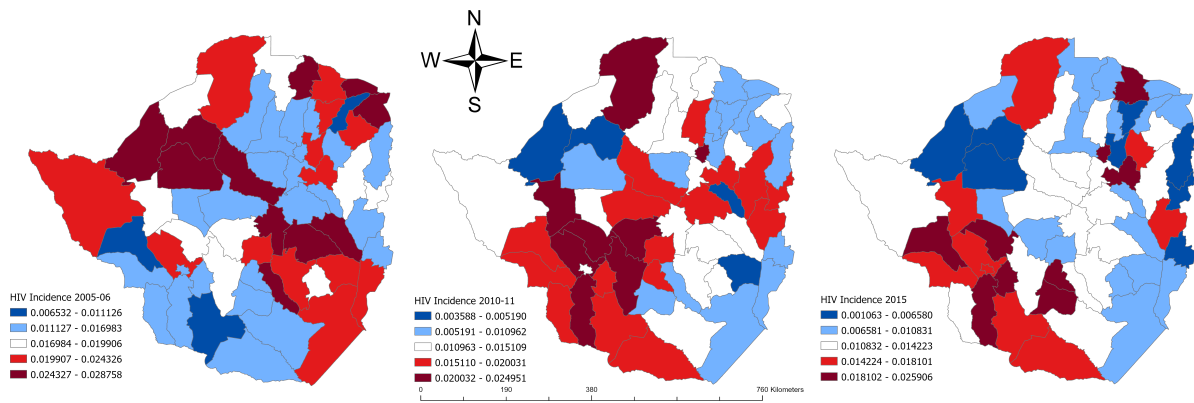


Figure 7.2: Spatial Variation of HIV incidence per survey year

7.3.1 Hot spot analyses – Spatial clustering of HIV incidence

The Getis-Ord G_i^* statistic was used to reveal spatial patterns and statistically significant clustering of districts of the HIV incidence into hot spots and cold spots in different survey years under consideration in this study. HIV hot spots were represented in the map with the colour red, while cold spots had the colour blue. Areas with no colour were not statistically significant based on the Getis-Ord G_i^* statistic. In 2005-06, hot spots for HIV incidence with 99% confidence were observed in 11 (18.3%) districts: Mount Darwin, Rushinga, Uzumba-Maramba-Pfungwe (UMP), Shamva, Mudzi, Mutoko, Chiredzi, Mwenezi, Beitbridge, Mberengwa and Tsholotsho (**Figure 7.3(a)**). No significant 99% cold spots were observed in 2005-06, rather just one 90% cold spot. In 2010-11, Bikita, Buhera, Chikomba, Gutu, Hwange and Tsholotsho were the observed 99% hot spots, down by five districts from the 2005-06 count. Positively, Chiredzi and Beitbridge which were 99% confidence hot spots in 2005-06, had become 90% confidence hot spots in 2010-11. Three 99% confidence cold spots were observed in 2010-11, including Mount Darwin, Mudzi and Nyanga districts. In 2015, there were three 99% confidence hot spot remaining, which were Tsholotsho, Bulilima and Mangwe districts all in Matabeleland

North Province. Tsholotsho district remained a 99% confidence hot spot throughout the decade under observation in this study. From the three 99% confidence cold spots observed in 2010-11, two more 99% confidence cold spots were added in 2015, which included Centenary/Muzarabani and Rushinga districts. Based on Figure 7.3(a-c) hot spot analysis, it can be noticed that the HIV incidence hot spots were significantly decreasing in Zimbabwe over time.

Figure 7.3(d-f) showed that the HIV incidence clusters were delimited with four zones. High-high clusters were represented with red colour. The high-high clusters were located in Lupane, Tsholotsho, Matobo, Gwanda, Beitbridge, Mberengwa and Zaka in 2005-06. In 2010-11, Hwange, Tsholotsho, Umguza, Bulawayo, Umzingwane and Beitbridge were high-high clusters. Lastly, Hwange, Tsholotsho, Umguza, Bulawayo, Matobo and Beitbridge were high-high clusters in 2015. In contrast to the high-high clusters, low-low clusters were represented by dark blue colour, meaning they were low HIV incidence risk surrounded by districts with similar risk. This low-low cluster was observed in Nyanga in 2005-06, Mount Darwin, Rushinga, Uzumba-Maramba-Pfungwe (UMP), Shamva, Mutoko and Nyanga in 2010-11, lastly, Guruve, Centenary, Mount Darwin, Rushinga, UMP, Shamva, Mutoko, Nyanga, Bindura and Goromonzi in 2015.

The interpolation of HIV incidence hot spots produced Figure 7.3(g-i) diagrams. Areas with high HIV incidence are shown in red (hot spots), while areas with low HIV incidence are shown in blue (cold spots). From Figure 7.3(g-i), continuous surfaces of HIV incidence also confirmed and spatially unpacked district and regional HIV incidence variations in Zimbabwe. Figure 7.3(g-i) captured the general intensification of HIV incidence and spatial extent from 2005 to 2015. It can be noticed that the southern and western parts of Zimbabwe have high HIV incidence rates, as shown by the red zones in

Figure 7.3(g-i), whereas the eastern and northern regions of Zimbabwe have cold spots of the HIV epidemic.

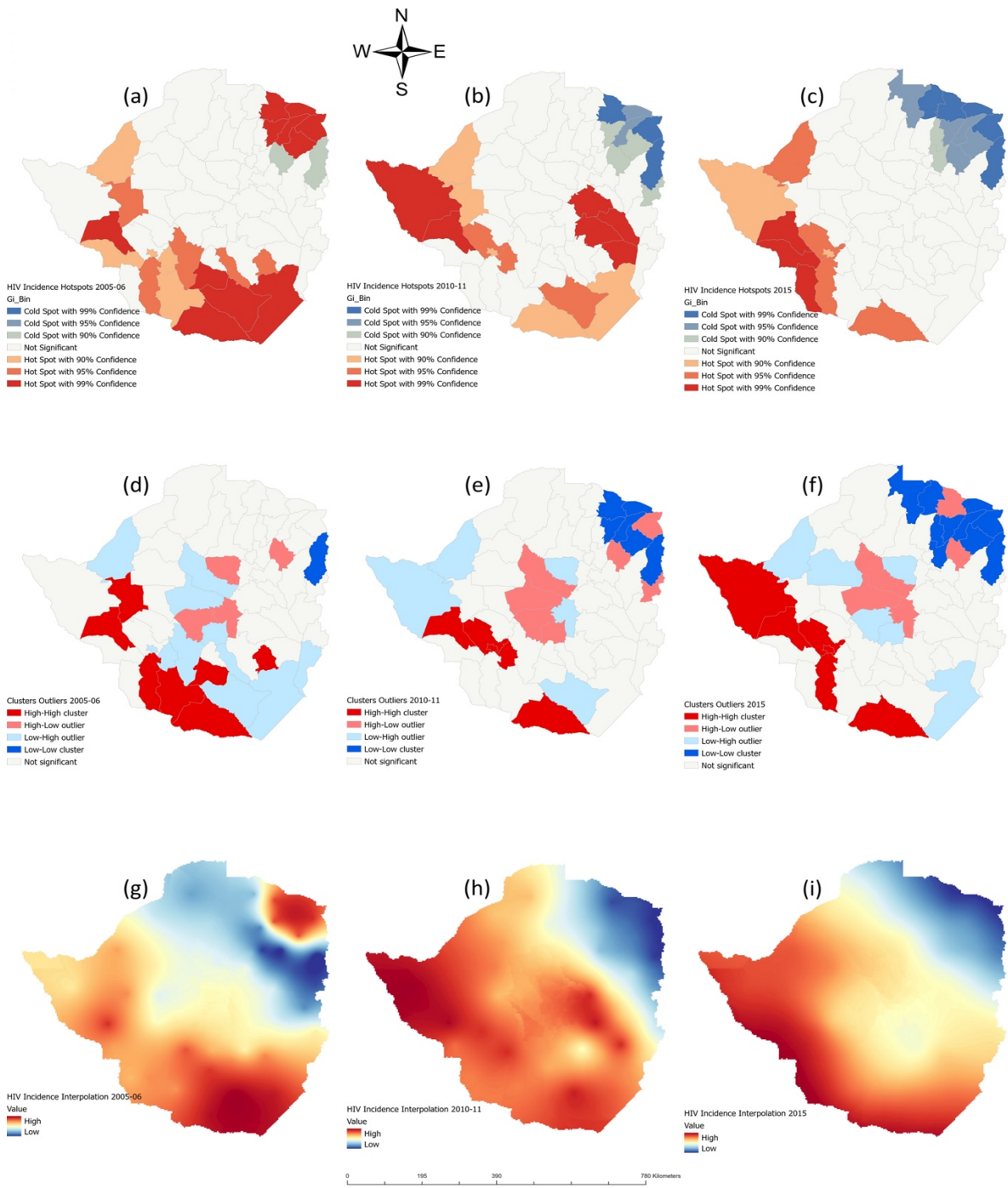


Figure 7.3: HIV incidence Hot spots (a-c), Clusters (d-f) and Interpolation (g-i) per survey year.

7.3.2 Ordinary Least Squares Regression Results

The Ordinary Least Squares regression (OLS) was used to determine the best combination of variables that optimise on predicting HIV incidence for each survey year. Therefore, OLS of estimated district-level HIV incidence produced three models for each survey year considered in this study. The final variables used in all three models were selected through a multi-stage statistical process from an initial pool of 14 variables drawn from literature and based on availability at the survey year and district level. The first step in the variable selection for the OLS included an Exploratory Regression (ER) done on all the 14 variables on the three survey years separately. The ER yielded three variables (STI, married and single) for the 2005-06 data, three variables (single, richer and older partner > 5 years) for the 2010-11 data, and four variables (Multiple sex partners, STI, richer and older partner > 5 years) for the 2015 data. The variables obtained through ER were then used for the OLS resulting in results displayed in **Table 7.1**.

It was observed that all variables for all the survey years were statistically significant as shown in **Table 7.1**. Based on the OLS results, we checked for multicollinearity, residual stationarity, and global spatial autocorrelation. Results from the spatial autocorrelation indicated that, all the three models' residuals were randomly distributed with Moran's I p-values of 0.59, 0.40 and 0.26 for the 2005-06, 2010-11 and 2015 dataset respectively. Since the p-values of the Moran's I statistic were not statistically significant, it meant that the structure of the residuals was random, and the model is properly specified. It was also observed that all the variables for all the survey years under study had VIF values less than 5, indicating that there was low multicollinearity in the models. The Jarque-Bera Statistic was not statistically significant for all the models meaning that the models were

not biased, and key variables were selected. Lastly, the Koenker (BP) statistic was significant for all models in all survey years with p-values of 0.002, 0.006 and 0.037 for the 2005-06, 2010-11 and 2015 respectively. This meant that the residuals of the errors from the OLS were non-stationary and the GWR or MGWR were the best models to model the spatial variability. Spatial variation of the selected variables are presented through choropleth maps in Supplementary Information.

Table 7.1: Ordinary Least Squares Regression (OLS) results.

2005-06							
Variable	Coefficient	Standard Error	P value	VIF	Moran's I (p-value)	Koenker (BP) (p-value)	Jarque-Bera (p-value)
Intercept	-0.025	0.003	<0.001		-0.01 (0.59)	14.33 (0.002)	0.91 (0.63)
STI	0.032	0.015	0.03	1.02			
No condom Use	0.020	0.005	0.02	1.43			
Married	0.039	0.005	<0.001	1.46			
2010-11							
Intercept	0.001	0.001	0.37		-0.03 (0.40)	12.31 (0.006)	0.29 (0.86)
Single	0.009	0.004	0.01	1.45			
Richer	0.019	0.004	<0.001	2.60			
Older partner > 5 years	0.011	0.003	0.01	2.17			
2015							
Intercept	-0.002	0.001	0.02		-0.03 (0.26)	10.20 (0.037)	4.07 (0.13)
Multiple Sex Partners	0.013	0.004	<0.001	4.17			
STI	0.043	0.019	<0.001	1.02			
Richer	0.008	0.003	0.03	1.06			
Older partner > 5 years	0.010	0.004	<0.001	4.17			

7.3.3 Geographical/Spatial Variation of HIV incidence Results

As observed in the results presented in **Table 7.2**, the MGWR model is a better fit based on the adjusted R^2 and AICc values. In addition, the MGWR model explained 81.5%, 86.5% and 80.2% of the total variation in HIV incidence in the 2005-06, 2010-11 and 2015 models, respectively, which was an improvement from the OLS and GWR. For

additional comparison between MGWR and GWR, see Supplementary Information with spatial maps for both models.

Table 7.2: Regression Results Comparison

Survey Year	Model	AICc	Adjusted R ²
2005-06	OLS	145.6	77.0%
	GWR	90.0	77.0%
	MGWR	82.7	81.5%
2010-11	OLS	135.2	72.5%
	GWR	103.6	73.2%
	MGWR	97.1	86.5%
2015	OLS	140.3	70.9%
	GWR	102.1	70.9%
	MGWR	99.3	80.2%

Figure 7.4 shows how the coefficients of STI, no condom use and married for the ZDHS 2005-06 survey year are associated with HIV incidence over space. The red areas represented in red show where the coefficients are large and locations where the explanatory variables are strong predictors of high HIV incidence. Having had an STI, using no condoms and married where all positive predictors of high HIV incidence in 2005-06, indicating that HIV incidence rates increases as the percentage of individuals with an STI, who do not use condoms and are married increased. As observed in **Figure 7.4**, the impact of STI, condom use and married on HIV incidence has been strong in the Eastern parts of Zimbabwe with Mashonaland Central, Mashonaland East and Manicaland provinces. If condoms are used properly and consistently, they are one of the most effective ways of reducing STIs and HIV transmission rates. Therefore, the promotion of condom use and continuous free distribution of condoms in the Eastern parts of Zimbabwe focusing on married individuals would decrease HIV transmission

based on **Figure 7.4(a-c)**. Additionally, areas with high burden of STIs, sexual reproductive health services should be free and easily accessible.

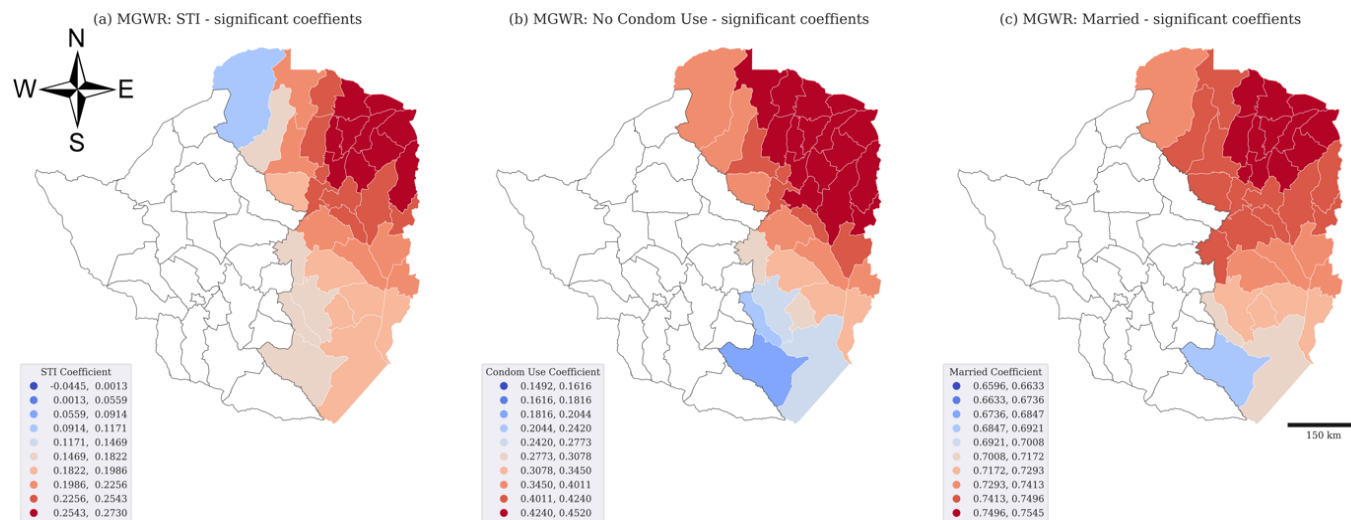


Figure 7.4: Coefficient estimates of (a) STI in the past 12 months; (b) No condom use; and (c) Married in describing HIV incidence using Multiscale Geographically Weighted Regression (MGWR) for the ZDHS 2005-06 survey year with ± 1.96 significant areas indicated in colour.

Figure 7.5 shows how the coefficients of richer, single and females who have an older partner > 5 years for the ZDHS 2010-11 survey year are associated with HIV incidence over space. As shown by districts in red, being richer according to the wealth index in districts situated in Matabeleland North was a strong predictor of HIV incidence. As the percentage of single individuals increased in Masvingo province, the HIV incidence also increased by 24-25% as shown in **Figure 7.5(b)**, indicating that being single was a strong predictor of HIV incidence in Masvingo province. Females having older partners was

strongly associated with high HIV incidence and the areas mostly affected were Binga, Kariba, Gokwe North and Gokwe South districts as shown in **Figure 7.5(c)**.

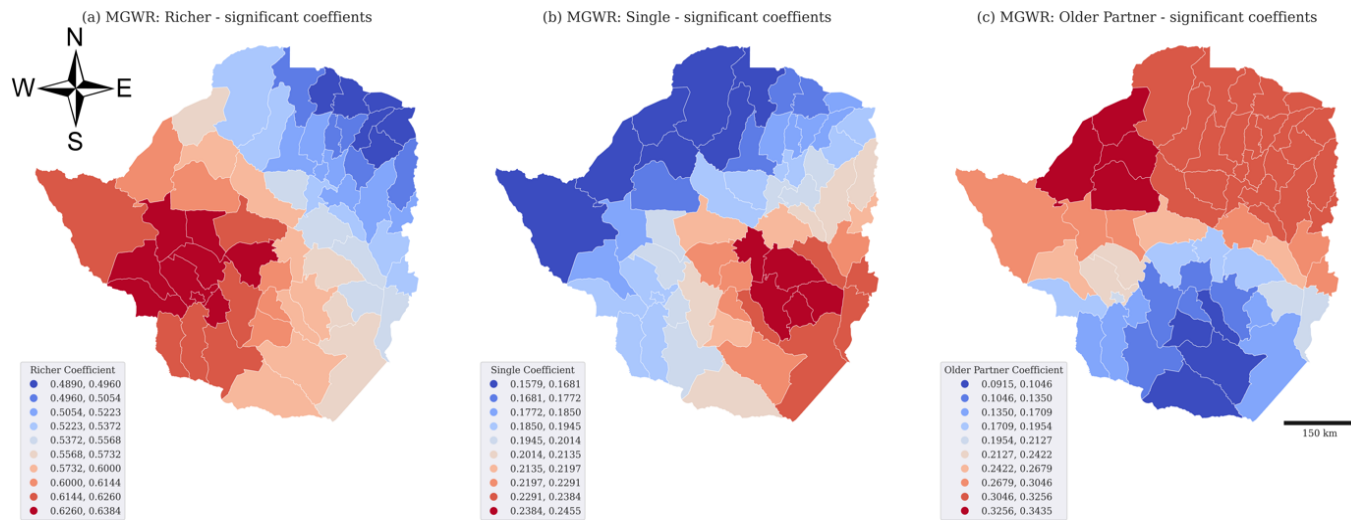


Figure 7.5: Coefficient estimates of (a) Richer; (b) Single; and (c) Older partner > 5 years in describing HIV incidence using Multiscale Geographically Weighted Regression (MGWR) for the ZDHS 2010-11 survey year with ± 1.96 significant areas indicated in colour.

Figure 7.6 shows how the coefficients of multiple sex partners, STI, richer and females who have an older partner > 5 years for the ZDHS 2015 survey year are associated with HIV incidence over space. The red areas represented in red show where the coefficients are large and locations where the explanatory variables are strong predictors of high HIV incidence. Tsholotsho, Lupane, Nkayi, Bubi and Umguza, all districts in Matabeleland North showed that a high percentage increase of individuals with multiple sex partners and STI had a very strong positive association with HIV incidence. Being richer had a strong impact on HIV incidence in districts in Masvingo province, mainly Chiredzi,

Mwenezi, Mberengwa, Zvishavane and Chivi. Females having an older partner was significantly associated with a high HIV incidence in districts in Lupane and Matabeleland North districts.

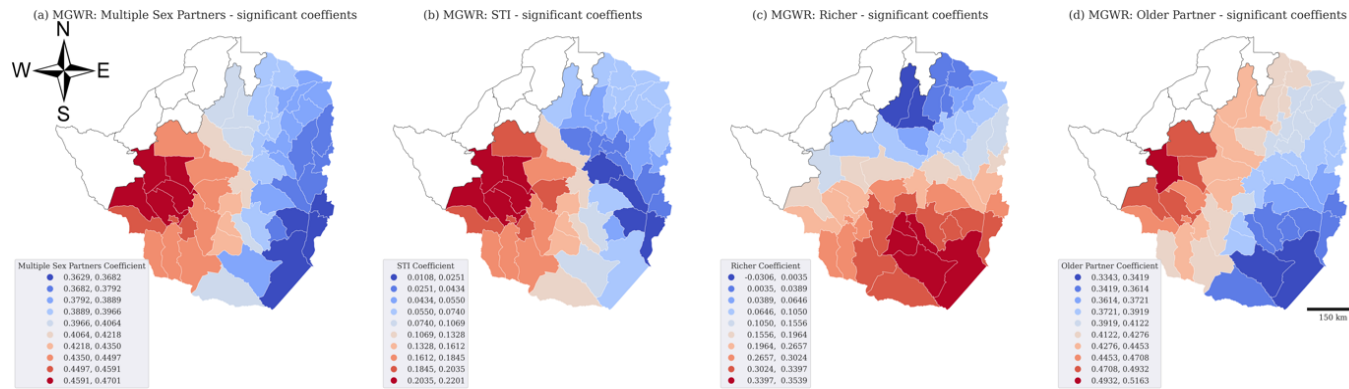


Figure 7.6: Coefficient estimates of (a) Multiple sex partners; (b) STI in the past 12 months; (c) Richer and (d) Older partner > 5 years in describing HIV incidence using Multiscale Geographically Weighted Regression (MGWR) for the ZDHS 2015 survey year with ± 1.96 significant areas indicated in colour.

7.4 Discussion

This study explored the spatial heterogeneity and spatial predictors of HIV incidence over a decade using spatial analytic methods. Notably, in 2005-06, the southern parts of Zimbabwe had many hot spot districts. These findings can be corroborated by reports that state that HIV prevalence and incidence is high in the southern parts of Zimbabwe ^{80, 263, 264}. Zimbabwe's southern region shares borders with Botswana and South Africa. Botswana and South Africa have been reported in literature to have the highest burden of HIV in Sub-Saharan Africa ²⁶⁵. Therefore, we can hypothesise that the districts in the southern parts of Zimbabwe have the highest burden of HIV since they border with

nations (South Africa and Botswana) that have also recorded high HIV incidences. One of the provinces situated in the Southern region is Matabeleland North. Matabeleland North constitutes of Tsholotsho, Lupane, Nkayi, Bubi and Umguza districts. From our findings from the MGWR, we observed that Matabeleland North's HIV incidence rates are driven by wealth index (**Figure 7.5(a)**), multiple sex partners (**Figure 7.6(a)**), STI (**Figure 7.6(b)**) and females with older partners (**Figure 7.6(d)**).

An example of one of the districts with the highest burden of HIV in the southern region of Zimbabwe is Tsholotsho ²⁶³. Tsholotsho is a district situated in Matabeleland North. Over the ten years considered in this analysis, Tsholotsho remained a constant hot spot area. It has been reported that in the southern parts of Zimbabwe, Tsholotsho has the highest burden ²⁶³. It has been reported that the key drivers of HIV in Tsholotsho include high Gender Based Violence (GBV), low condom use, low-risk perception and unequal power relations ²⁶⁶. In addition, because the district is near the border, high migrations are experienced, resulting in high levels of couples separating, which drives the HIV pandemic. Further research into why Tsholotsho has remained a hot spot area for over a decade is required in order to tailor interventions that effectively transform this district's HIV burden narrative.

This study also highlighted that the proportion of richer individuals was positively associated with high HIV incidence rates for all three survey years. The proportion of individuals richer was obtained from the variable wealth index from the DHS, which is a composite measure of a household's cumulative living standard ¹⁰⁴⁻¹⁰⁶. A study by Fox ²⁶⁷ indicated that in wealthier countries, poor individuals had a higher risk of HIV infection, whereas, in poorer countries, individuals who had more wealth were more likely to be HIV infected. Given Zimbabwe's economic situation, Zimbabwe cannot be

described as a wealthy nation; therefore, the conclusion that richer individuals are at a higher risk of HIV infection is applicable to Zimbabwe. Another study by Vinod et al. revealed that, in most cases, the HIV burden among adults increased monotonically with wealth ²⁶⁸. These two findings support our conclusion that areas with a higher proportion of richer individuals have a higher HIV incidence rate. However, another study by Gaumer et al. ²⁶⁹ suggested that the DHS survey samples may contain an inflated number of HIV positive individuals surviving in the wealthier group, thus living longer than poorer individuals, giving the impression that the HIV burden is higher in the richer group. Whether or not the DHS sampling introduces a bias, policies that target areas with high wealth index inequalities are required. To sum up, the higher HIV infection rates among wealthier individuals in Zimbabwe may be the result of complex and intersecting factors, including perceived low risk, greater economic and social power, and engagement in high-risk sexual behaviours. Addressing these factors through targeted HIV prevention and education interventions could help to reduce the spread of HIV in the country and may be particularly important for wealthier individuals who may have higher vulnerability to HIV infection.

This study observed that proportions of females with partners five years older than them were positively associated with an increase in HIV incidence in 2010-11 and 2015. Unfortunately, we could not test the same covariate in 2005-06 as this variable was not yet part of the questions asked in the DHS questionnaire. Age-disparate or age-mixing patterns refer to sexual relationships where the age difference between partners is five years or greater ²⁷⁰. A national population-based household survey conducted in South Africa noted that females who resided in rural areas were more likely to have a partner

who was five years or greater than them compared to females who resided in urban areas²⁷⁰. Our findings support this notion as districts with a high proportion of females with partners > 5 years are rural districts. Additionally, in rural areas, especially poor rural areas, lack of employment, education and general economic deprivation leads to women seeking older partners for financial security^{250, 271}. More investigations are required to understand the HIV transmission dynamics between living in rural areas and having partners > 5 years older. This investigation will warrant a tailored response to those specific areas thereby strengthening already existing HIV programmes.

In our study, we explored both the Multiscale geographically weighted regression (MGWR) and Geographically weighted regression (GWR). Multiscale geographically weighted regression (MGWR) is a spatial statistical technique that extends geographically weighted regression (GWR) by allowing for spatial heterogeneity to be detected at multiple spatial scales²⁶². Some of the strengths of the MGWR include identification of spatially varying relationships between variables more effectively than GWR because it models these relationships at different spatial scales, providing deeper insight into the underlying patterns of data²⁵⁴. Secondly, MGWR is highly flexible and can incorporate a wide range of spatial covariates, including local and global socio-economic variables, environmental factors, and other spatially related data sets²⁶². Lastly, MGWR can be used to project and forecast future spatial patterning, providing insight into the long-term consequences of policy decisions. However, as with GWR, MGWR can lead to overfitting and unstable estimates when there are too few observations relative to the number of explanatory variables. MGWR requires a relatively large amount of data, adding to the computational demands compared to GWR, and this can sometimes result in slow performance when working with large and complex data sets. MGWR deals with

nonstationarity by modelling relationships at different scales ²⁷¹. However, this can also be seen as a limitation as it can lead to difficulties in interpretation when attempting to integrate data across multiple scales or when discerning overall patterns in the data.

7.5 Study Limitations.

Although the study contributes significantly to advancing spatial analysis of the HIV epidemic in Zimbabwe, it has limitations. First, since this was a secondary data analysis from a cross-sectional study, our results are affected by recall bias and social desirability since the covariates were self-reported. Secondly, since this was a cross-sectional study design, we cannot infer causality.

Since we estimated HIV incidence for each district from HIV prevalence, we noted that the data was not geared for district-level analysis as some districts had very low HIV prevalence rates, which could have biased the final estimated value. However, since the results were corroborated with other studies, we believe that the HIV incidence estimates were not far off from the true estimates.

7.6 Conclusion

Regardless of the limitations, the study found relationships between covariates and HIV incidence over space. The study disseminated the message that different factors vary from district to district and over time. Findings from this study could be useful to policymakers for the purposes of resource allocation in the context of public health programmes. This study's findings also indicate the need to zoom in on districts like Tsholotsho, which consistently have a high burden of HIV over time. Interventions that promote partner

notification when one has an STI especially with multiple sex partners, are encouraged to help reduce asymptomatic STIs in the districts in turn, this is expected to reduce HIV incidence rates. This study also recommends policymakers to budget for more local/district-level analysis than the province to inform policies and strategies for resource allocation in areas of great need.

7.7 Supplementary Information

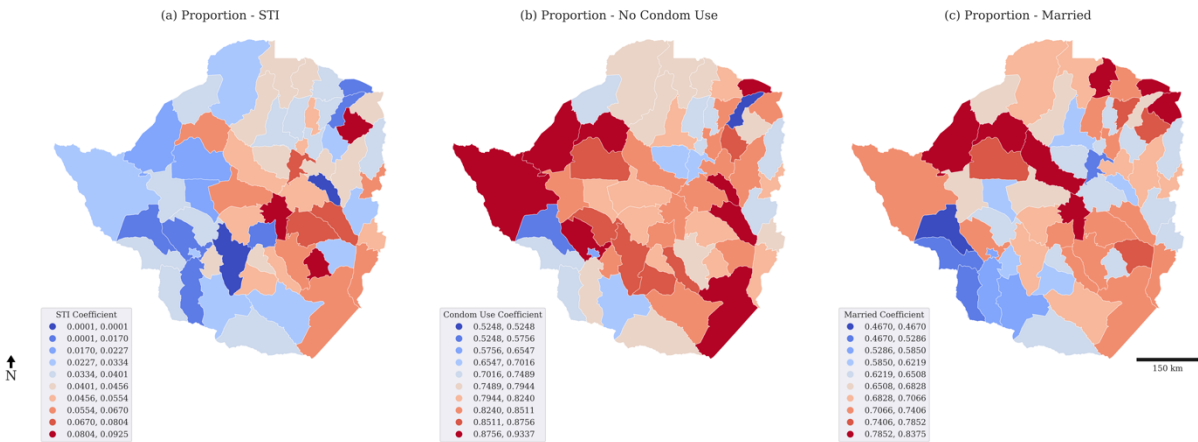


Fig S1: Choropleth maps for proportions for STI, No condom use and Married for the 2005-06 ZDHS.

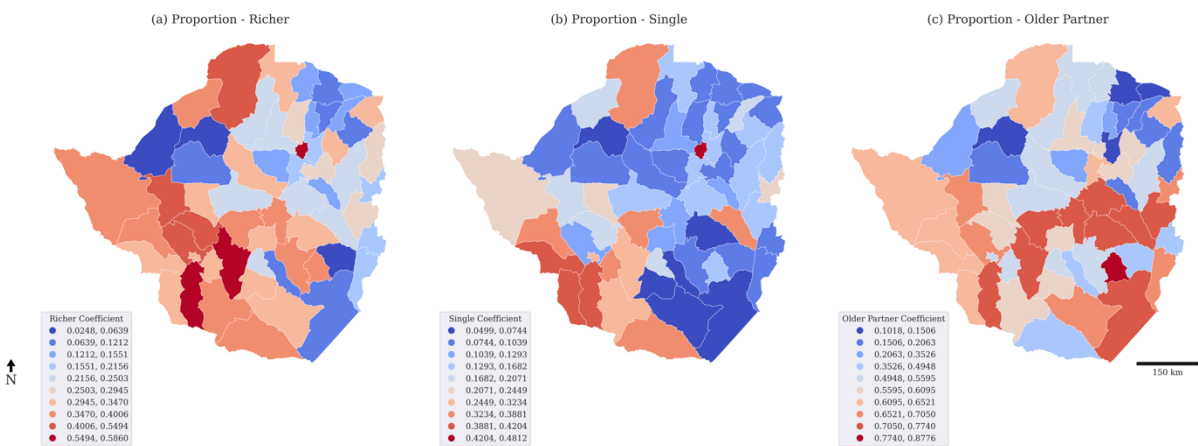


Fig S2: Choropleth maps for proportions for Richer, Single and Females with older partners > 5 years for the 2010-11 ZDHS.

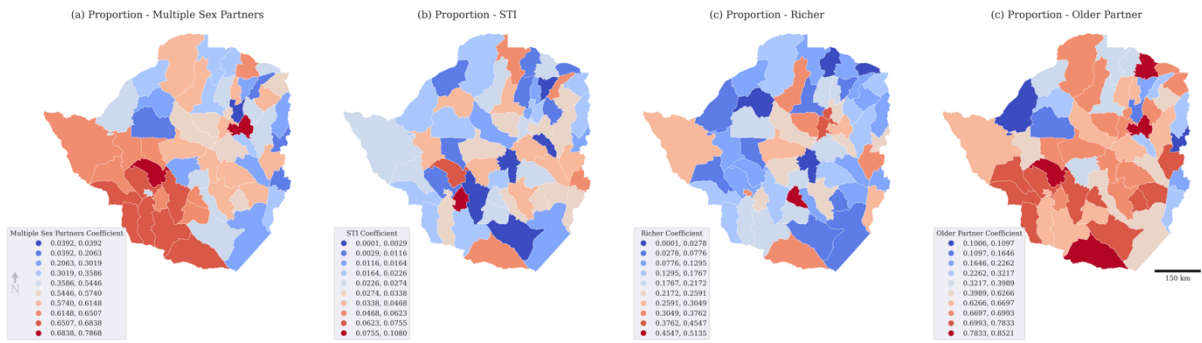


Fig S3: Choropleth maps for proportions for Multiple sex partners, STI, Richer and Females with older partners > 5 years for the 2015 ZDHS.

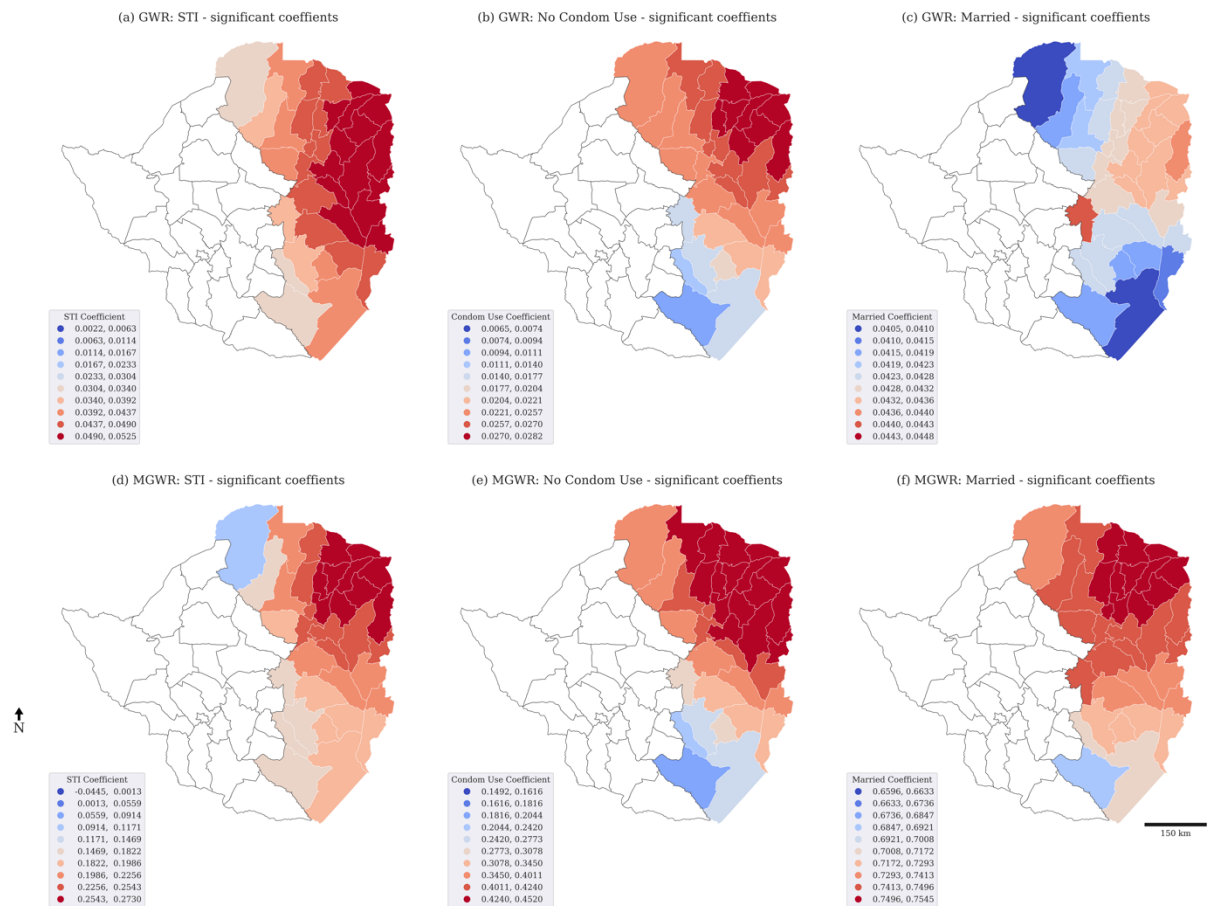


Fig S4: Coefficient estimates of (a,d) STI in the past 12 months; (b,e) No condom use; and (c,f) Married in describing HIV incidence using GWR & MGWR for the ZDHS 2005-06 survey year with ± 1.96 significant areas indicated in colour.

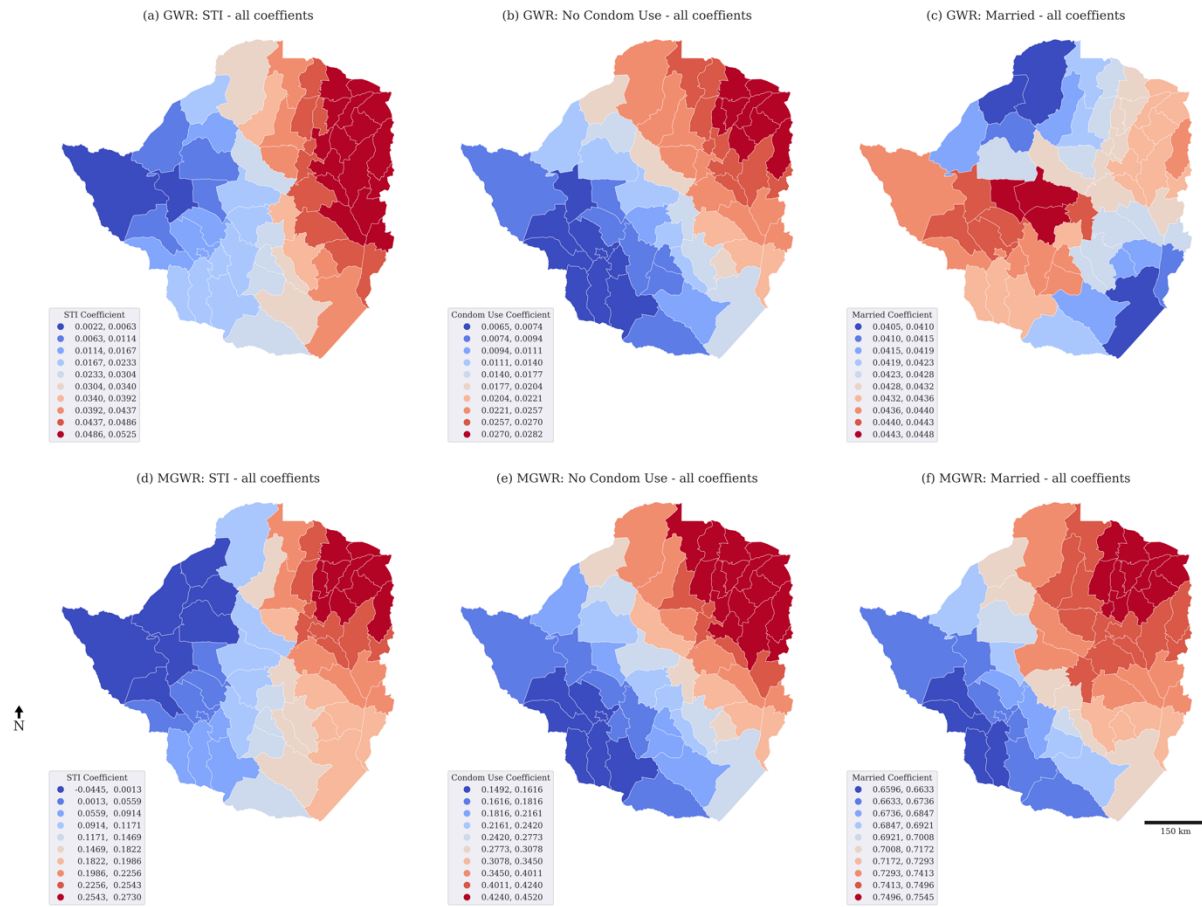


Fig S5: Coefficient estimates of (a,d) STI in the past 12 months; (b,e) No condom use; and (c,f) Married in describing HIV incidence using GWR & MGWR for the ZDHS 2005-06 survey year.

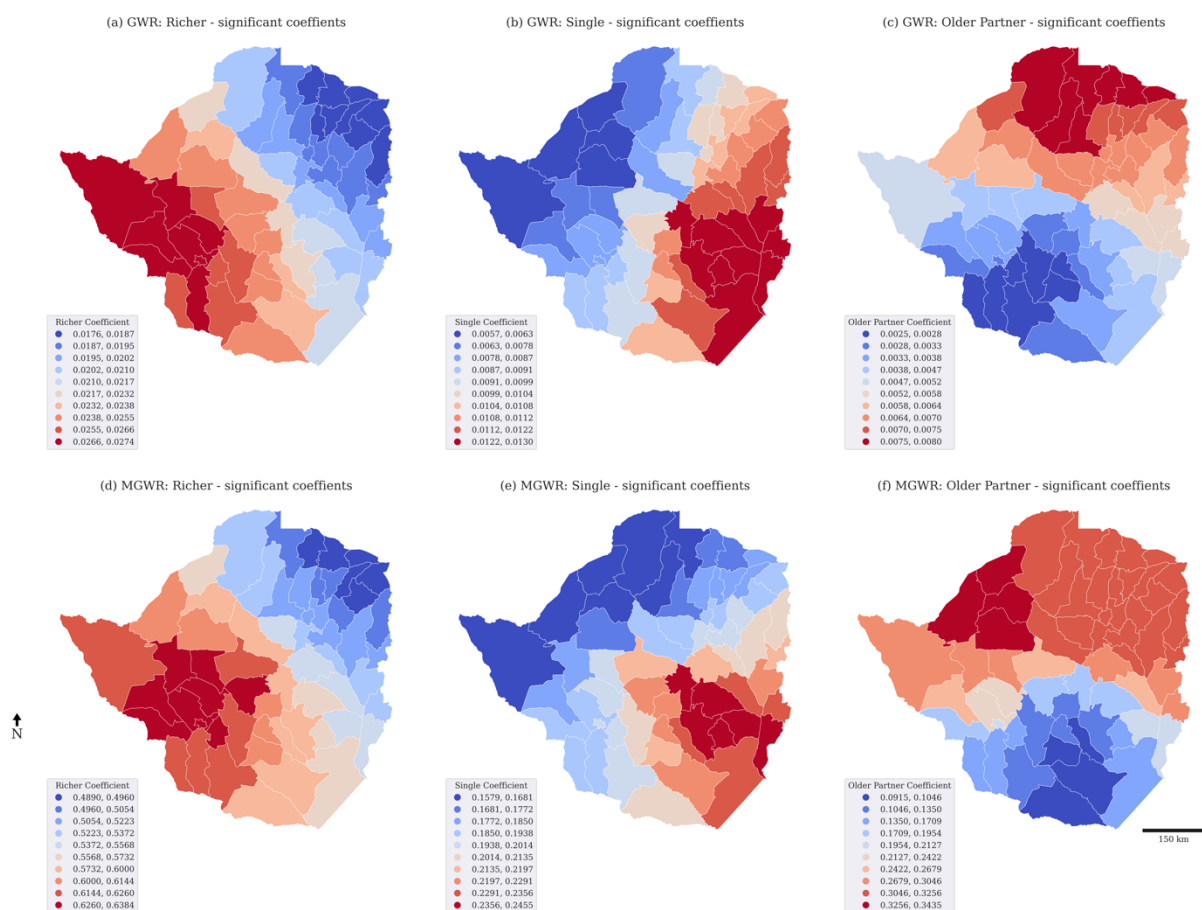


Fig S6: Coefficient estimates of (a,d) Richer; (b,e) Single; and (c,f) Older partner > 5 years in describing HIV incidence using GWR & MGWR for the ZDHS 2010-11 survey year

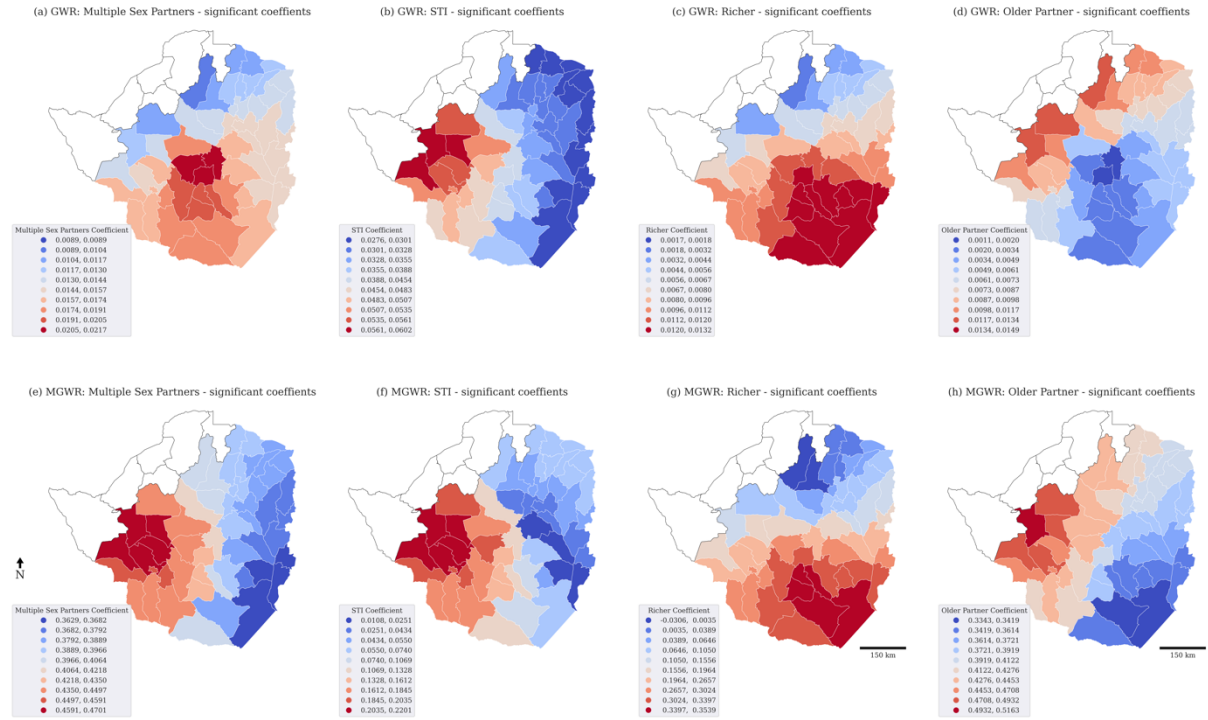


Fig S7: Coefficient estimates of (a,e) Multiple sex partners; (b,f) STI in the past 12 months; (c,g) Richer and (d,h) Older partner > 5 years in describing HIV incidence using GWR & MGWR for the ZDHS 2015 survey year with ± 1.96 significant areas indicated in colour

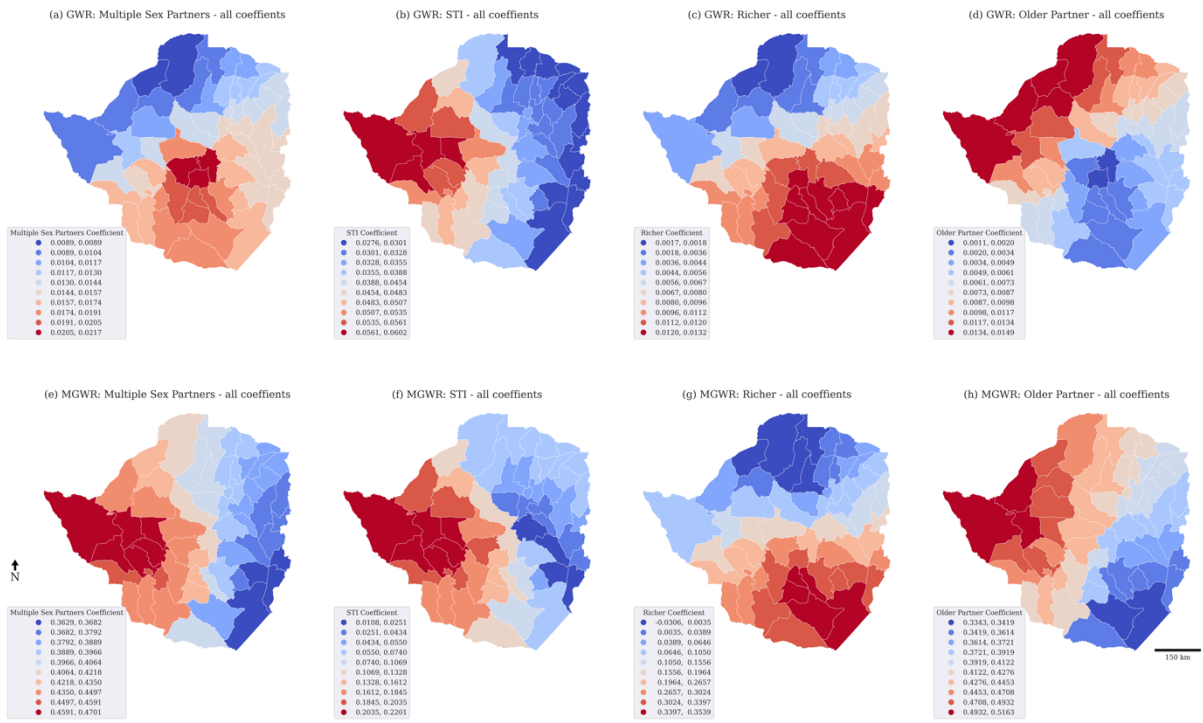


Fig S8: Coefficient estimates of (a,e) Multiple sex partners; (b,f) STI in the past 12 months; (c,g) Richer and (d,h) Older partner > 5 years in describing HIV incidence using GWR & MGWR for the ZDHS 2015 survey year.

CHAPTER 8

8.1 DISCUSSION AND CONCLUSIONS

The 90-90-90 UNAIDS fast track targets' aim was to have zero HIV incidence by 2030, and this thesis objectives were centred around these targets to help inform the progress and achieve the 2030 goals. This thesis looked at the 90-90-90 fast track changes instead of the 95-95-95 track changes, which are now being implemented by the time this thesis is being submitted. This is because the thesis used data from the DHS, which conducts its surveys every five years, and for Zimbabwe, the most recent DHS was in 2015, with the scheduled 2020 survey cancelled because of the COVID pandemic. This thesis objective was implemented different estimation methods to determine the risk and incidence of HIV using spatial, statistical and machine learning methods. These methods were applied to address crucial epidemiological questions such as 1) determining the risk factors of HIV in Zimbabwe over a decade, 2) quantifying and mapping the HIV hot spots in Zimbabwe, and 3) ascertaining statistical models which can be used to determine the HIV incidence using low-cost data. The dataset used in these analyses was from the 2005-2015 Demographic Health Survey. In each of the previous chapters, a comprehensive discussion of each of the findings was given. This chapter of the thesis focuses on the main findings and contributions of the research work. These main findings can be grouped into three thematic areas: HIV risk estimation: Determination of HIV positivity; HIV incidence modelling; and Spatial regression and HIV hotspot analysis (shown in Figure 1.4).

8.2 HIV RISK ESTIMATION: DETERMINATION OF HIV POSITIVITY

Theme 1: Key Summary Points

1. Interval Censoring method is better than the standard survival analysis in estimating HIV hazard risk
2. Machine learning algorithms have the potential to enhance the implementation of pre-exposure prophylaxis (PrEP)
3. whether beating was justified or not, if there was Intimate Partner Violence, there was a higher risk of HIV infection.
4. Being Female, single, living in urban areas, no education/primary level, had an STI, and being unemployed were risk factors for HIV infection.

In this thesis, through the manuscripts in Chapters 3 and 4, we identified risk factors for HIV infection. Similar determinants have been found in other studies, however, none of these studies have used interval-censoring and machine learning methods to determine the risk while using cross-sectional data. Chapter 3 highlights how interval-censoring is better at obtaining the hazard risk of HIV infection than the conventional survival analysis method. In this chapter, we observed that the mean survival age before contracting HIV was around the early twenties for females and late twenties for males in 2005. However, the mean survival age for both sexes increased over time (2010 and 2015). The two most important variables for determining the acquisition of HIV were marital status and sex. Because of the ability to automatically learn from data with complex structures, machine learning methods have been found to be of great use in the epidemiological world.

Utilising these machine learning algorithms in this thesis has allowed us to determine the prediction capability of machine learning models effectively. In Chapter 4, the XGBoost algorithm was able to identify the most common features associated with HIV positivity in both sexes. We also assessed the direction of association between HIV status and the predictor variables using SHAP plots. The XGBoost algorithm managed to predict one of the most common risk factors of HIV infection, which has been reported extensively in literature as having many sexual partners²⁷²⁻²⁷⁵. An interesting factor associated with HIV positivity was that of Intimate Partner Violence. Our results indicated that IPV was a risk factor for HIV positivity. We noted that by better identifying individuals who are at high risk of HIV infection, machine learning algorithms have the potential to enhance the implementation of pre-exposure prophylaxis (PrEP).

Factors like being female, single, living in urban areas, having no education or primary education, having had an STI and being unemployed were some of the risk factors observed in Chapter 3. These risk factors can be corroborated by other studies in Sub-Saharan Africa²⁷⁶⁻²⁸⁰.

8.3 HIV INCIDENCE MODELLING

Theme 2: Key Summary Points

1. At the national level comparison, the estimates of the Catalytic and Farrington models seemed to be similar to the EPP/SPECTRUM estimates.
2. At the sub-national level comparison, the estimates of the Catalytic and Farrington models seemed to be similar to the Manicaland cohort estimates.
3. A peak in the cohort-specific prevalence and FOI confirms findings of age-mixing patterns, with the rise of HIV incidence occurring approximately 15 years earlier in females than males.
4. The triangulation of multiple methods for incidence estimation and interpretation of results have the potential to improve programmatic and policy decisions in the national HIV response.

In Chapters 5 and 6, this thesis explored how we could estimate HIV incidence using serial prevalence data. Chapter 5 focused on age-specific HIV incidence, and Chapter 6 focused on age-specific HIV incidence however, we included birth cohorts in the analysis as well as time (2005-2015). This thesis wanted to achieve statistical models which can be used to report HIV incidence while using low-cost data and provide estimates which can be triangulated with other estimates from models already used in Zimbabwe. In Chapter 5, we explored whether a model that considered differential mortality (Catalytic model) produced similar estimates to a model that did not consider differential mortality (Farrington model). These two models were then compared to estimates from UNAIDS and a cohort in Manicaland. This comparison was done on a micro and macro bases, which gave us an idea of whether the models under investigation could estimate HIV incidence on a smaller or larger scale. Results showed the estimates of the two models

were substantially similar regardless of their difference in the differential mortality assumption. Again, at the national and sub-national levels, the estimates of the two models were similar to the UNAIDS and cohort estimates. Because of the similarities in these estimates, we concluded that triangulation of multiple methods for incidence estimation and interpretation of results was required to improve programmatic and policy decisions in the national HIV response.

Chapter 6 estimated the age- and time- HIV prevalence and force-of-infection. The inclusion of the time component ensured that we did not assume the force-of-infection to be stationary over time. Results from Chapter 6 indicated the issue of age-mixing patterns, where the HIV incidence peaked approximately 15 years earlier for females and later for males. This potentially can become a problem as HIV transmission can happen across birth cohorts, which can lead to persistence of the HIV epidemic in Zimbabwe.

8.4 SPATIAL REGRESSION AND HIV HOT SPOT ANALYSIS

Theme 3: Key Summary Points

1. Over the ten years considered in this analysis, Tsholotsho, a district situated in the southern region of Zimbabwe in Matabeleland North remained a constant hotspot area.
2. This study's findings also indicate the need to zoom in on districts like Tsholotsho, which consistently have a high burden of HIV over time.
3. Proportions of females with partners five years older than them were positively associated with an increase in HIV incidence in 2010-11 and 2015 especially in the rural areas.
4. This study also highlighted that the proportion of richer individuals was positively associated with high HIV incidence rates for all three survey years in 2010-11

In Chapter 7, we explored ways in which we could investigate the different hot spots and cold spots of HIV incidence in Zimbabwe. This was achieved by doing a hot spot analysis which has the capability to find clusters which influence the response variable greatly. One of the interesting findings was how Tsholotsho district remained a hot spot area in all the years under investigation in this thesis. One of the results from Chapter 7 was the issue of age-mixing patterns or age-disparate relationships which was also confirmed by results obtained in Chapter 6. Chapter 7 analysis and results uncovered that most of the age-mixing patterns were frequent in rural districts than urban districts. The outcry on age-mixing patterns is their influence in the fuelling the risk of HIV infection. This is mainly as mentioned by other studies ^{250, 271}, because of lack of education, employment and financial stability, females tend to seek older partners for financial security.

Therefore, one of the conclusions that can be made to reduce the incidence of age-mixing patterns that fuel the risk of HIV infection is to advocate for programmes that educate the girl child on financial emancipation as well as risks of HIV infection especially in rural areas.

The study also highlighted how richer individuals based on the DHS wealth index, were at an increased risk of HIV infection. It has been noted by Fox ²⁶⁷ that countries which are wealthier will have the poorer individuals at risk of HIV infection, however for the poorer countries, the flip side is true, as individuals who are wealthier are more at risk of HIV infection. This theory holds true for Zimbabwe, since the economic situation in Zimbabwe does not put it as one of the wealthier countries, hence you will notice, as supported by our findings that richer/wealthier individuals are at an increased risk of HIV infection.

8.5 STRENGTHS AND LIMITATIONS OF THE STUDY

The thesis had both strengths and limitations, which will be discussed in the following section under these four subheadings:

1. Variable recall and selection bias
2. DHS data, representativeness, and generalisability
3. Missing data handling
4. Analysis and Computing

8.5.1 Variable Recall and Selection Bias

Having an accurate assessment of HIV risk behaviours is essential to be able to describe and predict trends, identify population at risk, evaluate the effectiveness of interventions

and advocate for support to develop policies and programs. However, the assessment of the HIV risk behaviours is complex owing to the stigmatisation associated with the disease and sometimes the illegal nature of some of the sexual behaviours. This complexity, in turn, will influence the individual's self-reports as well as recall bias. Variables like time since last sex, age of last sexual partner, age at first sex, which were used in some of the analyses in Chapters 3, 4 and 7 and many more, are subject to recall bias. In general, virtually all information collected in the DHS surveys is subject to reporting and recall biases. However, because some of our conclusions (Chapters 3, 4 and 7) were similar to what was already in literature, we considered that this bias did not have that much of an effect on our results.

Regarding selection bias, it has been reported that the HIV biomarker response rate for the Demographic Health Survey has been high as 30%. Therefore, when participants and non-participants differ with respect to HIV status, naïve use of the prevalence data from the survey yields biased results. In most cases, individuals who decline to be tested are those who might be aware of their HIV status and would want to avoid unnecessary re-testing. Also, those who do not know their status but suspect they are positive might also decline being tested, fearing disclosure or stigmatisation. This participation selection might have caused overestimation or underestimation of the estimates (Chapters 5 and 6) reported in this study, however, again, our results were not significantly different from the UNAIDS estimates and the cohort estimates reported in Chapter 5.

8.5.2 DHS Data, Representativeness, and Generalisability

One of the main strengths of this study was the ability of the dataset set used to be representative of the general population. The way the DHS is carried out ensures that regions, districts and even wards can be directly comparable without any issues, given the standard sampling strategies. As a result, the DHS surveys provide a global, comparative, high-quality database on a wide range of health indicators. Furthermore, the period (2005-2015) covered in this study was very long, allowing us to evaluate changes over time exhibited in the results obtained in Chapters 3, 5, 6 and 7. In addition, using individual data enabled exploration and adjustment for individual-level covariates associated with HIV, which is impossible with aggregated data.

One of the most significant limitations of this study is the failure to model the different impacts associated with the introduction of ART and the changes in the guidelines on ART over time. In addition, DHS surveys have yet to incorporate the use of ART in their questionnaire, hence in all the analyses, the effect of ART was not included.

8.5.3 Missing Data Handling

Missing data was a major challenge in our analysis, and how we dealt with missing data in our analysis differed depending on the objective. However, in Chapter 4, we used the MICE imputation to impute any missing values. We assumed that the missingness was Missing at Random. Before the imputation procedure, variables with more than 30% missingness were excluded from the analysis. The way we handled missingness in Chapter 4 might have influenced the results we obtained. For the other Chapters, we only excluded those individuals with no HIV test results, which might have increased selection

bias, as discussed above. However, since we established that our results were similar to what was already in literature, we assumed that this bias had no significant repercussions on our results and conclusions.

8.5.4 Analysis and Computing

One of the strengths of this thesis was the ability to explore different statistical techniques which have not been used in HIV modelling. For example, in Chapter 3, interval censoring has not been used to the best of our knowledge to compute the hazard risk of HIV infection while using survey data. This thesis explored this modelling strategy, producing results significantly similar to what was reported in literature. In Chapter 4, the extensive coding required (from data pre-processing to evaluation) to estimate HIV positivity using machine learning algorithms was time intensive. Chapter 5 explored using the Farrington model to estimate HIV incidence. The Farrington model in literature has been used to model the force-of-infection of other infectious diseases other than HIV, and the results turned out to be similar to the Catalytic model, which has been used to estimate HIV incidence. This is an advantage since not many models are available in literature that estimates HIV incidence using survey data. Chapter 6 estimated the cohort-specific force-of-infection which made use of generalised additive models (GAM). Lastly, Chapter 7, explored the local/district spatial patterns of HIV incidence in Zimbabwe over the 10-year period between 2005-2015 using Multiscale Geographically Weighted Regression (MGWR) to account for nonstationarity between response (HIV incidence) and predictor variables.

The statistical analysis techniques used in this thesis were very rigorous and helped explore challenging statistical modelling fields beyond my comfort zone. The software

used in the thesis analysis included R software, STATA, Python, ArcGIS and WinBugs. In general, the design of our study, that is, the cross-sectional nature of our dataset, limited us to discuss associations without extending these to causality. In addition, we could only use variables provided in DHS and faced the problem of not being able to include important variables like ART initiation and usage to support our explanations.

8.6 CONCLUSION

Throughout the thesis, the most common conclusion in all the manuscripts was the issue of age-mixing patterns and the risk of HIV infection being high among females. In Chapters 5 and 6, we observed that HIV incidence peaks approximately 15 years earlier in females than males, suggesting a possibility of age-mixing patterns. This strengthens the need to explore these groups in-depth to establish causal effects that might better inform policy.

In Chapter 2, the use of machine learning methods will help with the detection of socio-behavioural HIV detection as well as improving screening procedures in limited resource settings using low-cost data. Also, if programmes targeted at HIV testing could incorporate machine learning approaches, high-risk individuals would be identified adequately and promote the use of PrEP more effectively. However, further research is required to integrate and implement the use of machine learning methods in a real-world primary-care context.

8.7 IMPLICATIONS FOR INTERVENTIONS AND POLICY

When reporting the progress of the HIV epidemic in Zimbabwe, only the estimates obtained from the UNAIDS EPP/SPECTRUM are used in the national report. However, to improve programmatic and policy decisions in the national HIV response, we recommend the triangulation of multiple methods for identifying groups and clusters at high risk as well as reporting the HIV prevalence and incidence. Reporting results from multiple estimation approaches in the national report should be considered to reduce uncertainty in the general interpretation of results, resulting in giving policy conclusions and recommendations that are generated from an extensive triangulation of different methods obtained using different datasets and assumptions.

Policy recommendations for each manuscript include: 1) recommending interventions targeted at females with less than primary level of education living in urban areas who are unemployed and have a history of STI (Chapter 3); 2) integrating machine learning approaches in primary health care will help in identifying at high-risk groups and subsequently the effective implementation of PrEP (Chapter 4); 3) 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 (Chapter 5); 4) introduce effective interventions that would include programmes that increase financial and female independence risk perception and engage men to change sex norms that favour dialogue between partners (Chapter 6); and the findings of this study also point to the need to focus on districts like Tsholotsho, which persistently had a high HIV burden throughout time. Interventions that promote partner notification when one has an STI especially with multiple sex partners, are encouraged to help reduce

asymptomatic STIs in the districts in turn, this is expected to reduce HIV incidence rates. This study also suggests that policymakers budget for more local/district-level analysis than the province to inform policies and strategies for resource distribution in high-need areas (Chapter 7).

In conclusion, this study was motivated by the need to attain zero HIV incidence by 2030 according to the UNAIDS 90-90-90 fast track targets. Therefore, the aim was to understand how HIV risk factors influence the epidemic and track how the epidemic has been improving over time by estimating the HIV incidence in different age groups. Despite all the efforts that have been made to achieve the UNAIDS targets at the country level, a comprehensive HIV response is still a requirement in the fight against HIV to attain the global target of zero HIV incidence by 2030.

8.8 SCOPE FOR FUTURE RESEARCH

The main limitation of this study was the minimum variables available in the DHS data to explain the risks and changes in HIV incidence. For example, this study did not incorporate a variety of factors, including ART initiation dates for those infected, policy changes on ART dose administration, ART adherence patterns, CD4 cell counts and viral load of those infected, HIV interventions implemented in different parts of the country, and so many other variables which could clearly explain the HIV epidemic over time. Therefore, going forward, further studies incorporating many datasets from different sources to estimate the HIV incidence will strengthen the evidence on the risks and trends found in this study, as well as subsidising data not available in DHS to explain the epidemic.

Modelling approaches did not fully account for mortality associated with HIV, therefore, further studies should include data with comprehensive mortality details specifically associated with HIV to improve the estimation of HIV prevalence and incidence.

REFERENCES

1. UNAIDS. Global HIV statistics - Fact Sheet, https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf (2022).
2. Blick G and Wraight JM. UNAIDS “90-90-90”: Fighting HIV Stigma and Discrimination Needed to End HIV as a Global Threat by 2030. *Journal of Infectious Diseases and Diagnosis* 2017; 2.
3. Sharp PM, Shaw GM and Hahn BH. Simian immunodeficiency virus infection of chimpanzees. *Journal of virology* 2005; 79: 3891-3902.
4. Sharp PM, Bailes E, Chaudhuri RR, et al. The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 2001; 356: 867-876.
5. Keele BF, Van Heuverswyn F, Li Y, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 2006; 313: 523-526.
6. Horsburgh JR CR, Jason J, Longini JR I, et al. Duration of human immunodeficiency virus infection before detection of antibody. *The Lancet* 1989; 334: 637-640.
7. Tindall B and Cooper D. Primary HIV infection: host responses and intervention strategies. *AIDS* 1991; 5: 1-14.
8. Cohen MS, Gay CL, Busch MP, et al. The detection of acute HIV infection. *The Journal of infectious diseases* 2010; 202: S270-S277.
9. Taylor D, Durigon M, Davis H, et al. Probability of a false-negative HIV antibody test result during the window period: a tool for pre-and post-test counselling. *International journal of STD & AIDS* 2015; 26: 215-224.

10. De Wolf F, Spijkerman I, Schellekens PT, et al. AIDS prognosis based on HIV-1 RNA, CD4+ T-cell count and function: markers with reciprocal predictive value over time after seroconversion. *AIDS* 1997; 11: 1799-1806.
11. Cheung TW. AIDS-Related Cancer in the Era of Highly Active Antiretroviral Therapy (HAART): A Model of the Interplay of the Immune System, Virus, and Cancer.—“On the Offensive—The Trojan Horse is Being Destroyed”—Part A: Kaposi's Sarcoma. *Cancer investigation* 2004; 22: 774-786.
12. Bacchetti P and Moss AR. Incubation period of AIDS in San Francisco. *Nature* 1989; 338: 251-253.
13. Worobey M, Gemmel M, Teuwen DE, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 2008; 455: 661-664.
14. Barre-Sinoussi F, Chermann J, Rey F, et al. Isolation of T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Revista de Investigación Clínica* 2004; 56: 126-129.
15. Levy JA, Hoffman AD, Kramer SM, et al. Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. *Science* 1984; 225: 840-842.
16. UNAIDS. Forty years of the AIDS response, <https://www.unaids.org/en/resources/infographics/40-years-of-the-AIDS-response> (2021, accessed 09-03-2023 2023).
17. Duri K, Stray-Pedersen B and Muller F. HIV/AIDS: The Zimbabwean situation and trends. *American Journal of Clinical Medicine Research* 2013; 15-22: 15-22.
18. Dehne K, Dhlakama D, Richter C, et al. Herpes zoster as an indicator of HIV infection in Africa. *Tropical Doctor* 1992; 22: 68-70.
19. Topley J. HIV infection in Zimbabwe. *Archives of disease in childhood* 1988; 63: 842-844.

20. Denhe K, Dhlakama D, Richter C, et al. Pattern of HIV-infection in Hurungwe district, Mashonaland West, Zimbabwe. *Cent Afr J Med* 1992; 38: 139-143.
21. Mertens T, Tondorf G, Siebolds M, et al. Epidemiology of HIV and hepatitis B virus (HBV) in selected African and Asian populations. *Infection* 1989; 17: 4-7.
22. McFarland W, Mvere D, Shamu R, et al. Risk factors for HIV seropositivity among first-time blood donors in Zimbabwe. *Transfusion* 1998; 38: 279-284.
23. McFarland W, Mvere D, Shandera W, et al. Epidemiology and prevention of transfusion-associated human immunodeficiency virus transmission in sub-Saharan Africa. *Vox sanguinis* 1997; 72: 85-92.
24. Tswana S, Nystrom L, Moyo S, et al. A sero-epidemiology cross-sectional nationwide study of the prevalence of human immunodeficiency virus in Zimbabwe 1989-1991. *African Journal of Health Sciences* 1996; 3: 96-100.
25. Fischl MA, Richman DD, Grieco MH, et al. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine* 1987; 317: 185-191.
26. Detels R, Munoz A, McFarlane G, et al. Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. *Jama* 1998; 280: 1497-1503.
27. Moore RD and Chaisson RE. Natural history of HIV infection in the era of combination antiretroviral therapy. *AIDS* 1999; 13: 1933-1942.
28. Tellalian D, Maznavi K, Bredeek UF, et al. Pre-exposure prophylaxis (PrEP) for HIV infection: results of a survey of HIV healthcare providers evaluating their knowledge, attitudes, and prescribing practices. *AIDS patient care and STDs* 2013; 27: 553-559.

29. Grant RM, Lama JR, Anderson PL, et al. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *New England Journal of Medicine* 2010; 363: 2587-2599.
30. Baeten JM, Donnell D, Ndase P, et al. Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *New England Journal of Medicine* 2012; 367: 399-410.
31. McCormack S, Dunn DT, Desai M, et al. Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. *The Lancet* 2016; 387: 53-60.
32. Brookmeyer R and Gail MH. *AIDS epidemiology: a quantitative approach*. Oxford University Press on Demand, 1994.
33. Foulkes MA. Advances in HIV/AIDS statistical methodology over the past decade. *Statistics in Medicine* 1998; 17: 1-25.
34. Becker NG and Marschner IC. A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data. *Biometrika* 1993; 80: 165-178.
35. Brookmeyer R and Gail M. Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. *The Lancet* 1986; 328: 1320-1322.
36. Brookmeyer R and Gail MH. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* 1988; 83: 301-308.
37. Brookmeyer R and Damiano A. Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine* 1989; 8: 23-34.
38. Aalen O, Farewell V, De Angelis D, et al. The use of human immunodeficiency virus diagnosis information in monitoring the acquired immune deficiency syndrome

- epidemic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 1994; 157: 3-16.
39. Marschner IC. Using time of first positive HIV test and other auxiliary data in back-projection of AIDS incidence. *Statistics in medicine* 1994; 13: 1959-1974.
 40. Farewell V, Aalen O, Angelis DD, et al. Estimation of the rate of diagnosis of HIV infection in HIV infected individuals. *Biometrika* 1994; 81: 287-294.
 41. Dietz K, Seydel J and Schwartländer B. Back-projection of German AIDS data using information on dates of tests. *Statistics in medicine* 1994; 13: 1991-2008.
 42. Raab GM, Fielding KL and Allardice G. Incorporating HIV test data into forecasts of the AIDS epidemic in Scotland. *Statistics in medicine* 1994; 13: 2009-2020.
 43. De Angelis D, Gilks W and Day N. Bayesian projection of the acquired immune deficiency syndrome epidemic. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1998; 47: 449-498.
 44. Bellocco R and Marschner IC. Joint analysis of HIV and AIDS surveillance data in back-calculation. *Statistics in medicine* 2000; 19: 297-311.
 45. Cui J and Becker NG. Estimating HIV incidence using dates of both HIV and AIDS diagnoses. *Statistics in medicine* 2000; 19: 1165-1177.
 46. Becker NG, Lewis JJ, Li Z, et al. Age-specific back-projection of HIV diagnosis data. *Statistics in medicine* 2003; 22: 2177-2190.
 47. Sommen C, Alioum A and Commenges D. A multistate approach for estimating the incidence of human immunodeficiency virus by using HIV and AIDS French surveillance data. *Statistics in medicine* 2009; 28: 1554-1568.
 48. An Q, Kang J, Song R, et al. A Bayesian hierarchical model with novel prior specifications for estimating HIV testing rates. *Statistics in medicine* 2016; 35: 1471-1487.

49. Aalen OO, Farewell VT, De Angelis D, et al. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in medicine* 1997; 16: 2191-2210.
50. Posner SJ, Myers L, Hassig SE, et al. Estimating HIV incidence and detection rates from surveillance data. *Epidemiology* 2004: 164-172.
51. Birrell PJ, Chadborn TR, Gill ON, et al. Estimating trends in incidence, time-to-diagnosis and undiagnosed prevalence using a CD4-based Bayesian back-calculation. *Statistical Communications in Infectious Diseases* 2012; 4.
52. Hall HI, Song R, Szwarcwald CL, et al. Time from infection with the human immunodeficiency virus to diagnosis, United States. *Journal of acquired immune deficiency syndromes (1999)* 2015; 69: 248.
53. Szwarcwald C, Pascom A and Souza Júnior P. Estimation of the HIV incidence and of the number of people living with HIV/AIDS in Brazil, 2012. *J Aids Clin Res* 2015; 6: 430.
54. Song R, Hall HI, Green TA, et al. Using CD4 data to estimate HIV incidence, prevalence, and percent of undiagnosed infections in the United States. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2017; 74: 3-9.
55. Xia Q, Teixeira-Pinto A, Forgione LA, et al. Estimated HIV incidence in the United States, 2003–2010. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2017; 74: 10-14.
56. Karon JM, Fleming PL, Steketee RW, et al. HIV in the United States at the turn of the century: an epidemic in transition. *American journal of public health* 2001; 91: 1060.
57. Brookmeyer R. Measuring the HIV/AIDS epidemic: approaches and challenges. *Epidemiologic reviews* 2010; 32: 26-37.

58. Ghys PD, Brown T, Grassly N, et al. The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics. *Sexually transmitted infections* 2004; 80: i5-i9.
59. Stover J. Projecting the demographic consequences of adult HIV prevalence trends: the Spectrum Projection Package. *Sexually transmitted infections* 2004; 80: i14-i18.
60. Stover J, Walker N, Grassly N, et al. Projecting the demographic impact of AIDS and the number of people in need of treatment: updates to the Spectrum projection package. *Sexually transmitted infections* 2006; 82: iii45-iii50.
61. Stover J, Johnson P, Zaba B, et al. The Spectrum projection package: improvements in estimating mortality, ART needs, PMTCT impact and uncertainty bounds. *Sexually Transmitted Infections* 2008; 84: i24-i30.
62. Stover J, Johnson P, Hallett T, et al. The Spectrum projection package: improvements in estimating incidence by age and sex, mother-to-child transmission, HIV progression in children and double orphans. *Sexually transmitted infections* 2010; 86: ii16-ii21.
63. Stover J, Andreev K, Slaymaker E, et al. Updates to the spectrum model to estimate key HIV indicators for adults and children. *AIDS (London, England)* 2014; 28: S427.
64. Brookmeyer R and Quinn TC. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *American journal of epidemiology* 1995; 141: 166-172.
65. Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *Jama* 1998; 280: 42-48.

66. Karon JM, Song R, Brookmeyer R, et al. Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results. *Statistics in medicine* 2008; 27: 4617-4633.
67. Sommen C, Commenges D, Vu SL, et al. Estimation of the distribution of infection times using longitudinal serological markers of HIV: implications for the estimation of HIV incidence. *Biometrics* 2011; 67: 467-475.
68. Sweeting MJ, De Angelis D, Parry J, et al. Estimating the distribution of the window period for recent HIV infections: a comparison of statistical methods. *Statistics in medicine* 2010; 29: 3194-3202.
69. Kassanjee R, De Angelis D, Farah M, et al. Cross-sectional HIV incidence surveillance: A benchmarking of approaches for estimating the ‘mean duration of recent infection’. *Statistical communications in infectious diseases* 2017; 9.
70. Koulai L, Presanis A, Murphy G, et al. Quantifying the recency of HIV infection using multiple longitudinal biomarkers. *arXiv preprint arXiv:170602508* 2017.
71. Bernoulli D and Blower S. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in medical virology* 2004; 14: 275.
72. Muench H. Catalytic models in epidemiology. *Catalytic Models in Epidemiology*. Harvard University Press, 2013.
73. Jewell NP and Laan Mvd. Generalizations of current status data with applications. *Lifetime data: models in reliability and survival analysis* 1996: 141-148.
74. Griffiths D. A catalytic model of infection for measles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1974; 23: 330-339.
75. Grenfell BT and Anderson R. The estimation of age-related rates of infection from case notifications and serological data. *Epidemiology & Infection* 1985; 95: 419-436.

76. Vandormael A, Dobra A, Bärnighausen T, et al. Incidence rate estimation, periodic testing and the limitations of the mid-point imputation approach. *International Journal of Epidemiology* 2018; 47: 236-245.
77. Hsu CH, Taylor JM, Murray S, et al. Multiple imputation for interval censored data with auxiliary variables. *Statistics in medicine* 2007; 26: 769-781.
78. Chen L and Sun J. A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. *Computational statistics & data analysis* 2010; 54: 1109-1116.
79. Birri Makota R and Musenge E. Factors associated with HIV infection in Zimbabwe over a decade from 2005 to 2015: an interval-censoring survival analysis approach. *Frontiers in public health* 2019; 7: 262.
80. Gonese E, Musuka G, Ruangtragool L, et al. Comparison of HIV Incidence in the Zimbabwe Population-Based HIV Impact Assessment Survey (2015–2016) with Modeled Estimates: Progress Toward Epidemic Control. *AIDS research and human retroviruses* 2020; 36: 656-662.
81. Ali MS, Wit MD, Chabata ST, et al. Estimation of HIV incidence from analysis of HIV prevalence patterns by age and years since starting sex work among female sex workers in Zimbabwe. *AIDS* 2022; 36: 1141-1150. DOI: 10.1097/QAD.0000000000003198.
82. Hargrove JW, Humphrey JH, Mahomva A, et al. Declining HIV prevalence and incidence in perinatal women in Harare, Zimbabwe. *Epidemics* 2011; 3: 88-94.
83. Lopman B and Gregson S. When did HIV incidence peak in Harare, Zimbabwe? Back-calculation from mortality statistics. *PLoS One* 2008; 3: e1711.

84. Munjoma MW, Mhlanga FG, Mapingure MP, et al. The incidence of HIV among women recruited during late pregnancy and followed up for six years after childbirth in Zimbabwe. *BMC Public Health* 2010; 10: 1-8.
85. Contag CH, Zejena L, Mbizvo M, et al. Primary Subtype C HIV-1 Infection in Harare, Zimbabwe. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 1999; 20.
86. Humphrey JH, Marinda E, Mutasa K, et al. Mother to child transmission of HIV among Zimbabwean women who seroconverted postnatally: prospective cohort study. *BMJ* 2010; 341.
87. Hallett TB, Gregson S, Mugurungi O, et al. Assessing evidence for behaviour change affecting the course of HIV epidemics: a new mathematical modelling approach and application to data from Zimbabwe. *Epidemics* 2009; 1: 108-117.
88. Lopman B, Lewis J, Nyamukapa C, et al. HIV incidence and poverty in Manicaland, Zimbabwe: is HIV becoming a disease of the poor? *AIDS (London, England)* 2007; 21: S57.
89. Schaefer R, Gregson S, Eaton JW, et al. Age-disparate relationships and HIV incidence in adolescent girls and young women: evidence from Zimbabwe. *AIDS (London, England)* 2017; 31: 1461.
90. Humphrey JH, Hargrove JW, Malaba LC, et al. HIV incidence among post-partum women in Zimbabwe: risk factors and the effect of vitamin A supplementation. *AIDS* 2006; 20: 1437-1446.
91. Corbett EL, Makamure B, Cheung YB, et al. HIV incidence during a cluster-randomized trial of two strategies providing voluntary counselling and testing at the workplace, Zimbabwe. *AIDS* 2007; 21: 483-489.

92. Sherr L, Lopman B, Kakowa M, et al. Voluntary counselling and testing: uptake, impact on sexual behaviour, and HIV incidence in a rural Zimbabwean cohort. *AIDS* 2007; 21: 851-860.
93. Kumwenda N, Hoffman I, Chirenje M, et al. HIV incidence among women of reproductive age in Malawi and Zimbabwe. *Sexually transmitted diseases* 2006: 646-651.
94. Mundandi C, Vissers D, Voeten H, et al. No difference in HIV incidence and sexual behaviour between out-migrants and residents in rural Manicaland, Zimbabwe. *Tropical Medicine & International Health* 2006; 11: 705-711.
95. Lopman B, Nyamukapa C, Mushati P, et al. HIV incidence in 3 years of follow-up of a Zimbabwe cohort—1998–2000 to 2001–03: contributions of proximate and underlying determinants to transmission. *International Journal of Epidemiology* 2008; 37: 88-105.
96. Lopman BA, Garnett GP, Mason PR, et al. Individual level injection history: a lack of association with HIV incidence in rural Zimbabwe. *PLoS Medicine* 2005; 2: e37.
97. Hensen B, Hargreaves JR, Chiyaka T, et al. Evaluating the impact of DREAMS on HIV incidence among young women who sell sex: protocol for a non-randomised study in Zimbabwe. *BMC Public Health* 2018; 18: 1-10.
98. Hargreaves JR, Mtetwa S, Davey C, et al. Implementation and operational research: cohort analysis of program data to estimate HIV incidence and uptake of HIV-related services among female sex workers in Zimbabwe, 2009–2014. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2016; 72: e1-e8.
99. Hargreaves JR, Mtetwa S, Davey C, et al. Cohort analysis of programme data to estimate HIV incidence and uptake of HIV-related services among female sex workers in Zimbabwe, 2009-14. *Journal of acquired immune deficiency syndromes (1999)* 2015.

100. Gregson S, Machekano R, Donnelly CA, et al. Estimating HIV incidence from age-specific prevalence data: comparison with concurrent cohort estimates in a study of male factory workers, Harare, Zimbabwe. *AIDS* 1998; 12: 2049-2058.
101. Minnis AM, van der Straten A, Gerdt C, et al. A comparison of four condom-use measures in predicting pregnancy, cervical STI and HIV incidence among Zimbabwean women. *Sexually transmitted infections* 2010; 86: 231-235.
102. Gregson S, Mugurungi O, Eaton J, et al. Documenting and explaining the HIV decline in east Zimbabwe: the Manicaland General Population Cohort. *BMJ open* 2017; 7: e015898.
103. Wikipedia. Map of Zimbabwe, https://en.wikipedia.org/wiki/Provinces_of_Zimbabwe#/media/File:Administrative_Divisions_of_Zimbabwe.svg (2008).
104. Zimbabwe Central Statistical Office and Macro International. *Zimbabwe Demographic and Health Survey 2005-06*. 2007. Calverton, Maryland, USA: Central Statistical Office/Zimbabwe and Macro International.
105. Zimbabwe National Statistics Agency and ICF International. *Zimbabwe Demographic and Health Survey 2015: Final Report*. 2016. Rockville, Maryland, USA: Zimbabwe National Statistics Agency (ZIMSTAT) and ICF International.
106. Zimbabwe National Statistics Agency - ZIMSTAT and ICF International. *Zimbabwe Demographic and Health Survey 2010-11*. 2012. Calverton, Maryland, USA: ZIMSTAT and ICF International.
107. The DHS Program. Demographic and Health Surveys [Online Database], https://dhsprogram.com/data/dataset_admin/login_main.cfm (2017).

108. Swain PK and Grover G. Determination of predictors associated with HIV/AIDS patients on ART using accelerated failure time model for interval censored survival data. *Am J Biostat* 2016; 6: 12-19.
109. Anderson-Bergman C. icenReg: regression models for interval censored data in R. *Journal of Statistical Software* 2017; 81: 1-23.
110. Leung K-M, Elashoff RM and Afifi AA. Censoring issues in survival analysis. *Annual review of public health* 1997; 18: 83-104.
111. Malley B, Ramazzotti D and Wu JT-y. Data pre-processing. *Secondary analysis of electronic health records* 2016: 115-141.
112. Probst P, Bischl B and Boulesteix A-L. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research* 2019; 20: 1-32.
113. Marcus JL, Hurley LB, Krakower DS, et al. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *The lancet HIV* 2019; 6: e688-e695.
114. Nohara Y, Matsumoto K, Soejima H, et al. Explanation of machine learning models using improved Shapley Additive Explanation. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 2019, pp.546-546.
115. Li J, Cheng K, Wang S, et al. Feature selection: A data perspective. *ACM computing surveys (CSUR)* 2017; 50: 1-45.
116. Miao J and Niu L. A survey on feature selection. *Procedia Computer Science* 2016; 91: 919-926.
117. Fonti V and Belitser E. Feature selection using lasso. *VU Amsterdam research paper in business analytics* 2017; 30: 1-25.

118. Muthukrishnan R and Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. In: *2016 IEEE international conference on advances in computer applications (ICACA)* 2016, pp.18-20. IEEE.
119. Setiawan D, Kusuma WA and Wigena AH. Sequential forward floating selection with two selection criteria. In: *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* 2017, pp.395-400. IEEE.
120. Hens N, Aerts M, Faes C, et al. Seventy-five years of estimating the force of infection from current status data. *Epidemiology & Infection* 2010; 138: 802-812.
121. Mossong J, Grapsa E, Tanser F, et al. Modelling HIV incidence and survival from age-specific seroprevalence after antiretroviral treatment scale-up in rural South Africa. *AIDS (London, England)* 2013; 27: 2471.
122. Hallett TB, Zaba B, Todd J, et al. Estimating incidence from prevalence in generalised HIV epidemics: methods and validation. *PLoS medicine* 2008; 5: e80.
123. West RM. Generalised additive models. *Modern methods for epidemiology*. Springer, 2012, pp.261-278.
124. Zuur AF, Ieno EN and Smith GM. Additive and generalised additive modelling. *Analysing Ecological Data* 2007: 97-124.
125. Ades A and Nokes D. Modeling age-and time-specific incidence from seroprevalence: toxoplasmosis. *American Journal of Epidemiology* 1993; 137: 1022-1034.
126. Nagelkerke N, Heisterkamp S, Borgdorff M, et al. Semi-parametric estimation of age–time specific infection incidence from serial prevalence data. *Statistics in Medicine* 1999; 18: 307-320.
127. Bain LE, Nkoke C and Noubiap JJN. UNAIDS 90-90-90 targets to end the AIDS epidemic by 2020 are not realistic: comment on "Can the UNAIDS 90-90-90 target be

- achieved? A systematic analysis of national HIV treatment cascades". *BMJ Glob Health* 2017; 2: e000227-e000227. DOI: 10.1136/bmjgh-2016-000227.
128. Zimbabwe Ministry of Health. *Zimbabwe Population-Based HIV Impact Assessment ZIMPHIA 2015–2016*. 2016. Harare
 129. Vandormael A, Dobra A, Bärnighausen T, et al. Incidence rate estimation, periodic testing and the limitations of the mid-point imputation approach. *International journal of epidemiology* 2017; 47: 236-245.
 130. Duri K, Stray-Pedersen B and Muller F. HIV/AIDS: The Zimbabwean situation and trends. *American Journal of Clinical Medicine Research* 2013; 1: 15-22.
 131. Jewell NP and van der Laan M. Generalizations of current status data with applications. *Lifetime Data Analysis* 1995; 1: 101-109. DOI: 10.1007/BF00985261.
 132. Dinse GE and Lagakos S. Regression analysis of tumour prevalence data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1983; 32: 236-248.
 133. Zhang Z and Sun J. Interval censoring. *Statistical methods in medical research* 2010; 19: 53-70. DOI: 10.1177/0962280209105023.
 134. The DHS Program. Demographic and Health Surveys [Online Database], <https://dhsprogram.com> (2005, accessed Accessed 26 May 2018).
 135. StataCorp. Stata Statistical Software: Release 15. *College Station, TX: StataCorp LLC* 2017.
 136. StataCorp. Stata 15 Base Reference Manual. *College Station, TX: Stata Press* 2017.
 137. Anderson-Bergman C. icenReg: Regression Models for Interval Censored Data in R. *Journal of Statistical Software* 2017; 81: 1-23. DOI: 10.18637/jss.v081.i12.
 138. Farrington CP. Residuals for proportional hazards models with interval-censored survival data. *Biometrics* 2000; 56: 473-482. 2000/07/06.

139. Lawless JF and Babineau D. Models for interval censoring and simulation-based inference for lifetime distributions. *Biometrika* 2006; 93: 671-686.
140. REN JJ. Goodness of fit tests with interval censored data. *Scandinavian Journal of Statistics* 2003; 30: 211-226.
141. Bogaerts K, Komarek A and Lesaffre E. *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS*. Chapman and Hall/CRC, 2017.
142. Dallabetta GA, Miotti PG, Chipangwi JD, et al. High socioeconomic status is a risk factor for human immunodeficiency virus type 1 (HIV-1) infection but not for sexually transmitted diseases in women in Malawi: implications for HIV-1 control. *J Infect Dis* 1993; 167: 36-42. 1993/01/01. DOI: 10.1093/infdis/167.1.36.
143. Kumwenda N, Hoffman I, Chirenje M, et al. HIV incidence among women of reproductive age in Malawi and Zimbabwe. *Sexually transmitted diseases* 2006; 33: 646-651.
144. Hargreaves JR, Bonell CP, Boler T, et al. Systematic review exploring time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa. *Aids* 2008; 22: 403-414.
145. Asiedu C, Asiedu E and Owusu F. The Socio-Economic Determinants of HIV/AIDS Infection Rates in Lesotho, Malawi, Swaziland and Zimbabwe. *Development Policy Review* 2012; 30: 305-326. DOI: 10.1111/j.1467-7679.2012.00578.x.
146. Gonese E, Mapako T, Dzangare J, et al. Within-Gender Changes in HIV Prevalence among Adults between 2005/6 and 2010/11 in Zimbabwe. *PloS one* 2015; 10: e0129611-e0129611. DOI: 10.1371/journal.pone.0129611.

147. Kembo J. Risk factors associated with HIV infection among young persons aged 15–24 years: Evidence from an in-depth analysis of the 2005–06 Zimbabwe Demographic and Health Survey. *SAHARA-J: Journal of Social Aspects of HIV/AIDS* 2012; 9: 54-63.
148. Mishra V, Hong R, Bignami-Van Assche S, et al. *The role of partner reduction and faithfulness in HIV prevention in Sub-Saharan Africa: Evidence from Cameroon, Rwanda, Uganda, and Zimbabwe*. 2009. Calverton, Maryland, USA: Macro International.
149. Gregson S, Mugurungi O, Eaton J, et al. Documenting and explaining the HIV decline in east Zimbabwe: the Manicaland General Population Cohort. *BMJ open* 2017; 7: e015898. 2017/10/11. DOI: 10.1136/bmjopen-2017-015898.
150. Balkus JE, Brown E, Palanee T, et al. An empiric HIV risk scoring tool to predict HIV-1 acquisition in African women. *Journal of acquired immune deficiency syndromes (1999)* 2016; 72: 333.
151. Oller R, Gómez G and Calle ML. Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics* 2004; 32: 315-326.
152. Population-based HIV Impact Project. *Zimbabwe Population-based HIV Impact Assessment*. 2020. ICAP at Columbia University.
153. Mutai CK, McSharry PE, Ngaruye I, et al. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. *BMC medical research methodology* 2021; 21: 1-11.
154. UNAIDS. Fast-track commitments to end AIDS by 2030. UNAIDS Geneva, Switzerland, 2016.

155. Jewell BL, Mudimu E, Stover J, et al. Potential effects of disruption to HIV programmes in sub-Saharan Africa caused by COVID-19: results from multiple mathematical models. *The lancet HIV* 2020; 7: e629-e640.
156. Dorward J, Khubone T, Gate K, et al. The impact of the COVID-19 lockdown on HIV care in 65 South African primary care clinics: an interrupted time series analysis. *The Lancet HIV* 2021; 8: e158-e165.
157. Zhu Q, JiKe C, Xu C, et al. A New Strategy to Quantitatively Identify Hot-Spot Areas in Growth of New HIV Infections for Targeted Interventions. *Frontiers in Public Health* 2021; 9: 680867.
158. Young SD and Zhang Q. Using search engine big data for predicting new HIV diagnoses. *PloS one* 2018; 13: e0199527.
159. Young SD, Rivers C and Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive medicine* 2014; 63: 112-115.
160. Xiao J, Ding R, Xu X, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine* 2019; 17: 1-13.
161. Roy A, Nikolitch K, McGinn R, et al. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine* 2020; 3: 1-12.
162. Whiting D and Fazel S. How accurate are suicide risk prediction models? Asking the right questions for clinical practice. *Evidence-based mental health* 2019; 22: 125-128.
163. Farran B, AlWotayan R, Alkandari H, et al. Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: a retrospective cohort study of health data from Kuwait. *Frontiers in endocrinology* 2019; 10: 624.

164. Park JH, Cho HE, Kim JH, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ digital medicine* 2020; 3: 1-7.
165. Kwiecinski J, Tzolos E, Meah MN, et al. Machine learning with 18F-sodium fluoride PET and quantitative plaque analysis on CT angiography for the future risk of myocardial infarction. *Journal of Nuclear Medicine* 2022; 63: 158-165.
166. Gruber S, Krakower D, Menchaca JT, et al. Using electronic health records to identify candidates for human immunodeficiency virus pre-exposure prophylaxis: An application of super learning to risk prediction when the outcome is rare. *Statistics in medicine* 2020; 39: 3059-3073.
167. Bao Y, Medland NA, Fairley CK, et al. Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. *Journal of Infection* 2021; 82: 48-59.
168. Xu X, Ge Z, Chow EP, et al. A Machine-Learning-Based Risk-Prediction Tool for HIV and Sexually Transmitted Infections Acquisition over the Next 12 Months. *Journal of clinical medicine* 2022; 11: 1818.
169. Chingombe I, Dzinamarira T, Cuadros D, et al. Predicting HIV status among men who have sex with men in Bulawayo & Harare, Zimbabwe using bio-behavioural data, recurrent neural networks, and machine learning techniques. *Tropical Medicine and Infectious Disease* 2022; 7: 231.
170. Chingombe I, Musuka G, Mbunge E, et al. Predicting HIV Status Using Machine Learning Techniques and Bio-Behavioural Data from the Zimbabwe Population-Based HIV Impact Assessment (ZIMPHIA15-16). In: *Computer Science On-line Conference* 2022, pp.247-258. Springer.

171. ZIMSTAT. 2022 *Population and Housing Census Preliminary Results*. 22 September 2022. UNFPA ZIMBABWE.
172. The DHS Program. Demographic and Health Surveys [Online Database], https://dhsprogram.com/data/dataset_admin/login_main.cfm (2005).
173. Saito T and Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 2015; 10: e0118432.
174. Simelane MS, Vermaak K, Zwane E, et al. Individual and community-level factors associated with lifetime number of sexual partners among women aged 15–49 in Eswatini. *Plos one* 2021; 16: e0246100.
175. Mishra V, Thaddeus S, Kafuko J, et al. *Fewer Lifetime Sexual Partners and Partner Faithfulness Reduce Risk of HIV Infection: Evidence from a National Sero-Survey in Uganda*. 2009.
176. Epstein M, Bailey JA, Manhart LE, et al. Understanding the link between early sexual initiation and later sexually transmitted infection: test and replication in two longitudinal studies. *Journal of Adolescent Health* 2014; 54: 435-441. e432.
177. Harling G, Newell M-L, Tanser F, et al. Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in rural KwaZulu-Natal, South Africa. *Journal of acquired immune deficiency syndromes (1999)* 2014; 66: 443.
178. Armstrong HL, Roth EA, Rich A, et al. Associations between sexual partner number and HIV risk behaviors: implications for HIV prevention efforts in a Treatment as Prevention (TasP) environment. *AIDS care* 2018; 30: 1290-1297.
179. Dunkle KL, Stephenson R, Karita E, et al. New heterosexually transmitted HIV infections in married or cohabiting couples in urban Zambia and Rwanda: an analysis of survey and clinical data. *The Lancet* 2008; 371: 2183-2191.

180. Hattori MK and Dodoo FN-A. Cohabitation, marriage, and 'sexual monogamy' in Nairobi's slums. *Social Science & Medicine* 2007; 64: 1067-1078.
181. Kaiser R, Bunnell R, Hightower A, et al. Factors associated with HIV infection in married or cohabitating couples in Kenya: results from a nationally representative study. *PLoS one* 2011; 6: e17842.
182. Maharaj P and Cleland J. Risk perception and condom use among married or cohabiting couples in KwaZulu-Natal, South Africa. *International family planning perspectives* 2005: 24-29.
183. Matovu JK. Preventing HIV transmission in married and cohabiting HIV-discordant couples in sub-Saharan Africa through combination prevention. *Current HIV research* 2010; 8: 430-440.
184. Kelso B. AIDS: Orphans of the Storm. *Africa Report* 1994; 39: 50.
185. Mabala R. From HIV prevention to HIV protection: addressing the vulnerability of girls and young women in urban areas. *Environment and Urbanization* 2006; 18: 407-432.
186. Van Breda AD. The phenomenon and concerns of child-headed households in Africa. *Sozialarbeit des südens* 2010; 3: 259-279.
187. Chigwenya A, Chuma M and Nyanga T. Trapped in the Vicious Circle: An Analysis of the Sustainability of the Child-Headed Households' Livelihoods in Ward 30, Gutu District. 2008.
188. Kazerooni PA, Mousavi M, Khazaei Z, et al. The prevalence of HIV among female head of household in Shiraz in 2015. *Immunopathologia Persa* 2018; 4: e29-e29.
189. Schatz E and Gilbert L. "My legs affect me a lot.... I can no longer walk to the forest to fetch firewood": Challenges related to health and the performance of daily tasks

for older women in a high HIV context. *Health Care for Women International* 2014; 35: 771-788.

190. Madise NJ, Ziraba AK, Inungu J, et al. Are slum dwellers at heightened risk of HIV infection than other urban residents? Evidence from population-based HIV prevalence surveys in Kenya. *Health & place* 2012; 18: 1144-1152.

191. Magadi MA. The disproportionate high risk of HIV infection among the urban poor in sub-Saharan Africa. *AIDS and Behavior* 2013; 17: 1645-1654.

192. Hajizadeh M, Sia D, Heymann SJ, et al. Socioeconomic inequalities in HIV/AIDS prevalence in sub-Saharan African countries: evidence from the Demographic Health Surveys. *International journal for equity in health* 2014; 13: 1-22.

193. Magure T, Wekwete N and Mpofu A. Why do urban black Zimbabweans have a lower HIV sero-prevalence than their rural counterparts? Results from a population-based survey. *Journal of Acquired Immune Deficiency Syndromes* 2015, p.e70-e77.

194. Henderson L, Zerai A and Morrow RL. Intimate partner violence and HIV status among ever-married and cohabiting Zimbabwean women: an examination of partners' traits. *African Journal of Reproductive Health* 2017; 21: 45-54.

195. Nyamayemombe C, Mishra V, Rusakaniko S, et al. The Association between Violence against Women and HIV: Evidence from a National Population-Based Survey in Zimbabwe. *Zimbabwe Working Papers, No4* 2010; Calverton, Maryland, USA: ICF Macro.

196. Orel E, Esra R, Estill J, et al. Prediction of HIV status based on socio-behavioural characteristics in East and Southern Africa. *PloS one* 2022; 17: e0264429.

197. Fellows IE, Shiraishi RW, Cherutich P, et al. A new method for estimating HIV incidence from a single cross-sectional survey. *Plos one* 2020; 15: e0237221.

198. Busza J, Dauya E, Bandason T, et al. The role of community health workers in improving HIV treatment outcomes in children: lessons learned from the ZENITH trial in Zimbabwe. *Health policy and planning* 2018; 33: 328-334.
199. Mahomva A, Greby S, Dube S, et al. HIV prevalence and trends from data in Zimbabwe, 1997–2004. *Sexually transmitted infections* 2006; 82: i42-i47.
200. Sun X, Nishiura H and Xiao Y. Modeling methods for estimating HIV incidence: a mathematical review. *Theoretical Biology and Medical Modelling* 2020; 17: 1-14.
201. Artzrouni M, Brown T, Feeney G, et al. Improved methods and assumptions for estimation of the HIV/AIDS epidemic and its impact: Recommendations of the UNAIDS Reference Group on Estimates, Modelling and Projections. *AIDs* 2002; 16.
202. Rajan SS and Sokal D. A simpler tool for estimation of HIV incidence from cross-sectional, age-specific prevalence data. *Journal of Epidemiology & Community Health* 2011; 65: 111-115.
203. Hallett TB. Estimating the HIV incidence rate—recent and future developments. *Current Opinion in HIV and AIDS* 2011; 6: 102.
204. Europe WGoEoHPi. HIV in hiding: methods and data requirements for the estimation of the number of people living with undiagnosed HIV. *Aids* 2011; 25: 1017-1023.
205. Marschner IC. A method for assessing age-time disease incidence using serial prevalence data. *Biometrics* 1997: 1384-1398.
206. Gregson S, Donnelly CA, Parker CG, et al. Demographic approaches to the estimation of incidence of HIV-1 infection among adults from age-specific prevalence data in stable endemic conditions. *AIDS (London, England)* 1996; 10: 1689-1697.

207. Ades AE and Nokes DJ. Modeling Age- and Time-Specific Incidence from Seroprevalence: Toxoplasmosis. *American Journal of Epidemiology* 1993; 137: 1022-1034. DOI: 10.1093/oxfordjournals.aje.a116758.
208. Hens N, Shkedy Z, Aerts M, et al. *Modeling infectious disease parameters based on serological and social contact data: a modern statistical perspective*. Springer Science & Business Media, 2012.
209. Hens N, Aerts M, Shkedy Z, et al. Modelling multisera data: the estimation of new joint and conditional epidemiological parameters. *Statistics in medicine* 2008; 27: 2651-2664.
210. Case KK, Johnson LF, Mahy M, et al. Summarizing the results and methods of the 2019 Joint United Nations Programme on HIV/AIDS HIV estimates. *AIDS (London, England)* 2019; 33: S197.
211. Mahy M, Brown T, Stover J, et al. Producing HIV estimates: from global advocacy to country planning and impact measurement. *Global health action* 2017; 10: 1291169.
212. Silhol R, Gregson S, Nyamukapa C, et al. Empirical validation of the UNAIDS Spectrum model for subnational HIV estimates: case-study of children and adults in Manicaland, Zimbabwe. *Aids* 2017; 31: S41-S50.
213. Brown T, Le Bao JWE, Hogan DR, et al. Improvements in prevalence trend fitting and incidence estimation in EPP 2013. *AIDS (London, England)* 2014; 28: S415.
214. Gregson S, Garnett GP, Nyamukapa CA, et al. HIV decline associated with behavior change in eastern Zimbabwe. *Science* 2006; 311: 664-666.
215. Hallett TB, Zaba B, Todd J, et al. Estimating incidence from prevalence in generalised HIV epidemics: methods and validation. *PLoS Med* 2008; 5: e80.

216. Padian NS, McLOY SI, Balkus JE, et al. Weighing the gold in the gold standard: challenges in HIV prevention research. *AIDS (London, England)* 2010; 24: 621.
217. Batter V, Matela B, Nsuami M, et al. High HIV-1 incidence in young women masked by stable overall seroprevalence among childbearing women in Kinshasa, Zaïre: estimating incidence from serial seroprevalence data. *AIDS (London, England)* 1994; 8: 811-817.
218. Bucyendore A, Van de Perre P, Karita E, et al. Estimating the seroincidence of HIV-1 in the general adult population in Kigali, Rwanda. *AIDS (London, England)* 1993; 7: 275-277.
219. Miotti PG, Dallabetta G, Ndovi E, et al. HIV-1 and pregnant women: associated factors, prevalence, estimate of incidence and role in fetal wastage in central Africa. *AIDS (London, England)* 1990; 4: 733-736.
220. Kim AA, Hallett T, Stover J, et al. Estimating HIV incidence among adults in Kenya and Uganda: a systematic comparison of multiple methods. *PloS one* 2011; 6: e17535.
221. Mapingure MP, Msuya S, Kurewa NE, et al. Sexual behaviour does not reflect HIV-1 prevalence differences: a comparison study of Zimbabwe and Tanzania. *Journal of the International AIDS Society* 2010; 13: 1-9.
222. Tanser F, Bärnighausen T, Grapsa E, et al. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science* 2013; 339: 966-971.
223. Bärnighausen T, Tanser F, Gqwede Z, et al. High HIV incidence in a community with high HIV prevalence in rural South Africa: findings from a prospective population-based study. *Aids* 2008; 22: 139-144.

224. Ott MQ, Bärnighausen T, Tanser F, et al. Age-gaps in sexual partnerships: seeing beyond ‘sugar daddies’. *AIDS (London, England)* 2011; 25: 861.
225. Chapman R, White R, Shafer L, et al. Do behavioural differences help to explain variations in HIV prevalence in adolescents in sub-Saharan Africa? *Tropical Medicine & International Health* 2010; 15: 554-566.
226. Akullian A, Vandormael A, Miller JC, et al. Large age shifts in HIV-1 incidence patterns in KwaZulu-Natal, South Africa. *Proceedings of the National Academy of Sciences* 2021; 118.
227. Zaidi J, Grapsa E, Tanser F, et al. Dramatic increases in HIV prevalence after scale-up of antiretroviral treatment: a longitudinal population-based HIV surveillance study in rural kwazulu-natal. *AIDS (London, England)* 2013; 27: 2301.
228. Negin J, Gregson S, Eaton JW, et al. Rising levels of HIV infection in older adults in Eastern Zimbabwe. *PLoS One* 2016; 11: e0162967.
229. Kimanga DO, Ogola S and Umuro M. Prevalence and incidence of HIV infection, trends, and risk factors among persons aged 15–64 years in Kenya: results from a nationally representative study. *Journal of acquired immune deficiency syndromes (1999)* 2014; 66: S13.
230. Vollmer S, Harttgen K, Alfven T, et al. The HIV epidemic in sub-Saharan Africa is aging: evidence from the demographic and health surveys in sub-Saharan Africa. *AIDS and Behavior* 2017; 21: 101-113.
231. Balakrishna Y. *Estimating the force of infection from prevalence data: infectious disease modelling*. 2013.
232. Hallett TB, Stover J, Mishra V, et al. Estimates of HIV incidence from household-based prevalence surveys. *AIDS (London, England)* 2010; 24: 147-152. DOI: 10.1097/QAD.0b013e32833062dc.

233. Boerma JT, Ghys PD and Walker N. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *The Lancet* 2003; 362: 1929-1931.
234. Garcia-Calleja J, Gouws E and Ghys P. National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates. *Sexually transmitted infections* 2006; 82: iii64-iii70.
235. Hallett TB, Aberle-Grasse J, Bello G, et al. Declines in HIV prevalence can be associated with changing sexual behaviour in Uganda, urban Kenya, Zimbabwe, and urban Haiti. *Sexually transmitted infections* 2006; 82: i1-i8.
236. Nagelkerke NJ, de Vlas SJ, Jha P, et al. Heterogeneity in host HIV susceptibility as a potential contributor to recent HIV prevalence declines in Africa. *AIDS (London, England)* 2009; 23: 125.
237. Batter V, Matela B, Nsuami M, et al. High HIV-1 incidence in young women masked by stable overall seroprevalence among childbearing women in Kinshasa, Zaire: estimating incidence from serial seroprevalence data. *AIDS (London, England)* 1994; 8: 811-817.
238. Nagelkerke N, Heisterkamp S, Borgdorff M, et al. Semi-parametric estimation of age-time specific infection incidence from serial prevalence data. *Statistics in Medicine* 1999; 18: 307-320. DOI: 10.1002/(sici)1097-0258(19990215)18:3<307::Aid-sim15>3.0.Co;2-z.
239. Ades A and Medley G. Estimates of disease incidence in women based on antenatal or neonatal seroprevalence data: HIV in New York City. *Statistics in medicine* 1994; 13: 1881-1894.
240. Keiding N. Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 1991; 154: 371-396.

241. Marschner IC. Fitting a multiplicative incidence model to age-and time-specific prevalence data. *Biometrics* 1996; 492-499.
242. Podgor MJ and Leske MC. Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Statistics in medicine* 1986; 5: 573-578.
243. !!! INVALID CITATION !!! .
244. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>. 2017.
245. Hontelez JA, de Vlas SJ, Baltussen R, et al. The impact of antiretroviral treatment on the age composition of the HIV epidemic in sub-Saharan Africa. *AIDS (London, England)* 2012; 26.
246. Negin J, Mills EJ and Albone R. Continued neglect of ageing of HIV epidemic at UN meeting. *The Lancet* 2011; 378: 768.
247. Schaefer R, Gregson S, Eaton JW, et al. Age-disparate relationships and HIV incidence in adolescent girls and young women: evidence from Zimbabwe. *AIDS* 2017; 31: 1461-1470. DOI: 10.1097/qad.0000000000001506.
248. Shisana O, Rehle T, Simbayi L, et al. South African national HIV prevalence, HIV incidence, behavior and communication survey, 2005. *Cape Town, South Africa* 2005.
249. Gregson S, Nyamukapa CA, Garnett GP, et al. Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. *The Lancet* 2002; 359: 1896-1903.
250. Leclerc-Madlala S. Age-disparate and intergenerational sex in southern Africa: the dynamics of hypervulnerability. *Aids* 2008; 22: S17-S25.
251. Pettifor AE, Levandowski BA, MacPhail C, et al. Keep them in school: the importance of education as a protective factor against HIV infection among young South

- African women. *International Journal of Epidemiology* 2008; 37: 1266-1273. DOI: 10.1093/ije/dyn131.
252. Hallett T, Gregson S, Lewis J, et al. Behaviour change in generalised HIV epidemics: impact of reducing cross-generational sex and delaying age at sexual debut. *Sexually transmitted infections* 2007; 83: i50-i54.
253. Schaefer R, Gregson S, Takaruzza A, et al. Spatial patterns of HIV prevalence and service use in East Zimbabwe: implications for future targeting of interventions. *Journal of the International AIDS Society* 2017; 20: 21409.
254. Oseni Z, Seedat F and Kandala N-B. HIV EPIDEMIC HETEROGENEITY IN ZIMBABWE: EVIDENCE FROM SUCCESSIVE DEMOGRAPHIC AND HEALTH SURVEYS. *Journal of Biosocial Science* 2018; 50: 840-852. 2018/01/15. DOI: 10.1017/S0021932017000657.
255. Gwitira I, Murwira A, Mberikunashe J, et al. Spatial overlaps in the distribution of HIV/AIDS and malaria in Zimbabwe. *BMC infectious diseases* 2018; 18: 1-10.
256. Cuadros DF, Li J, Mukandavire Z, et al. Towards UNAIDS Fast-Track goals: targeting priority geographic areas for HIV prevention and care in Zimbabwe. *AIDS* 2019; 33: 305-314. DOI: 10.1097/qad.0000000000002052.
257. Manyangadze T, Chimbari MJ and Mavhura E. Spatial Heterogeneity Association of HIV Incidence with Socio-economic Factors in Zimbabwe. *Journal of Geographical Research* 2021; 4.
258. Birri Makota RB and Musenge E. Estimating Age specific Prevalence and Force of Infection in Zimbabwe using Combined cross-sectional surveys from 2005-2015. *Frontiers in Epidemiology*: 59.

259. Manepalli UR, Bham GH and Kandada S. Evaluation of hotspots identification using kernel density estimation (K) and Getis-Ord (Gi*) on I-630. In: *3rd International Conference on Road Safety and Simulation* 2011, pp.14-16.
260. Gribov A and Krivoruchko K. Empirical Bayesian kriging implementation and usage. *Science of The Total Environment* 2020; 722: 137290. DOI: <https://doi.org/10.1016/j.scitotenv.2020.137290>.
261. Wheeler DC and Páez A. Geographically weighted regression. *Handbook of applied spatial analysis*. Springer, 2010, pp.461-486.
262. Fotheringham AS, Yang W and Kang W. Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers* 2017; 107: 1247-1265.
263. Moyo N, Maharaj P and Mambondiani L. Food challenges facing people living with HIV/AIDS in Zimbabwe. *African Journal of AIDS Research* 2017; 16: 225-230.
264. Makurumidze R, Decroo T, Lynen L, et al. District-level strategies to control the HIV epidemic in Zimbabwe: a practical example of precision public health. *BMC Research Notes* 2020; 13: 1-6.
265. Dwyer-Lindgren L, Cork MA, Sligar A, et al. Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* 2019; 570: 189-193.
266. Lumbidzani D. Why Tsholotsho has the highest HIV burden. *Chronicles*, Date published: 21 March 2022 2022.
267. Fox AM. THE HIV-POVERTY THESIS RE-EXAMINED: POVERTY, WEALTH OR INEQUALITY AS A SOCIAL DETERMINANT OF HIV INFECTION IN SUB-SAHARAN AFRICA? *Journal of Biosocial Science* 2012; 44: 459-480. 2012/01/04. DOI: 10.1017/S0021932011000745.

268. Mishra V, Assche SB-V, Greener R, et al. HIV infection does not disproportionately affect the poorer in sub-Saharan Africa. *AIDS* 2007; 21: S17-S28. DOI: 10.1097/01.aids.0000300532.51860.2a.
269. Gaumer G, Sherafat-Kazemzadeh R, Jordan M, et al. Wealth and wealth inequality in adult HIV prevalence. *Journal of Global Health Reports* 2021; 4: e2020105.
270. Mabaso M, Mlangeni L, Makola L, et al. Factors associated with age-disparate sexual partnerships among males and females in South Africa: a multinomial analysis of the 2012 national population-based household survey data. *Emerging Themes in Epidemiology* 2021; 18: 3. DOI: 10.1186/s12982-021-00093-5.
271. Wabiri N, Shisana O, Zuma K, et al. Assessing the spatial nonstationarity in relationship between local patterns of HIV infections and the covariates in South Africa: a geographically weighted regression analysis. *Spatial and spatio-temporal epidemiology* 2016; 16: 88-99.
272. Kalichman SC, Ntseane D, Nthomang K, et al. Recent multiple sexual partners and HIV transmission risks among people living with HIV/AIDS in Botswana. *Sexually transmitted infections* 2007; 83: 371-375.
273. Townsend L, Jewkes R, Mathews C, et al. HIV risk behaviours and their relationship to intimate partner violence (IPV) among men who have multiple female sexual partners in Cape Town, South Africa. *AIDS and Behavior* 2011; 15: 132-141.
274. Nkomazana N and Maharaj P. Perception of risk of HIV infections and sexual behaviour of the sexually active university students in Zimbabwe. *SAHARA-J: Journal of Social Aspects of HIV/AIDS* 2014; 11: 42-50.
275. Mutenheri E. Correlates of multiple sexual partnerships among formerly married men in Zimbabwe. *International journal of research in medical and health sciences* 2014; 3: 2307-2083.

276. Gavin L, Galavotti C, Dube H, et al. Factors associated with HIV infection in adolescent females in Zimbabwe. *Journal of Adolescent Health* 2006; 39: 596. e511-596. e518.
277. Pettifor AE, Van der Straten A, Dunbar MS, et al. Early age of first sex: a risk factor for HIV infection among women in Zimbabwe. *Aids* 2004; 18: 1435-1442.
278. Gregson S, Mason PR, Garnett GP, et al. A rural HIV epidemic in Zimbabwe? Findings from a population-based survey. *International journal of STD & AIDS* 2001; 12: 189-196.
279. Clark S. Early marriage and HIV risks in sub-Saharan Africa. *Studies in family planning* 2004; 35: 149-160.
280. Mermin J, Musinguzi J, Opio A, et al. Risk factors for recent HIV infection in Uganda. *Jama* 2008; 300: 540-549.

Appendix 1: Papers Under Review-Reviewer Comments

PAPER 2

Dear Editor,

This is the response to reviewers addressing each reviewer's comments. These are presented in a tabular format. I have added more information to what has been submitted now. I have included analysis of the best algorithm per survey year to get a picture of how factors associated with HIV positivity change over time.

Yours sincerely,

Rutendo

Review report 1

General review

The authors researched a very relevant and novel domain. The content reads logically and it kept the reader's attention very easily. One of the strong points of this paper is the authors' ability to critically analyse and consolidate arguments and viewpoints. Objectivity towards all viewpoints was visible, but the perceptive insight with which literature was discussed was quite noticeable. The paper is assembled and presented in an outstanding structure, that exhibits an overall coherence and the rationale for the chosen structure is articulated exceptionally well and is sustained throughout the remainder of the work. The language register employed in the manuscript is excellent for this level of study, and language usage is therefore on a high scientific standard.

Points of concern and suggestions are listed below:

#	Comment	Line number comment effected
1.	Keywords should be in alphabetic order	Keywords sorted alphabetically L42-43
2.	In the abstract, the authors state that "Furthermore, compared to traditional regression approaches, machine learning uncovered patterns in predicting HIV infection with comparatively reduced uncertainty and, therefore, crucial for effective decisionmaking." I suggest removing the term regression approaches to statistical approaches.	Term replaced with statistical approaches L39
3.	Be consistent with the use of the term percentage or "%".	This has been rectified
4.	Authors need to specifically mention that this was a supervised machine learning binary classification problem.	Please see L144
5.	The methodology section should state that this was a retrospective cross sectional study that employed secondary data analysis. One suggestion would be to have sub sections like study design, population, ML models etc. in the methodology.	Study design L108 ML models L143 Variable Selection L162
6.	Line 141, the authors stated that "Due to the unequal class distribution of the classification variable..." I suggest that the authors mention the class imbalance in ratio terms of majority:	Class imbalance ratio has been included L156

	minority and explore the use of data balancing strategies to address such, using strategies like SMOTE or ROSE. I will be interested in getting the new results.	
7.	Also if the authors used age as a predictor, how was it scaled to be within the same magnitude with the outcome variables, thus between 0 and 1 inclusively. They have to apply feature scaling to avoid bias in the results.	This is already specified in L132 & L141-142
8.	The authors state that the 20% of the data was used for testing purposes. What percentage was for validation e.g. testing 10% and validation 10%?	This is also stated in L139-140
9.	The authors can include a summary table of the selected hyper parameters for each model.	Table included in supplementary data

NB: Additionally, authors can explore with geographical mapping to investigate if certain areas are prone to HIV or not, obviously driven by the socio-economic status.

This might be an interesting investigation.

Review Report 2

Title: Predicting HIV infection in the decade (2005- 2015) pre-COVID-19 in Zimbabwe:

A supervised classification-based machine learning approach.

Submission ID: PDIG-D-22-00273

This manuscript presented a research study by utilizing data from three ZDHS HIV test results and the datasets from adult interviews and claimed to determine common risk factors of HIV positivity (to identify individuals with a higher likelihood of HIV infection) in Zimbabwe between the decade 2005 to 2015 and found that the ML algorithm XGBoost has the best classification performance. This manuscript is well written and presented except few things. Following are my comments to consider and address:

#	Comment	Line number comment effected
1.	“Introduction” section should be increased and previous research findings should be discussed in more detail, followed by mentioning previous research limitations and then possible solution or improvement proposed in this study.	This has been effected, see L87-104
2.	Line 88-90: “Although a few studies have been conducted on future HIV prediction....” and Line 252-254: “Although there have been a few studies utilizing machine learning techniques”— The cited references [2], [15], [16], [17], [18], etc. have used machine learning to predict HIV, so it is not clear that have the researchers done something different or achieved some better results compared to the results of the cited references? If they have achieved better results, then this sentence should be modified accordingly. Also, “routinely collected survey data” should be explained with respect to the process, frequency, etc.	This has been added and explained in L87-104
3.	Despite a number of other researchers also have published work using machine learning methods to predict HIV using Demographic Health Survey (DHS) data in Zimbabwe, why authors claim to be first in this study, please justify the claim?	This has been added and explained in L87-104
4.	Train, test and validation procedure: this section should include about the basis of data split, training dataset, validation dataset and testing dataset, how models were trained and other related information. Also, how data (information) leakage was avoided during training and validation phase?	Data split, training, testing and validation has been explained under subsection: Train, Test and Validation Procedure. See L134-1142. Data information leakage was avoided by performing k-fold cross-validation, which has been recommended as one of the best ways to get rid of data leakage. L141
5.	Line 137-138: “Each of the best three models by....”— It is not clear what does the author mean by averaging the best 3 models? Wouldn’t it compromise the best performing model?	The three models mentioned are for each survey year. The sentenced has been further clarified. See L152
6.	This work seems highly motivated from the references [2] and [38] with respect to the methodology and their results also found XGBoost algorithm performed better. Author must discuss about the differences and compare the results.	This has been added, see L306-312

7.	Authors should also discuss about other studies which have concluded that child head families are at the more risk of HIV infection. Did the authors also consider other causes of HIV infections such as blood transfusion?	References to that statement have been inserted. See L267
8	However, the figures are exported with 300dpi, figures quality can be further improved. Please use high resolution clear images and to insert, follow the journal's guidelines	I have ran the figures using the journal transformation application (PACE). I also noticed that if you download the manuscript the figures inserted at the end are blurry, therefore there is need to download the Figure on the manuscript using the download icon at the top of the inserted figure.
9.	Line 283-288: Authors should clearly explain and discuss in more detail about how patriarchal and hypermasculist culture, physical violence, etc. attributed as a valid cause of HIV infection?	This statement was a reference from a previous study which was cited in text.

Overall, this manuscript is well written and presented but still require some improvement to clearly express the research findings and justify their claim.

PAPER 3

Dear Editor,

Below in tabular form are the reviewer comments and how they were addressed. In your email, you stated why we did not use the ZIMPHIA data. ZIMPHIA is also a cross-sectional study just like the DHS and we felt that including a survey similar to what we already had was redundant. The ZIMPHIA data available is for year 2015-16 and we already have the 2015 DHS dataset, again, we felt this is more or less similar data obtained from the same population. However, information regarding ZIMPHIA data has been included in the introduction and discussion sections.

Reviewer #1: Reviewer's Comments	How Reviewer's Comment was addressed.
Summary I appreciate the opportunity I have to review for this HIV incidence modeling article in Zimbabwe. Comparing different modelling methods and comparing the variability is a good study which will enhance the validity of the models used in estimations like this. The authors' results and conclusions are quite informative and good for decision making by estimators and implementers of HIV intervention in Zimbabwe	No address needed on this comment.
Cover page • The title clearly depicts the research that was carried out • Each author's contribution was clearly stated, and corresponding author's contact is clear.	No address needed on this comment.
Abstract • The abstract showed clearly the work done briefly and concisely.	No address needed on this comment.
Introduction • The introduction was well written and introduced the readers to the need to want to find out what the authors found out. • The objectives are well stated at the end of the introduction	No address needed on this comment.
Methods and Materials • This is a well-designed method and clearly written for others to replicate • Definitions of each of the 2 predictive models are well clearly stated	No address needed on this comment.

Results • Results are well written but Figure 4 is not clear: the sizes are small and the picture is pixelating	Figure 4 was redone and run through PACE.
Discussion • The discussion is well written and provides the necessary reasoning behind the results of the study	No address needed on this comment.
Conclusion • Provides answers to the objectives • The recommendations made by the authors is appropriate for the study	No address needed on this comment.
References • The references are complete and well cited	No address needed on this comment.
Recommendation • Accept	
Reviewer #2: Reviewer's Comments	
The authors have not taken into account the latest Literature on HIV Incidence in Zimbabwe. There is lack of clarity around Prevalence and Incidence- sweeping statements in introduction highlight this lack of clarity. Lack of Reference to ZIMPHIA data and other work done in Zimbabwe on estimation of HIV Incidence is particularly worrisome.	A paragraph was added to this effect. Please see line 48-60. ZIMPHIA data has been referenced in the paragraph.
The Discussion and conclusion present themselves as an academic exercise rather than a publication presenting new and useful data	The Discussion section attended to so as to incorporate reviewer comments. It was presented in a way that, we firstly discussed the validation procedure see line, then moved on to some of the limitations and strengths of our models, thirdly, we discussed potential biases, lastly, we contextualised the results. The conclusion included recommendations to policy makers based on the models as well as contextual. If there is any other information the reviewer wants us to remove or add to strengthen the paper, may they please advise and we will do as they recommend.

Appendix 2: Papers Analysis Scripts/Codes

PAPER 1

RCode

```
setwd("/Users/rutendozivai/Desktop/Frontiers Public health")
read.csv("/Users/rutendozivai/Desktop/fontiersdAta_interval1.csv")
library(icenReg)
library(foreach)
library(doParallel)
attach(fontiersdata_interval1)
##### Inspecting model fit#####
#####
##### 1. Examining Baseline Distribution#####
quartz(width=10, height=10)
par(mfrow=c(2,2))

diag_baseline(cbind(left, right) ~ year, model = "po", data = fontiersdata_interval1, dists =
c("exponential", "weibull", "loglogistic", "gamma"), lgdLocation = "topright")
diag_baseline(cbind(left, right) ~ year + sex + maritalstatus + sti + v025 + education + v714 +
sexpartners, model = "ph", data = fontiersdata_interval1, dists = c("exponential", "weibull",
"loglogistic", "gamma"), lgdLocation = "topright")

#####Fitting Models using icenReg #####
#####
#####1. Non-parametric models#####
#####non-parametric maximum likelihood estimator (NPMLE) can be fit using
ic_np#####

np_fit = ic_np(cbind(left, right) ~ year, data = fontiersdata_interval2_males)
plot(np_fit, col = c('blue', 'orange', 'red'), xlab = 'Time', ylab = 'Estimated Survival')

#####3. Parametric Models #####
#####We can fit parametric models in icenReg using the ic_par function
fontiersdata<- read.csv("/Users/rutendozivai/Desktop/ /fontiersdata_interval1_2005.csv")
library(icenReg)
library(foreach)
library(doParallel)
attach(fontiersdata_interval1_2005)
mle_fit <- ic_par(cbind(left, right) ~ relevel(sex, ref="female") + relevel(maritalstatus, ref
="never married") + relevel(sti, ref="no") + relevel(v025, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(v714, ref="no"), model = "ph",
dist = "weibull", data = fontiersdata_interval1_2005)
summary(mle_fit)
mle_fit
coef(mle_fit)
exp(confint(mle_fit))

quartz(width=10, height=10)
par(mfrow=c(2,3))

fontiersdata_interval1_2005$sex <-as.factor(fontiersdata_interval1_2005$sex)
cols <- c("maritalstatus", "sti", "v025", "education", "sex", "sexpartners", "v714")
fontiersdata_interval1_2005[cols] <- lapply(fontiersdata_interval1_2005[cols], factor)
sapply(fontiersdata_interval1_2005, class)
newdata <- data.frame(sex=c("female", "male"))
rownames(newdata) <- c('female', 'male')

plot(mle_fit, newdata, lgdLocation = "topright")

#####3. Parametric Models 2010#####
#####We can fit parametric models in icenReg using the ic_par function
rm(list=ls(all=TRUE))
detach(fontiersdata_interval1)
```

```

detach(fontiersdata_interval1_females)
detach(fontiersdata_interval1_males)
detach(fontiersdata_interval2)
detach(fontiersdata_interval2_females)
detach(fontiersdata_interval2_males)
detach(fontiersdata_interval1_2005)
detach(fontiersdata_interval1_2010)
detach(fontiersdata_interval1_2015)
rm(list=ls(all=TRUE))

fontiersdata_interval2_2010<- read.csv("/Users/rutendozivai/Desktop/Frontiers Public
health/fontiersdata_interval2_2010.csv")
library(icenReg)
library(foreach)
library(doParallel)
attach(fontiersdata_interval2_2010)
mle_fit <- ic_par(cbind(left, right) ~ relevel(sex, ref="female") + relevel(maritalstatus, ref
="never married") + relevel(sti, ref="no") + relevel(v025, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(v714, ref="no"), model = "ph",
dist = "weibull", data = fontiersdata_interval2_2010)
summary(mle_fit)
mle_fit
coef(mle_fit)
exp(confint(mle_fit))
plot(mle_fit, newdata, lgdLocation = "topright")

#####3. Parametric Models 2015#####
####We can fit parametric models in icenReg using the ic_par function
rm(list=ls(all=TRUE))
detach(fontiersdata_interval1)
detach(fontiersdata_interval1_females)
detach(fontiersdata_interval1_males)
detach(fontiersdata_interval2)
detach(fontiersdata_interval2_females)
detach(fontiersdata_interval2_males)
detach(fontiersdata_interval1_2005)
detach(fontiersdata_interval1_2010)
detach(fontiersdata_interval1_2015)
rm(list=ls(all=TRUE))

fontiersdata_interval2_2015<- read.csv("/Users/rutendozivai/Desktop/Frontiers Public
health/fontiersdata_interval2_2015.csv")
library(icenReg)
library(foreach)
library(doParallel)
attach(fontiersdata_interval2_2015)
mle_fit <- ic_par(cbind(left, right) ~ relevel(sex, ref="female") + relevel(maritalstatus, ref
="never married") + relevel(sti, ref="no") + relevel(v025, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(v714, ref="no"), model = "ph",
dist = "weibull", data = fontiersdata_interval2_2015)
summary(mle_fit)
mle_fit
coef(mle_fit)
exp(confint(mle_fit))
plot(mle_fit, newdata, lgdLocation = "topright")

##### CHANGING TO INFINITE

fontiersdata_interval1$right[is.na(fontiersdata_interval1$right)]<-Inf
#####
#####
#####
#####model inspection#####
rm(list=ls(all=TRUE))
detach(fontiersdata_interval2_2005)
library(icenReg)
library(foreach)
library(doParallel)
fontiersdata_interval2_2005 <- read.csv("/Users/rutendozivai/Desktop/
/fontiersdata_interval2_2005.csv")

```

```

attach(fontiersdata_interval2_2005)
fontiersdata_interval2_2005$right[is.na(fontiersdata_interval2_2005$right)]<-Inf
fontiersdata_interval2_2005$sex <- as.factor(fontiersdata_interval2_2005$sex)
cols <- c("maritalstatus", "sti", "v025", "education", "sex", "sexpartners", "v714")
fontiersdata_interval2_2005[cols] <- lapply(fontiersdata_interval2_2005[cols], factor)
sapply(fontiersdata_interval2_2005, class)
newdata <- data.frame(sex=c("female", "male"))
rownames(newdata) <- c('female', 'male')
quartz(width=8, height=8)
par(mfrow=c(1,3))
#####model inspection#####

##### 1. Examining Baseline Distribution#####
#diag_baseline(cbind(left, right) ~ year, model = "ph", data = combinedHIVfinal, dists =
c("exponential", "weibull", "loglogistic", "gamma"), lgdLocation = "topright")
diag_baseline(cbind(left, right) ~ sex + maritalstatus + sti + v025 + education + v714 +
sexpartners, model = "ph", data = fontiersdata_interval2_2005, dists = c("exponential",
"weibull", "loglogistic", "gamma"), lgdLocation = "topright")

#####
#####Survival Curves#####
library(survival)
library(ranger)
library(ggplot2)
library(dplyr)
library(ggfortify)

km_trt_fit <- survfit(Surv(time, status) ~ sex, data=fontiersdata_diagvisit_2005)
A=autoplot(km_trt_fit, xlim=c(14,50),ylim=c(0,1))
detach(fontiersdata_diagvisit_2005)
library(survival)
library(survminer)
library(SurvRegCensCov)
km_trt_fit1 <- survfit(Surv(diagtime, status) ~ year, data=data1)
km_trt_fit2 <- survfit(Surv(midtime, status) ~ year, data=data1)
#data1$right[is.na(data1$right)]<-Inf
km_trt_fit3 <- survfit(Surv(left, right, type="interval2")~ year, data =data1)
# Combine on the same plot
fit <- list(SurveyVisit = km_trt_fit1, Midpoint = km_trt_fit2, Interval = km_trt_fit3)
quartz(width=15, height=15)
ggsurvplot_combine(fit, xlim=c(14,50),ylim=c(0,1), data1)

library(survival)
library(SurvRegCensCov)
data1=read.csv("/Users/rutendozivai/Desktop/Frontiers Public health/fontiersdata_alltimes.csv")
data1$right[is.na(data1$right)]<-Inf
quartz(width=15, height=15)
par(mfrow=c(1,3))
a=WeibullDiag(Surv(diagtime, status) ~ year, data=data1)
b= WeibullDiag(Surv(midtime, status) ~ year, data=data1)
c=WeibullDiag(Surv(left, right, type="interval2") ~ year, data=data1)
quartz(width=15, height=15)
par(mfrow=c(1,3))
plot(a)
plot(b)
plot(c)

fontiersdata_interval2_2005$right[is.na(fontiersdata_interval2_2005$right)]<-Inf
fontiersdata_interval2_2010$right[is.na(fontiersdata_interval2_2010$right)]<-Inf
fontiersdata_interval2_2015$right[is.na(fontiersdata_interval2_2015$right)]<-Inf
quartz(width=15, height=15)
par(mfrow=c(3,3))

WeibullDiag <- function(formula, data = parent.frame(),
labels = names(m$strata))
{

```

```

m <- survival::survfit(formula, data = data)
if(!"strata" %in% names(m)) stop("No strata detected. No plot will be produced.")
cp <- 1:length(m$strata)
plot(m, fun = "cloglog",
      col = cp, lty = cp, lwd = 2, mark.time = FALSE,
      xlab = "Survival Time (log scale)", ylab = "Log Cumulative Hazard",
      main = "Weibull Regression\nDiagnostic Plot", xlim = c(15,50), ylim = c(-10, 2)
)
legend("topleft", legend = labels, col = cp,
      lwd = 2, lty = cp,
      bty = "n")
pCH <- list(x = log(m$time), y = log(-log(m$surv)),
            strata = rep(names(m$strata), m$strata))
}

WeibullDiag(Surv(time, status) ~ sex, data=fontiersdata_diagvisit_2005)
WeibullDiag(Surv(time, status) ~ sex, data=fontiersdata_diagvisit_2010)
WeibullDiag(Surv(time, status) ~ sex, data=fontiersdata_diagvisit_2015)

WeibullDiag(Surv(time, status) ~ sex, data=fontiersdata_midpoint_2005)
WeibullDiag(Surv(time, status) ~ sex, data=fontiersdata_midpoint_2010)
WeibullDiag(Surv(time, status) ~ sex, data=fontiersdata_midpoint_2015)

WeibullDiag(Surv(left, right, type="interval2") ~ sex, data=fontiersdata_interval2_2005)
WeibullDiag(Surv(left, right, type="interval2") ~ sex, data=fontiersdata_interval2_2010)
WeibullDiag(Surv(left, right, type="interval2") ~ sex, data=fontiersdata_interval2_2015)

library(survival)
library(MASS)
data1=read.csv("/Users/rutendozivai/Desktop/Frontiers Public health/fontiersdata_alltimes.csv")
plot(survfit(Surv(left, right, type="interval2")~ year, data =data1, fun="cloglog"))

fit.weib= survreg(Surv(midtime,status) ~ year, dist="weibull", data=data1)
qq.reg.resid.s(data1,data1$midtime,data1$status,fit.weib,"qweibull","standard extreme value
quantiles")

#####MODEL GOODNESS OF FIT#####
#####
#####
library(survival)
quartz(width=12, height=12)
par(mfrow=c(3,3))
fontiersdata_interval2_2005$right[is.na(fontiersdata_interval2_2005$right)]<-Inf
fontiersdata_interval2_2010$right[is.na(fontiersdata_interval2_2010$right)]<-Inf
fontiersdata_interval2_2015$right[is.na(fontiersdata_interval2_2015$right)]<-Inf
## Overplot estimation from Cox regression by Efron method
cox.null <- coxph(Surv(time, status) ~ 1, data=fontiersdata_diagvisit_2005)
plot(survfit(cox.null), xlim=c(15,50), ylab="S(t)", xlab = "time")

## Parametric estimation with Weibull distribution
weibull.null <- survreg(Surv(time, status) ~ 1, data=fontiersdata_diagvisit_2005, dist =
"weibull")
lines(x = predict(weibull.null, type = "quantile", p = seq(0.01, 0.99, by=.01))[1,],
      y = rev(seq(0.01, 0.99, by = 0.01)),
      col = "red")
legend(x = "topright", legend = c("Semi-parametric", "weibull"),
      col=c("black", "red"), lty=1:1)

fontiersdata_interval2_2005$right[is.na(fontiersdata_interval2_2005$right)]<-Inf
fontiersdata_interval2_2010$right[is.na(fontiersdata_interval2_2010$right)]<-Inf
fontiersdata_interval2_2015$right[is.na(fontiersdata_interval2_2015$right)]<-Inf

fit1=psm(Surv(time, status)~relevel(sex, ref="female") + relevel(maritalstatus, ref = "never
married") + relevel(sti, ref="no") + relevel(residence, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(employed, ref="no"),
dist="weibull", data = fontiersdata_diagvisit_2005)

```

```

fit2=psm(Surv(time, status)~relevel(sex, ref="female") + relevel(maritalstatus, ref = "never
married") + relevel(sti, ref="no") + relevel(residence, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(employed, ref="no"),
dist="weibull", data = fontiersdata_diagvisit_2010)
fit3=psm(Surv(time, status)~relevel(sex, ref="female") + relevel(maritalstatus, ref = "never
married") + relevel(sti, ref="no") + relevel(residence, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(employed, ref="no"),
dist="weibull", data = fontiersdata_diagvisit_2015)

plot(anova(fit1), margin=c("chisq", "d.f.", "p"))
plot(anova(fit2), margin=c("chisq", "d.f.", "p"))
plot(anova(fit3), margin=c("chisq", "d.f.", "p"))
fit1=survreg(Surv(time, status)~relevel(sex, ref="female") + relevel(maritalstatus, ref = "never
married") + relevel(sti, ref="no") + relevel(residence, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(employed, ref="no"),
dist="weibull", data = fontiersdata_diagvisit_2005)
fit2=survreg(Surv(time, status)~relevel(sex, ref="female") + relevel(maritalstatus, ref = "never
married") + relevel(sti, ref="no") + relevel(residence, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(employed, ref="no"),
dist="weibull", data = fontiersdata_diagvisit_2010)
fit3=survreg(Surv(time, status)~relevel(sex, ref="female") + relevel(maritalstatus, ref = "never
married") + relevel(sti, ref="no") + relevel(residence, ref="urban") + relevel(education,
ref="no edu/primary") + relevel(sexpartners, ref="0") + relevel(employed, ref="no"),
dist="weibull", data = fontiersdata_diagvisit_2015)

CSRplot(fit1)
CSRplot(fit2)
CSRplot(fit3)

```

STATA DO-FILE

```

*****DIAGNOSTIC 2005*****
*****
clear
use "/Users/rutendozivai/Desktop/resubmission/dat2005.dta"
encode residence, gen(residence1)
encode employed , gen(employed1)
encode sex, gen(sex1)
encode maritalstatus , gen(maritalstatus1)
encode education, gen(education1)
encode sexpartners, gen(sexpartners1)
encode sti, gen(sti1)
stset time, failure(status)
streg i.sex1 i.maritalstatus1 i.sti1 i.residence1 i.education1 i.employed1 i.sexpartners1,
distribution(weibull)
estat ic
predict double cs, csnell
stset cs, failure(status)
sts generate km=s
generate double H=-ln(km)
line H cs cs, sort
graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/csdiagvisit_2005",
replace

*****DIAGNOSTIC 2010*****
*****
streg i.sex1 i.maritalstatus1 i.sti1 i.residence1 i.education1 i.employed1 i.sexpartners1,
distribution(weibull)
estat ic
predict double cs, csnell
stset cs, failure(status)
sts generate km=s
generate double H=-ln(km)
line H cs cs, sort
graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/csdiagvisit_2010",
replace
qui streg i.sex i.maritalstatus i.sti i.residence i.education i.employed i.sexpartners,
distribution(weibull)
predict dev, deviance
scatter dev time, yline(0) m(0)

```

```

graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/mgdiagvisit_2010",
replace

*****DIAGNOSTIC 2015*****
*****
qui stset time, failure(status) streg i.sex1 i.maritalstatus1 i.sti1 i.residence1 i.education1
i.employed1 i.sexpartners1, distribution(weibull)
estat ic
predict double cs, csnell
stset cs, failure(status)
sts generate km=s
generate double H=-ln(km)
line H cs cs, sort
graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/csdiagvisit_2015",
replace
qui streg i.sex i.maritalstatus i.sti i.residence i.education i.employed i.sexpartners,
distribution(weibull)
predict dev, deviance
scatter dev time, yline(0) m(0)
graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/mgdiagvisit_2015",
replace

*****INTERVAL1 2005*****
*****
stintreg i.sex1 i.maritalstatus1 i.sti1 i.residence1 i.education1 i.employed1
i.sexpartners1, interval(left right) distribution(weibull)
estat ic
**estat gofplot, outfile("/Users/rutendozivai/Desktop/Frontiers Public health/dat2005gof")
graph twoway (line cumhaz csresid) (function y = x, range(csresid))
graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/csinterval_2005",
replace

*****INTERVAL1 2010*****
*****
stintreg i.sex1 i.maritalstatus1 i.sti1 i.residence1 i.education1 i.employed1
i.sexpartners1, interval(left right) distribution(weibull)
estat ic
**estat gofplot, outfile("/Users/rutendozivai/Desktop/Frontiers Public health/dat2010gof")
graph twoway (line cumhaz csresid) (function y = x, range(csresid))
graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/csinterval_2010",
replace
*****INTERVAL1 2015*****
*****
stintreg i.sex1 i.maritalstatus1 i.sti1 i.residence1 i.education1 i.employed1
i.sexpartners1, interval(left right) distribution(weibull)
estat ic
**estat gofplot, outfile("/Users/rutendozivai/Desktop/Frontiers Public health/dat2015gof")
graph twoway (line cumhaz csresid) (function y = x, range(csresid))
graph save "/Users/rutendozivai/Desktop/Frontiers Public health/resubmission/csinterval_2015",
replace

```

PAPER 2

XGBOOST Python code

```

# import libraries
import os
import pandas as pd
import numpy as np
import pickle
from joblib import dump, load

from sklearn.model_selection import StratifiedKFold, RandomizedSearchCV

```

```

from sklearn.metrics import f1_score, recall_score, precision_score
from xgboost import XGBClassifier
from matplotlib import pyplot
import shap
from sklearn.metrics import confusion_matrix, f1_score, recall_score, precision_score,
precision_recall_curve, auc
random_seed = 5

# set working directory
working_directory = '/Users/rutendozivai/Desktop/ML_reanalysis'
os.chdir(working_directory)
# code for uploading pickled data
M_X_train_ready =
pickle.load(open("year_original_data/2005zw/Train_samples/M_X_train_ready.pkl", 'rb'))
M_Y_train_ready =
pickle.load(open("year_original_data/2005zw/Train_samples/M_Y_train_ready.pkl", 'rb'))
M_X_train_ready = pd.concat(M_X_train_ready.values(), join='inner', ignore_index=True)
M_Y_train_ready = pd.concat(M_Y_train_ready.values(), join='inner', ignore_index=True)

F_X_train_ready =
pickle.load(open("year_original_data/2005zw/Train_samples/F_X_train_ready.pkl", 'rb'))
F_Y_train_ready =
pickle.load(open("year_original_data/2005zw/Train_samples/F_Y_train_ready.pkl", 'rb'))
F_X_train_ready = pd.concat(F_X_train_ready.values(), join='inner', ignore_index=True)
F_Y_train_ready = pd.concat(F_Y_train_ready.values(), join='inner', ignore_index=True)

M_X_test_ready =
pickle.load(open("year_original_data/2005zw/Test_samples/M_X_test_ready.pkl", 'rb'))
M_Y_test_ready =
pickle.load(open("year_original_data/2005zw/Test_samples/M_Y_test_ready.pkl", 'rb'))
M_X_test_ready = pd.concat(M_X_test_ready.values(), join='inner', ignore_index=True)
M_Y_test_ready = pd.concat(M_Y_test_ready.values(), join='inner', ignore_index=True)

F_X_test_ready =
pickle.load(open("year original data/2005zw/Test samples/F X test ready.pkl", 'rb'))
F_Y_test_ready =
pickle.load(open("year_original_data/2005zw/Test_samples/F_Y_test_ready.pkl", 'rb'))
F_X_test_ready = pd.concat(F_X_test_ready.values(), join='inner', ignore_index=True)
F_Y_test_ready = pd.concat(F_Y_test_ready.values(), join='inner', ignore_index=True)

CM_test_ready =
pickle.load(open("year_original_data/2005zw/Left_one_out_samples/CM_test_ready.pkl",
'rb'))
CM_Y_test =
pickle.load(open("year_original_data/2005zw/Left_one_out_samples/CM_Y_test.pkl", 'rb'))
CM_test_ready = pd.concat(CM_test_ready.values(), join='inner', ignore_index=True)
CM_Y_test = pd.concat(CM_Y_test.values(), join='inner', ignore_index=True)

CF_test_ready =
pickle.load(open("year_original_data/2005zw/Left_one_out_samples/CF_test_ready.pkl",
'rb'))
CF_Y_test =
pickle.load(open("year_original_data/2005zw/Left_one_out_samples/CF_Y_test.pkl", 'rb'))
CF_test_ready = pd.concat(CF_test_ready.values(), join='inner', ignore_index=True)
CF_Y_test = pd.concat(CF_Y_test.values(), join='inner', ignore_index=True)

```



```

M_selected_features = ['Type of place of residence',
                        'Highest educational level',
                        'Religion',
                        'Number of household members',
                        'Age of household head',
                        'Times away from home in last 12 months',
                        'Current contraceptive method',
                        'Number of injections in last 12 months',
                        'Respondent circumcised',
                        'Currently/formerly/never in union',
                        'Cohabitation duration (grouped)',
                        'Currently working',
                        'Beating justified',
                        'Reduce risk of getting HIV',
                        'Had any STI in last 12 months',
                        'Wife justified asking husband to use condom if he has STI',
                        'Total lifetime number of sex partners']

F_selected_features = ['Type of place of residence',
                        'Religion',
                        'Number of household members',
                        'Relationship to household head',
                        'Sex of household head',
                        'Age of household head',
                        'Times away from home in last 12 months',
                        'Wealth index',
                        'Current contraceptive method',
                        'Current contraceptive by method type',
                        'Currently/formerly/never in union',
                        'Number of unions',
                        'Cohabitation duration (grouped)',
                        'Time since last sex (in days)',
                        'Recent sexual activity',
                        'Beating justified',
                        'Ever heard of AIDS',
                        'Reduce risk of getting HIV',
                        'Had any STI in last 12 months',
                        'Ever been tested for HIV',
                        'Wife justified asking husband to use condom if he has STI',
                        'Total lifetime number of sex partners',
                        'Can ask partner to use condom']

M_X_train_ready_selected = M_X_train_ready[M_selected_features]
F_X_train_ready_selected = F_X_train_ready[F_selected_features]
M_X_test_ready_selected = M_X_test_ready[M_selected_features]
F_X_test_ready_selected = F_X_test_ready[F_selected_features]
CM_test_ready_selected = CM_test_ready[M_selected_features]
CF_test_ready_selected = CF_test_ready[F_selected_features]

#=====
# parameters space, models definition and random grid search
#=====

# stratified k-fold and model initialization
kfold = StratifiedKFold(n_splits=5, shuffle=True, random_state=random_seed)
scoring = ['f1', 'recall', 'precision']

```

```

# XGBoost space
params_xgb = {'learning_rate': np.linspace(0, 1, 11),
              'min_split_loss': np.linspace(0, 1, 6),
              'max_depth': np.linspace(2, 10, 5, dtype=int),
              'min_child_weight': np.linspace(1, 20, 20, dtype=int),
              'colsample_bytree': np.linspace(0.5, 1, 6),
              'reg_alpha': np.linspace(0, 1, 11),
              'scale_pos_weight': np.linspace(4, 50, 24, dtype=int),
              'n_estimators': np.linspace(10, 450, 12, dtype=int),
              'reg_lambda': np.linspace(0, 10, 11),
              }

M_xgb_sample = RandomizedSearchCV(estimator=XGBClassifier(booster='gbtree',

objective='binary:logistic',

random_state=random_seed,

nthread=1,
verbosity=1),

param_distributions=params_xgb,
n_iter=50,
scoring=scoring,
cv=kfold,
n_jobs=-1,
random_state=random_seed,
verbose=2,
refit='f1')

F_xgb_sample = RandomizedSearchCV(estimator=XGBClassifier(booster='gbtree',

objective='binary:logistic',

random_state=random_seed,

nthread=1,
verbosity=1),

param_distributions=params_xgb,
n_iter=50,
scoring=scoring,
cv=kfold,
n_jobs=-1,
random_state=random_seed,
verbose=2,
refit='f1')

M_xgb_sample.fit(M_X_train_ready_selected, M_Y_train_ready)
dump(M_xgb_sample, "scripts_original_2005/M_xgb_sample_data.joblib")

F_xgb_sample.fit(F_X_train_ready_selected, F_Y_train_ready)
dump(F_xgb_sample, "scripts_original_2005/F_xgb_sample_data.joblib")
M_xgb_selected = load("scripts_original_2005/M_xgb_sample_data.joblib")
F_xgb_selected = load("scripts_original_2005/F_xgb_sample_data.joblib")
CM_xgb_selected = load("scripts_original_2005/M_xgb_sample_data.joblib")
CF_xgb_selected = load("scripts_original_2005/F_xgb_sample_data.joblib")

M_best_index = M_xgb_selected.best_index_

```

```

F_best_index = F_xgb_selected.best_index_

M_Y_train_selected_pred = M_xgb_selected.predict(M_X_train_ready_selected)
# M_Y_train_selected_pred_thresh =
(M_xgb_selected.predict_proba(M_X_train_selected)[: ,1] >= 0.3).astype(bool) # set
threshold as 0.3
F_Y_train_selected_pred = F_xgb_selected.predict(F_X_train_ready_selected)
from sklearn.metrics import classification_report
report = classification_report(M_Y_test_ready,
M_xgb_selected.best_estimator_.predict(M_X_test_ready_selected))
print(report)

report = classification_report(F_Y_test_ready,
F_xgb_selected.best_estimator_.predict(F_X_test_ready_selected))
print(report)
print('Confusion matrix on the M train dataset:', confusion_matrix(M_Y_train_ready,
M_Y_train_selected_pred))
print('Average F1 score on the M train sample',
"{:.1%}".format(M_xgb_selected.cv_results_['mean_test_f1'][M_best_index]))
print('CI F1 score on the M train sample',
"{:.1%}".format(2*M_xgb_selected.cv_results_['std_test_f1'][M_best_index]))
print('Average Sensitivity on the M train sample',
"{:.1%}".format(M_xgb_selected.cv_results_['mean_test_recall'][M_best_index]))
print('CI Sensitivity on the M train sample',
"{:.1%}".format(2*M_xgb_selected.cv_results_['std_test_recall'][M_best_index]))
print('Average Positive Predictive Value on the M train sample',
"{:.1%}".format(M_xgb_selected.cv_results_['mean_test_precision'][M_best_index]))
print('CI Positive Predictive Value on the M train sample',
"{:.1%}".format(2*M_xgb_selected.cv_results_['std_test_precision'][M_best_index]))
print('Parameters', M_xgb_selected.best_estimator_)
print('Confusion matrix on the F train dataset:', confusion_matrix(F_Y_train_ready,
F_Y_train_selected_pred))
print('Average F1 score on the F train sample',
"{:.1%}".format(F_xgb_selected.cv_results_['mean_test_f1'][F_best_index]))
print('CI F1 score on the F train sample',
"{:.1%}".format(2*F_xgb_selected.cv_results_['std_test_f1'][F_best_index]))
print('Average Sensitivity on the F train sample',
"{:.1%}".format(F_xgb_selected.cv_results_['mean_test_recall'][F_best_index]))
print('CI Sensitivity on the F train sample',
"{:.1%}".format(2*F_xgb_selected.cv_results_['std_test_recall'][F_best_index]))
print('Average Positive Predictive Value on the F train sample',
"{:.1%}".format(F_xgb_selected.cv_results_['mean_test_precision'][F_best_index]))
print('CI Positive Predictive Value on the F train sample',
"{:.1%}".format(2*F_xgb_selected.cv_results_['std_test_precision'][F_best_index]))
print('Parameters', F_xgb_selected.best_estimator_)
from matplotlib import pyplot
M_Y_test_selected_pred = M_xgb_selected.predict(M_X_test_ready_selected)
F_Y_test_selected_pred = F_xgb_selected.predict(F_X_test_ready_selected)
CM_Y_test_selected_pred = M_xgb_selected.predict(CM_test_ready_selected)
CF_Y_test_selected_pred = F_xgb_selected.predict(CF_test_ready_selected)

print('Confusion matrix on the M test dataset:', confusion_matrix(M_Y_test_ready,
M_Y_test_selected_pred))

```

```

print('F1 score on the M test dataset:', "{:.1%}".format(f1_score(M_Y_test_ready,
M_Y_test_selected_pred)))

print('Sensitivity on the M test dataset:', "{:.1%}".format(recall_score(M_Y_test_ready,
M_Y_test_selected_pred)))
print('Positive Predictive Value on the M test dataset:',
"{:.1%}".format(precision_score(M_Y_test_ready, M_Y_test_selected_pred)))

print('Confusion matrix on the F test dataset:', confusion_matrix(F_Y_test_ready,
F_Y_test_selected_pred))
print('F1 score on the F test dataset:', "{:.1%}".format(f1_score(F_Y_test_ready,
F_Y_test_selected_pred)))
print('Sensitivity on the F test dataset:', "{:.1%}".format(recall_score(F_Y_test_ready,
F_Y_test_selected_pred)))
print('Positive Predictive Value on the F test dataset:',
"{:.1%}".format(precision_score(F_Y_test_ready, F_Y_test_selected_pred)))

print('Confusion matrix on the CF test dataset:', confusion_matrix(CF_Y_test,
CF_Y_test_selected_pred))
print('F1 score on the CF test dataset:', "{:.1%}".format(f1_score(CF_Y_test,
CF_Y_test_selected_pred)))
print('Sensitivity on the CF test dataset:', "{:.1%}".format(recall_score(CF_Y_test,
CF_Y_test_selected_pred)))
print('Positive Predictive Value on the CF test dataset:',
"{:.1%}".format(precision_score(CF_Y_test, CF_Y_test_selected_pred)))

print('Confusion matrix on the CM test dataset:', confusion_matrix(CM_Y_test,
CM_Y_test_selected_pred))
print('F1 score on the CM test dataset:', "{:.1%}".format(f1_score(CM_Y_test,
CM_Y_test_selected_pred)))
print('Sensitivity on the CM test dataset:', "{:.1%}".format(recall_score(CM_Y_test,
CM_Y_test_selected_pred)))
print('Positive Predictive Value on the CM test dataset:',
"{:.1%}".format(precision_score(CM_Y_test, CM_Y_test_selected_pred)))
from matplotlib import pyplot
def pr_f1_curve(model, dx_test, dy_test):
    dy_test_predict_proba = model.predict_proba(dx_test)[: ,1]
    dy_test_predict = model.predict(dx_test)
    ppv_curve, sensitivity_curve, _ = precision_recall_curve(dy_test,
dy_test_predict_proba)
    F1_score, AUC_score = f1_score(dy_test, dy_test_predict), auc(sensitivity_curve,
ppv_curve)
    print(': F1 score = %.2f - Area under Curve = %.2f' % (F1_score, AUC_score))
    prevalence = np.sum(dy_test)/len(dy_test)
    pyplot.plot([0, 1], [prevalence, prevalence], linestyle='--', label='Random
Testing')
    pyplot.plot(sensitivity_curve, ppv_curve, marker='.', label='Algorithmic Testing')
    pyplot.plot(sensitivity_curve,
2*(sensitivity_curve*ppv_curve)/(sensitivity_curve+ppv_curve), marker='', label='F1')
    pyplot.xlabel('Sensitivity')
    pyplot.ylabel('Positive Predicted Value')
    pyplot.legend(bbox_to_anchor=(0.45, 0.75))
    pyplot.title('Precision-Recall Curve - Males')
    pyplot.text(0.1, 0, "F1 score = 0.86 - Area under Curve = 0.93")

```

```

    pyplot.show()
pr_fl_curve(M_xgb_selected, M_X_test_ready_selected, M_Y_test_ready)
def pr_fl_curve(model, dx_test, dy_test):
    dy_test_predict_proba = model.predict_proba(dx_test)[: ,1]
    dy_test_predict = model.predict(dx_test)
    ppv_curve, sensitivity_curve, _ = precision_recall_curve(dy_test,
dy_test_predict_proba)
    F1_score, AUC_score = f1_score(dy_test, dy_test_predict), auc(sensitivity_curve,
ppv_curve)
    print(': F1 score = %.2f - Area under Curve = %.2f' % (F1_score, AUC_score))
    prevalence = np.sum(dy_test)/len(dy_test)
    pyplot.plot([0, 1], [prevalence, prevalence], linestyle='--', label='Random
Testing')
    pyplot.plot(sensitivity_curve, ppv_curve, marker='.', label='Algorithmic Testing')
    pyplot.plot(sensitivity_curve,
2*(sensitivity_curve*ppv_curve)/(sensitivity_curve+ppv_curve), marker='', label='F1')
    pyplot.xlabel('Sensitivity')
    pyplot.ylabel('Positive Predicted Value')
    pyplot.legend(bbox_to_anchor=(0.45, 0.75))
    pyplot.title('Precision-Recall Curve - Females')
    pyplot.text(0.1, 0, "F1 score = 0.81 - Area under Curve = 0.90")
    pyplot.show()
pr_fl_curve(F_xgb_selected, F_X_test_ready_selected, F_Y_test_ready)
# Explain model predictions using shapley value
M_explainer = shap.TreeExplainer(M_xgb_selected.best_estimator_)
M_shap_values = M_explainer.shap_values(M_X_train_ready_selected)
F_explainer = shap.TreeExplainer(F_xgb_selected.best_estimator_)
F_shap_values = F_explainer.shap_values(F_X_train_ready_selected)
# Plot summary_plot
M_shap = shap.summary_plot(M_shap_values, M_X_train_ready_selected, max_display=32)
F_shap = shap.summary_plot(F_shap_values, F_X_train_ready_selected, max_display=32)

```

ROC and Precision-recall curves Python Code

```

# import libraries
import os
import pandas as pd
import numpy as np
import pickle
from joblib import dump, load

from sklearn.model_selection import StratifiedKFold, RandomizedSearchCV
from sklearn.metrics import f1_score, recall_score, precision_score
from xgboost import XGBClassifier
from matplotlib import pyplot
import shap
from sklearn.metrics import confusion_matrix, f1_score, recall_score, precision_score,
precision_recall_curve, auc
random_seed = 5
#load models males
M_dt_selected = load("scripts_original_2005/M_dt_sample_data.joblib")
M_en_selected = load("scripts_original_2005/M_en_sample_data.joblib")
M_knn_selected = load("scripts_original_2005/M_knn_sample_data.joblib")
M_rfc_selected = load("scripts_original_2005/M_rfc_sample_data.joblib")

```

```
#M_svm_selected = load("scripts_original_2005/M_svm_sample_data.joblib")
M_lgb_selected = load("scripts_original_2005/M_lgb_sample_data.joblib")
M_xgb_selected = load("scripts_original_2005/M_xgb_sample_data.joblib")
```

```
#load models females
```

```
F_dt_selected = load("scripts_original_2005/F_dt_sample_data.joblib")
F_en_selected = load("scripts_original_2005/F_en_sample_data.joblib")
F_knn_selected = load("scripts_original_2005/F_knn_sample_data.joblib")
F_rfc_selected = load("scripts_original_2005/F_rfc_sample_data.joblib")
#F_svm_selected = load("scripts_original_2005/F_svm_sample_data.joblib")
F_lgb_selected = load("scripts_original_2005/F_lgb_sample_data.joblib")
F_xgb_selected = load("scripts_original_2005/F_xgb_sample_data.joblib")
```

```
pred_prob1 = M_dt_selected.predict_proba(M_X_test_ready_selected)
pred_prob2 = M_en_selected.predict_proba(M_X_test_ready_selected)
pred_prob3 = M_knn_selected.predict_proba(M_X_test_ready_selected)
pred_prob4 = M_rfc_selected.predict_proba(M_X_test_ready_selected)
#pred_prob5 = M_svm_selected.predict_proba(M_X_test_ready_selected)
pred_prob6 = M_lgb_selected.predict_proba(M_X_test_ready_selected)
pred_prob7 = M_xgb_selected.predict_proba(M_X_test_ready_selected)
```

```
from sklearn.metrics import roc_curve
```

```
# roc curve for models
```

```
fpr1, tpr1, thresh1 = roc_curve(M_Y_test_ready, pred_prob1[:,1], pos_label=1)
fpr2, tpr2, thresh2 = roc_curve(M_Y_test_ready, pred_prob2[:,1], pos_label=1)
fpr3, tpr3, thresh3 = roc_curve(M_Y_test_ready, pred_prob3[:,1], pos_label=1)
fpr4, tpr4, thresh4 = roc_curve(M_Y_test_ready, pred_prob4[:,1], pos_label=1)
#fpr5, tpr5, thresh5 = roc_curve(M_Y_test_ready, pred_prob5[:,1], pos_label=1)
fpr6, tpr6, thresh6 = roc_curve(M_Y_test_ready, pred_prob6[:,1], pos_label=1)
fpr7, tpr7, thresh7 = roc_curve(M_Y_test_ready, pred_prob7[:,1], pos_label=1)
```

```
# roc curve for tpr = fpr
```

```
random_probs = [0 for i in range(len(M_Y_test_ready))]
p_fpr, p_tpr, _ = roc_curve(M_Y_test_ready, random_probs, pos_label=1)
```

```
from sklearn.metrics import roc_auc_score
```

```
# auc scores
```

```
auc_score1 = roc_auc_score(M_Y_test_ready, pred_prob1[:,1])
auc_score2 = roc_auc_score(M_Y_test_ready, pred_prob2[:,1])
auc_score3 = roc_auc_score(M_Y_test_ready, pred_prob3[:,1])
auc_score4 = roc_auc_score(M_Y_test_ready, pred_prob4[:,1])
#auc_score5 = roc_auc_score(M_Y_test_ready, pred_prob5[:,1])
auc_score6 = roc_auc_score(M_Y_test_ready, pred_prob6[:,1])
auc_score7 = roc_auc_score(M_Y_test_ready, pred_prob7[:,1])
```

```
print(auc_score1, auc_score2, auc_score3, auc_score4, auc_score6, auc_score7)
```

```
# matplotlib
```

```
import matplotlib.pyplot as plt
```

```

plt.style.use('seaborn')
# plot roc curves
plt.plot(fpr1, tpr1, linestyle='--',color='orange', label='DT (area = %0.2f)' %
auc_score1)
plt.plot(fpr2, tpr2, linestyle='--',color='green', label='EN (area = %0.2f)' %
auc_score2)
plt.plot(fpr3, tpr3, linestyle='--',color='red', label='KNN (area = %0.2f)' %
auc_score3)
plt.plot(fpr4, tpr4, linestyle='--',color='purple', label='RFC (area = %0.2f)' %
auc_score4)
#plt.plot(fpr5, tpr5, linestyle='--',color='yellow', label='SVM (area = %0.2f)' %
auc_score5)
plt.plot(fpr6, tpr6, linestyle='--',color='black', label='LIGHTGBM (area = %0.2f)' %
auc_score6)
plt.plot(fpr7, tpr7, linestyle='--',color='blue', label='XGBOOST (area = %0.2f)' %
auc_score7)
plt.plot(p_fpr, p_tpr, linestyle='--', color='grey')
# title
plt.title('ROC curve Males - 2005-06 Original Data')
# x label
plt.xlabel('False Positive Rate')
# y label
plt.ylabel('True Positive rate')
plt.legend(loc='best')
#plt.savefig('ROC',dpi=300)
plt.show();

```

```

precision1, recall1, thresholds1 = precision_recall_curve(M_Y_test_ready,
pred_prob1[:,1], pos_label=1)
precision2, recall2, thresholds2 = precision_recall_curve(M_Y_test_ready,
pred_prob2[:,1], pos_label=1)
precision3, recall3, thresholds3 = precision_recall_curve(M_Y_test_ready,
pred_prob3[:,1], pos_label=1)
precision4, recall4, thresholds4 = precision_recall_curve(M_Y_test_ready,
pred_prob4[:,1], pos_label=1)
#precision5, recall5, thresholds5 = precision_recall_curve(M_Y_test_ready,
pred_prob5[:,1], pos_label=1)
precision6, recall6, thresholds6 = precision_recall_curve(M_Y_test_ready,
pred_prob6[:,1], pos_label=1)
precision7, recall7, thresholds7 = precision_recall_curve(M_Y_test_ready,
pred_prob7[:,1], pos_label=1)

```

```

auc1 = auc(recall1, precision1)
auc2 = auc(recall2, precision2)
auc3 = auc(recall3, precision3)
auc4 = auc(recall4, precision4)
#auc5 = auc(recall5, precision5)
auc6 = auc(recall6, precision6)
auc7 = auc(recall7, precision7)

```

```

# matplotlib
import matplotlib.pyplot as plt
plt.style.use('seaborn')
# plot p-r curves

```

```

plt.plot(recall1, precision1, linestyle='--',color='orange', label='DT (area = %0.2f)' %
auc1)
plt.plot(recall2, precision2, linestyle='--',color='green', label='EN (area = %0.2f)' %
auc2)
plt.plot(recall3, precision3, linestyle='--',color='red', label='KNN (area = %0.2f)' %
auc3)
plt.plot(recall4, precision4, linestyle='--',color='purple', label='RFC (area = %0.2f)'
% auc4)
#plt.plot(recall5, precision5, linestyle='--',color='yellow', label='SVM (area = %0.2f)'
% auc5)
plt.plot(recall6, precision6, linestyle='--',color='black', label='LIGHTGBM (area =
%0.2f)' % auc6)
plt.plot(recall7, precision7, linestyle='--',color='blue', label='XGBOOST (area =
%0.2f)' % auc7)
#plt.plot(p_fpr, p_tpr, linestyle='--', color='grey')
# title
plt.title('Precision-Recall curves Males - 2005-06 Original Data')
# x label
plt.xlabel('Recall')
# y label
plt.ylabel('Precision')
plt.legend(loc='best')
#plt.savefig('ROC',dpi=300)
plt.show();

```

PAPER 3

Incidence Estimation

```

rm(list = ls())
library(foreign)
library(stats4)
library(stats)
library(gplots)
library(deSolve)
library(lattice)
setwd("/Users/rutendozivai/Desktop/data")

##READ DATA
dt<-read.csv("dt2.csv")
data1females<-read.csv("data2005females.csv",header=T)
data1males<-read.csv("data2005males.csv",header=T)
data2females<-read.csv("data2010females.csv",header=T)
data2males<-read.csv("data2010males.csv",header=T)
data3females<-read.csv("data2015females.csv",header=T)
data3males<-read.csv("data2015males.csv",header=T)
#ATTACH DATA
attach(dt)
attach(data1females)
attach(data2females)
attach(data3females)
attach(data1males)
attach(data2males)
attach(data3males)

tot2005females<-neg2005females+pos2005females
tot2010females<-neg2010females+pos2010females

```



```

tot2015females<-neg2015females+pos2015females
tot2005males<-neg2005males+pos2005males
tot2010males<-neg2010males+pos2010males
tot2015males<-neg2015males+pos2015males

### Farrington's Model ###
#####
a=15:49
farrington1females=function(alpha,beta,gamma)
{
  p=1-exp((alpha/beta)*a*exp(-beta*a)
    +(1/beta)*((alpha/beta)-gamma)*(exp(-beta*a)-1)-gamma*a)
  ll=pos2005females*log(p)+(tot2005females-pos2005females)*log(1-p)
  #alternative definition of the log-likelihood
  #ll=sum(log(dbinom(Pos,Tot,prob=p)))
  return(-sum(ll))
}
farrington1males=function(alpha,beta,gamma)
{
  p=1-exp((alpha/beta)*a*exp(-beta*a)
    +(1/beta)*((alpha/beta)-gamma)*(exp(-beta*a)-1)-gamma*a)
  ll=pos2005males*log(p)+(tot2005males-pos2005males)*log(1-p)
  #alternative definition of the log-likelihood
  #ll=sum(log(dbinom(Pos,Tot,prob=p)))
  return(-sum(ll))
}

farrington2females=function(alpha,beta,gamma)
{
  p=1-exp((alpha/beta)*a*exp(-beta*a)
    +(1/beta)*((alpha/beta)-gamma)*(exp(-beta*a)-1)-gamma*a)
  ll=pos2010females*log(p)+(tot2010females-pos2010females)*log(1-p)
  #alternative definition of the log-likelihood
  #ll=sum(log(dbinom(Pos,Tot,prob=p)))
  return(-sum(ll))
}
farrington2males=function(alpha,beta,gamma)
{
  p=1-exp((alpha/beta)*a*exp(-beta*a)
    +(1/beta)*((alpha/beta)-gamma)*(exp(-beta*a)-1)-gamma*a)
  ll=pos2010males*log(p)+(tot2010males-pos2010males)*log(1-p)
  #alternative definition of the log-likelihood
  #ll=sum(log(dbinom(Pos,Tot,prob=p)))
  return(-sum(ll))
}

farrington3females=function(alpha,beta,gamma)
{
  p=1-exp((alpha/beta)*a*exp(-beta*a)
    +(1/beta)*((alpha/beta)-gamma)*(exp(-beta*a)-1)-gamma*a)
  ll=pos2015females*log(p)+(tot2015females-pos2015females)*log(1-p)
  #alternative definition of the log-likelihood
  #ll=sum(log(dbinom(Pos,Tot,prob=p)))
  return(-sum(ll))
}
farrington3males=function(alpha,beta,gamma)
{
  p=1-exp((alpha/beta)*a*exp(-beta*a)
    +(1/beta)*((alpha/beta)-gamma)*(exp(-beta*a)-1)-gamma*a)
  ll=pos2015males*log(p)+(tot2015males-pos2015males)*log(1-p)
  #alternative definition of the log-likelihood
  #ll=sum(log(dbinom(Pos,Tot,prob=p)))
  return(-sum(ll))
}

### Farrington's model assuming gamma=0 ###
#####
model60females=mle(farrington1females,fixed=list(gamma=0),

```

```

        start=list(alpha=0.05,beta=0.1))
summary(model60females)
AIC(model60females)
model60males=mle(farrington1males,fixed=list(gamma=0),
                start=list(alpha=0.05,beta=0.1))
summary(model60males)
AIC(model60males)

model61females=mle(farrington2females,fixed=list(gamma=0),
                start=list(alpha=0.05,beta=0.11))
summary(model61females)
AIC(model61females)
model61males=mle(farrington2males,fixed=list(gamma=0),
                start=list(alpha=0.1,beta=0.1))
summary(model61males)
AIC(model61males)

model62females=mle(farrington3females,fixed=list(gamma=0),
                start=list(alpha=0.05,beta=0.15))
summary(model62females)
AIC(model62females)
model62males=mle(farrington3males,fixed=list(gamma=0),
                start=list(alpha=0.08,beta=0.2))
summary(model62males)
AIC(model62males)

# likelihood Based on Grenfell and Anderson, J Hyg Camb, 1985
# Differential Mortality on Gregson AIDS 1996

# Use lognormal functional form for force of infection lamda
lamda <- function(a, mu, sigma, theta)
{res=numeric(length(a))
for (i in 1:length(a))
{if (a[i]<=11) {res[i]=0} else {res[i]=theta*dlnorm(a[i]-12, meanlog = mu, sdlog = sigma, log =
FALSE)}}}
res}

# Integral Lamda of force of infection
Lamda<-function(a, mu, sigma, theta)
{res=numeric(length(a))
for (i in 1:length(a))
{if (a[i]<=11) {res[i]=0} else {res[i]=theta*plnorm(a[i]-12, meanlog = mu, sdlog = sigma, log =
FALSE)}}}
res
}

# Calculate F(a) = proportion of susceptibles having experienced infection
# WITHOUT WEIBULL mortality
#F <- function(a, mu, sigma, theta)
#1-exp(-Lamda(a,mu,sigma,theta))

# Survival function. Note that Gregsons 1996 terminology lamda corresponds to our beta & theta
to our alpha
survival <- function(x, alpha, beta)
exp(-(x/beta)^alpha)

death <- function(x, alpha,beta)
dweibull(x,shape=alpha,scale=beta)

sumint <- function(a,mu,sigma,theta,alpha,beta)
{res=numeric(a)
for (i in 1:a)
{res[i]=lamda(i,mu,sigma,theta)*exp(-Lamda(i,mu,sigma,theta))*survival(a-i,alpha,beta)
}
sum(res)
}

sumint_d <- function(a,mu,sigma,theta,alpha,beta)
{res=numeric(a)
for (i in 1:a)

```

```

{res[i]=lamda(i,mu,sigma,theta)*exp(-Lamda(i,mu,sigma,theta))*death(a-i,alpha,beta)
}
sum(res)
}

dead<- function(a,mu,sigma,theta,alpha,beta)
{res=numeric(length(a))
for (i in 1:length(a))
{res[i]=sumint_d(a[i], mu, sigma, theta, alpha, beta)}
res
}

infect<- function(a,mu,sigma,theta,alpha,beta)
{res=numeric(length(a))
for (i in 1:length(a))
{res[i]=sumint(a[i], mu, sigma, theta, alpha, beta)}
res
}

sus<-function(x,mu,sigma,theta)
exp(-Lamda(x,mu,sigma,theta))

# mortality at age a, mort(a)
mort <- function(x,mu,sigma,theta,alpha,beta)
dead(x,mu,sigma,theta,alpha,beta)/(infect(x,mu,sigma,theta,alpha,beta)+sus(x,mu,sigma,theta))

#Prevalence at age a F
F <- function(x,mu,sigma,theta,alpha,beta)

infect(x,mu,sigma,theta,alpha,beta)/(infect(x,mu,sigma,theta,alpha,beta)+sus(x,mu,sigma,theta))

# Tabulate number of negatives and positives by yearly age group

# Define the likelihood with lognormal foi WITH weibull mortality
mlogl <- function(mu,sigma,theta,alpha,beta) {
res=sum(-(log(F(a,mu,sigma,theta,alpha,beta))*Rjij+log(1-
F(a,mu,sigma,theta,alpha,beta))*(Njij-Rjij)))
res
}

a=15:49
Nj=table(dt$Age,dt$year,dt$Sex)
Rj=table(dt$Age[dt$hivpos==1],dt$year[dt$hivpos==1],dt$Sex[dt$hivpos==1])
p=Rj/Nj
results=vector("list",6)
results1=vector("list",6)
deviance=numeric(6)
for (j in 1:2){
for (i in 1:3){
Njij=Nj[,i,j]
Rjij=Rj[,i,j]
if (j==1) {alpha0=8} else {alpha0=8}
fit0<-mle(mlogl,start=list(mu=3.2,sigma=0.55,theta=1.52,beta=10.6),fixed=list(alpha=alpha0))
results[(j-1)*3+i]=fit0
results1[(j-1)*3+i]=AIC(fit0)
}
}
results
results1

mlmu=numeric(6)
mlsigma=numeric(6)
mltheta=numeric(6)
mlalpha=numeric(6)
mlbeta=numeric(6)
mlmu_se=numeric(6)
mlsigma_se=numeric(6)
mltheta_se=numeric(6)
mlalpha_se=numeric(6)
mlbeta_se=numeric(6)

# Now extract coefficients and their standard error from

```

```

# fit objects
for (k in 1:2)
{
  for (j in 1:3)
  {
    i=(k-1)*3+j
    m1mu[i]=coef(results[[i]])[[1]]
    m1sigma[i]=coef(results[[i]])[[2]]
    m1theta[i]=coef(results[[i]])[[3]]
    m1alpha[i]=coef(results[[i]])[[4]]
    m1beta[i]=coef(results[[i]])[[5]]

    m1mu_se[i]=sqrt(vcov(results[[i]])[[1,1]])
    m1sigma_se[i]=sqrt(vcov(results[[i]])[[2,2]])
    m1theta_se[i]=sqrt(vcov(results[[i]])[[3,3]])
    m1beta_se[i]=sqrt(vcov(results[[i]])[[4,4]])

  }
}

quartz(width=8, height=8)
par(mfrow=c(2,3))
#age2=seq(1.5,18.5, by=0.5)
age2=seq(1.5,35.5, by=1)
age2
###INCIDENCE PER YEAR 2005
##FEMALES
alpha0females<-coef(model60females)[1]
beta0females<-coef(model60females)[2]
gamma0females<-0
foi0females<-(alpha0females*age2-gamma0females)*exp(-beta0females*age2)+gamma0females

plot(a,lamda(a,m1mu[1],m1sigma[1],m1theta[1]-0.3),ylab="Force of
Infection",xlim=c(15,50),ylim=c(0,0.04), type="l",lty=1, xlab="")
lines(age,foi0females,lty=2)
abline(h=0)

title(main="(FEMALES 2005-06 ZDHS)")
legend("topright", legend=c("Catalytic", "Farrington"),
      lty=1:3, cex=0.5, col = c("black","black"))

###INCIDENCE PER YEAR 2010
##FEMALES
alpha1females<-coef(model61females)[1]
beta1females<-coef(model61females)[2]
gamma1females<-0
foi1females<-(alpha1females*age2-gamma1females)*exp(-beta1females*age2)+gamma1females

plot(a,lamda(a,m1mu[2],m1sigma[2],m1theta[2]-0.3),ylab="Force of
Infection",xlim=c(15,50),ylim=c(0,0.04), type="l",lty=1, xlab="")
lines(age,foi1females,lty=2)
abline(h=0)

title(main="(FEMALES 2010-11 ZDHS)")
legend("topright", legend=c("Catalytic", "Farrington"),
      lty=1:3, cex=0.5, col = c("black","black"))

###INCIDENCE PER YEAR 2015
##FEMALES
alpha2females<-coef(model62females)[1]
beta2females<-coef(model62females)[2]
gamma2females<-0
foi2females<-(alpha2females*age2-gamma2females)*exp(-beta2females*age2)+gamma2females

plot(a,lamda(a,m1mu[3],m1sigma[3],m1theta[3]-0.3),ylab="Force of
Infection",xlim=c(15,50),ylim=c(0,0.04), type="l",lty=1, xlab="")
lines(age,foi2females, lty=2)
abline(h=0)

title(main="(FEMALES 2015 ZDHS)")
legend("topright", legend=c("Catalytic", "Farrington"),
      lty=1:3, cex=0.5, col = c("black","black"))

```

```

age2=seq(1.5,18.5, by=0.5)
age2
##INCIDENCE PER YEAR 2005
##MALES
alpha0males<-coef(model60males)[1]
beta0males<-coef(model60males)[2]
gamma0males<-0
foi0males<-(alpha0males*age2-gamma0males)*exp(-beta0males*age2)+gamma0males

plot(a,lamda(a,m1mu[4],mlsigma[4],mltheta[4]-1.5), ylim=c(0,0.04), xlim=c(15,50), ylab="Force of
Infection", type="l", lty=1, xlab="")
lines(age,foi0males,lty=2)
abline(h=0)

title(main="(MALES 2005-06 ZDHS)")
legend("topright", legend=c("Catalytic", "Farrington"),
      lty=1:3, cex=0.5, col = c("black","black"))

###INCIDENCE PER YEAR 2010
##MALES
alpha1males<-coef(model61males)[1]
beta1males<-coef(model61males)[2]
gamma1males<-0
foi1males<-(alpha1males*age2-gamma1males)*exp(-beta1males*age2)+gamma1males

plot(a,lamda(a,m1mu[5],mlsigma[5],mltheta[5]-1.5),ylab="Force of
Infection",xlim=c(15,50),ylim=c(0,0.04), type="l",lty=1, xlab="")
lines(age,foi1males,lty=2)
abline(h=0)

title(main="(MALES 2010-11 ZDHS)")
legend("topright", legend=c("Catalytic", "Farrington"),
      lty=1:3, cex=0.5, col = c("black","black"))

###INCIDENCE PER YEAR 2015
##MALES
alpha2males<-coef(model62males)[1]
beta2males<-coef(model62males)[2]
gamma2males<-0
foi2males<-(alpha2males*age2-gamma2males)*exp(-beta2males*age2)+gamma2females

plot(a,lamda(a,m1mu[6],mlsigma[6],mltheta[6]-1.5),ylab="Force of
Infection",xlim=c(15,50),ylim=c(0,0.04), type="l",lty=1, xlab="")
lines(age,foi2males, lty=2)
abline(h=0)

title(main="(MALES 2015 ZDHS)")
legend("topright", legend=c("Catalytic", "Farrington"),
      lty=1:3, cex=0.5, col = c("black","black"))

```

Blandr Plots

```

library(blandr)
library(ggplot2)
setwd("/Users/rutendozivai/Desktop/FOI paper/DATA")
data<-read.csv("incidencedata.csv")
attach(data)
blandr.statistics (cohort2005females , foi0females , sig.level=0.95 )
stats.data1females <- blandr.statistics (cohort2005females , foi0females , sig.level=0.95 )
summary(stats.data1females)
stats.data1females
quartz(width=8, height=8)
par(mfrow=c(2,2))
blandr.output.text (cohort2005females , foi0females , sig.level=0.95 )
blandr.draw(cohort2005females , foi0females)
blandr.draw(cohort2005females , foi0females , ciDisplay = FALSE , ciShading = FALSE )
library(ggpubr)
stats.cat0females <- blandr.statistics (cohort2005females , cat0females , sig.level=0.95 )
stats.cat1females <- blandr.statistics (cohort2010females , cat1females , sig.level=0.95 )

```

```

stats.cat0males <- blandr.statistics (cohort2005males , cat0males , sig.level=0.95 )
stats.cat1males <- blandr.statistics (cohort2010males , cat1males , sig.level=0.95 )
stats.foi0females <- blandr.statistics (cohort2005females , foi0females , sig.level=0.95 )
stats.foi1females <- blandr.statistics (cohort2010females , foi1females , sig.level=0.95 )
stats.foi0males <- blandr.statistics (cohort2005males , foi0males , sig.level=0.95 )
stats.foi1males <- blandr.statistics (cohort2010males , foi1males , sig.level=0.95 )

a=blandr.plot.ggplot(stats.cat0females, plotTitle = "Catalytic 2005-06 females", ciShading =
FALSE )
b=blandr.plot.ggplot(stats.cat1females, plotTitle = "Catalytic 2010-11 females", ciShading =
FALSE )
c=blandr.plot.ggplot(stats.cat0males, plotTitle = "Catalytic 2005-06 males", ciShading = FALSE )
d=blandr.plot.ggplot(stats.cat1males, plotTitle = "Catalytic 2010-11 males", ciShading = FALSE )

e=blandr.plot.ggplot(stats.foi0females, plotTitle = "Farrington 2005-06 females", ciShading =
FALSE )
f=blandr.plot.ggplot(stats.foi1females, plotTitle = "Farrington 2010-11 females", ciShading =
FALSE )
g=blandr.plot.ggplot(stats.foi0males, plotTitle = "Farrington 2005-06 males", ciShading = FALSE
)
h=blandr.plot.ggplot(stats.foi1males, plotTitle = "Farrington 2010-11 males", ciShading = FALSE
)

quartz(width=15, height=15)
par(mfrow=c(2,4))

ggarrange(a,b,c,d,e,f,g,h, ncol = 4, nrow = 2)

```

PAPER 4

###Age-Time analysis

####FIGURE 1

setwd("/Users/rutendozivai/Desktop/paper 2 data")

DEFINE FUNCTIONS

library(mgcv)

```

pavit<- function(pos=pos,tot=rep(1,length(pos)))
{
  gi<- pos/tot
  pai1 <- pai2 <- gi
  N <- length(pai1)
  ni<-tot
  for(i in 1:(N - 1)) {
    if(pai2[i] > pai2[i + 1]) {
      pool <- (ni[i]*pai1[i] + ni[i+1]*pai1[i + 1])/(ni[i]+ni[i+1])
      pai2[i:(i + 1)] <- pool
      k <- i + 1
      for(j in (k - 1):1) {
        if(pai2[j] > pai2[k]) {
          pool.2 <- sum(ni[j:k]*pai1[j:k])/(sum(ni[j:k]))
          pai2[j:k] <- pool.2
        }
      }
    }
  }
  return(list(pai1=pai1,pai2=pai2))
}

```

foi.num<-function(x,p)

```

{
  grid<-sort(unique(x))
  pgrid<-(p[order(x)])[duplicated(sort(x))=F]
  dp<-diff(pgrid)/diff(grid)
  foi<-approx((grid[-1]+grid[-length(grid)])/2,dp,grid[c(-1,-length(grid))])$y/(1-pgrid[c(-1,-
length(grid))])
}

```

```

    return(list(grid=grid[c(-1,-length(grid))],foi=foi))
  }
#-----
data<-read.csv("/Users/rutendozivai/Desktop/paper 2 data/females.csv",header=T)
attach(data)
y<-status
a<-age
x<-birth_cohort
ts=year
# Seroprevalence plot
quartz(width=15, height=5)
par(mfrow=c(2,3),lwd=3,las=1,cex.axis=1.1,cex.lab=1.1,lwd=3,mgp=c(2,
0),mar=c(4.1,4.1,4.1,3))

resp1<-y[ts==2005]
age1<-a[ts==2005]
pos1<-table(age1,resp1)[,2]
neg1<-table(age1,resp1)[,1]
tot1<-neg1+pos1
plot(unique(age1),pos1/tot1,cex=0.002*tot1,ylab="prevalence",xlab="age",xlim=c(14,50))

resp2<-y[ts==2010]
age2<-a[ts==2010]
pos2<-table(age2,resp2)[,2]
neg2<-table(age2,resp2)[,1]
tot2<-neg2+pos2
plot(unique(age2),pos2/tot2,cex=0.002*tot2,ylab="prevalence",xlab="age",xlim=c(14,50))

resp3<-y[ts==2015]
age3<-a[ts==2015]
pos3<-table(age2,resp2)[,2]
neg3<-table(age2,resp2)[,1]
tot3<-neg2+pos2
plot(unique(age2),pos2/tot2,cex=0.002*tot2,ylab="prevalence",xlab="age",xlim=c(14,50))

##### MODEL SELECTION FIGURE 2

# Analyses
#-----
data<-read.csv("/Users/rutendozivai/Desktop/paper 2 data/chapter13datafemales.csv",header=T)
attach(data)
y<-status
a<-age
x<-birth_cohort
ts=year

x=x-min(x)
xf=as.factor(x)

#####PARAMETRICS MODELS#####

gamfit1=gam(y~a+x,family=binomial(link="cloglog"))
c(sum(gamfit1$edf),AIC(gamfit1),AIC(gamfit1, k = log(length(y))))

#####SEMI-PARAMETRICS MODELS#####

gamfit2=gam(y~s(a)+x,family=binomial(link="cloglog"))
c(sum(gamfit2$edf),AIC(gamfit2),AIC(gamfit2, k = log(length(y))))

gamfit3=gam(y~a+s(x),family=binomial(link="cloglog"))
c(sum(gamfit3$edf),AIC(gamfit3),AIC(gamfit3, k = log(length(y))))

#####NON-PARAMETRICS MODELS#####

gamfit4=gam(y~s(a)+s(x),family=binomial(link="cloglog"))
c(sum(gamfit4$edf),AIC(gamfit4),AIC(gamfit4, k = log(length(y))))

#####NON-PROPORTIONAL HAZARDS MODELS#####

```

```

gamfit5=gam(y~s(a,x),family=binomial(link="cloglog"))
c(sum(gamfit5$edf),AIC(gamfit5),AIC(gamfit5, k = log(length(y))))

# best fitting model

gamfit4=gam(y~s(a)+s(x),family=binomial(link="cloglog"))
c(sum(gamfit4$edf),AIC(gamfit4),AIC(gamfit4, k = log(length(y))))

# Testing for the proportional hazards assumption
anova(gamfit4,gamfit5,test="Chisq")

##### BEST MODEL - FUGURE 2

##### FIGURE 3 - COHORT SPECIFIC PREVALENCE

resp1<-y[ts==2005]
age1<-a[ts==2005]
pos1<-table(age1,resp1)[,2]
neg1<-table(age1,resp1)[,1]
tot1<-neg1+pos1

resp2<-y[ts==2010]
age2<-a[ts==2010]
pos2<-table(age2,resp2)[,2]
neg2<-table(age2,resp2)[,1]
tot2<-neg2+pos2

resp3<-y[ts==2015]
age3<-a[ts==2015]
pos3<-table(age2,resp2)[,2]
neg3<-table(age2,resp2)[,1]
tot3<-neg2+pos2

#####NON-PARAMETRICS MODELS#####
gamfit4=gam(y~s(a)+s(x),family=binomial(link="cloglog"))

quartz(width=10, height=5)
par(mfrow=c(1,2),lwd=3,las=1,cex.axis=1.1,cex.lab=1.1,lwd=3,mgp=c(2,
0.5,
0),mar=c(4.1,4.1,4.1,3))
agrid=seq(range(unique(a))[1],range(unique(a))[2],length.out=100)

x1<-age1
pos1<-table(age1,resp1)[,2]
neg1<-table(age1,resp1)[,1]
tot1<-neg1+pos1
plot(unique(x1),pos1/tot1,cex=0.002*tot1,ylab="prevalence",xlab="age",xlim=c(14,50),type="n")
points(unique(x1),pos1/tot1,cex=0.002*tot1,ylab="prevalence",xlab="age",xlim=c(14,50))

x2<-age2
pos2<-table(age2,resp2)[,2]
neg2<-table(age2,resp2)[,1]
tot2<-neg2+pos2
points(unique(x2),pos2/tot2,cex=0.002*tot2,ylab="prevalence",xlab="age",xlim=c(14,50),col="darkgrey")

x3<-age3
pos3<-table(age3,resp3)[,2]
neg3<-table(age3,resp3)[,1]
tot3<-neg3+pos3
points(unique(x3),pos3/tot3,cex=0.002*tot3,ylab="prevalence",xlab="age",xlim=c(14,50),col="blue")
)

gamfitplot=gamfit4

for (xsel in c(10,20,30,40,50)){
  agel=2005-1956-xsel
  ager=2010-1956-xsel
  agesel=(agrid>=agel)&(agrid<=ager)
  newd <- data.frame(a=agrid,x=xsel)

```



```

ypred=predict.gam(gamfitplot,newdata=newd,type="terms")
con=ypred[1,2]+attributes(ypred)$constant
s1=ypred[,1]+con
pi0=1-exp(-exp(s1))
pi0m=pavit(1-exp(-exp(s1)))$pai1
lines(agrid[agesel],pi0m[agesel],lwd=2, col="red")
lines(agrid[!agesel],pi0m[!agesel],lty=3)
}
for (xsel in c(10,20,30,40,50)){
  agel=2010-1956-xsel
  ager=2015-1956-xsel
  agesel=(agrid>=agel)&(agrid<=ager)
  newd <- data.frame(a=agrid,x=xsel)
  ypred=predict.gam(gamfitplot,newdata=newd,type="terms")
  con=ypred[1,2]+attributes(ypred)$constant
  s1=ypred[,1]+con
  pi0=1-exp(-exp(s1))
  pi0m=pavit(1-exp(-exp(s1)))$pai1
  lines(agrid[agesel],pi0m[agesel],lwd=2, col="red")
  lines(agrid[!agesel],pi0m[!agesel],lty=3)
}

legend("topleft",legend=c("Prevalence from data","Prevalence extrapolated"),lty=c(1,3),
col=c("red","black"),cex=.5)
legend("bottomright",legend=c("Observed prevalence 2005-06", "Observed prevalence 2010-11",
"Observed prevalence 2015"), pch=c(1,1,1), col=c("black", "grey", "blue"),cex=.5)

##### FIGURE 4 - COHORT SPECIFIC FOI

#####NON-PARAMETRICS MODELS#####
gamfit4=gam(y~s(a)+s(x), family=binomial(link="cloglog"))

agrid=seq(range(unique(a))[1],range(unique(a))[2],length.out=85)

age_x=5:87
#age_x2=15:111
#age_x3=15:110

gamfitplot=gamfit4
quartz(width=10, height=5)
par(mfrow=c(1,2),lwd=3,las=1,cex.axis=1.1,cex.lab=1.1,lwd=3,mgp=c(3, 0.5, 0))
plot(c(min(agrid),max(agrid),min(agrid),max(agrid)),c(0,0.1,0,0.1),type="n",xlab="age",ylab="for
ce of infection",ylim=c(0,0.04), xlim=c(15,49))
for (xsel in seq(0,50,by=10)){
  agel=2005-1955-xsel
  ager=2010-1955-xsel
  agesel=(agrid>=agel)&(agrid<=ager)
  newd <- data.frame(a=agrid,x=xsel)
  ypred=predict.gam(gamfitplot,newdata=newd,type="terms")
  con=ypred[1,2]+attributes(ypred)$constant
  s1=ypred[,1]+con
  pi0=1-exp(-exp(s1))
  pi0m=pavit(1-exp(-exp(s1)))$pai1
  agridfoi=foi.num(agrid,pi0m)$grid
  ypredfoi=foi.num(agrid,pi0m)$foi
  if (xsel==14) lines(age_x,ypredfoi,col="grey",lwd=3)
  if (xsel!=14) {
    lines(age_x[agesel],ypredfoi[agesel],lwd=2, col="red")
    lines(age_x[!agesel],ypredfoi[!agesel],lty=3)
  }
}
for (xsel in seq(0,50,by=10)){
  agel=2010-1955-xsel
  ager=2015-1955-xsel
  agesel=(agrid>=agel)&(agrid<=ager)
  newd <- data.frame(a=agrid,x=xsel)
  ypred=predict.gam(gamfitplot,newdata=newd,type="terms")
  con=ypred[1,2]+attributes(ypred)$constant
  s1=ypred[,1]+con
  pi0=1-exp(-exp(s1))

```

```

pi0m=pavit(1-exp(-exp(s1)))$pai1
agridfoi=foi.num(agrid,pi0m)$grid
ypredfoi=foi.num(agrid,pi0m)$foi
if (xsel==0) lines(age_x,ypredfoi,col="grey",lwd=3)
if (xsel!=0) {
  lines(age_x[agesel],ypredfoi[agesel],lwd=2, col="red")
  lines(age_x[!agesel],ypredfoi[!agesel],lty=3)
}
}

legend("topright",legend=c("FOI from data","FOI extrapolated", "FOI from baseline birth-
cohort"),lty=c(1,3,1), col=c("red","black", "grey"),cex=.5)

```

Appendix 3: Ethics Certificate



R14/49 Ms Birri Rutendo Beauty

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M151154

NAME: Ms Birri Rutendo Beauty
(Principal Investigator)

DEPARTMENT: Epidemiology and Biostatistics
University of the Witwatersrand

PROJECT TITLE: Bayesian Modeling of HIV Incidence in Zimbabwe:
A Locally Adapted Estimation Projection Package
Approach

DATE CONSIDERED: 27/11/2015

DECISION: Approved unconditionally

CONDITIONS:

SUPERVISOR: Dr Eustasius Musenge

APPROVED BY: 
Professor P Cleaton-Jones, Chairperson, HREC (Medical)

DATE OF APPROVAL: 30/11/2015

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Secretary in Room 10004, 10th floor, Senate House, University.
I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report.**

Principal Investigator Signature _____

Date _____

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 4: Published papers

The following are the papers for Chapters 3, 4 and 6 as published; as such page numbers and other detail are according to the specific journals.



Factors Associated With HIV Infection in Zimbabwe Over a Decade From 2005 to 2015: An Interval-Censoring Survival Analysis Approach

Rutendo Birri Makota* and Eustasius Musenge

Division of Epidemiology and Biostatistics, Faculty of Health Sciences, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa

OPEN ACCESS

Edited by:

Michael Johnson Mahande,
Kilimanjaro Christian Medical
University College, Tanzania

Reviewed by:

Jenny Renju,
University of London, United Kingdom
Henry Mwambi,
University of KwaZulu-Natal,
South Africa

*Correspondence:

Rutendo Birri Makota
rutendobirri@gmail.com

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 22 January 2019

Accepted: 27 August 2019

Published: 18 September 2019

Citation:

Birri Makota R and Musenge E (2019)
Factors Associated With HIV Infection
in Zimbabwe Over a Decade From
2005 to 2015: An Interval-Censoring
Survival Analysis Approach.
Front. Public Health 7:262.
doi: 10.3389/fpubh.2019.00262

Objectives: The main objective of this study was to compare results from two approaches for estimating the effect of different factors on the risk of HIV infection and determine the best fitting model.

Study design: We performed secondary data analysis on cross-sectional data which was collected from the Zimbabwe Demographic Health Survey (ZDHS) from 2005 to 2015.

Methods: Survey and cluster adjusted logistic regression was 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 10 years. The data for the survival analysis were modeled assuming age at survey imputation (Model 1) and interval-censoring (Model 2).

Results: Model goodness of fit test based on the Cox-Snell residuals against the cumulative hazard indicated that Model 1 was the best model. On the contrary, the Akaike Information Criterion (AIC) indicated that Model 2 was the best model. Factors associated with a high risk of HIV infection were being female, number of sexual partners, and having had an STI in the past year prior to the survey.

Conclusion: The difference between the results from the Cox-Snell residuals graphical method and the model estimates and AIC value maybe due to the lack of adequate 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 literature. Even though we consider the interval-censoring model as the superior model with regards to our specific data, the method had its own set of limitations.

Keywords: interval-censoring, HIV, survival, prevalence, Zimbabwe

INTRODUCTION

The 90–90–90 targets was 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 (1).

In Zimbabwe, a population based survey carried out in 2016 reported that 74.2% of people living with HIV (PLHIV) aged 15–64 years knew their HIV status. Amongst the PLHIV who knew their status, 86.8% self-reported current use of Antiretroviral treatment (ART), with 86.5% of those who self-reported, are virally suppressed (2). In order for these 90–90–90 targets to be met, prevalence, and incidence rates estimates are crucial in understanding the current status of the HIV epidemic and determine whether the trends are improving in a manner which can facilitate to achieve the 2020 target.

The gold standard for estimating the HIV incidence is to test uninfected individuals for new infections periodically, however this method is feasible though costly and time-consuming. In addition, even if an HIV negative cohort is followed over-time, the exact date of infection is rarely observed (3). In this scenario, an interval can be determined between the latest negative and the earliest positive test dates. Taking into consideration the issue of cost and time, cohort analysis for estimating the HIV incidence rate in a general epidemic will not produce estimates which are representative of the whole population. Due to these reasons, sentinel surveillance systems have been set up to monitor the spread of the pandemic (4). In addition, population-based

surveys, in which HIV tests are performed, are carried after every 5-years in Zimbabwe. The advantage of the population-based survey is that, data is more representative of the population than a cohort. On the other hand, the same data does not provide the exact date of the infection but rather provide what is called “current status data.”

Current status data occurs when an individual is observed at one single point, and the only information obtained is whether the event of interest has occurred (5). An example of current status data includes information collected during a demographic health survey, in which an individual is tested whether they are HIV positive or negative. If an individual were found to be HIV positive, the individual was recorded as left censored at the time the test was done. If an individual were HIV negative, they were recorded to be right censored. Sometimes current status data can be referred to as case interval-censored data, with case II interval-censored data referred to as the general case (6). Interval censoring takes into account the range, that is, an interval inside of which one can say the outcome of interest has occurred (7). Given that we would want to determine the factors associated with the hazard of infection using data from these surveys, then survival analysis can be implemented.

In the setting of standard survival analysis, modeling the hazard rate of HIV infection can be achieved by imputing the time-to-onset of disease as the time at diagnostic visit (3). Modeling current status data using the mentioned two approaches may overestimate the hazard rate. However, analyzing this type of data using interval censoring will be a better approach. Although they are documented advantages

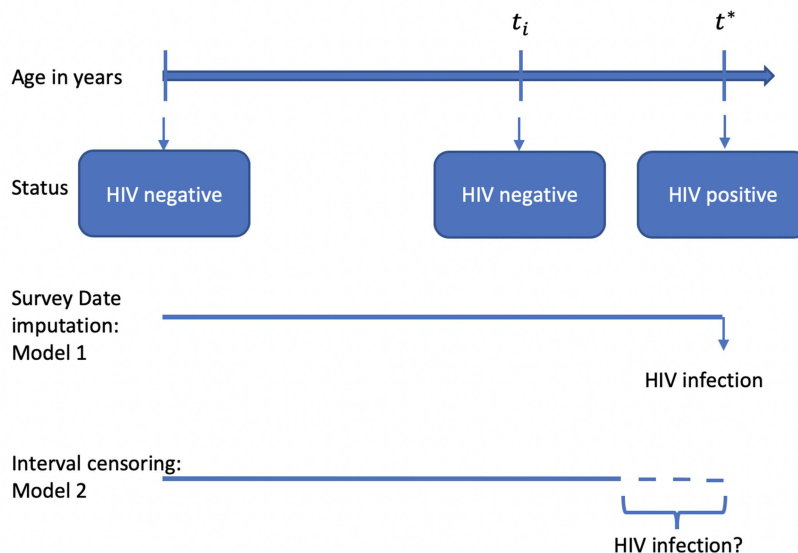


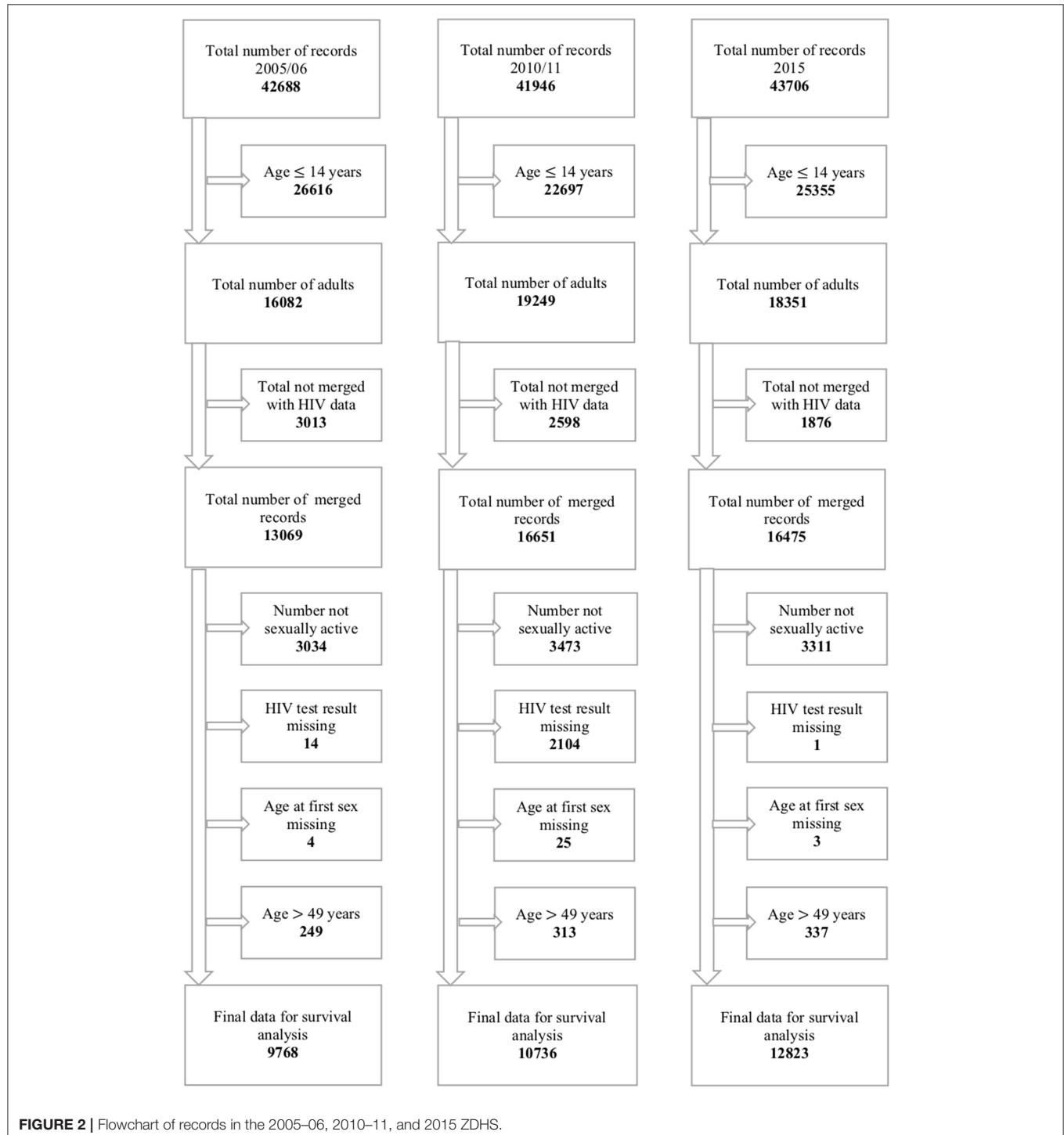
FIGURE 1 | Illustration of different modeling strategies when investigating the factors associated with HIV infection.

of interval censoring compared to the standard cause-specific survival model, the superiority of the interval-censoring remains unclear in estimating the effect of different exposures on the risk of HIV infection. With this argument in mind, the main objective of this study was to compare results from these different approaches for estimating the effect of different factors on the risk of HIV infection and determine the best fitting model.

MATERIALS AND METHODS

Study Design and Area

This study used data from Zimbabwe, a landlocked country, bordered by Mozambique on the East, South Africa on the South, Botswana on the West, and Zambia on the North and Northwest. Zimbabwe is sub-divided into 10 Provinces which are: Matabeleland South, Matabeleland North, Mashonaland



East, Mashonaland Central, Mashonaland West, Midlands, Masvingo, Manicaland, Harare, and Bulawayo. Each province is subdivided into districts, and each district is made up of wards. The designs for the three 5 yearly surveys were cross-sectional.

Zimbabwe Demographic Health Survey Data and Sampling

Demographic and Health Surveys (DHS) are nationally-representative household surveys that have been implemented in approximately 70 countries since 1984 (8–10). They provide data for a wide range of monitoring and impact evaluation in

the areas of population, health and nutrition. Data used for the analysis were obtained from 2005–06, 2010–11, and 2015 ZDHS and were retrieved from the DHS programme website (<https://dhsprogram.com>) (11). A representative probability sample of 10,800, 10,828, and 11,196 households were selected for the 2005–06, 2010–11 and 2015 ZDHS, respectively. A two-stage cluster sampling technique was used to select the households. The first stage selected 400, 406, and 400 enumeration areas (EAs) for 2005–06, 2010–11, and 2015 ZDHS, respectively. At the second stage, using a complete listing of households in the selected EAs, a fixed number of households were randomly chosen. This

TABLE 1 | HIV prevalence in Zimbabwe and changes in HIV prevalence (weighted) from the Zimbabwe DHS surveys 2005–06, 2010–11, and 2015.

Survey year	2005/6	2010/11	2015	p-value
Characteristic	Prevalence (CI)	Prevalence (CI)	Prevalence (CI)	
National summary	22.4 (21.6, 23.2)	19.6 (18.8, 20.3)	17.7 (17.0, 18.4)	<0.001
Sex				
Male	18.2 (16.7, 19.4)	16.2 (15.1, 17.3)	13.6 (12.7, 14.5)	<0.001
Female	25.1 (24.0, 26.2)	21.8 (20.8, 22.8)	20.7 (19.8, 21.6)	<0.001
Marital status				
Never married	10.8 (9.36, 12.3)	9.66 (8.29, 11.0)	7.86 (6.78, 8.95)	0.001
Married/cohabiting	20.5 (19.5, 21.5)	17.8 (16.9, 18.7)	16.8 (16.0, 17.6)	<0.001
Separated/divorced/widowed	45.6 (43.0, 48.2)	42.2 (39.6, 44.8)	38.3 (35.9, 40.8)	<0.001
Place of residence				
Urban	23.7 (22.2, 25.3)	22.0 (20.5, 23.4)	18.3 (17.3, 19.3)	<0.001
Rural	21.8 (20.8, 22.8)	18.5 (17.7, 19.4)	17.3 (16.4, 18.1)	<0.001
Education level				
No education/primary	22.6 (21.3, 23.9)	21.1 (19.7, 22.4)	20.7 (19.3, 22.1)	0.05
Secondary	22.6 (21.6, 23.7)	19.4 (18.5, 20.3)	17.4 (16.5, 18.2)	<0.001
Higher	16.3 (12.6, 20.0)	12.6 (9.86, 15.4)	12.1 (10.3, 13.9)	0.05
STI treatment in past 12 months				
No	21.4 (20.7, 22.2)	18.9 (18.2, 19.7)	17.2 (16.5, 17.8)	<0.001
Yes	46.2 (41.2, 51.2)	39.0 (34.0, 44.0)	37.2 (32.0, 42.3)	0.01
No. of sexual partners				
0	31.3 (28.9, 33.6)	28.2 (25.9, 30.5)	24.5 (22.2, 26.9)	<0.001
1	20.8 (19.9, 21.7)	18.2 (17.4, 19.1)	17.0 (16.2, 17.7)	<0.001
2+	19.5 (16.1, 22.9)	17.7 (14.8, 20.7)	15.3 (13.2, 17.4)	0.03
Currently employed				
No	22.7 (21.5, 23.9)	18.9 (17.9, 20.0)	17.9 (16.9, 19.0)	<0.001
Yes	22.1 (21.0, 23.2)	20.2 (19.2, 21.3)	17.5 (16.7, 18.4)	<0.001

*p-values below 0.05 are considered statistically significant.

TABLE 2 | Akaike information criterion and bayesian information criterion values.

Year	Model	Obs.	ll(null)	ll(model)	df	AIC	BIC
2005–06	1	9768	–2995.77	–2605.36	14	5238.7	5339.3
	2	9768	–2177.37	–1912.55	14	3853.09	3953.71
2010–11	1	10734	–2767.33	–2516.62	14	5061.23	5163.17
	2	10734	–2080.36	–1777.98	14	3583.95	3685.89
2015	1	12822	–2887.20	–2580.70	14	5189.41	5293.83
	2	12822	–2141.63	–1759.41	14	3546.83	3651.25

allowed the use of EAs specific weights to be assigned in the design (8–10).

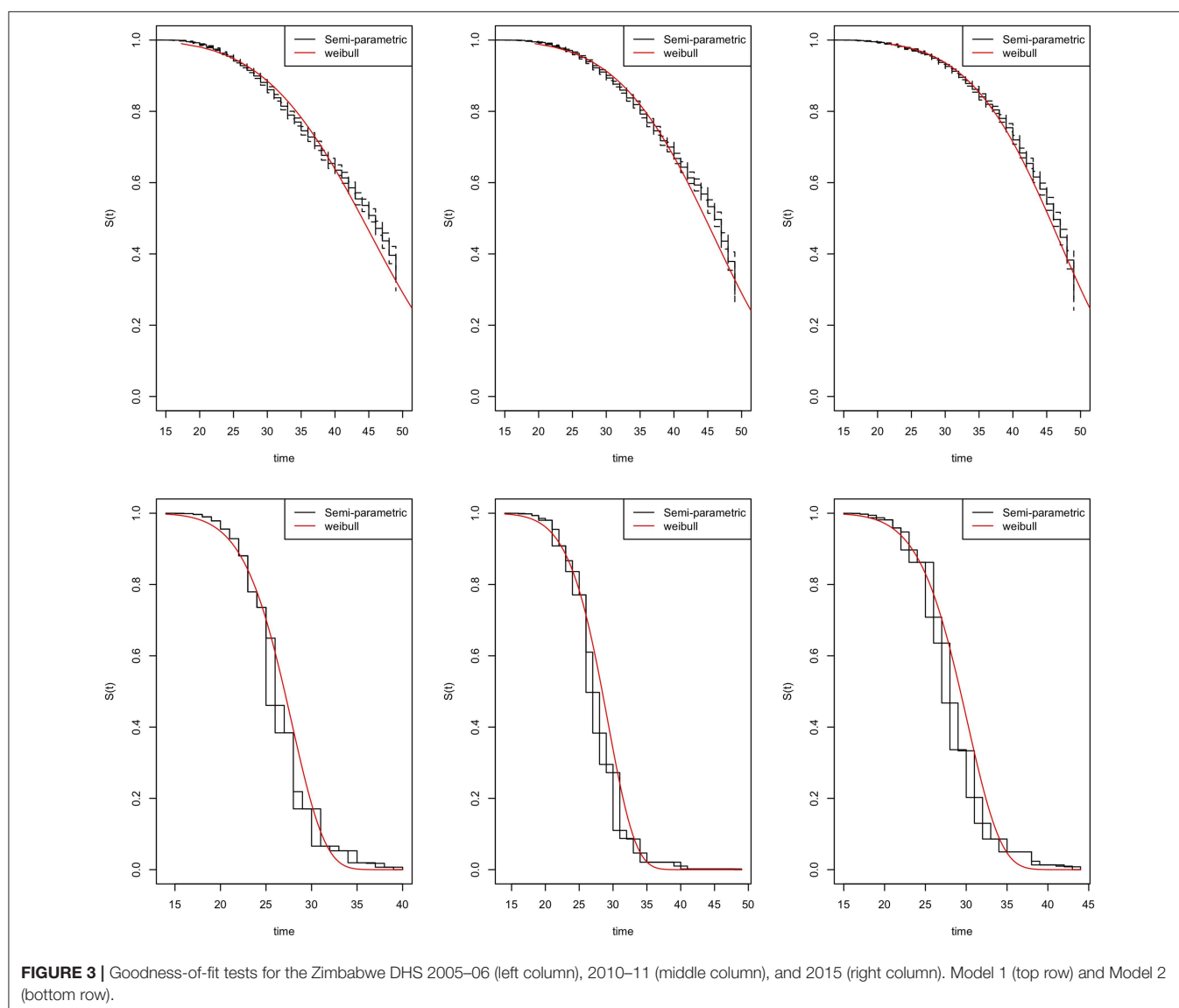
Measurement of the Outcome (HIV Status) and Explanatory Variables

With consent from the respondent or parent/guardian (for minors), blood samples were collected in all households for HIV testing in the laboratory for females aged 0–49 and males aged 0–54. Blood spots were collected on filter paper from a finger prick and transported to a laboratory for testing. An initial ELISA test was performed, and then retesting of all positive and 5–10 % of the negative tests with a second ELISA was done. If they were discordant results on the two ELISA tests, a new ELISA or a Western Blot was performed. The data used for this study was obtained from the DHS Data Archives (11) and only included individuals aged 15–49 years. The following explanatory variables were extracted from the dataset: sex, marital status,

education level, religion, currently employed, place of residence, STI treatment in the past 12 months and number of sexual partners (8–10).

Ethical Considerations

The ZDHS HIV testing protocol for all the three surveys was reviewed and approved by the ethical review boards Medical Research Council of Zimbabwe (MRCZ) in Harare, Zimbabwe; the ORC Macro Institutional Review Board in Calverton, Maryland, USA; and the Centers for Diseases Control (CDC) in Atlanta, Georgia, USA (8–10). This work was granted ethical clearance by the University of Witwatersrand's Human Research Ethics Committee (Medical) (No. M151154). The dataset used in this study was obtained through an application made to Measure DHS program, which was approved on the 16th of May 2017. The DHS Program is authorized to distribute, at no cost, unrestricted



survey data files for legitimate academic research. Registration was required for access to data.

Statistical Methodology

Application to Zimbabwe Demographic Health Survey

The socio-demographic datasets for men and women records were appended to provide a single analysis dataset for all the three surveys. The appended dataset was then merged using the unique combination of the individual line number, household line number, and cluster (EA) number to the HIV prevalence dataset. All individuals without an HIV test result, never been sexually active, and individuals who did not have an age at first intercourse were excluded from the analysis. In the case of individuals who were HIV positive when the survey was conducted, the age at HIV infection was defined as age at survey date and for individuals who were HIV negative, the age at HIV infection was right-censored at the date of survey in Models 1 (**Figure 1**). Accounting for interval-censoring, the age at HIV infection was interval-censored between age at first sexual intercourse and age at date of survey, but right-censored at the age at survey, for individuals who were HIV negative for Model 2. All the models assumed a parametric Weibull distribution for the baseline hazard $\lambda_0(t)$ which allowed estimation of β which is

the vector of regression coefficient. Model Specification for all the models, refer to **Supplementary Data**.

Statistical Analysis

In this study, the outcome or response variable was the HIV status, a binary variable. The study investigated the socio-cultural, socio-economic, behavioral, and demographic factors, which are associated with HIV. Trends in HIV prevalence were assessed using the non-parametric trend test in STATA. A stepwise logistic regression approach was adopted in STATA SE version 15.1 statistical software (12) using the command `svy: swaic` (13) on some selected explanatory variables highlighted earlier. Factors that were significantly associated with HIV from the stepwise survey logistic regression were then considered for the parametric survival analysis. The most suitable baseline hazard function was investigated using the package `icenReg` (14) in R software. The data were modeled assuming age at survey imputation and interval-censoring (**Figure 1**). The model goodness-of-fit (GOF) test was assessed using the Akaike Information Criterion (AIC) and two graphical methods which included the Cox-Snell residuals. All analysis were performed in STATA SE version 15.1 and R statistical software.

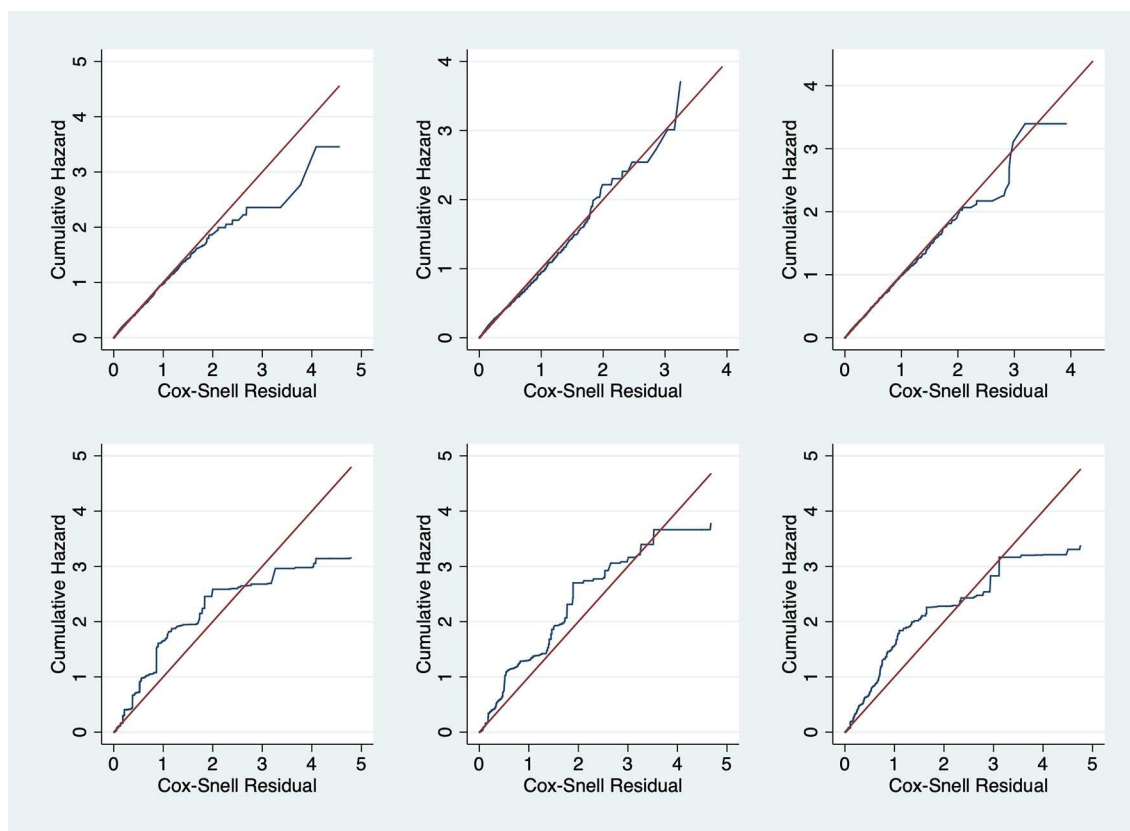


FIGURE 4 | Cox-Snell residual against cumulative hazard goodness-of-fit tests for the Zimbabwe DHS 2005–06 (left column), 2010–11 (middle column) and 2015 (right column). Model 1 (top row) and Model 2 (bottom row).

RESULTS

The 2005–06 ZDHS database had 10,800 households in which 42,698 records were retrieved. Of the 42,698 records, 16,082 records were for individuals aged 15–49 years with 7,175 (44.6%) males and 8,097 (55.4%) females. The 2010–11 ZDHS database had 10,828 households in which 41,946 records were retrieved. Of the 41,946 records, 16,651 records were for individuals aged 15–49 years with 7,480 (44.9%) males, and 9,171 (55.1%) females. The 2015 ZDHS database had 11,196 households in which 43,706 records were retrieved. Of the 43,706 records, 18,351 records were for individuals aged 15–49 years with 8,396 (45.8%) males and 9,955 (54.2%) females. The above information is represented in **Figure 2**.

Nationally, the non-parametric trend test ($p < 0.001$) showed a significant decline of HIV prevalence from 22.4 to 19.6 to 17.7% for 2005–06, 2010–11, and 2015 ZDHS, respectively (see **Table 1**). A similar decline trend was observed for gender, marital status, place of residence, education level, current employment status, STI in the past year preceding the survey and number of sexual partners.

The mean survival time for age at HIV infection for Model 1 was 41.1 years for females, 42.8 years for males in 2005/06 ZDHS; 41.9 years for females, 43.4 years for males in 2010/11 ZDHS and 42 years for females, 44 years for males in 2015 ZDHS. The mean survival time for age at HIV infection for Model 2 was 24.9 years for females, 28.3 years for males in 2005/06 ZDHS; 25.7 years for females, 29.4 years for males in 2010/11 ZDHS and 26.5 years for females, 30.5 years for males in 2015 ZDHS. According to the

survival times, Model 2 produced lower times of survival before HIV infection.

The parametric Weibull distribution was used to investigate the factors associated with HIV infection. The Weibull parametric function was suitable to model the baseline hazard distribution, as shown in **Figure 4**. Model goodness of fit test indicated that Model 2 was the best model based on the Akaike Information Criterion (AIC) presented in **Table 2**. A graphical goodness of fit test was performed, where the semi-parametric model was compared to the parametric model. According to the results in **Figure 3**, Model 1 fits the data better than Model 1 and Model 2. Based on **Figure 3**, Model 1 overestimates the survival rates between 15 and 35 years and underestimates the survival rates between 35 and 49 years. However, **Figure 4** with the Cox-Snell residuals shows that Model 1 is better than Model 2 as the estimated cumulative hazards are close to the reference line which is formed by the Cox-Snell residuals. Furthermore, dot charts depicting the importance of variables in the three models were plotted (**Figure 5**). Place of residence was the least important variable for all the three models, while marital status and sex where the most important variables for Model 1 and Model, respectively (**Figure 5**).

The risk of HIV infection was lower in males than females, as shown in **Table 3** for all the models. The risk for HIV infection had a slight decrease from 2005–06 to 2010–11 to 2015 (HR=0.26, 95% CI: 0.23, 0.31), (HR=0.25, 95% CI:0.21,0.29), (HR=0.22, 95% CI:0.19,0.26), respectively with reference to Model 2. Individuals who were married or cohabiting had a lower risk of HIV infection as compared to those who were single. These results were consistent for all three models. However,

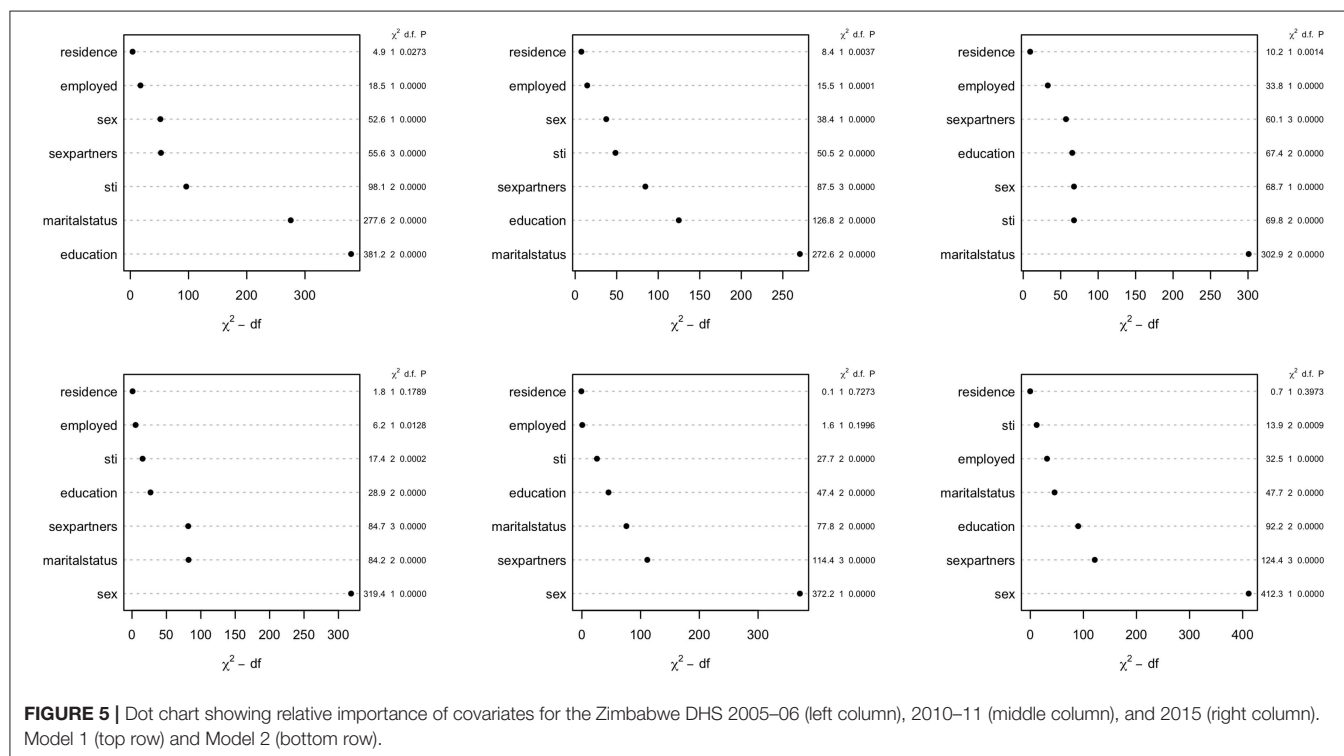


TABLE 3 | Estimated effects of covariates at baseline on the risk of HIV infection based on different survival models, Zimbabwe Demographic Health Survey (ZDHS) 2005/06, 2010/11, and 2015.

Model	MODEL 1			MODEL 2		
	2005/06	2010/11	2015	2005/06	2010/11	2015
Variable	HR (CI)	HR (CI)	HR (CI)	HR (CI)	HR (CI)	HR (CI)
Sex						
Female	1	1	1	1	1	1
Male	0.68 (0.62–0.76)	0.73 (0.66–0.80)	0.66 (0.60–0.73)	0.26 (0.23–0.31)	0.25 (0.21–0.29)	0.22 (0.19–0.26)
Marital status						
Single	1	1	1	1	1	1
Married/cohabiting	0.30 (0.25–0.35)	0.32 (0.27–0.38)	0.29 (0.25–0.34)	0.40 (0.33–0.50)	0.45 (0.37–0.55)	0.59 (0.48–0.71)
Separated/divorced/widowed	0.64 (0.54–0.76)	0.70 (0.59–0.84)	0.60 (0.50–0.71)	0.69 (0.55–0.87)	0.85 (0.67–1.08)	1.08 (0.84–1.38)
Place of residence						
Urban	1	1	1	1	1	1
Rural	1.11 (1.01–1.22)	0.87 (0.79–0.95)	0.86 (0.79–0.94)	1.10 (0.96–1.27)	1.03 (0.89–1.19)	1.06 (0.92–1.22)
Education level						
No education/primary	1	1	1	1	1	1
Secondary	2.53 (2.29–2.79)	1.62 (1.47–1.79)	1.05 (0.95–1.15)	1.12 (0.97–1.29)	0.91 (0.79–1.07)	0.63 (0.53–0.74)
Higher	1.32 (1.01–1.72)	0.75 (0.58–0.97)	0.51 (0.43–0.62)	0.43 (0.30–0.62)	0.30 (0.21–0.43)	0.25 (0.19–0.33)
STI in the last 12 months						
No	1	1	1	1	1	1
Yes	2.19 (1.88–2.56)	1.87 (1.57–2.22)	2.15 (1.79–2.58)	1.73 (1.33–2.25)	2.09 (1.58–2.76)	1.80 (1.31–2.46)
No. of sexual partners						
0	1	1	1	1	1	1
1	1.62 (1.42–1.86)	1.84 (1.61–2.12)	1.64 (1.43–1.89)	2.12 (1.71–2.64)	2.32 (1.86–2.89)	2.26 (1.77–2.88)
2+	1.95 (1.54–2.47)	2.19 (1.70–2.67)	2.06 (1.68–2.54)	4.91 (3.39–7.10)	5.58 (3.95–7.89)	6.88 (4.84–9.76)
Currently employed						
No	1	1	1	1	1	1
Yes	0.82 (0.75–0.90)	0.83 (0.76–0.91)	0.77 (0.70–0.84)	0.84 (0.74–0.96)	0.92 (0.80–0.99)	0.66 (0.57–0.76)

HR, hazard ratio; CI, confidence interval.

Model 1 was the weibull parametric survival model with survey age imputation as the time and Model 2 was a weibull interval censoring model.

the risk of HIV infection for all the survey years was lower in individuals who were separated/divorced/widowed in reference to Model 1, but Model 2 results indicated that the risk was higher for that particular group of individuals as compared to those who were single (Table 3). According to the results in Model 1, the risk of HIV infection was almost the same for those who had one or more than two sexual partners, as compared to those who did not have any sexual partners in the past 12 months prior to the survey. However, Model 2, the risk was more than four times for those who had more than two sexual partners, as compared to those who did not have any sexual partners in the past 12 months prior to the survey. Of interest, the risk of HIV infection for those who had more than two sexual partners increased over time for Model 2, with Model 1 having a decreasing trend (Table 3).

DISCUSSION

Frequently, researchers are interested in using standard survival models in determining the failure times, however interval censoring has become increasingly common. The purpose of this study was to identify risk factors for HIV infection using three different models and determine the best fitting model. To

identify the best fitting model, we utilized the Akaike Information Criterion (AIC), where the model with the least AIC value was the best fitting model. We also used graphical methods to ascertain the best fitting model, however graphical goodness-of-fit test for interval-censored data is rare with the available methods are still lacking in implementation (15–18). We managed to check the goodness-of-fit for all the models by overlaying the semi-parametric model with the fitted parametric model of the survival function and also plotting the Cox-Snell residuals against the cumulative hazard.

The model with interval censored data resulted in better estimates of the risk of HIV infection as compared to the standard survival model i.e., Models 1 based on the AIC value and Figure 3. The superiority of Model 2 was due to the ability to precisely mimic some of the results in literature from previous studies which also used population-based HIV surveys or specific cohorts in Zimbabwe. For example, a study in Zimbabwe conducted between 1999 and 2001 reported that they observed an increasing trend of HIV incidence among the educated individuals, which was rather unexpected; however, this might have been due to a higher socio-economic status, a factor reported to be associated with HIV infection in the region of

Africa (19, 20). However, a systematic review which explored time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa, reported a shift in the HIV epidemic from educated to the uneducated (21). According to a study in which they used the 2005/06 Zimbabwe Demographic Health Survey (ZDHS) to determine the relationship between HIV status and the demographic and socio-economic characteristics among adults in Zimbabwe by construction the risk profile of the average adult, they concluded that there was a significant negative association between HIV infection and education (22). They further clarified that an extra year of schooling to an average of 8 years (i.e., secondary education and above) was associated with a 0.5 percent point decrease in the probability of HIV infection for Zimbabwe (22). Another study which used the 2010/11 ZDHS also reported a lower risk of HIV infection of individuals with secondary level education and above (23). Based on these findings from literature, and the trend observed, the model with interval-censoring was consistent with the reported results, however, the other two models reported a higher risk of HIV infection for individuals with secondary level education, which was contrary to the findings from literature. according to an in-depth analysis of the 2005/06 ZDHS on the risk factors associated with HIV infection, it was reported that individuals who never had any sexually transmitted infection 12 months prior the survey were significantly associated with a 0.437 times lower risk of HIV infection compared with their counterparts who had a sexually transmitted infection during the same period (24). This result was close to the results obtained for the hazard of HIV infection in the model with interval-censoring. Another example was a study of the analysis of 2005/06 ZDHS, which reported that the likelihood of being HIV infected increased with the number of sexual partners and decreased with the level of faithfulness to a spousal partner. In the same study, it was reported that the odds of being HIV infected were 3 to 4 times greater among those who had two more sexual partner (25). These results were again consistent with results from the model with interval-censoring rather than the other two models for the 2005/06 ZDHS. On a general note, a 2013 study noted that having a large number of life partners increased HIV infection in a cohort from Manicaland (26).

Comparisons of the three models used in this study reveal a consistent match on the factors associated with HIV infection estimated from ZDHS data and the results obtained from previous studies using population-based HIV surveys or specific cohorts in Zimbabwe. For example, in all the three models, and based on all the three surveys, i.e., ZDHS 2005/06, 2010/11 and 2015, females were more at risk of HIV infection than men. Similarly, studies which determined the factors associated with HIV infection using the 2005/06 and 2010/11 ZDHS reported the same findings (22–24). However, though the surveys were the same, they did not use the same analytical methods to reach the same conclusions. Results suggest that marriage was associated with a lower risk of infection based on all the models. This is further supported by a study which analyzed the 2005/06 and 2010/11 ZDHS data (23). The study reported that marital union was positively associated with the decline of HIV infection for both men and women (23). Another study to determine the baseline predictions of HIV-1 acquisition among

women reported that being unmarried was the strongest risk factor for HIV-1 acquisition (27). Results from these studies are again consistent with what all the models in our studies. Even though the models reported results similar to what had already been reported in literature, the precision of the model with interval-censoring in explaining some of the covariates is what stood out the most. However, the Cox-Snell residuals clearly showed that Model 1 was the best fitting model. The difference between the Cox-Snell residuals graphical method and the model estimates maybe due to the lack of adequate methods to test the goodness-of-fit of interval-censored data as cited by other authors (15–18).

The main strength of this study dependent on the quality of the data obtained from the surveys. These data were derived from population-based surveys, which in reality provides more reliable and robust data. Another strength of this study was due to the fact that we did not restrict our analysis to one method, however, we had the opportunity to determine the best model to fit the hazard of infection by comparing two different scenarios. For instance, if the median survival time for HIV infection was 5 years given the type of data we had, and the intervals were about 3 to 6 months wide, then we would have no reason to complicate the analysis by considering interval censoring. On the other hand, if the intervals were about 1 year or longer, then accounting for uncertainty in the analysis was necessary, which we did when we implemented the interval-censoring approach. Another reason for concluding that interval-censoring gave better estimates was due to its consistency with the published results from literature. Even though we consider the interval-censoring model as the superior model with regards to our specific data, the method had its own set of limitations. These limitations included the wide range of intervals used, which could have underestimated or overestimated the effect of other factors on the risk of HIV infection. Inclusion of competing risks factors in the model would have greatly improved the modeling approach. Further studies can be done on imputation models, which imputes an estimated time of HIV infection based on the data.

AUTHOR CONTRIBUTIONS

RB and EM conceived of the presented idea and verified the analytical methods and models. RB developed the theory and performed the data analysis. EM supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported through the Developing Excellence in Leadership Training and Science Africa (DELTA) initiative. The DELTA's Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating (NEPAD) Agency with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa

Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the United Kingdom (UK) government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, or the UK government.

REFERENCES

- Bain LE, Nkoke C, Noubiap JN. UNAIDS 90-90-90 targets to end the AIDS epidemic by 2020 are not realistic: comment on “Can the UNAIDS 90-90-90 target be achieved? A systematic analysis of national HIV treatment cascades”. *BMJ Glob Health*. (2017) 2:e000227-e. doi: 10.1136/bmjgh-2016-000227
- Zimbabwe Ministry of Health. *Zimbabwe Population-Based HIV Impact Assessment ZIMPHIA 2015–2016*. Harare: Zimbabwe Ministry of Health (2016).
- Vandormael A, Dobra A, Bärnighausen T, de Oliveira T, Tanser F. Incidence rate estimation, periodic testing and the limitations of the mid-point imputation approach. *Int J Epidemiol*. (2017) 47:236–45. doi: 10.1093/ije/dyx134
- Duri K, Stray-Pedersen B, Muller F. HIV/AIDS: the Zimbabwean situation and trends. *Am J Clin Med Res*. (2013) 1:15–22. doi: 10.12691/ajcmr-1-1-5
- Jewell NP, van der Laan M. Generalizations of current status data with applications. *Lifetime Data Anal*. (1995). 1:101–9. doi: 10.1007/BF00985261
- Dinse GE, Lagakos S. Regression analysis of tumour prevalence data. *J R Stat Soc Ser C*. (1983) 32:236–48. doi: 10.2307/2347946
- Zhang Z, Sun J. Interval censoring. *Stat Methods Med Res*. (2010) 19:53–70. doi: 10.1177/0962280209105023
- Zimbabwe Central Statistical Office, Macro International. *Zimbabwe Demographic and Health Survey 2005–06*. Calverton, MD: Central Statistical Office/Zimbabwe and Macro International (2007).
- Zimbabwe National Statistics Agency, ICF International. *Zimbabwe Demographic and Health Survey 2015: Final Report*. Rockville, MD: Zimbabwe National Statistics Agency (ZIMSTAT) and ICF International (2016).
- Zimbabwe National Statistics Agency–ZIMSTAT, ICF International. *Zimbabwe Demographic and Health Survey 2010–11*. Calverton, MD: ZIMSTAT and ICF International (2012).
- The DHS Program. *Demographic and Health Surveys*. (2005). Available online at: <https://dhsprogram.com> (accessed May 26, 2018).
- StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC (2017).
- StataCorp. *Stata 15 Base Reference Manual*. College Station, TX: Stata Press (2017).
- Anderson-Bergman C. icenReg: regression models for interval censored data in R. *J Stat Softw*. (2017) 81:1–23. doi: 10.18637/jss.v081.i12
- Farrington CP. Residuals for proportional hazards models with interval-censored survival data. *Biometrics*. (2000) 56:473–82. doi: 10.1111/j.0006-341X.2000.00473.x
- Lawless JF, Babineau D. Models for interval censoring and simulation-based inference for lifetime distributions. *Biometrika*. (2006) 93:671–86. doi: 10.1093/biomet/93.3.671
- REN JJ. Goodness of fit tests with interval censored data. *Scandinav J Stat*. (2003) 30:211–26. doi: 10.1111/1467-9469.00327
- Bogaerts K, Komarek A, Lesaffre E. *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS, 1st Edn*. New York, NY: Chapman and Hall/CRC (2017). doi: 10.1201/9781315116945
- Dallabetta GA, Miotto PG, Chipangwi JD, Saah AJ, Liomba G, Odaka N, et al. High socioeconomic status is a risk factor for human immunodeficiency virus type 1 (HIV-1) infection but not for sexually transmitted diseases in women in Malawi: implications for HIV-1 control. *J Infect Dis*. (1993) 167:36–42. doi: 10.1093/infdis/167.1.36
- Kumwenda N, Hoffman I, Chirenje M, Kelly C, Coletti A, Ristow A, et al. HIV incidence among women of reproductive age in Malawi and Zimbabwe. *Sex Trans Dis*. (2006) 33:646–51. doi: 10.1097/01.olq.0000223283.27142.9f
- Hargreaves JR, Bonell CP, Boler T, Boccia D, Birdthistle I, Fletcher A, et al. Systematic review exploring time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa. *AIDS*. (2008) 22:403–14. doi: 10.1097/QAD.0b013e3282f2aac3
- Asiedu C, Asiedu E, Owusu F. The socio-economic determinants of HIV/AIDS Infection Rates in Lesotho, Malawi, Swaziland and Zimbabwe. *Dev Policy Rev*. (2012) 30:305–26. doi: 10.1111/j.1467-7679.2012.00578.x
- Gonese E, Mapako T, Dzangare J, Rusakaniko S, Kilmarx PH, Postma MJ, et al. Within-Gender Changes in HIV Prevalence among Adults between 2005/6 and 2010/11 in Zimbabwe. *PLoS ONE*. (2015) 10:e0129611-e. doi: 10.1371/journal.pone.0129611
- Kembo J. Risk factors associated with HIV infection among young persons aged 15–24 years: evidence from an in-depth analysis of the 2005–06 Zimbabwe Demographic and Health Survey. *J Soc Aspects HIV/AIDS*. (2012) 9:54–63. doi: 10.1080/17290376.2012.683579
- Mishra V, Hong R, Bignami-Van Assche S, Barrere B. *The role of partner reduction and faithfulness in HIV prevention in Sub-Saharan Africa: Evidence from Cameroon, Rwanda, Uganda, and Zimbabwe*. Calverton, MD: Macro International (2009).
- Gregson S, Mugurungi O, Eaton J, Takaruzza A, Rhead R, Maswera R, et al. Documenting and explaining the HIV decline in east Zimbabwe: the Manicaland General Population Cohort. *BMJ Open*. (2017) 7:e015898. doi: 10.1136/bmjopen-2017-015898
- Balkus JE, Brown E, Palanee T, Nair G, Gafoor Z, Zhang J, et al. An empiric HIV risk scoring tool to predict HIV-1 acquisition in African women. *J Acquir Immune Def Syndr*. (2016) 72:333–43. doi: 10.1097/QAI.0000000000000974

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2019.00262/full#supplementary-material>

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Birri Makota and Musenge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

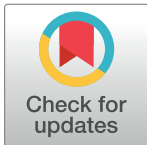
RESEARCH ARTICLE

Predicting HIV infection in the decade (2005–2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach

Rutendo Beauty Birri Makota^{*}, Eustasius Musenge

Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

* rutendobbirri@gmail.com



OPEN ACCESS

Citation: Birri Makota RB, Musenge E (2023) Predicting HIV infection in the decade (2005–2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. PLOS Digit Health 2(6): e0000260. <https://doi.org/10.1371/journal.pdig.0000260>

Editor: Heather Mattie, Harvard University T H Chan School of Public Health, UNITED STATES

Received: September 23, 2022

Accepted: April 24, 2023

Published: June 7, 2023

Copyright: © 2023 Makota, Musenge. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The datasets used for this study can be found on the Demographic Health Survey Program website, which requires that interested researchers register for an account in order to request access to survey data files: <https://dhsprogram.com/data/new-user-registration.cfm>. The data files underlying this study are included in surveys Zimbabwe 2005-06 (<https://dhsprogram.com/methodology/survey/survey-display-260.cfm>), Zimbabwe 2010-11 (<https://dhsprogram.com/methodology/survey/survey-display-367.cfm>), and Zimbabwe 2015 (<https://dhsprogram.com/methodology/survey/survey-display-367.cfm>).

Abstract

The burden of HIV and related diseases have been areas of great concern pre and post the emergence of COVID-19 in Zimbabwe. Machine learning models have been used to predict the risk of diseases, including HIV accurately. Therefore, this paper aimed to determine common risk factors of HIV positivity in Zimbabwe between the decade 2005 to 2015. The data were from three two staged population five-yearly surveys conducted between 2005 and 2015. The outcome variable was HIV status. 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. Feature selection was done using Lasso regression, and the best combination of selected features was determined using Sequential Forward Floating Selection. We compared six algorithms in both sexes based on the F1 score, which is the harmonic mean of precision and recall. The overall HIV prevalence for the combined dataset was 22.5% and 15.3% for females and males, respectively. The best-performing algorithm to identify individuals with a higher likelihood of HIV infection was XGBoost, with a high F1 score of 91.4% for males and 90.1% for females based on the combined surveys. The results from the prediction model identified six common features associated with HIV, with total number of lifetime sexual partners and cohabitation duration being the most influential variables for females and males, respectively. In addition to other risk reduction techniques, machine learning may aid in identifying those who might require Pre-exposure prophylaxis, particularly women who experience intimate partner violence. Furthermore, compared to traditional statistical approaches, machine learning uncovered patterns in predicting HIV infection with comparatively reduced uncertainty and, therefore, crucial for effective decision-making.

Author summary

The Joint United Nations Programme (UNAIDS) set up fast-track targets to reach HIV epidemic control by 2030, where it is expected that 95% of people living with HIV know



OPEN ACCESS

EDITED BY

Farrokh Habibzadeh,
Global Virus Network, Iran

REVIEWED BY

Kristen Stafford,
University of Maryland, United States
Lucie Leon,
Santé Publique France, France

*CORRESPONDENCE

Rutendo Birri Makota
rutendobirri@gmail.com

SPECIALTY SECTION

This article was submitted to
Research Methods and Advances in
Epidemiology,
a section of the journal
Frontiers in Epidemiology

RECEIVED 27 August 2022

ACCEPTED 14 November 2022

PUBLISHED 06 December 2022

CITATION

Birri Makota R and Musenge E (2022)
Estimating age specific prevalence and
force of infection in Zimbabwe using
combined cross-sectional surveys
from 2005 to 2015.
Front. Epidemiol. 2:1029583.
doi: 10.3389/fepid.2022.1029583

COPYRIGHT

© 2022 Birri Makota and Musenge.
This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Estimating age specific prevalence and force of infection in Zimbabwe using combined cross-sectional surveys from 2005 to 2015

Rutendo Birri Makota* and Eustasius Musenge

Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences,
University of the Witwatersrand, Johannesburg, South Africa

Objective: Age structured sexual mixing patterns have been noted to be associated with HIV prevalence and force of infection. Therefore, this study aimed to estimate the age dependent HIV force of infection using survey cross-sectional data from Zimbabwe.

Methods: We fit generalized additive models namely; linear, semi-parametric, non-parametric and non-proportional hazards models. Using the 2005–06, 2010–11 and 2015 Zimbabwe Demographic Health Surveys data. The Akaike Information Criteria was used to select the best model. The best model was then used to estimate the age dependent HIV prevalence and force-of-infection.

Results: Based on birth year cohort-specific prevalence, the female HIV prevalence reaches the highest peak at around 29 years of age, then declines thereafter. Males have a lower cohort specific prevalence between 15 and 30 years than females. Male cohort-specific prevalence slightly decreases between the ages of 33 and 39, then peaks around the age of 40. The cohort-specific FOI is greater in females than in males throughout all age categories. In addition, the cohort-specific HIV FOI peaked at ages 22 and 40 for females and males, respectively. The observed 18-year age difference between the HIV FOI peaks of males and females.

Conclusion: Our model was appealing because we did not assume that the FOI is stationary over time; however, we used serological survey data to distinguish the FOI's age-and-time effect. The cohort-specific FOI peaked 18 years earlier in females than males, indicative of age-mixing patterns. We recommend interventions that target younger females so as to reduce HIV transmission rates.

KEYWORDS

force of infection, prevalence, Demographic Health Survey (DHS), generalized additive models (GAMs), hazards models

Introduction

There has been a decline in HIV incidence in adults in sub-Saharan (SSA) (1). The prevalence of HIV has increased in those aged 25 and older despite a decrease in HIV incidence due to increasing life expectancy brought on by the advent of antiretroviral treatment (ART) (2–5). A cohort study in Manicaland has highlighted that the cohort of individuals living with HIV in Zimbabwe is aging such that the HIV prevalence is increasing among individuals aged 45 and older and decreasing among younger age groups (3). Due to the high survival rates of those who are HIV infected earlier in life, a shift in the relative share of infections to older individuals has been observed (1). Age-structured sexual mixing patterns have been noted to have an association with HIV force-of-infection (FOI) (6). Force-of-infection (FOI) is defined as the rate at which susceptible individuals acquire an infectious disease, a key parameter that specifies the current state of an epidemic (7). Even though the overall HIV incidence has declined in the era of ART, older cohorts (40–49 years) still account for a growing proportion of HIV incidence (1). Over time, understanding how the HIV FOI rates have shifted across age groups is needed to realign prevention targets.

Given the information about FOI in a population, interventions can be targeted to those at the most significant risk of infection. Subsequently, information on trends in the FOI over time can be used to evaluate the impact of programmes on the rate of new infections so that resources can be directed to the most effective interventions and plan future health care needs. Additionally, understanding the HIV incidence disparities across ages can help prioritize groups where more effort is needed to lower infection rates.

The most direct method of estimating the FOI is using cohort studies. However, these studies are costly; in most cases, representative samples are not attained (8). By contrast, cross-sectional surveys are less expensive, more straightforward, quicker to organize and can be carried out on a mass scale. In addition, serological testing has been more prevalent recently and is now a part of household surveys like the Demographic and Health Survey (DHS), which estimates the prevalence of HIV in the general adult population (9, 10). Given the ease of getting HIV prevalence data from household surveys, mathematical models can help untangle the complicated relationship between HIV FOI and prevalence (11, 12).

One of the modeling approaches to estimating FOI from prevalence data has come from Demography (13, 14). Here, serial cross-sectional population survey estimates of prevalence are used to generate an estimate of the FOI based on the principle that a sample of individuals of age a at time t is under some conditions representative of a group of individuals aged $a - \tau$ at time $t - \tau$, where τ is the interval between these cross-sectional population surveys (15).

Understanding age-specific differences in the infection rates are essential in predicting the magnitude of the FOI, thus, targeting age-groups for different intervention programs. However, temporal trends and estimates of age-specific FOI have to be recovered from age and time-specific prevalence data (16). In cross-sectional surveys, the information obtained when an individual has been tested is whether that individual has already experienced an infection before the age (a) of testing or not. This is commonly known as “current status” data. If the FOI does not change with time, then a single cross-sectional survey of individuals at different ages is sufficient to estimate age-specific infection rates. While this assumption is reasonable for some infections, this assumption is untenable in the case of HIV infection.

This study aimed to estimate the age- and time-dependent prevalence and HIV FOI using current status data from the Zimbabwe Demographic Health Survey (ZDHS) collected in 2005–06, 2010–11 and 2015. This was achieved by developing parametric, semi-parametric and non-parametric models; and determining which models best fit the data.

Methods

Zimbabwe survey cross-sectional data

Data from 2005–06, 2010–11, and 2015 ZDHS were used in this analysis and random samples of 10,800, 10,828 and 11,196 households were chosen, respectively. The households were selected using a two-stage cluster sampling method. At the first stage, random samples of 400, 406, and 400 enumeration areas (EAs), were selected for the 2005–06, 2010–11, and 2015 ZDHS, respectively. At the second stage, households were randomly picked using a comprehensive list of all the households in the specified EAs. This made it possible to apply specific weights to EAs in the design (17–19).

Outcome variable measurement: HIV testing procedure

Blood samples from all households were taken with the respondent's consent, or that of the respondent's parent or guardian in the case of minors, for HIV testing in the laboratory on females aged 0–49 years and males aged 0–54 years. Finger-prick blood samples were collected on filter paper and taken to a laboratory for analysis. All HIV-positive and 5–10% of the negative tests were retested with a second ELISA after an initial ELISA (enzyme-linked immunosorbent assay) test. A second ELISA or a Western Blot was conducted if the two ELISA tests yielded inconsistent findings. The DHS Data Archive (20) provided the data for this study, which only included people

between the ages of 15 and 49. We took into account the sampling design by including survey weights in the analysis.

Statistical model

Over the years, methods have been proposed that estimate the age-specific force-of-infection using seroprevalence data (9, 13–16, 21–28). However, these methods are primarily based on an infection acting on a set of susceptible individuals. Given that $\lambda_A(a)$ is the force-of-infection at age a , if we assume time homogeneity (stationarity) and lifelong immunity, then the proportion susceptible, $p_A(a)$ is:

$$p_A(a) = \exp \left[- \int_0^a \lambda_A(z) dz \right] \quad (1)$$

In reality, time homogeneity assumed in Equation 1 is not feasible as over the years, the introduction of ART, the evolution of cohorts and mutation of the pathogen has resulted in the change of the force-of-infection over time (15). Therefore, the time-specific force of infection $\lambda_T(t)$ can be estimated instead. By integrating exposure experience between the date of birth, $t - a$, and the survey date t results in the proportion susceptible $p_T(a)$ (15):

$$p_T(a) = \exp \left[- \int_0^a \lambda_T(t - a + z) dz \right] \quad (2)$$

Considering that in this analysis, we deal with three-time points, estimating the force of infection $\lambda_T(t)$, will not be ideal (21). Therefore, while estimating the force-of-infection, especially when working with many time points, it is crucial to model the impacts of age and time effects. The relation between age- and time-specific seroprevalence data and age-time force-of-infection (15), $\lambda(a, t)$, in an individual of age a at time t , results in the proportion susceptible $p(a, t)$ (29):

$$p(a, t) = \exp \left[- \int_0^a \lambda(z, t - a + z) dz \right] \quad (3)$$

This paper proposed five models to model the relation between age- and time-specific seroprevalence data and $\lambda(a, t)$ in Equation 3. These models are model 1, representing a parametric model, models 2 and 3, representing semi-parametric models, model 4, representing a non-parametric model, and model 5, representing the non-proportional hazards model. The first is a flexible semi-parametric method suitable for simple exploratory analysis. Then, these methods are applied to age- and time-specific ZDHS HIV seroprevalence data.

The parametric model can capture reasonably complex patterns of age- and time-dependence to determine the degree of curvature required by the data (15). Model 1 constrained the force-of-infection to be positive and is part

of the Exponential polynomial (EP) functions which have the following properties (29):

$$\lambda(a, t) = \exp \left(\mu_0 + \sum \mu_i a^i + \sum \theta_j t^j \right), \quad i, j = 1, 2, \dots \quad (4)$$

where i represents the time spent by an individual in age band i , and j represents the time spent by an individual in time band j . We can estimate the age-time dependence of the FOI in Equation 4 by setting all μ_i and θ_j to zero, thus resulting in Equation 5 (15):

$$\lambda(a, t) = \exp(\mu a) \quad (5)$$

Equation 5 represents model 1 and will be referred to in terms of their time/age parameterisation which is model EP1/EP1 being an exponential polynomial of degree 1 in time and degree 1 in age.

Assuming that the FOI changes exponentially with time represented by the following identity (for two calendar times t_1 and t_2) (29) in Equation 6:

$$\lambda(a, t_2) = \exp(\beta(t_2 - t_1)) \lambda(a, t_1) \quad (6)$$

We applied the methods of Nagelkerke et al. (12) to estimate age- and time-specific FOI without any assumption about age dependency (16). To be able to assess the cohort experience, a cohort born at calendar time b and has age a at calendar time $a + b$, then the age-specific FOI for that particular cohort is defined by $\lambda_b(a)$ in Equation 7 (29):

$$\lambda_b(a) = \lambda(a, a + b) \quad (7)$$

The proportional hazards assumption in Equation (6) then translates into Equation 8, which leads to models 2 and 3 in this study, which are proportional hazards, semi-parametric models (29).

$$\lambda_b(a) = \exp(\beta(b - b_0)) \lambda_{b0}(a) \quad (8)$$

where the hazard at baseline year of birth b_0 , is defined by $\lambda_{b0}(a)$, which can also be expressed as $\lambda(a, a + b_0)$ (29).

The semi-parametric model in Equation (8) can be reformulated as a fully non-parametric proportional hazards model by:

$$\lambda(a, t_2) = \exp(g(t_2 - t_1)) \lambda(a, t_1) \quad (9)$$

where $g(t)$ is a smooth function with the constraint that $g(0) = 0$.

Substituting Equation (9) in the expression given in Equation (3) leads to:

$$p_b(a) = p(a, b + a) = \exp \{ \exp(g(b - b_0)) \log p_{b0}(a) \} \quad (10)$$

where; with $\lambda_{b0}(a) = \lambda(a, b_0 + a)$,

$$p_{b0}(a) = \exp \left\{ - \int_0^a \lambda_{b0}(u) du \right\}.$$

Equation (10) which is Model 4, may be expressed as a Generalized Additive Model (GAM) for binary data with complementary log-log link function (29) (Equation 11), with $\pi_b(a) = 1 - p(a)$ representing the proportion of persons infected at age a or before from the cohort with year of birth b (29):

$$\log(-\log(1 - \pi_b(a))) = f_1(a) + f_2(b), \quad (11)$$

where

$$f_2(b) = g(b - b_0),$$

and

$$f_1(a) = \log(-\log(\pi_{b0}(a))).$$

The corresponding FOI for this cohort then equals to (29):

$$\lambda_b(a) \frac{\pi'_b(a)}{1 - \pi_b(a)} = \exp(f_2(b)) \lambda_{b0}(a), \quad (12)$$

with

$$\lambda_{b0}(a) \frac{\pi'_{b0}(a)}{1 - \pi_{b0}(a)} = \frac{d}{da} \exp(f_1(a)) \quad (13)$$

The proportional hazards model translates into an additive *cloglog* model for the seroprevalence status (29), as shown by Equation (11), represented in this study as Model 4. In addition to the year of birth, this GAM technique may be generalized to include additional subject-specific characteristics, such as gender (29).

A non-proportional hazards model (Model 5) was obtained by modeling the interaction of age a with year of birth b . Taking the GLM model with a two-dimensional smoother (29):

$$\log(-\log(1 - \pi_b(a))) = f(a, b), \quad (14)$$

some straightforward calculus shows that:

$$\frac{\lambda_b(a)}{\lambda_{b0}(a)} = (\exp(f(a, b) - f(a, 0))) \left(\frac{\frac{\partial f(a, b)}{\partial a}}{\frac{\partial f(a, 0)}{\partial a}} \right) \quad (15)$$

The right-hand side of Equation (15) is no longer independent of age a . Note that it reduces to (12) in case $f(a, b) = f_1(a) + f_2(b)$. The proportional hazards model was tested by comparing a model with a (parametric or non-parametric) (age $\times$ year of birth) interaction term with the corresponding model with main effects only (29).

Statistical analysis

This study estimated age- and time-dependent prevalence and FOI from serial cross-sectional population survey prevalence data. Our analysis was based on comparing a generalized linear additive model (model 1), generalized semi-parametric additive models (models 2 and 3) and generalized non-parametric additive models (models 4 and 5). All the models were proportional hazards, except for model 5, a non-proportional hazard model. Model 1 adapted Ades and Nokes' (15) exponential polynomial model (EP) of type EP₁/EP₁.

Models 2 and 3 were modified from the semi-parametric model of Nagarkerke et al. (16) by employing a spline smoother instead of the isotonic step function for the baseline hazard function $\lambda_{b0}(a)$. The age-specific FOI, $\lambda_b(a)$, for that particular cohort, was estimated in a semi-parametric way using an iterative MM algorithm with a parametric proportionality factor $\exp[\beta(b - b_0)]$ and a non-parametric isotonic stepwise estimate for the baseline hazard $\lambda_{b0}(a)$ as shown in Equation (6).

Model 4 was reformulated to a fully non-parametric model using the semi-parametric model of Nagarkerke et al. (16) as shown in section Outcome variable measurement: HIV testing procedure. Model 5 was obtained by an interaction of age (a) with the year of birth (b), resulting in a non-proportional hazards model. Models 1–5 are fully explained in section Outcome variable measurement: HIV testing procedure.

Model 1–5 specification and goodness-of-fit

We were interested in modeling some response variable y (force-of-infection) which follows an exponential family distribution using predictor variables a which is age and b being the year of birth. Let $\mu = E(y)$, then models 1–5 are:

$$\begin{aligned} \text{Model 1 (linear additive model)} : h(\mu) \\ = \beta_0 + \beta_1 a + \beta_2 b \end{aligned} \quad (16)$$

$$\begin{aligned} \text{Model 2 (semi-parametric additive model)} : h(\mu) \\ = \beta_0 + \beta_1 b + f_1 a \end{aligned} \quad (17)$$

$$\begin{aligned} \text{Model 3 (semi-parametric additive model)} : h(\mu) \\ = \beta_0 + \beta_1 a + f_1 b \end{aligned} \quad (18)$$

$$\begin{aligned} \text{Model 4 (non-parametric additive model)} : h(\mu) \\ = \beta_0 + f_1 a + f_2 b \end{aligned} \quad (19)$$

$$\begin{aligned} \text{Model 5 (non-proportional hazards additive model)} : h(\mu) \\ = \beta_0 + f_1(a, b) \end{aligned} \quad (20)$$

where h is a smooth monotonic “link” function of mean μ and f_1 and f_2 are smooth functions of the covariates considered. In order for these models to be distinguishable, the smooth functions were restricted to have a mean of zero using the gam function. Model specifications for models 1–5 and link

function were done using the family argument to generalized additive model (GAM) framework (26) in the R software package *mgcv* (30) with the *gam* function. The Pool-Adjacent-Violators Algorithm (PAVA) application ensured monotonicity in the age dimension. The *mgcv* package's estimation of GAMs used a penalized likelihood method. Finally, we used the Akaike Information Criteria (AIC) to choose the best model for the model comparison.

$$AIC = \underbrace{-2\log p(a)}_{\text{goodness of fit}} + \underbrace{2\tau}_{\text{model complexity}}$$

where τ is the effective degrees of freedom (edf). Therefore, if $AIC_{m1} < AIC_{m2}$ choose model 1. The effective degrees of freedom (edf) showed the complexity of the model. The 95% confidence interval of the best model was obtained using the R *confit.gam* package which calculates the point-wise confidence intervals for the smooth terms of a fitted gam model.

Results

The dataset was stratified by gender, and all subsequent analysis was done according to the stratification. Generalized non-parametric additive model (model 4) was selected as the best model according to the AIC criteria for females and males (Table 1 and Figure 1). Furthermore, since model 4 was a proportional hazards model, using analysis-of-deviation, the proportionality assumption was determined to be valid for both the female and male models, with p -values of 0.71 and 0.50, respectively.

Figure 2 shows the age-specific fitted and observed HIV prevalence estimates for the three calendar periods 2005–06, 2010–11, and 2015, for both females and males. Model 4 showed a good model fit for the observed data. Comparing the fitted HIV prevalence model, we detect a change of high HIV prevalence among 33-year-old females in 2005–06, 36-year-old females in 2010–11, and 40-year-old females in 2015. In

2005–06, the prevalence of HIV was highest among 42-year-old males; however, in 2010–11, the prevalence declined in the same age group. In 2015, the prevalence of HIV steadily rose with age among males. In general, the prevalence of HIV was greater among females than males over all three time periods.

Figure 3 illustrates the fitted cohort-specific prevalence curves for females and males as a function of age, according to the best model (model 4). Prevalence was estimated for each birth cohort for all survey years combined. The solid red component of each prevalence curve corresponds to the proportion of the age range (birth cohort) for which data were available, while the dotted portion of each curve was extrapolated using the model. Figure 3 demonstrates that data for both models were available for the age range considered in this study. Based on cohort-specific prevalence, the female HIV prevalence reaches the highest peak at around 29 years of age, then declines up to the age of 49 years (these values can be observed by examining the red lines/curves). Males, have a lower cohort-specific prevalence between 15 and 30 years than females. Male cohort-specific prevalence slightly decreases between the ages of 33 and 39. The highest peak of HIV prevalence for males is around the age of 40.

Figure 4 depicts the fitted cohort-specific FOI curves as a function of age for females and males, as predicted by Model 4. The FOI was estimated for every birth cohort over the combined survey years. The gray FOI curve displays the baseline FOI $\lambda b_0(a)$ for an assumed $b_0 = 1,955$, 1,955 being the minimum time (birth year) in the cohort. The black dotted curves are generated from the proportionality factor $\exp(g(b - b_0))$ in Equation 11 resulting in the red curves. Figure 4 indicates that, where data is available, the cohort-specific FOI is greater in females than males throughout all age categories, as indicated by the red curve. The cohort-specific HIV FOI peaked at ages 22 and 40, respectively, for females and males, demonstrating more than 15-year age difference between the HIV FOI peaks of males and females. Nevertheless, based on the extrapolated cohort-specific FOI, males tend to have a greater FOI than females in older age groups.

TABLE 1 The Akaike Information Criterion (AIC), Schwarz's Bayesian Criterion (BIC), and effective degrees of freedom (edf) for completely parametric (model 1), semi-parametric (models 2 and 3), and non-parametric (models 4 and 5) models.

Model components	Females					Males				
	edf	AIC	Rank	BIC	Rank	edf	AIC	Rank	BIC	Rank
Model 1: $a + b$	3	22,571.69	5	22,596.00	5	3	12,196.91	5	12,220.32	5
Model 2: $f(a) + b$	6.99	21,841.76	3	21,897.11	2	8.25	11,968.76	3	12,033.11	1
Model 3: $a + f(b)$	6.61	21,945.97	4	21,999.54	4	8.03	12,003.26	4	12,065.93	4
Model 4: $f(a) + f(b)$	9.93	21,779.61	1	21,860.10	1	12.5	11,938.44	1	12,035.99	2
Model 5: $f(a, b)$	15.7	21,785.40	2	21,912.22	3	15.5	11,939.81	2	12,060.90	3

a, age; b, year of birth; f, smooth function.

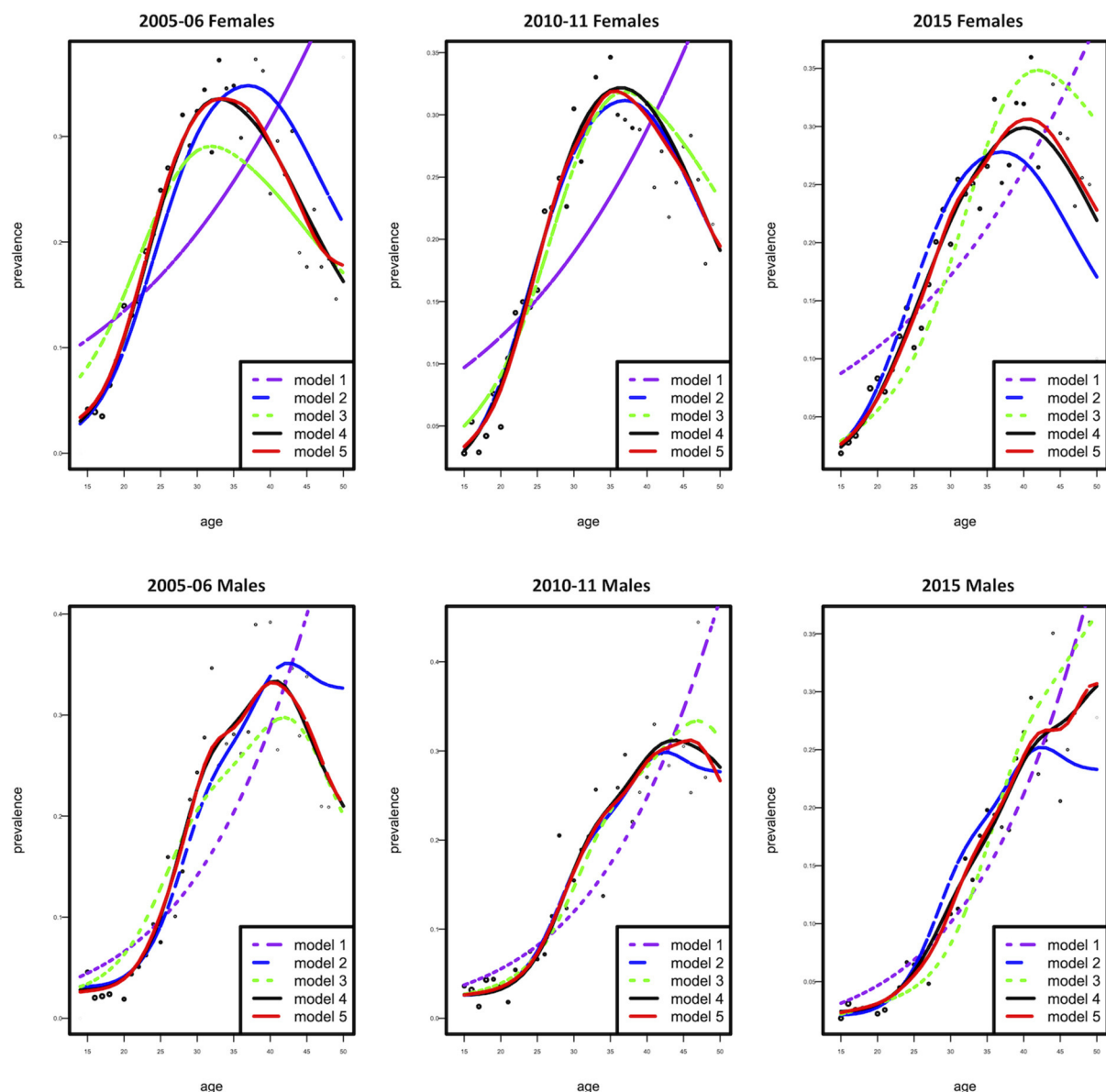


FIGURE 1
Observed (dotted curve) and fitted prevalence (five solid curves) as a function of age according to females (top) and males (bottom).

Discussion

This study estimated the age- and time-dependent HIV prevalence and FOI. Our model was appealing because we did not assume that the FOI is stationary over time; however, we used serological survey data to distinguish the FOI's age-and-time effect. It has been reported that the estimation of HIV FOI might lead to over-interpretation of results because of selection bias, partially due to lower consent rates (27, 31). However, recent work has confirmed that non-participation in surveys

does not result in significant biases in estimating cross-sectional HIV prevalence (2, 27).

Generally, HIV prevalence decreased for both males and females over the years, with HIV prevalence higher in females than males. Of interest, we observed the peak HIV prevalence for each successive survey year was lower but shifted with an increase in age, which can be explained by either that people are now being infected at older ages or that these are people who were diagnosed with HIV many years after being infected. Unfortunately, data on HIV testing patterns in different age

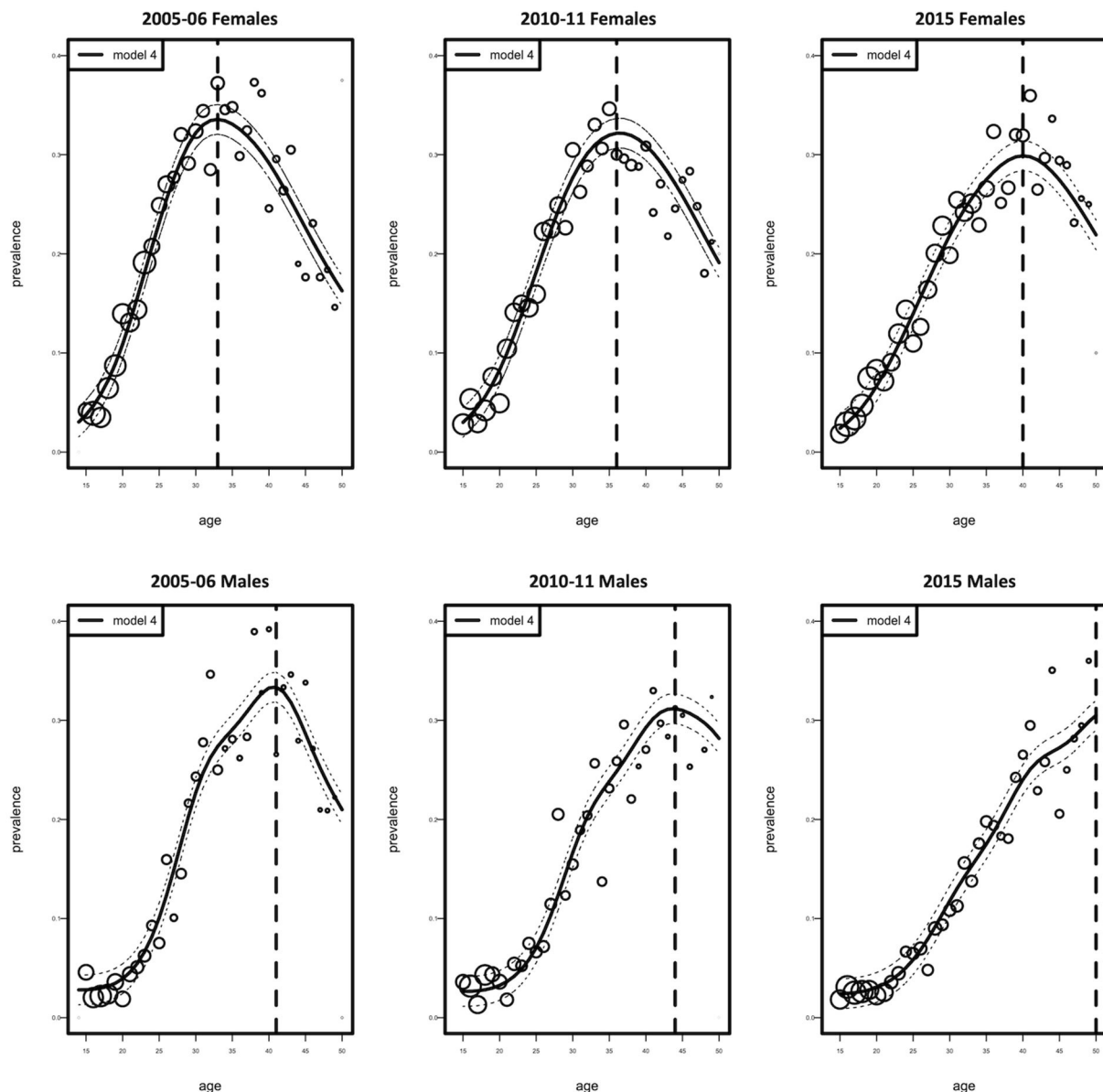


FIGURE 2
Observed (circles) and fitted (solid curve with 95% Confidence Interval, dotted curve) HIV prevalence as a function of age according to females and males. Circles represent the estimated prevalence by age. Their size is proportional to the number of individuals.

groups are unavailable, and the observed age increase may have resulted from increased testing in cohorts born earlier. In another school of thought, it is believed that due to the scale-up of antiretroviral treatment, there has been an increase in life expectancy; therefore, the prevalence in the population for the cohorts born earlier will increase (32). However, the shift toward increased HIV prevalence in older birth cohorts is typical of populations of people living with HIV (PLHIV) across the world (33). Our model, especially for males, projects similar age profiles (Figure 1).

A sharp increase in cohort-specific HIV prevalence is observed in our results to occur approximately a decade apart in females (at 29 years) than males (at 40 years). Cohort-specific FOI peaks ~15 years earlier in females (at 22 years) than males (at 40 years). These findings confirm the issue of age-mixing patterns which can be corroborated by other studies (6, 31, 34). A study in Zimbabwe found that age-disparate sexual relationships are associated with increased HIV incidence among young women, mainly when partners are ten or more years older (35). In a 2005 national

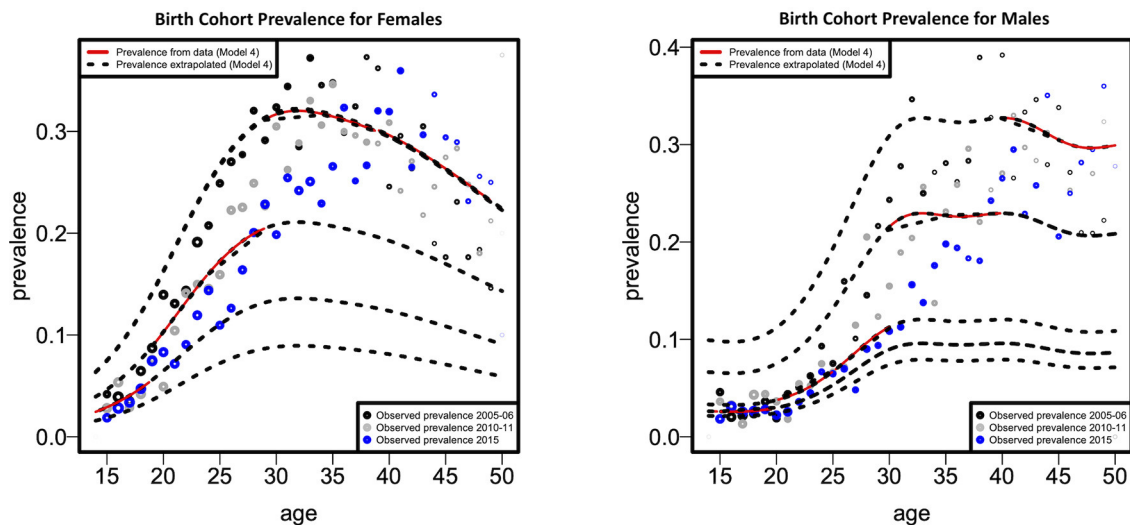


FIGURE 3
Fitted cohort-specific prevalence curves as a function of age according to females and males.

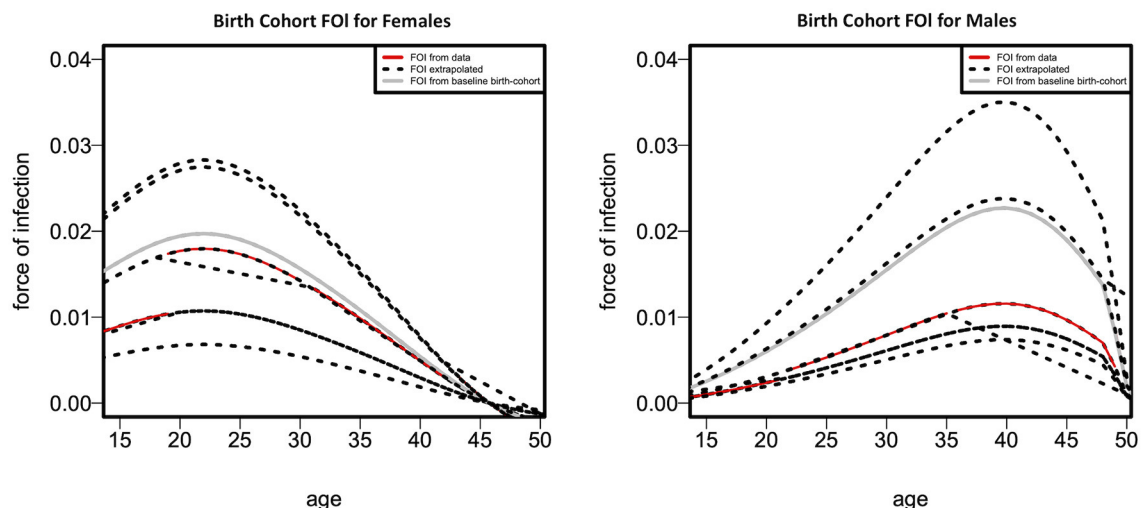


FIGURE 4
Fitted cohort-specific force-of-infection curves as a function of age according to females (left panel) and males (right panel).

household survey done in South Africa, it was determined that younger females had partners who were at least 5 years older than them, and it was hypothesized that these connections were linked to an increased HIV risk (36). Gregson et al. offered empirical evidence that age-disparate sexual interactions were responsible for the observed variations in the epidemiological patterns of young women and men in rural Zimbabwe (37). Therefore, age-disparate sexual relationships expose individuals to an increased risk of HIV infection. It has also been noted that age-mixing in sexual relationships is likely to reduce the younger person's ability to negotiate safe

sex successfully (38). Increasing female education regarding factors that increase HIV transmission rates might reduce age-disparate relationships, thereby decreasing infection rates in this age group. Pettifor et al. observed that education might lower young women's susceptibility to HIV infection, suggesting that interventions that target structural and partner-level risk factors, such as keeping young women in school, are necessary to lessen this vulnerability (39). This notion is supported by a study that found that higher female education was associated with a lower occurrence of age-disparate relationships (40).

A significant limitation of our study was that we did not address the issue of differential selection. This is when disease-related mortality affects the interpretation of observed serological profiles. Instead, we assumed that the excess mortality of individuals testing positive was negligible. For obvious reasons, this is a crude assumption, and other authors have proposed models which address this issue (13, 21, 26). However, our main advantage was combining three cross-sectional surveys from different time points, thereby jointly modeling age and epidemic effects. Since these surveys are conducted after every 5 years, the next one was scheduled for 2020 but because of COVID, no recent DHS survey has been conducted for Zimbabwe.

Combined cross-sectional surveys should be used to evaluate control measures and monitor the trends in HIV FOI. The same age groups should also be adopted for each successive survey, assuming that the temporal trend in the FOI can be monitored from the changes in prevalence. Future work would significantly be important to include differential selection and refine model estimates within a Bayesian framework by integrating other data sources.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Demographic Health Survey Program website [<https://dhsprogram.com>].

Ethics statement

The studies involving human participants were reviewed and approved by the Human Research Ethics Committee (Medical) of the University of Witwatersrand (No. M151154). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

RB and EM conceived of the presented idea and verified the analytical methods and models. RB developed the theory and

performed the data analysis. EM supervised the findings of this work. Both authors discussed the results and contributed to the final manuscript.

Funding

This work was supported through the Developing Excellence in Leadership Training and Science Africa (DELTA) initiative. The DELTA's Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating (NEPAD) Agency with funding from the Wellcome Trust [Grant 107754/Z/15/Z—DELTA's Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the United Kingdom (UK) government.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

References

1. Akullian A, Vandormael A, Miller JC, Bershteyn A, Wenger E, Cuadros D, et al. Large age shifts in HIV-1 incidence patterns in KwaZulu-Natal, South Africa. *Proc Natl Acad Sci*. (2021) 118:e2013164118. doi: 10.1073/pnas.2013164118
2. Zaidi J, Grapsa E, Tanser F, Newell M-L, Barnighausen T. Dramatic increases in HIV prevalence after scale-up of antiretroviral treatment: a longitudinal population-based HIV surveillance study in rural kwazulu-natal. *AIDS (London, England)*. (2013) 27:2301. doi: 10.1097/QAD.0b013e328362e832
3. Negin J, Gregson S, Eaton JW, Schur N, Takaruza A, Mason P, et al. Rising levels of HIV infection in older adults in Eastern Zimbabwe. *PLoS ONE*. (2016) 11:e0162967. doi: 10.1371/journal.pone.0162967
4. Kimanga DO, Ogola S, Umuro M. Prevalence and incidence of HIV infection, trends, and risk factors among persons aged 15–64 years in Kenya: results from a nationally representative study. *J Acquir Immune Defic Syndr*. (1999). 66:S13. doi: 10.1097/QAI.0000000000000124

5. Vollmer S, Harttgen K, Alfven T, Padayachy J, Ghys P, Bärnighausen T. The HIV epidemic in sub-Saharan Africa is aging: evidence from the demographic and health surveys in sub-Saharan Africa. *AIDS Behav.* (2017) 21:101–13. doi: 10.1007/s10461-016-1591-7
6. Ott MQ, Bärnighausen T, Tanser F, Lurie MN, Newell M-L. Age-gaps in sexual partnerships: seeing beyond 'sugar daddies'. *AIDS (London, England)*. (2011) 25:861. doi: 10.1097/QAD.0b013e32834344c9
7. Balakrishna Y. *Estimating the Force of Infection from Prevalence Data: Infectious Disease Modelling (Doctoral Dissertation)* (2013).
8. Hallett TB, Stover J, Mishra V, Ghys PD, Gregson S, Boerma T. Estimates of HIV incidence from household-based prevalence surveys. *AIDS (London, England)*. (2010) 24:147–52. doi: 10.1097/QAD.0b013e32833062dc
9. Boerma JT, Ghys PD, Walker N. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *Lancet.* (2003) 362:1929–31. doi: 10.1016/S0140-6736(03)14967-7
10. Garcia-Calleja J, Gouws E, Ghys P. National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates. *Sex Transm Infect.* (2006) 82:iii64–70. doi: 10.1136/sti.2006.019901
11. Hallett TB, Aberle-Grasse J, Bello G, Boulos L, Cayemittes M, Cheluget B, et al. Declines in HIV prevalence can be associated with changing sexual behaviour in Uganda, urban Kenya, Zimbabwe, and urban Haiti. *Sex Transm Infect.* (2006) 82:i1–8. doi: 10.1136/sti.2005.016014
12. Nagelkerke NJ, de Vlas SJ, Jha P, Luo M, Plummer FA, Kaul R. Heterogeneity in host HIV susceptibility as a potential contributor to recent HIV prevalence declines in Africa. *AIDS (London, England)*. (2009) 23:125. doi: 10.1097/QAD.0b013e3283177f20
13. Batter V, Matela B, Nsuami M, Manzila T, Kamenga M, Behets F, et al. High HIV-1 incidence in young women masked by stable overall seroprevalence among childbearing women in Kinshasa, Zaire: estimating incidence from serial seroprevalence data. *AIDS (London, England)*. (1994) 8:811–7. doi: 10.1097/00002030-199406000-00014
14. Gregson S, Donnelly CA, Parker CG, Anderson RM. Demographic approaches to the estimation of incidence of HIV-1 infection among adults from age-specific prevalence data in stable endemic conditions. *AIDS (London, England)*. (1996) 10:1689–97. doi: 10.1097/00002030-199612000-00014
15. Ades AE, Nokes DJ. Modeling age- and time-specific incidence from seroprevalence: toxoplasmosis. *Am J Epidemiol.* (1993) 137:1022–34. doi: 10.1093/oxfordjournals.aje.a116758
16. Nagelkerke N, Heisterkamp S, Borgdorff M, Broekmans J, Van Houwelingen H. Semi-parametric estimation of age–time specific infection incidence from serial prevalence data. *Stat Med.* (1999) 18:307–20. doi: 10.1002/(SICI)1097-0258(19990215)18:3<307::AID-SIM15>3.0.CO;2-Z
17. Zimbabwe Central Statistical Office, Macro International. *Zimbabwe Demographic and Health Survey 2005-06*. Calverton, Maryland, USA: Central Statistical Office/Zimbabwe and Macro International (2007).
18. Zimbabwe National Statistics Agency, ICF International. *Zimbabwe Demographic and Health Survey 2015: Final Report*. Rockville, Maryland, USA: Zimbabwe National Statistics Agency (ZIMSTAT) and ICF International (2016).
19. Zimbabwe National Statistics Agency—ZIMSTAT, ICF International. *Zimbabwe Demographic and Health Survey 2010-11*. Calverton, Maryland, USA: ZIMSTAT and ICF International (2012).
20. The DHS Program. *Demographic and Health Surveys [Online Database]*. (2005). Available online at: https://dhsprogram.com/data/dataset_admin/login_main.cfm (accessed on updated May 2018).
21. Ades A, Medley G. Estimates of disease incidence in women based on antenatal or neonatal seroprevalence data: HIV in New York City. *Stat Med.* (1994) 13:1881–94. doi: 10.1002/sim.4780131809
22. Hallett TB, Zaba B, Todd J, Lopman B, Mwita W, Biraro S, et al. Estimating incidence from prevalence in generalised HIV epidemics: methods and validation. *PLoS Med.* (2008) 5:e80. doi: 10.1371/journal.pmed.0050080
23. Keiding N. Age-specific incidence and prevalence: a statistical perspective. *J Royal Stat Soc Ser A (Stat Soc)*. (1991) 154:371–96. doi: 10.2307/2983150
24. Kim AA, Hallett T, Stover J, Gouws E, Musinguzi J, Mureithi PK, et al. Estimating HIV incidence among adults in Kenya and Uganda: a systematic comparison of multiple methods. *PLoS ONE.* (2011) 6:e17535. doi: 10.1371/journal.pone.0017535
25. Marschner IC. Fitting a multiplicative incidence model to age-and time-specific prevalence data. *Biometrics.* (1996) 1:492–9. doi: 10.2307/2532889
26. Marschner IC. A method for assessing age-time disease incidence using serial prevalence data. *Biometrics.* (1997) 1:1384–98. doi: 10.2307/2533505
27. Mossong J, Grapsa E, Tanser F, Bärnighausen T, Newell M-L. Modelling HIV incidence and survival from age-specific seroprevalence after antiretroviral treatment scale-up in rural South Africa. *AIDS (London, England)*. (2013) 27:2471. doi: 10.1097/01.aids.0000432475.14992.da
28. Podgor MJ, Leske MC. Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Stat Med.* (1986) 5:573–8. doi: 10.1002/sim.4780050604
29. Hens N, Shkedy Z, Aerts M, Faes C, Van Damme P, Beutels P. Modeling infectious disease parameters based on serological and social contact data: a modern statistical perspective. Berlin, Germany: Springer Science & Business Media (2012).
30. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (2017). Available online at: <https://www.R-project.org>.
31. Bärnighausen T, Tanser F, Gqwede Z, Mbizana C, Herbst K, Newell M-L. High HIV incidence in a community with high HIV prevalence in rural South Africa: findings from a prospective population-based study. *Aids.* (2008) 22:139–44. doi: 10.1097/QAD.0b013e3282f2ef43
32. Hontelez JA, de Vlas SJ, Baltussen R, Newell M-L, Bakker R, Tanser F, et al. The impact of antiretroviral treatment on the age composition of the HIV epidemic in sub-Saharan Africa. *AIDS (London, England)*. (2012) 26:S19–30. doi: 10.1097/QAD.0b013e3283558526
33. Negin J, Mills EJ, Albane R. Continued neglect of ageing of HIV epidemic at UN meeting. *Lancet.* (2011) 378:768. doi: 10.1016/S0140-6736(11)61373-1
34. Tanser F, Bärnighausen T, Grapsa E, Zaidi J, Newell M-L. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science.* (2013) 339:966–71. doi: 10.1126/science.1228160
35. Schaefer R, Gregson S, Eaton JW, Mugurungi O, Rhead R, Takaruzza A, et al. Age-disparate relationships and HIV incidence in adolescent girls and young women: evidence from Zimbabwe. *AIDS.* (2017) 31:1461–70. doi: 10.1097/QAD.0000000000001506
36. Shisana O, Rehle T, Simbayi L, Parker W, Zuma K, Bhana A, et al. *South African national HIV prevalence, HIV incidence, behavior and communication survey, 2005*. Cape Town, South Africa (2005).
37. Gregson S, Nyamukapa CA, Garnett GP, Mason PR, Zhuwau T, Caraël M, et al. Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. *Lancet.* (2002) 359:1896–903. doi: 10.1016/S0140-6736(02)08780-9
38. Leclerc-Madlala S. Age-disparate and intergenerational sex in southern Africa: the dynamics of hypervulnerability. *AIDS.* (2008) 22:S17–25. doi: 10.1097/01.aids.0000341774.86500.53
39. Pettifor AE, Levandowski BA, MacPhail C, Padian NS, Cohen MS, Rees HV. Keep them in school: the importance of education as a protective factor against HIV infection among young South African women. *Int J Epidemiol.* (2008) 37:1266–73. doi: 10.1093/ije/dyn131
40. Hallett T, Gregson S, Lewis J, Lopman B, Garnett GP. Behaviour change in generalised HIV epidemics: impact of reducing cross-generational sex and delaying age at sexual debut. *Sex Transm Infect.* (2007) 83:i50–i4. doi: 10.1136/sti.2006.023606

[com/methodology/survey/survey-display-475.cfm](https://doi.org/10.1371/journal.pdig.0000260)).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

their status, and of those, 95% should be on treatment and of those on treatment 95% should have reached viral suppression. In Zimbabwe, by 2020, it was found that 86.8% of adults living with HIV were aware of their status and of those aware of their status, 97.0% were on antiretroviral treatment. Furthermore, among those on treatment, 90.3% achieved viral load suppression. In order to achieve these targets, modern predictive algorithms using machine learning have the power to enhance HIV prevention and prediction capability. Furthermore, studies have reported that machine learning could accurately predict future HIV infection. This paper, therefore, aims to use these machine learning tools to predict key HIV populations based on socio-behavioural characteristics obtained from survey data.

Introduction

In the era before and after COVID-19, HIV ranks amongst the most serious infectious diseases globally. Despite the tremendous advancements in diagnosis and access to antiretroviral therapy (ART), Zimbabwe Population-based HIV Impact Assessment survey (ZIMPHIA 2020) predictions suggest that 1.23 million adults are living with HIV and that the incidence rate is about 0.21% annually [1]. HIV/AIDS as a threat to the public's health was to be eradicated by 2030, according to the Joint United Nations Programme (UNAIDS) [2,3]. The COVID-19 pandemic, however, is already reversing the gains obtained, and it may have a negative impact by increasing the number of AIDS-related fatalities in sub-Saharan Africa [4,5]. Furthermore, the HIV epidemic is unevenly spread throughout several geographic locations, with higher prevalence in some areas and among particular populations. Focused interventions are necessary for this situation to manage the HIV pandemic effectively since they have been found to maximise the prophylactic benefit at the lowest possible cost [6]. Therefore, it is essential to understand where and among which groups new infections develop to inform targeted interventions.

Monitoring and surveillance techniques, such as behavioural risk assessments, interviews, and laboratory test results, should form the basis of a nation's HIV response [7]. Unfortunately, most nations' surveillance systems, including structured government surveillance and ad hoc surveys, are insufficient to accurately monitor epidemic and risk patterns [7]. In addition, there is frequently a reporting lag before data are made public, and data collecting and aggregation techniques are time- and resource-intensive [8].

Non-requirement of statistical inferences or assumptions is one advantage of machine learning algorithms for developing predictive models. Since machine learning algorithms are data-driven, their greatest benefit is their ability to automatically learn from data that identifies complex nonlinear patterns and exploits complex interactions between risk factors. Machine learning models have been used to predict the future risk of other conditions [9–14], including HIV. Studies have reported that machine learning could accurately predict future HIV infection [2,15–18].

Three primary studies have been published that used machine learning methods to predict HIV status in Zimbabwe [2,19,20]. First, Mutai et al. [2] predicted HIV status using Demographic Health Survey data from Sub-Saharan Africa. The results obtained from the study by Mutai et al. [2] were not specific to Zimbabwe but sub-Saharan Africa. Second, using data from the Zimbabwe Ministry of Health and Child Care, Chingombe et al. [19] predicted HIV status among men who had sex with men in Zimbabwe's two major cities, Bulawayo and Harare. The findings of this study were limited to men who had sex with men in Bulawayo and Harare and could not be generalised to other cities in Zimbabwe or the general population. Third, using

nationally representative data from a cross-sectional survey from the Zimbabwe Population-Based HIV Impact Assessment (ZIMPHIA15-16), again, Chingombe et al. [20] predicted the HIV status using machine learning techniques. Based on the three primary studies that employed machine learning techniques to estimate the HIV status in Zimbabwe, our study and findings will add to the existing body of knowledge. In contrast to Mutai et al.'s [2] research, the findings of our study will be more applicable to Zimbabwe. In addition, unlike the two studies by Chingombe et al. [19,20], we have access to a vast quantity of data from 2005–2015, i.e., ten years of DHS data. Therefore, the primary aim of this paper was to determine common risk factors of HIV positivity in Zimbabwe. Secondly with the aid of machine learning algorithms, these risk factors were used to formulate a model that predicts HIV positivity.

Methodology

Data and study design

Situated in the southern parts of Africa, Zimbabwe is a landlocked nation with a population of 5.1 million people based on the 2022 census [21]. Three nationally representative surveys were held in 2005–06, 2010–11, and 2015 and all named Zimbabwe Demographic Health Survey (ZDHS). The data utilised for analysis in this paper can be obtained from the Demographic Health Survey (DHS) programme website (<https://dhsprogram.com>) [22]. Since 1984, around 70 nations have conducted DHS, which are nationally representative household surveys [23–25]. In the fields of population, health, and nutrition, they offer data for various monitoring and impact evaluation. Accordingly, blood samples were taken in all households with the consent of the respondent or parent/guardian (for minors) for HIV testing in the lab for females aged 0–49 and males aged 0–54. This was a retrospective cross-sectional study design that employed secondary data analysis.

Data pre-processing

The datasets from the three ZDHS HIV test results and the datasets from adult interviews were merged, and records without an HIV test result were excluded from the analysis. The outcome variable, HIV status, was divided into two categories (0 for HIV negative and 1 for HIV positive). To account for non-response, non-coverage, and population total adjustment weights, data were resampled utilising sample weights of HIV test results for each survey year. Variables with a correlation coefficient of 80% or higher, non-unique columns, arbitrary features, more than 30% missing values, and uninformative features were all eliminated. The label-code and one-hot encoding methods were used to encode nominal and ordinal variables derived from survey data. As part of step 1, represented in Fig 1, missing values were imputed using Multiple Imputations with Chained Equations (MICE), and in each of these categories, we assumed that missing was at random. The data were further normalised and scaled.

Train, test and validation procedure

The data were stratified by sex, and subsequent data training, testing and validation were done following the stratification. Following step 1 from Fig 1, one survey year was left out, with all the survey years being rotated and left out to produce three distinct datasets for each sex, each with only two surveys. The primary goal of the rotation was to evaluate the generalisation of the models separately for males and females in later testing and validation. Of the three newly constructed datasets for each sex as shown in Fig 1 step 1, 80% were selected for training, and 20% were utilised as test and validation samples. To prevent the test dataset from being contaminated, MICE imputation and data standardisation were carried out independently for the training and test datasets.

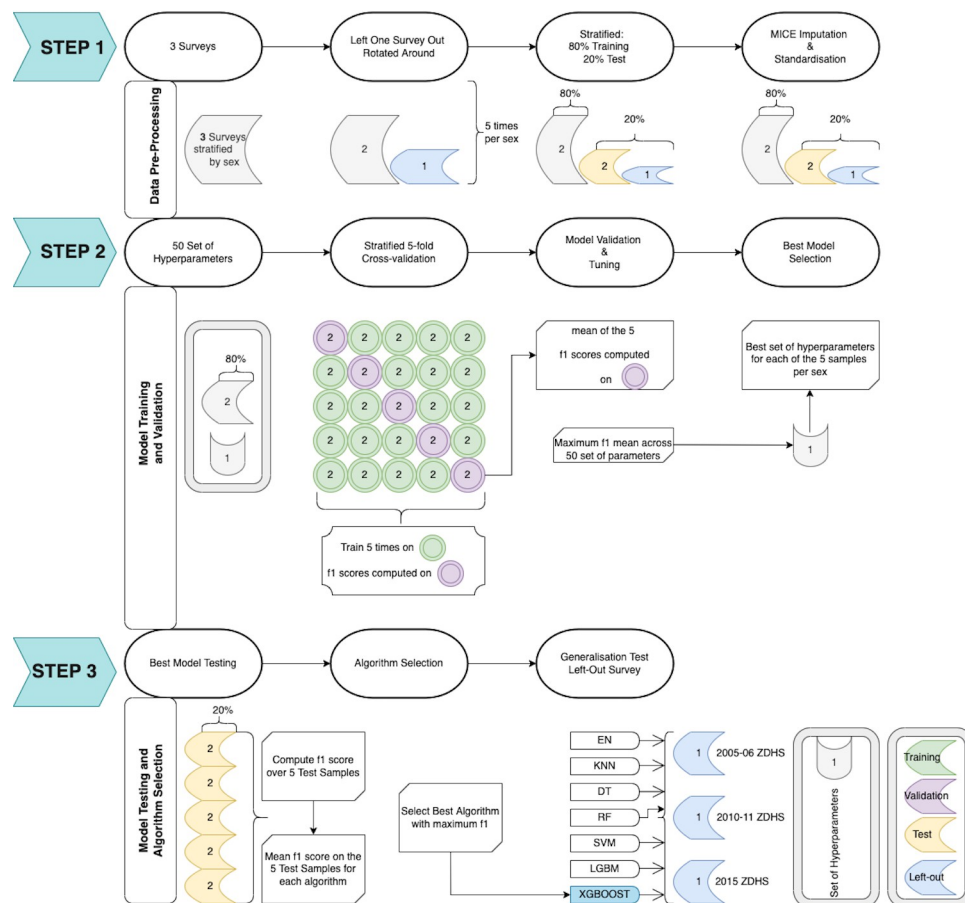


Fig 1. Flowchart of the methods and steps used.

<https://doi.org/10.1371/journal.pdig.0000260.g001>

Feature selection

Feature selection was done in two stages. We first used LASSO (Least Absolute Shrinkage and Selection Operator) regression to select features important in determining HIV status using combined dataset from the three surveys. Features selected through LASSO regression were then used to fit the algorithms under investigation. On variables selected through LASSO regression, subsequent analysis utilised the method with the algorithm with the highest F1 score and analysis was done separately for each survey year stratified by sex. We utilised the sequential forward floating selection (SFFS) approach with 80% of the training samples to select the final features to use in best algorithm among the six algorithms. Feature selection using the SFFS was implemented by considering features whose F1 score plateaued from the saturation point. Using SHapley Additive exPlanations (SHAP) [26], the contribution of each feature selected through SFFS to the probability of being HIV-positive was then examined.

Data balancing

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. This was achieved by generating synthetic data using the nearest neighbour's algorithm to balance out the classes.

Original dataset without implementing the SMOTE procedure was also analysed to compare results.

Machine learning models

A supervised machine learning binary classification was implemented. The following machine learning algorithms were compared: a penalised logistic regression (Elastic Net), k-nearest neighbour algorithm (KNN), Random Forest Classifier (RFC), Decision Tree approach (DT), Light Gradient Boosting Model (LightGBM) and the XGBoost model. To achieve step 2, training datasets were used, and hyperparameter tuning from a grid of 50 sets was performed using a randomised grid search. Then, using the stratified five-fold cross-validation technique over the validated samples, the average F1 scores for each of these 50 sets was calculated, and the most powerful set of hyperparameters was chosen.

Each of the best three models (for each survey year) by sex and algorithm was then run on the corresponding test dataset described in Step 3, Fig 1, and the resulting metric scores were averaged. Next, the algorithm with the best average F1 score was selected from the three test datasets. Finally, each selected model was applied to the survey left-out dataset.

Algorithm evaluation

Due to the unequal class distribution of the classification variable with a ratio of 4.2:1 in the original dataset, the Precision-Recall curve is recommended over the Receiver Operating Characteristic (ROC) curve [27] due to its insensitivity to imbalanced datasets. Other metrics were also considered to evaluate the algorithm's performance: accuracy, F1, precision and recall scores. To determine the ratio of correct classification, accuracy score was used. To determine the prediction of true positive cases, precision was used. To determine the proportion of positive cases that are correctly predicted, recall score was used. Lastly, the F1 score which is the weighted harmonic mean of recall and precision was used to determine the predictive power of the algorithm. An algorithm with the highest F1 score was considered to have the best predictive power. Finally, each evaluation used the same segmentation and repetition of data to ensure a fair comparison of models.

Statistical analysis

The features that were chosen using the SFFS XGBoost procedure were subjected to a logistic regression analysis. This step was added to confirm the outcomes that the SHAP plots yielded. In addition to the results of the SHAP plots, the logistic regression also provided the magnitude and direction of the risk of contracting HIV.

Results

Summary statistics

The overall HIV prevalence for the combined dataset, as shown in Table 1, was 22.5% and 15.3% for females and males, respectively. Those who were not currently employed had a higher prevalence (54.1%) than those who were currently employed (45.9%) for females. However, the opposite relationship is seen in males, where those currently employed had a higher HIV prevalence (74.2%) than those not currently employed (25.8%). Adults had a higher HIV prevalence than youths for both sexes. Of interest was a higher HIV prevalence noticed in those who had the highest level of education as secondary for both sexes.

Table 1. Background characteristics of ZDHS data.

	FEMALES	MALES
Characteristic	HIV Positive n(%)	HIV Positive n(%)
Overall summary	4457(22.5)	2215(15.3)
Survey year: 2005–06	1604(36.0)	760(34.3)
2010–11	1332(29.9)	733(33.1)
2015	1521(34.1)	722(32.6)
Current age: youths (15–24 years)	730(16.4)	179(8.1)
Adults (25–54 years)	3727(83.6)	2036(91.9)
Marital Status: single	302(6.8)	183(8.3)
cohabiting/married	2595(58.2)	1723(77.8)
Divorced/separated/widowed	1560(35.0)	309(14.0)
Place of Residence: urban	1625(36.5)	877(39.6)
rural	2832(63.5)	1338(60.4)
Education level: no education	123(2.8)	23(1.0)
primary	1510(33.9)	612(27.6)
secondary	2674(60.0)	1431(64.6)
higher	150(3.4)	149(6.7)
STI treatment in past 12 months: no	4127(92.7)	2102(94.9)
yes	325(7.3)	113(5.1)
Number of unions: once	2942(70.9)	1059(52.2)
more than once	1210(29.1)	970(47.8)
Currently employed: no	2410(54.1)	571(25.8)
yes	2047(45.9)	1644(74.2)
Wealth index: poorest	781(17.5)	366(16.5)
poorer	727(16.3)	411(18.6)
middle	929(20.8)	350(15.8)
richer	1198(26.9)	603(27.2)
richest	822(18.4)	485(21.9)

<https://doi.org/10.1371/journal.pdig.0000260.t001>

Feature selection

LASSO regression was performed to determine the possible predictors of HIV infection. A total of 32 predictors were considered for both males and females (see [S1 Table](#)). Out of the 32 predictors, type of place of residence, religion, number of household members, relationship to household head, sex of household head, age of household head, times away from home in last 12 months, wealth index, current contraceptive method, current contraceptive by method type, currently/formerly/never in union, number of unions, cohabitation duration (grouped), time since last sex (in days), recent sexual activity, beating justified, ever heard of AIDS, reduce risk of getting HIV, had any STI in last 12 months, ever been tested for HIV, wife justified asking husband to use condom if he has STI, total lifetime number of sex partners, can ask partner to use condom, were selected as predictors of HIV acquisition for females. While, type of place of residence, highest educational level, religion, number of household members, age of household head, times away from home in last 12 months, current contraceptive method, number of injections in last 12 months, respondent circumcised, currently/formerly/never in union, cohabitation duration (grouped), currently working, beating justified, reduce risk of getting HIV, had any STI in last 12 months, wife justified asking husband to use condom if he has STI, total lifetime number of sex partners, were selected as predictors of HIV acquisition for males.

Algorithm comparison

Model/algorithm's performance comparison was made based on the LASSO selected features. The comparison was done in two-phases, 1) using the original data and 2) using the SMOTE processed data. Table 2 describes the original and SMOTE data samples for the training, testing and validation sets for the three survey years stratified by sex.

Six algorithms were trained and tested using year-specific datasets on the original and SMOTE processed data. Table 3 gives results of the performance of the prediction capabilities of the algorithms. The SMOTE processed data performed better in all metrics than the original data. The precision and the F1 score were extremely low for most of the algorithms in the original data. This means that the original data had many false positives from the low precision and low accuracy based on the low F1 score. Overall, the XGBoost was best performing algorithm in both the original data and SMOTE processed data for all the survey year for each sex.

ROC curves and Precision-Recall curves of the six algorithms for all survey years per sex were shown in Figs 2 and 3. The results showed that the XGBoost model had better performance compared to the other 5 models for both the original data and SMOTE processed data. The Elastic Net model performed poorly compared to the other models. Additional ROC and Precision-Recall curves for the six algorithms based on combined survey years can be found in S1 Fig and S2 Fig.

Subsequent results were obtained utilising the best algorithm model which was the XGBoost. From the features obtained through LASSO, SFFS procedure was implemented on XGBoost to further determine features associated with HIV positivity. Fig 4 shows the saturation limit obtained by the SFFS procedure for selecting variables based on F1 scoring, with 17, 17 and 14 most influential features selected for females, in the 2005–06, 2010–11 and 2015 ZDHS respectively. Whereas, 13, 13, 11 most influential features were selected for males, in the 2005–06, 2010–11 and 2015 ZDHS respectively. The F1 score plateaued at 99.9% for both males and females for all survey years.

Table 2. Description of original data and SMOTE-processed data.

		MALES DATA				FEMALES DATA			
Dataset	Survey Year	Majority class	Minority Class	Total Samples	Ratio	Majority class	Minority Class	Total Samples	Ratio
Training	2005–06	7124	1164	8288	6:1	8809	2282	11091	4:1
	2010–11	6684	1186	7870	5:1	8176	2500	10676	3:1
	2015	5739	1194	6933	5:1	7583	2349	9932	3:1
Training-smote	2005–06	7124	7124	14248	1:1	8809	8809	17618	1:1
	2010–11	6684	6684	13368	1:1	8176	8176	16352	1:1
	2015	5739	5739	11478	1:1	7583	7583	15166	1:1
Testing	2005–06	1782	291	2073	6:1	2202	571	2773	4:1
	2010–11	1672	296	1968	5:1	2045	625	2670	3:1
	2015	1435	299	1734	5:1	1897	587	2484	3:1
Testing-smote	2005–06	1782	1782	3564	1:1	2202	2202	4404	1:1
	2010–11	1672	1672	3344	1:1	2045	2045	4090	1:1
	2015	1435	1435	2870	1:1	1897	1897	3794	1:1
Validation	2005–06	3312	760	4072	4:1	4345	1604	5949	3:1
	2010–11	3862	733	4595	5:1	5135	1332	6467	4:1
	2015	5044	722	5766	7:1	5876	1521	7397	4:1
Validation-smote	2005–06	3312	3312	6624	1:1	4345	4345	8690	1:1
	2010–11	3862	3862	7724	1:1	5135	5135	10270	1:1
	2015	5044	5044	10088	1:1	5876	5876	11752	1:1

<https://doi.org/10.1371/journal.pdig.0000260.t002>

Table 3. Results of classification models in original imbalanced and SMOTE-processed data.

		SMOTE-processed data				Original unbalanced data			
	MALES								
Survey Year	Models	Accuracy %	Precision %	Recall %	F-Score %	Accuracy %	Precision %	Recall %	F-Score %
2005–06	DT	83	87%	78	83	65	25	77	38
2010–11		82	88	75	81	67	29	80	42
2015		81	83	77	80	68	32	74	44
2005–06	EN	64	59	98	71	63	22	65	33
2010–11		64	59	95	72	60	23	69	34
2015		63	59	80	68	57	25	73	37
2005–06	KNN	86	92	78	84	93	79	65	72
2010–11		86	91	80	85	93	81	69	75
2015		82	88	75	81	90	77	63	77
2005–06	RFC	83	79	90	84	70	29	78	42
2010–11		81	78	87	82	66	28	80	42
2015		83	80	86	83	69	32	71	44
2005–06	LIGHTGBM	94	98	90	93	84	45	77	57
2010–11		94	98	89	94	82	45	75	56
2015		92	98	86	91	81	46	71	56
2005–06	XGBOOST	95	95	96	95	93	75	74	74
2010–11		94	95	94	94	92	71	76	73
2015		94	93	94	94	90	72	71	71
	FEMALES								
2005–06	DT	83	86	79	82	72	40	74	52
2010–11		81	84	76	80	70	42	72	53
2015		81	83	79	81	72	45	70	54
2005–06	EN	63	58	88	70	72	39	62	48
2010–11		61	58	80	67	67	38	60	46
2015		59	55	90	69	67	38	63	47
2005–06	KNN	83	90	74	81	89	79	66	72
2010–11		83	89	75	81	89	80	68	74
2015		81	88	72	80	89	81	68	74
2005–06	RFC	84	84	84	84	75	44	70	54
2010–11		81	81	81	81	73	45	68	54
2015		83	83	83	83	75	47	70	56
2005–06	LIGHTGBM	91	96	86	91	82	55	73	63
2010–11		90	94	84	89	79	54	72	62
2015		90	95	85	89	79	54	72	61
2005–06	XGBOOST	93	93	93	93	90	74	76	75
2010–11		92	92	93	92	88	73	77	75
2015		92	92	92	92	88	75	76	76

<https://doi.org/10.1371/journal.pdig.0000260.t003>

Figs 5 and 6 displays the subset of features most important to predicting an individual's HIV status following the SFFS procedure. As seen in Figs 5 and 6, the variables are ordered from most important to least important (from highest to lowest Shapley value). According to Fig 5, features for predicting HIV status which appeared throughout for the three survey years for females were: total lifetime number of sex partners, cohabitation duration (grouped), number of household members, age of household head, time since last sex (in days), times away

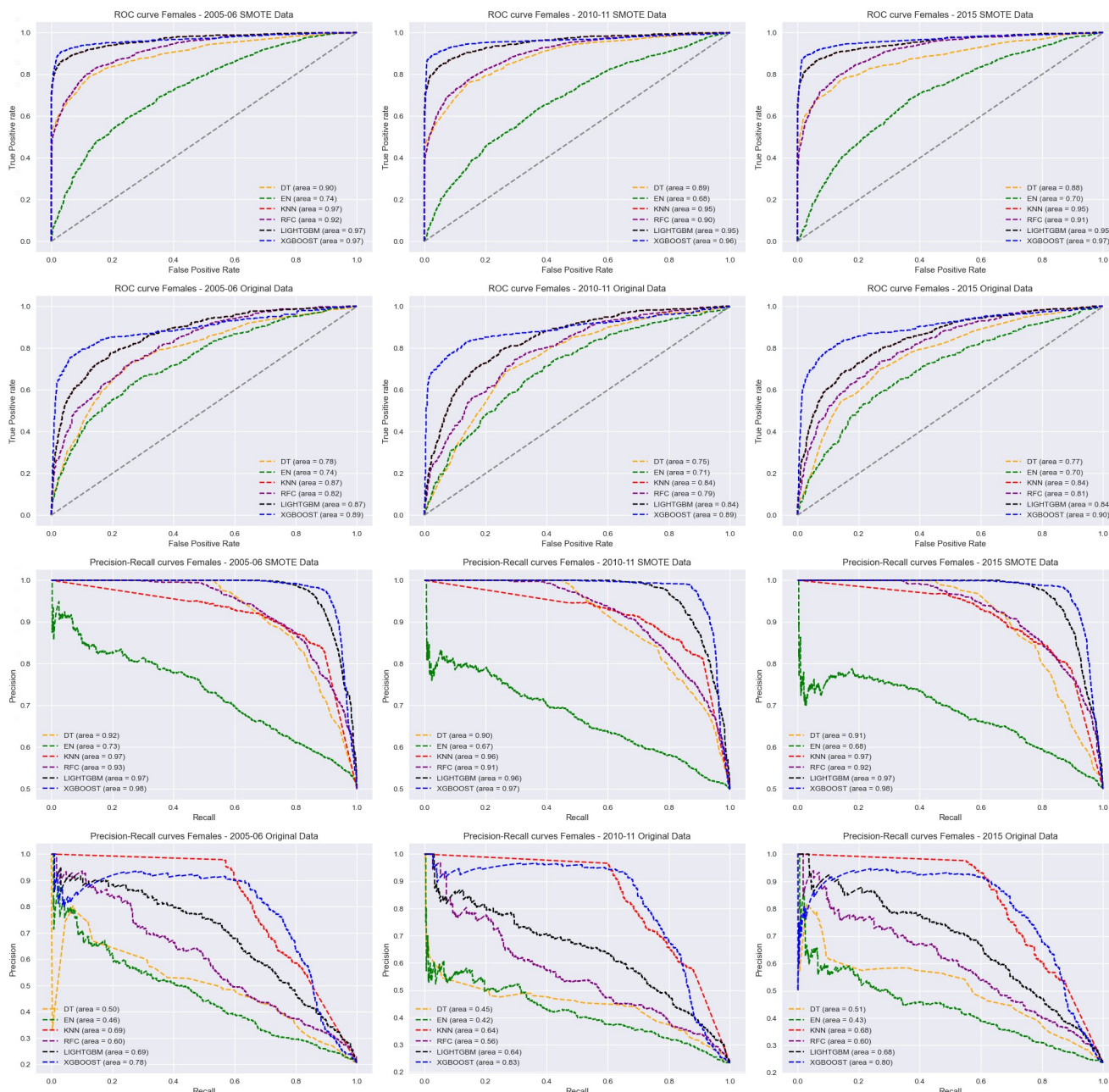


Fig 2. ROC curves (1st and 2nd row) and Precision-Recall curves (3rd and 4th row) of the six algorithms for 2005-06 (1st column), 2010-11 (2nd column) and 2015 (3rd column) ZDHS survey for females.

<https://doi.org/10.1371/journal.pdig.0000260.g002>

from home in last 12 months, beating justified, can ask partner to use condom, wealth index, reduce risk of getting HIV, wife justified asking husband to use condom if he has STI and religion.

Features for predicting HIV status which appeared throughout for the three survey years for males were: cohabitation duration (grouped), total lifetime number of sex partners, number of household members, age of household head, times away from home in last 12 months, type of place of residence, beating justified, currently working and religion. Identical variables for both sexes throughout the three survey years for predicting HIV status were: total lifetime

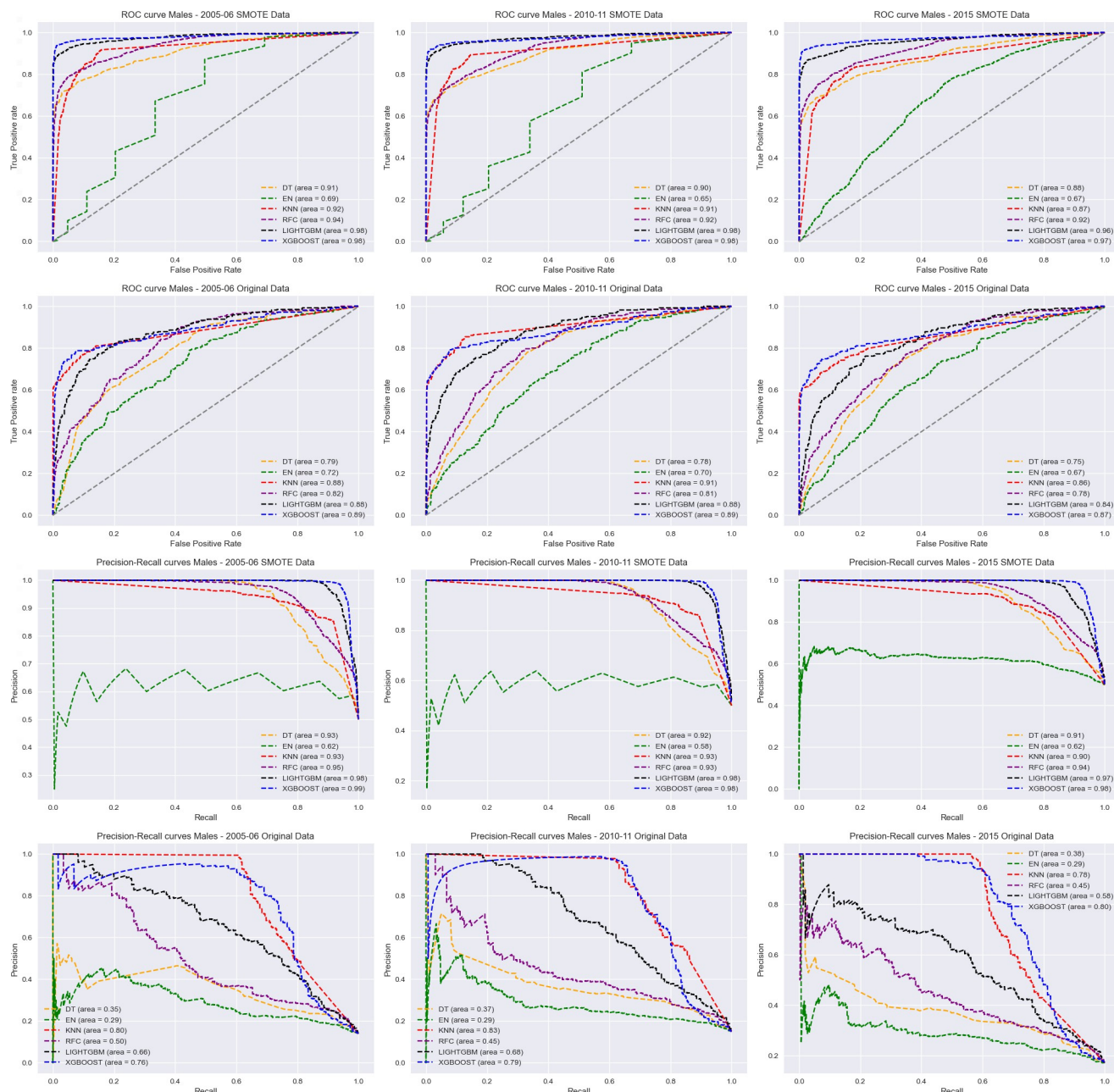


Fig 3. ROC curves (1st and 2nd row) and Precision-Recall curves (3rd and 4th row) of the six algorithms for 2005–06 (1st column), 2010–11 (2nd column) and 2015 (3rd column) ZDHS survey for males.

<https://doi.org/10.1371/journal.pdig.0000260.g003>

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).

A higher number of total lifetime number of sex partners, more years cohabitating, fewer household members, older age of household head, higher number of days since last sex, fewer times away from home, weather beating was justified or not, living in the rural areas, wearing a condom as current contraceptive method, having a female household head, many unions, can

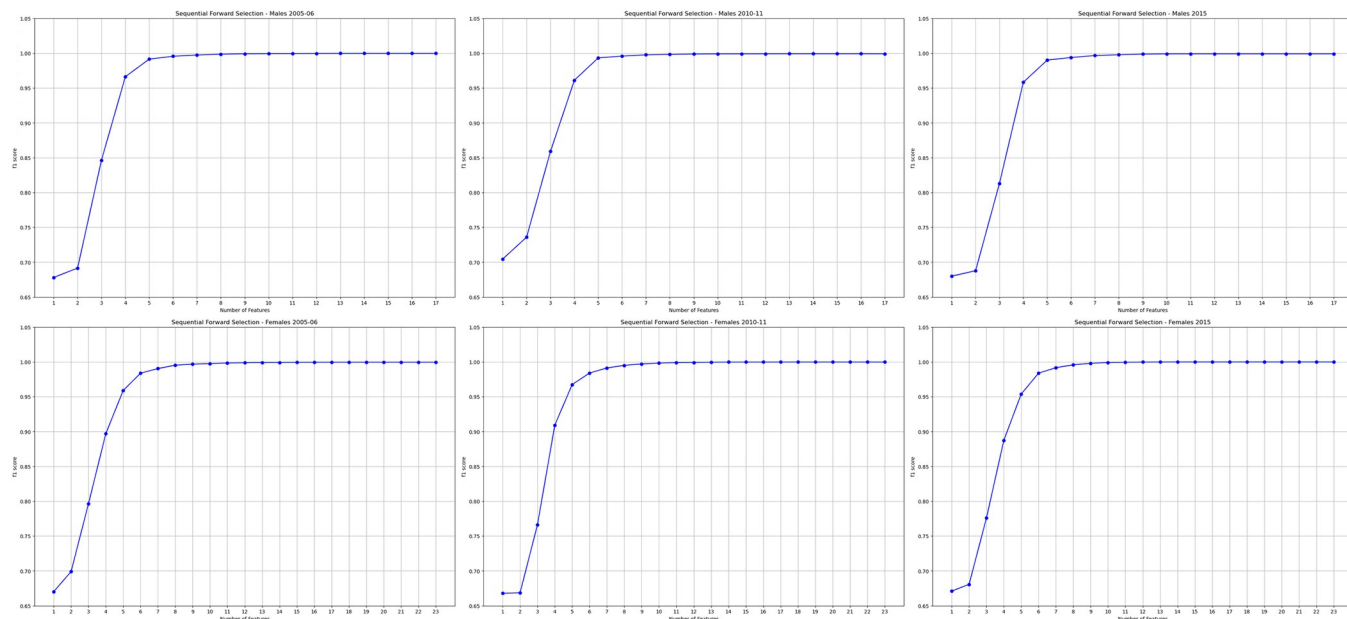


Fig 4. Sequential floating forward selection (SFFS) for males and females.

<https://doi.org/10.1371/journal.pdig.0000260.g004>

ask partner to use condom, higher wealth index, whether knowing the risk of getting HIV or not, ever been tested, wife justified asking husband to use condom if he has STI and being affiliated to a religion were factors associated with HIV positivity for females in the 2005–06 ZDHS survey. A higher number of total lifetime number of sex partners, more years cohabitating, fewer household members, higher number of days since last sex, younger age of household head, fewer times away from home, never been tested, weather beating was justified or not, living in the rural areas, not knowing the risk of getting HIV, having a female household head, can ask partner to use condom, wearing a condom as current contraceptive method, relationship to household head, higher wealth index, not affiliated to a religion and wife justified asking husband to use condom if he has STI were factors associated with HIV positivity for females in the 2010–11 ZDHS survey. A higher number of total lifetime number of sex

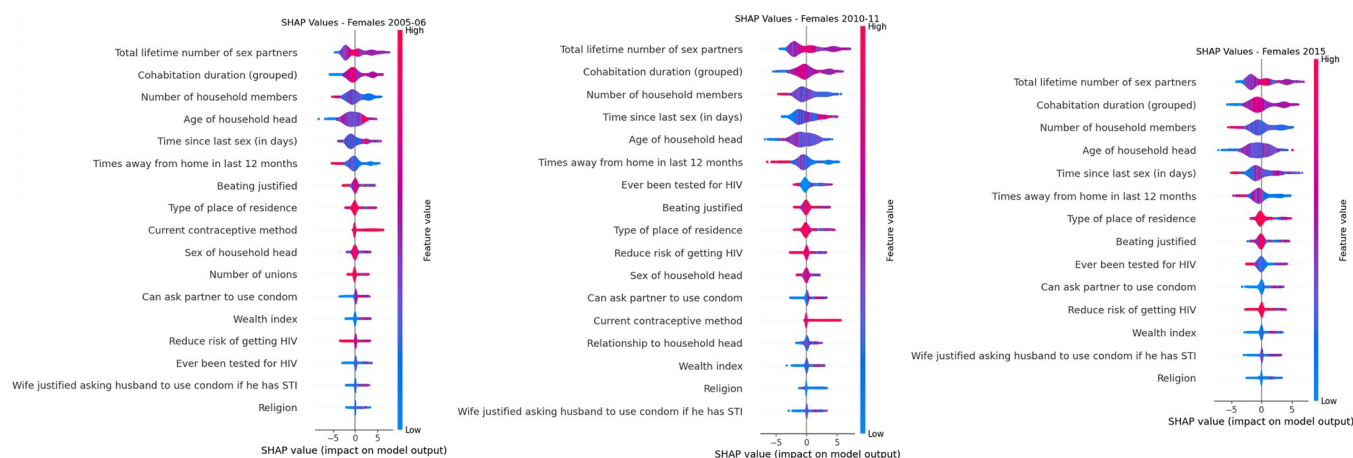


Fig 5. Shapley values for 2005–06, 2010–11 and 2015 Females ZDHS data.

<https://doi.org/10.1371/journal.pdig.0000260.g005>

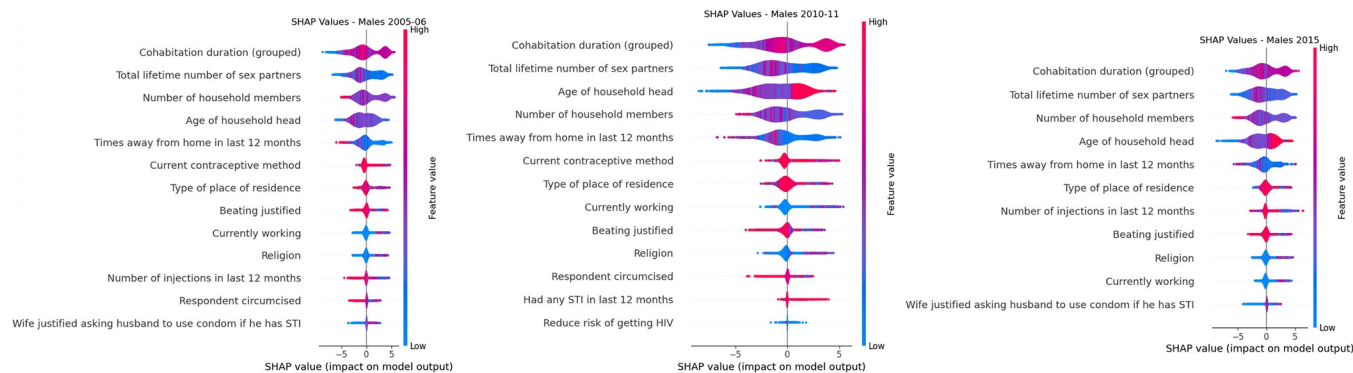


Fig 6. Shapley values for 2005–06, 2010–11 and 2015 Males ZDHS data.

<https://doi.org/10.1371/journal.pdig.0000260.g006>

partners, more years cohabitating, fewer household members, younger age of household head, fewer number of days since last sex, fewer times away from home, living in the rural areas, beating was justified, never been tested, can ask partner to use condom, not knowing the risk of getting HIV, higher wealth index and wife justified asking husband to use condom if he has STI were factors associated with HIV positivity for females in the 2015 ZDHS survey.

Based on Fig 6, a higher number of years cohabitating, fewer number of total lifetime number of sex partners, fewer household members, younger age of household head, fewer times away from home, not wearing a condom as current contraceptive method, living in the urban areas, beating was justified, currently working, being affiliated to a religion, less number of injections in the last 12 months, respondent not circumcised and wife justified asking husband to use condom if he has STI and were factors associated with HIV positivity for males in the 2005–06 ZDHS survey. A higher number of years cohabitating, fewer number of total lifetime number of sex partners, fewer household members, older age of household head, fewer times away from home, not wearing a condom as current contraceptive method, living in the urban areas, beating was justified, being affiliated to a religion, respondent not circumcised and had an STI in the last 12 months were factors associated with HIV positivity for males in the 2010–11 ZDHS survey. A higher number of years cohabitating, fewer number of total lifetime number of sex partners, fewer household members, older age of household head, fewer times away from home, living in the urban areas, fewer number of injections in the last 12 months, beating not justified, being affiliated to a religion, currently working and wife justified asking husband to use condom if he has STI were factors associated with HIV positivity for males in the 2015 ZDHS survey.

Table 4 shows the odds ratio of HIV infection performed on the selected features obtained through SFFS for females. The results in Table 4 agree with the results shown in Fig 5. For instance, each additional increase in the number of sex partners is associated with a 32% (OR: 1.32, $p < 0.001$), 11% (OR: 1.11, $p = 0.04$) and 41% (OR: 1.11, $p < 0.001$) increase in the odds of one being HIV infected for 2005–06, 2010–11 and 2015 ZDHS females data, which is corroborated by Fig 5 were total lifetime number of sex partners is associated with HIV positivity as the number of total lifetime sex partners increases for females in all survey years data. Table 4 also shows a protective effect of HIV infection in individuals who live in rural areas than those who reside in urban areas. This, again, is corroborated by Fig 5, were living in urban areas is associated with HIV positivity.

Interestingly, for the 2005–06 and 2010–11 survey data for females, individuals who have been cohabiting for 5–19 years were at higher risk of HIV infection than those who never cohabited. On the other hand, individuals who have more than 19 years of cohabitation have a

Table 4. Odds Ratio of female features selected in XGBoost using SFFS procedure.

Survey	2005–06		2010–11		2015	
Variable	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
Total lifetime number of sex partners	1.315912	<0.001	1.109939	0.037	1.410084	<0.001
Cohabitation duration (grouped)						
Never Married	ref		ref		ref	
5–9	1.340407	0.011	1.287969	0.034	1.474305	0.002
10–14	1.390041	0.012	1.459357	0.004	1.673866	<0.001
15–19	1.333851	0.032	1.425452	0.018	2.222664	<0.001
20–24	0.8778976	0.046	0.8974349	0.471	1.840237	<0.001
25–29	0.5542526	0.003	0.718883	0.011	1.408705	0.041
30+	0.4943325	0.031	0.5959646	0.048	1.368872	0.031
Number of household members	0.9596885	0.02	0.880098	<0.001	0.8523343	<0.001
Age of household head	1.105237	0.012	1.126954	<0.001	1.225387	<0.001
Time since last sex (in days)	1.002203	<0.001	1.000348	0.287	1.00133	0.001
Intimate Partner Violence (IPV)						
no	ref		ref		ref	
yes	1.5837278	0.153	1.614516	0.0466	1.9680338	0.0298
Type of place of residence						
urban	ref		ref		ref	
rural	0.8375157	0.251	0.6298531	<0.001	0.6740229	0.013
Current contraceptive method						
condom use	ref		ref			
No condom use	2.021359	0.011	5.635275	<0.001		
Sex of household head						
male	ref		ref			
female	0.9317121	0.466	0.9885284	0.955		
Number of unions						
once	ref					
more than once	2.259846	<0.001				
Can ask partner to use condom						
no	ref		ref		ref	
yes	1.042492	0.671	1.294777	0.007	2.029347	<0.001
Wealth index						
poor	ref		ref		ref	
poorer	1.315844	0.032	0.8705808	0.285	0.9036212	0.441
middle	1.171341	0.245	0.94288	0.652	1.054575	0.685
richer	1.432668	0.022	0.8276257	0.201	0.7554493	0.092
richest	0.9685689	0.878	0.5113931	<0.001	0.5450553	0.002
Reduce risk of getting HIV						
no	ref		ref		ref	
yes	0.917144	0.33	0.8602807	0.018	0.8791833	0.025
Ever been tested for HIV						
no	ref		ref		ref	
yes	1.060744	0.509	1.112168	0.273	1.498843	0.045
Wife justified asking husband to use condom if he has STI						
no	ref		ref		ref	
yes	0.8672589	0.275	1.159646	0.22	1.029157	0.843
Religion						

(Continued)

Table 4. (Continued)

Survey	2005–06		2010–11		2015	
Variable	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
No religion						
Christian	0.995277	0.974	1.033	0.844	0.7605706	0.101
Apostolic	0.8597532	0.313	0.9342901	0.681	0.7902854	0.156
Other	0.5309193	0.025	0.6997551	0.378	0.5492018	0.162
Times away from home in last 12 months			0.999514	0.916	0.9919302	0.06
Relationship to household head						
head			ref			
wife			0.7331381	0.166		
daughter			0.7313408	0.283		
daughter-in-law			0.4126077	0.001		
grand-daughter			1.098993	0.864		
sister			1.388966	0.453		
other relative			1.032241	0.904		
not related			1.122641	0.893		

<https://doi.org/10.1371/journal.pdig.0000260.t004>

protective effect of HIV infection than those who have not cohabited. However, this phenomenon changes in the 2015 survey data for females as all individuals who are cohabiting are more at risk of HIV infection than those who are not. HIV positivity was associated with whether one had the knowledge of reducing the risk of HIV transmission or not, according to Fig 5, 2005–06 survey data for females. However, for the 2010–11 and 2015 surveys, HIV positivity was associated with not having any knowledge of reducing the risk of getting HIV. This is similar to the logistic regression in Table 4 results, where having knowledge of reducing the risk of HIV transmission was a protective effect against HIV infection.

Table 5 shows the odds ratio of HIV infection performed on the selected features obtained through SFFS for males. Males who are currently working are less likely to be HIV infected compared to those not working with an odds ratio (OR): 0.75 ($p = 0.01$) and (OR): 0.71 ($p < 0.001$) for the 2005–06 and 2015 survey years, respectively. If there was intimate partner violence (IPV), the risk of HIV infection was two times more than when there was no intimate partner violence for both males and females, as shown in Tables 4 and 5. As the age of household head increases, the risk of infection within the household increases by 11–20% between 2005–2015 for males and females, as shown in Tables 4 and 5. Males cohabiting are more than 3–9 times at risk of HIV infection than those not cohabiting. This is true for all survey years; as shown in Table 5, the risk increases over the years. The risk of HIV transmission being higher in males is also exhibited in Fig 6, which shows that cohabitation is the most influential feature for males in all survey years associated with HIV positivity.

Discussion

Although there have been a few studies utilising machine learning techniques to predict HIV in the generalised HIV pandemic, this is, to our knowledge, the first one in Zimbabwe using routinely collected survey data. This study's primary goal was to determine the most prevalent risk factors for HIV infection and the predicted accuracy of machine learning models based on these risk factors.

By using socio-demographic factors obtained from three ZDHS, this study was able to predict the HIV status of individuals. After comparison of different algorithms, XGBoost was the

Table 5. Odds Ratio of male features selected in XGBoost using SFFS procedure.

Survey	2005–06		2010–11		2015	
Variable	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
Cohabitation duration (grouped)						
Never Married	ref					
0–4	3.347549	<0.001	4.232994	<0.001	4.377312	<0.001
5–9	5.874514	<0.001	7.047588	<0.001	6.530351	<0.001
10–14	7.124908	<0.001	7.865179	<0.001	7.520853	<0.001
15–19	6.792356	<0.001	9.919705	<0.001	8.423026	<0.001
20–24	7.735013	<0.001	9.479581	<0.001	9.519249	<0.001
25–29	3.733829	<0.001	7.960904	<0.001	7.066936	<0.001
30+	6.441015	<0.001	5.548076	<0.001	6.918233	<0.001
Total lifetime number of sex partners	1.012441	0.002	1.013705	<0.001	1.014312	<0.001
Number of household members	0.9135443	<0.001	0.9221065	<0.001	0.9093683	<0.001
Age of household head	1.114831	<0.001	1.117423	<0.001	1.200836	<0.001
Times away from home in last 12 months	1.002773	0.425	1.000984	0.747	0.9966006	0.234
Current contraceptive method						
condom use	ref		ref			
no condom use	1.237349	0.155	2.370684	<0.001		
Type of place of residence						
urban	ref		ref		ref	
rural	0.9331089	0.484	0.9045717	0.324	0.871984	0.119
Intimate Partner Violence (IPV)						
no	ref		ref		ref	
yes	1.613896	0.288	1.882546	0.0236	1.780742	0.01
Currently working						
no	ref		ref		ref	
yes	0.7523402	0.006	0.959925	0.678	0.7097387	<0.001
Religion						
No religion	ref		ref		ref	
Christian	0.7597485	0.01	0.8094892	0.055	0.7136371	0.002
Apostolic	0.710367	0.008	0.8097574	0.081	0.8003416	0.048
Other	0.9252984	0.592	0.9331523	0.703	1.322868	0.126
Number of injections in last 12 months	1.170525	<0.001			1.007539	0.252
Respondent circumcised						
no	ref		ref			
yes	1.156246	0.258	0.9399439	0.663		
Wife justified asking husband to use condom if he has STI						
no	ref				ref	
yes	1.564861	0.003			1.424217	0.017
Had any STI in last 12 months						
no			ref			
yes			2.305245	<0.001		
Reduce risk of getting HIV						
no			ref			
yes			0.8435852	0.167		

<https://doi.org/10.1371/journal.pdig.0000260.t005>

best algorithm to predict HIV status. The algorithm was able to determine the most predictive features/variables of HIV infection common to both sexes. Using SHAP plots, this study additionally evaluated the direction of the relationship between HIV infection and the predictive variables.

The total number of sexual lifetime partners was the main influential feature for females, and cohabitation duration for males in all survey years. Previous studies have shown that the total number of sexual partners is a risk factor for HIV infection [28–31]. This was in agreement with our results which indicated that the higher the total number of sexual partners an individual has, the more at risk of HIV infection. To add to the same evidence, a study by Armstrong et al. [32] indicated that sexual partner number was an important HIV risk measure. Our findings also showed that cohabitation was a risk factor of HIV positivity. This finding can be corroborated by other studies which reported that most heterosexual rates of HIV transmission take place within cohabiting or married couples [33–37].

Numerous studies have shown that child-head families are more at risk of HIV infection [38–41]. This is contrary to our findings, as our results indicated that households headed by older individuals are more at risk of HIV infection. This may be attributed to the fact that the older the household head is, the less financially stable they are, hence their only source of income for that family might be from their pension fund, which might not be adequate, resulting in the other household members seeking financial assistance from risky behaviours [42,43].

In addition, living in urban areas was found to increase the probability of HIV positivity. This was corroborated by other studies [44–46], which reported that people in urban settings had a greater chance of contracting HIV than people in rural areas.

Our results indicated a higher risk of HIV infection if there was pro-intimate partner violence. To support this evidence, research done among ever-married and cohabiting women in Zimbabwe by Henderson et al. [47] studied the relationship between intimate partner violence and HIV status, and they found out that women who had been victims of any kind of intimate partner violence were more likely to be HIV positive. The study also commented that the patriarchal and hypermasculist culture in Zimbabwe contributed to the likelihood of HIV infection. Further research utilising the ZDHS 2005–06 indicated that approximately six out of every ten women who reported experiencing some form of violence in their lifetimes had a much higher risk of being HIV-positive than women who had not suffered any physical or sexual abuse [48].

To ensure that a substantial percentage of persons tested are HIV positive, a high yield is necessary when targeted HIV case-finding procedures are used to boost testing's cost-effectiveness. The potential for further behavioural-based case-finding techniques to improve or supplement current focused case-finding techniques like index testing is still unknown. For some limited resource conditions and the planned test coverage, it may be necessary to modify the acceptable cut-offs for sensitivity and PPV thresholds.

By better identifying those at high risk of contracting HIV, machine learning algorithms have the potential to enhance the implementation of pre-exposure prophylaxis (PrEP). As part of an inclusive approach to PrEP, which has gone from development to implementation, programs can utilise these algorithms to spark discussions about PrEP. However, the limited number of variables accessible to us for this study constituted one of its drawbacks. We were unable to determine the impact of variables, including viral load, health care spending, HIV-risk groups, and other HIV-related interventions. The data also contained missing values, which necessitated making assumptions about their unpredictable nature and applying intrinsically flawed imputation techniques. Finally, several variables were self-reported and, as a result, were vulnerable to recall bias and social desirability.

Our technique for predictor identification primarily adapts from Orel's [49] and Mutai's [2] methodology, which also chose the XGBoost algorithm as the best. Contrarily, our findings demonstrate distinct predictors from those discovered in Orel's and Mutai's study. Considering that our study was more specific to Zimbabwe, this might have been the reason why results from Mutai's study were not similar to our findings. Additionally, Orel's population was based on Eastern Africa, which might have different epidemiological factors from the sub-Saharan Africa, therefore making the results statistically incomparable.

Conclusion

Our findings may help with social-behavioural HIV detection and improve screening procedures in limited resource settings. In addition, the amount of information needed to identify key populations in Zimbabwe will be significantly reduced by features/variables which have been identified in this study through machine learning. Adaptation of HIV screening methods that more effectively target the adult population, those with multiple partners, those who are frequently away from home, those who reside in urban areas, those who are not currently working and other risk factors associated with HIV positivity are needed. Programmes targeted at HIV testing could incorporate machine learning approaches to adequately and effectively identify high-risk individuals. However, to improve the machine learning approach, further research is required to integrate and implement them in a real-world primary care context. In addition to other risk reduction techniques, machine learning may aid in identifying those who might require PrEP.

Supporting information

S1 Table. Male and Female predictors variables considered in variable selection.
(DOCX)

S1 Fig. ROC Curves for SMOTE and ORIGINAL data of males and females combined survey years data.
(TIF)

S2 Fig. Precision-Recall Curves for SMOTE and ORIGINAL data of males and females combined survey years data.
(TIF)

Author Contributions

Conceptualization: Rutendo Beauty Birri Makota.

Data curation: Rutendo Beauty Birri Makota.

Formal analysis: Rutendo Beauty Birri Makota.

Methodology: Rutendo Beauty Birri Makota.

Resources: Eustasius Musenge.

Supervision: Eustasius Musenge.

Writing – original draft: Rutendo Beauty Birri Makota.

Writing – review & editing: Rutendo Beauty Birri Makota, Eustasius Musenge.

References

1. Population-based HIV Impact Project. Zimbabwe Population-based HIV Impact Assessment. ICAP at Columbia University, 2020.
2. Mutai C, McSharry P, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. *BMC medical research methodology*. 2021; 21(1):1–11.
3. UNAIDS. Fast-track commitments to end AIDS by 2030. UNAIDS Geneva, Switzerland; 2016.
4. Jewell B, Mudimu E, Stover J, Ten Brink D, Phillips A, Smith J, et al. Potential effects of disruption to HIV programmes in sub-Saharan Africa caused by COVID-19: results from multiple mathematical models. *The lancet HIV*. 2020; 7(9):e629–e40. [https://doi.org/10.1016/S2352-3018\(20\)30211-3](https://doi.org/10.1016/S2352-3018(20)30211-3) PMID: 32771089
5. Dorward J, Khubone T, Gate K, Ngobese H, Sookrajh Y, Mkhize S, et al. The impact of the COVID-19 lockdown on HIV care in 65 South African primary care clinics: an interrupted time series analysis. *The Lancet HIV*. 2021; 8(3):e158–e65. [https://doi.org/10.1016/S2352-3018\(20\)30359-3](https://doi.org/10.1016/S2352-3018(20)30359-3) PMID: 33549166
6. Zhu Q, JiKe C, Xu C, Liang S, Yu G, Wang J, et al. A New Strategy to Quantitatively Identify Hot-Spot Areas in Growth of New HIV Infections for Targeted Interventions. *Frontiers in Public Health*. 2021; 9:680867. <https://doi.org/10.3389/fpubh.2021.680867> PMID: 34322472
7. Young S, Zhang Q. Using search engine big data for predicting new HIV diagnoses. *PloS one*. 2018; 13(7):e0199527. <https://doi.org/10.1371/journal.pone.0199527> PMID: 30001360
8. Young S, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive medicine*. 2014; 63:112–5. <https://doi.org/10.1016/j.ypmed.2014.01.024> PMID: 24513169
9. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*. 2019; 17(1):1–13.
10. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky Z. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine*. 2020; 3(1):1–12. <https://doi.org/10.1038/s41746-020-0287-6> PMID: 32509975
11. Whiting D, Fazel S. How accurate are suicide risk prediction models? Asking the right questions for clinical practice. *Evidence-based mental health*. 2019; 22(3):125–8. <https://doi.org/10.1136/ebmental-2019-300102> PMID: 31248976
12. Farran B, AlWotayan R, Alkandari H, Al-Abdulrazzaq D, Channanath A, Thanaraj T. Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: a retrospective cohort study of health data from Kuwait. *Frontiers in endocrinology*. 2019; 10:624. <https://doi.org/10.3389/fendo.2019.00624> PMID: 31572303
13. Park J, Cho H, Kim J, Wall M, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ digital medicine*. 2020; 3(1):1–7. <https://doi.org/10.1038/s41746-020-0256-0> PMID: 32258428
14. Kwiecinski J, Tzolos E, Meah M, Cadet S, Adamson P, Grodecki K, et al. Machine learning with 18F-sodium fluoride PET and quantitative plaque analysis on CT angiography for the future risk of myocardial infarction. *Journal of Nuclear Medicine*. 2022; 63(1):158–65. <https://doi.org/10.2967/jnumed.121.262283> PMID: 33893193
15. Gruber S, Krakower D, Menchaca J, Hsu K, Hawrusik R, Maro J, et al. Using electronic health records to identify candidates for human immunodeficiency virus pre-exposure prophylaxis: An application of super learning to risk prediction when the outcome is rare. *Statistics in medicine*. 2020; 39(23):3059–73. <https://doi.org/10.1002/sim.8591> PMID: 32578905
16. Marcus J, Hurley L, Krakower D, Alexeeff S, Silverberg M, Volk J. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *The lancet HIV*. 2019; 6(10):e688–e95. [https://doi.org/10.1016/S2352-3018\(19\)30137-7](https://doi.org/10.1016/S2352-3018(19)30137-7) PMID: 31285183
17. Bao Y, Medland N, Fairley C, Wu J, Shang X, Chow E, et al. Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. *Journal of Infection*. 2021; 82(1):48–59. <https://doi.org/10.1016/j.jinf.2020.11.007> PMID: 33189772
18. Xu X, Ge Z, Chow E, Yu Z, Lee D, Wu J, et al. A Machine-Learning-Based Risk-Prediction Tool for HIV and Sexually Transmitted Infections Acquisition over the Next 12 Months. *Journal of clinical medicine*. 2022; 11(7):1818. <https://doi.org/10.3390/jcm11071818> PMID: 35407428
19. Chingombe I, Dzinamarira T, Cuadros D, Mapingure M, Mbunge E, Chaputsira S, et al. Predicting HIV status among men who have sex with men in Bulawayo & Harare, Zimbabwe using bio-behavioural data, recurrent neural networks, and machine learning techniques. *Tropical Medicine and Infectious Disease*. 2022; 7(9):231.

20. Chingombe I, Musuka G, Mbunge E, Chemhaka G, Cuadros D, Murewanhema G, et al., editors. Predicting HIV Status Using Machine Learning Techniques and Bio-Behavioural Data from the Zimbabwe Population-Based HIV Impact Assessment (ZIMPHIA15-16). Computer Science On-line Conference; 2022: Springer.
21. ZIMSTAT. 2022 Population and Housing Census Preliminary Results. UNFPA ZIMBABWE: 2022 22 September. Report No.
22. The DHS Program. Demographic and Health Surveys [Online Database] 2005 [updated May 2018]. Available from: https://dhsprogram.com/data/dataset_admin/login_main.cfm.
23. Zimbabwe Central Statistical Office, Macro International. Zimbabwe Demographic and Health Survey 2005–06. Calverton, Maryland, USA: Central Statistical Office/Zimbabwe and Macro International, 2007.
24. Zimbabwe National Statistics Agency, ICF International. Zimbabwe Demographic and Health Survey 2015: Final Report. Rockville, Maryland, USA: Zimbabwe National Statistics Agency (ZIMSTAT) and ICF International, 2016.
25. Zimbabwe National Statistics Agency—ZIMSTAT, ICF International. Zimbabwe Demographic and Health Survey 2010–11. Calverton, Maryland, USA: ZIMSTAT and ICF International, 2012.
26. Nohara Y, Matsumoto K, Soejima H, Nakashima N, editors. Explanation of machine learning models using improved Shapley Additive Explanation. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; 2019.
27. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*. 2015; 10(3):e0118432. <https://doi.org/10.1371/journal.pone.0118432> PMID: 25738806
28. Simelane M, Vermaak K, Zwane E, Masango S. Individual and community-level factors associated with lifetime number of sexual partners among women aged 15–49 in Eswatini. *Plos one*. 2021; 16(1): e0246100. <https://doi.org/10.1371/journal.pone.0246100> PMID: 33497398
29. Mishra V, Thaddeus S, Kafuko J, Opio A, Hong R, Kirungi W, et al. Fewer Lifetime Sexual Partners and Partner Faithfulness Reduce Risk of HIV Infection: Evidence from a National Sero-Survey in Uganda 2009.
30. Epstein M, Bailey J, Manhart L, Hill K, Hawkins JD, Haggerty K, et al. Understanding the link between early sexual initiation and later sexually transmitted infection: test and replication in two longitudinal studies. *Journal of Adolescent Health*. 2014; 54(4):435–41. e2. <https://doi.org/10.1016/j.jadohealth.2013.09.016> PMID: 24280303
31. Harling G, Newell M, Tanser F, Kawachi I, Subramanian S, Barnighausen T. Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in rural KwaZulu-Natal, South Africa. *Journal of acquired immune deficiency syndromes (1999)*. 2014; 66(4):443. <https://doi.org/10.1097/QAI.000000000000198> PMID: 24815854
32. Armstrong H, Roth E, Rich A, Lachowsky N, Cui Z, Sereda P, et al. Associations between sexual partner number and HIV risk behaviors: implications for HIV prevention efforts in a Treatment as Prevention (TasP) environment. *AIDS care*. 2018; 30(10):1290–7.
33. Dunkle K, Stephenson R, Karita E, Chomba E, Kayitenkore K, Vwalika C, et al. New heterosexually transmitted HIV infections in married or cohabiting couples in urban Zambia and Rwanda: an analysis of survey and clinical data. *The Lancet*. 2008; 371(9631):2183–91. [https://doi.org/10.1016/S0140-6736\(08\)60953-8](https://doi.org/10.1016/S0140-6736(08)60953-8) PMID: 18586173
34. Hattori M, Dodoo FN-A. Cohabitation, marriage, and 'sexual monogamy' in Nairobi's slums. *Social Science & Medicine*. 2007; 64(5):1067–78. <https://doi.org/10.1016/j.socscimed.2006.10.005> PMID: 17123680
35. Kaiser R, Bunnell R, Hightower A, Kim AA, Cherutich P, Mwangi M, et al. Factors associated with HIV infection in married or cohabitating couples in Kenya: results from a nationally representative study. *PLoS one*. 2011; 6(3):e17842. <https://doi.org/10.1371/journal.pone.0017842> PMID: 21423615
36. Maharaj P, Cleland J. Risk perception and condom use among married or cohabiting couples in Kwa-Zulu-Natal, South Africa. *International family planning perspectives*. 2005:24–9. <https://doi.org/10.1363/3102405> PMID: 15888406
37. Matovu J. Preventing HIV transmission in married and cohabiting HIV-discordant couples in sub-Saharan Africa through combination prevention. *Current HIV research*. 2010; 8(6):430–40. <https://doi.org/10.2174/157016210793499303> PMID: 20636280
38. Kelso B. AIDS: Orphans of the Storm. *Africa Report*. 1994; 39(1):50.
39. Mabala R. From HIV prevention to HIV protection: addressing the vulnerability of girls and young women in urban areas. *Environment and Urbanization*. 2006; 18(2):407–32.

40. Van Breda A. The phenomenon and concerns of child-headed households in Africa. *Sozialarbeit des südens*. 2010; 3:259–79.
41. Chigwenya A, Chuma M, Nyanga T. Trapped in the Vicious Circle: An Analysis of the Sustainability of the Child-Headed Households' Livelihoods in Ward 30, Gutu District. 2008.
42. Kazerooni P, Mousavi M, Khazaei Z, Esfahani M, Mohseni S, Nejat M, et al. The prevalence of HIV among female head of household in Shiraz in 2015. *Immunopathologia Persa*. 2018; 4(2):e29–e.
43. Schatz E, Gilbert L. "My legs affect me a lot... I can no longer walk to the forest to fetch firewood": Challenges related to health and the performance of daily tasks for older women in a high HIV context. *Health Care for Women International*. 2014; 35(7–9):771–88.
44. Madise N, Ziraba A, Inungu J, Khamadi S, Ezech A, Zulu E, et al. Are slum dwellers at heightened risk of HIV infection than other urban residents? Evidence from population-based HIV prevalence surveys in Kenya. *Health & place*. 2012; 18(5):1144–52. <https://doi.org/10.1016/j.healthplace.2012.04.003> PMID: 22591621
45. Magadi M. The disproportionate high risk of HIV infection among the urban poor in sub-Saharan Africa. *AIDS and Behavior*. 2013; 17(5):1645–54. <https://doi.org/10.1007/s10461-012-0217-y> PMID: 22660933
46. Hajizadeh M, Sia D, Heymann S, Nandi A. Socioeconomic inequalities in HIV/AIDS prevalence in sub-Saharan African countries: evidence from the Demographic Health Surveys. *International journal for equity in health*. 2014; 13(1):1–22. <https://doi.org/10.1186/1475-2876-13-18> PMID: 24533509
47. Henderson L, Zerai A, Morrow R. Intimate partner violence and HIV status among ever-married and cohabiting Zimbabwean women: an examination of partners' traits. *African Journal of Reproductive Health*. 2017; 21(4):45–54. <https://doi.org/10.29063/ajrh2017/v21i4.5> PMID: 29624950
48. Nyamayemombe C, Mishra V, Rusakaniko S, Benedikt C, Gwazane M, Mukweza P. The Association between Violence against Women and HIV: Evidence from a National Population-Based Survey in Zimbabwe. *Zimbabwe Working Papers, No4*. 2010; Calverton, Maryland, USA: ICF Macro.
49. Orel E, Esra R, Estill J, Thiabaud A, Marchand-Maillet S, Merzouki A, et al. Prediction of HIV status based on socio-behavioural characteristics in East and Southern Africa. *PloS one*. 2022; 17(3): e0264429. <https://doi.org/10.1371/journal.pone.0264429> PMID: 35239697