UNIVERSITY OF CAPE TOWN

Clustering of longitudinal viral loads in the Western Cape

Eke Nnanna Arua

ARXEKE001

Dissertation submitted in partial fulfilment of the requirements for the degree

MASTER OF PUBLIC HEALTH in Epidemiology

in the

School of Public Health & Family Medicine

Supervisors: Associate Professor Maia Lesosky

Co-Supervisor: Professor Landon Myer

February 2020

PREAMBLE

1. Declaration

I, Eke Nnanna Arua (ARXEKE001), hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 10 February 2020

2. Acknowledgements

I would like to thank my supervisor Associate Professor Maia Lesosky for her invaluable guidance and encouragement throughout this dissertation and the MPH in general. You continue to be a great source of inspiration for me.

To my co-supervisor Professor Landon Myer, I appreciate your input and your constant words of encouragement as we approached the final stages of this dissertation.

To Lilian, Jeff, Anathi, Phindi and my better half Ndiphe; I would not have made it this far without your words of motivation and support.

Finally, to my Family, this is one is for you!

3. Thesis abstract

Introduction: Routine viral load (VL) monitoring is important for assessing the effectiveness of ART in South Africa. There is little information however, on how the longitudinal VL patterns change for subgroups of persons living with HIV (PLHIV) who have experienced at least one elevated VL. We investigated the possible longitudinal VL patterns that may exist among this unique population.

Methods: This mini-dissertation offers three components; a research protocol (Section A), a literature review (Section B) and a journal ready manuscript (Section C). We examined HIV VL data for the Western Cape from 2008 to 2018, taken from the National Health Laboratory Services (NHLS). Using <1000 copies/mL as a threshold for viral suppression, we identified 109092 individuals who had at least one instance of an elevated VL. A nonparametric (KML-Shape) and a model-based (LCMM) clustering technique were used to identify latent subgroups of longitudinal VL trajectories among these individuals.

Results: Both the KML-Shape and LCMM clustering techniques identified five latent viral load trajectory subgroups. KML-Shape found majority of individuals' trajectories belonged to clusters that had a decreasing longitudinal VL trend (76.6% of individuals), while LCMM found a smaller proportion of individuals' trajectories belonged to clusters that had a decreasing longitudinal trend (52.5% of individuals). Most of the trajectory subgroups identified had long periods of low-level viremia.

Conclusion: Although majority of individuals belonged to clusters that had downward trends, further research is needed to better understand factors contributing to membership of clusters that did not have a downward longitudinal trend. Understanding these factors may help in the development of targeted HIV prevention programs for these individuals.

4. List of abbreviations

AIC:	Akaike information criterion
ART:	Antiretroviral therapy
BIC:	Bayesian information criterion
EAC:	Enhanced adherence counselling
EHR:	Electronic health records
HIV:	Human immunodeficiency virus
KML:	K-means for longitudinal data
KML-Shape:	K-means for longitudinal data by shape
LCMM:	Latent class mixture model
NHLS:	National health laboratory service
PLHIV:	Persons living with HIV
RIC:	Retention in care
UNAIDS:	The Joint United Nations Programme on HIV/AIDS
VL:	Viral load
WHO:	World Health Organization

Table of Contents

5. List of Tables

A. Protocol

C. Manuscript

6. List of figures

C. Manuscript

A. PROTOCOL

1. Introduction

Technological advancements have led to collection of large amounts of data as part of routine health care. Commonly, these data sources are often re-purposed for use in medical and health research. These data are usually stored in data sets that either contain; a large number of rows (observations), a large number of columns (variables), or both. For such large datasets, exploratory data analysis is an important approach for discovering relationships and patterns in the data. It is common for these data sets to contain multivariate data. Analyzing multivariate data in large quantities often present difficulties in analytic methodology and data interpretation.[1] Categorizing and labeling clusters of similar objects is important for summarizing large datasets so that a concise description of patterns of similarities and differences in the data can be provided.[2]

Cluster analysis is an important data exploration tool that can be used for determining patterns in multivariate data and is used in many disciplines ranging from biology, statistics, economics and marketing.[3] Cluster analysis attempts to separate objects into homogenous groups so that objects within groups are similar to each other but different to objects in other groups. In the field of health, it can be used to discover patient subgroups defined by differing characteristics. This may have several advantages such as; the development or improvement of diagnostic criteria, better understanding of heterogeneous outcomes, and adapting treatments to better suit patient profiles.[4]

Cluster analysis can and has been applied to human immunodeficiency virus (HIV) research in various ways.[5-7] The chronic nature of HIV in the era of effective antiretroviral treatment means that there can be considerable heterogeneity in disease management and progression amongst persons living with HIV (PLHIV). Long term suppression of HIV viral load (VL) is known to help reduce HIV incidence by preventing transmission of HIV from person to person, as well as being associated with better health outcomes.[8] In many health systems, VL is monitored routinely, and it is of interest, at both a population and individual level, to understand the trajectories of viral suppression over time. Cluster analysis can be applied to HIV VL data specifically and has previously provided insights in grouping patients according to their longitudinal viral load patterns.[9] There are several methods available to cluster longitudinal series which can be applied to repeated VL measurements. From a public health standpoint, it may be pertinent to identify different longitudinal VL patterns amongst PLHIV, especially those related to adverse health outcomes and to understand risk factors that are associated with these patterns. Understanding how and why PLHIV VLs change over time could lead to improved targeted HIV programs.

This work will apply and compare existing methods for clustering longitudinal trajectories to a data set of routine VL measurements from the Western Cape, South Africa.

1.1 Background of study

The high burden of HIV/AIDs remains a major public health issue in South Africa. As of 2017, the prevalence of HIV in the country was estimated to be one of the highest in the world at approximately 14%.[10] In 2011, the estimated proportion of direct and indirect deaths due to HIV/AIDS was 43.6%; which was a decline from previous years but remains unacceptably high.[11] Interventions aimed at reducing HIV-related mortality are crucial and their successful implementation will not only ease the direct burden caused by HIV but also improve the prevention and management of colliding epidemics, such as rising rates of diabetes.

With respect to tackling HIV-related mortality, antiretroviral therapy (ART) is the primary intervention for HIV infected persons. South Africa's treatment programme is the largest globally, accounting for 20% of people on ART worldwide.[12] Since the advent of ART in 2004, there has been a great improvement in the mortality rate of HIV infected persons in the country.[13,14] The life expectancies of South African adults living with HIV are comparable to those who do not have HIV provided they initiate ART early.[15] The Joint United Nations Programme on HIV/AIDS (UNAIDS) 2020 targets require that 90% of all HIV-infected persons be diagnosed, 90% of those diagnosed should be on treatment, with at least 90% achieving and maintaining viral suppression (90-90-90 targets).[16] There has been steady progress towards reaching these targets with approximately between 38-54% of all infected persons in the country being virally suppressed by 2015, however this is well short of the required 73%.[17]

Routine VL monitoring is important for measuring people's initial and continual response to ART. It is recommended by the World Health Organization (WHO) as the principal method for diagnosis and confirmation of treatment failure or virologic suppression for people who are on ART.[18] The data that is collected from routine VL monitoring is crucial in efforts towards understanding the population viral load suppression rates in the country. For example, routine VL data can facilitate the identification of groups of unsuppressed individuals that may require additional resources such as enhanced adherence counselling (EAC).[18]

South African routine VL data – which is longitudinal in nature – is often used to estimate or offer insights on mean population level proportions of ART adherence and virologic suppressions as well as their respective longitudinal trends. These analytic approaches are relevant and often yield important results

and insights, but they rely on pre-specified subpopulation of individuals such as age groups, race, geographic location and sex. Over-reliance on these sub-populations when investigating longitudinal VL trajectories hinder our ability to account for unobserved heterogeneity and fully understand both the longitudinal pattern and determinants of VL suppression within the population. There may exist unobserved sub-populations or categories of longitudinal profiles whose detection could lead to more effective and efficient treatment by adapting healthcare and health promotion to better suit peoples' profiles. Within South Africa's setting, there has been little research investigating whether there exists the presence of distinct latent VL trajectories amongst people who have initiated ART.

1.2 Rationale of the study

Because of the slow progress towards a 73% VL suppression of all HIV infected persons. South Africa is unlikely to achieve its 90-90-90 targets on schedule. Achieving a 90% viral suppression amongst HIV infected persons has proven a difficult landmark to reach. To understand the reasons why adequate viral suppression rates have not been obtainable so far, novel approaches in the analysis of VL data are necessary. In particular, the identification of clinically important sub-populations that have undesirable VL trajectories (i.e. trajectories that show an increasing trend after ART initiation) may offer important insights for targeted interventions – such as specific health promotion campaigns – and further research. The Western Cape routine VL monitoring data affords a unique opportunity to get deeper understandings of how different groups of individuals VLs change over time using a more objective approach compared to methods that are normally used for longitudinal data.

2. Study aims and objectives

2.1 Aims

This study has two aims. The first aim is to apply and compare methods to cluster longitudinal VL load trajectories using routine VL data collected in the Western Cape - a large province in South Africa. The second aim is to determine which factors are associated with membership to the identified clustered trajectories; particularly, trajectories that are not considered desirable in a public health context.

2.2 Objectives

Primary objective: To identify and apply appropriate methods for clustering and classification of longitudinal VL trajectories using routine VL data.

Secondary objectives:

1. Describe the clusters of VL trajectory identified and compare them by clustering methods used
2. Identify and compare for each clustering method used, the risk factors associated with cluster membership.

3. Methodology

3.1 Study design

To identify clustered VL trajectories, a secondary data analysis of anonymized individuals' repeated VL measurements from the National Health Laboratory Services (NHLS) in the Western Cape will be used. The NHLS is the largest diagnostic pathology service in South Africa. It has laboratories in every province in the country and covers 80% of the population.[19] In the Western Cape, the NHLS data covers all public sector ART services. Records obtained will span a period of 10 years from the year 2008 to 2018.

3.2 Sampling and Study population

The nature of this research implies that a sample size calculation is unnecessary as all records available from the 10-year period will be used for analysis. Participants to be included in this study are HIV infected individuals in the Western Cape who have initiated ART before or during the 10-year period for which data was collected and who have had at least one routine viral load monitoring test.

3.3 Data management

The data will be anonymized by the parent study data manager so that all identifiable information will be removed prior to analysis. Direct identifiers such as names and patient IDs will be removed from the data set. A unique identifier that cannot be linked with the original direct identifiers will be generated from the data set to assist with analysis. The relevant data will be kept on a password protected computer in the school of Public Health and Family Medicine at the University of Cape Town. When not in use, the computer will be kept in a locked office with limited access. The statistical software package that will be used for this research is R (R Core Team (2018), Vienna, Australia).

3.4 Data Analysis

3.4.1 Clustering

There are numerous approaches to clustering of longitudinal trajectories. Methods that will be applied are summarized below.

Broadly, most clustering methods can be categorized as either being parametric (model-based) or nonparametric. The non-parametric approach makes few assumptions on the distribution of the data while the model-based approach assumes data vectors are generated from a finite mixture of specified distributions.

Nonparametric clustering

This approach to clustering makes few assumptions on the distribution of the data. The number of categories, the choice of similarity (or dissimilarity) measure and the clustering algorithm are the major components needed for this method.[2,20]

Similarity/dissimilarity measure

Often called a "distance measure", the similarity or dissimilarity measure is any function *d,* used to measure the distinctiveness between two individuals or objects. The choice of distance measure requires careful consideration as it can have a large impact on the clustering results. The underlying logic in cluster analysis is that objects belonging to the same cluster are more similar than objects in different clusters.

Nonparametric clustering algorithms

These clustering algorithms can generally be grouped as either hierarchical or partitioning algorithms depending on the search strategy that is used.[3] Partitioning algorithms split the data into $k \geq 2$ distinct clusters. These clusters must be mutually exclusive and their objects collectively exhaustive. This means that there must be at least one object in each cluster and an object cannot belong to more than one cluster. The number of clusters, k, is often user-defined (although, there exist computer-based methods for determining it).[3] Hierarchical clustering techniques rely on constructing trees of clusters, in which the root node is a single cluster with all objects ($k = 1$) and the leaves are clusters that each contain a single object ($k = n$).[2,21] There are two types of ways hierarchical clustering can be carried out; in a bottom-up manner (agglomerative) or in a top-down manner (divisive). As the names suggest, the two methods construct their trees in opposite directions. The divisive method starts when all objects are in a single

cluster and are split up in each step until there are *n* of them. The agglomerative method starts when there are *n* clusters and each step two clusters are merged until there is only one cluster remaining. Hierarchical clustering should be used with an a priori scientific understanding of the hierarchical nature of the data.[2]

A nonparametric approach that will be used in this research is KML-Shape; an extension of KML (k-means for longitudinal data) by Genolini and Falissard.[22,23] It offers an efficient way to cluster longitudinal data on the basis of the shape of their VL trajectories and allows individuals who have different VL trajectory forms that have shifted locations in time to be grouped together. It uses a measure of distinctiveness between individuals' trajectories that is based on their shape rather than their level.

Model-based clustering

Under the nonparametric clustering approach, each object can only belong to one cluster. This is also known as "hard clustering". With model-based clustering, sometimes referred to as "soft" or "fuzzy" clustering, objects do not have this restriction. An object can belong to several clusters with differing probabilities of cluster membership. This mainly affects objects near the edge between two clusters. Hard clustering would require them to be placed into one group, whereas model-based clustering means that objects can have varying contributions to more than one cluster.[2] A priori assumptions on the distribution of data provide a framework to facilitate the estimation of cluster membership probabilities.

The Latent class mixture model (LCMM) is a popular model-based clustering approach that will be used in this study. VL measurements will be modelled using mixed effect models with fractional polynomials to get a sense of the mean population trajectory. These models allow for differences in individual VL trends over time because they utilize a random (growth) slope and intercept that vary across individuals, which in turn can yield individual trajectories. The use of fractional polynomials for the models is appropriate because they can account for non-linear trajectories and may fit the data better at extreme values compared to higher order polynomials.[24] The fractional polynomial terms that provide the best fit to the mixed effect model will be using in the LCMM to identify distinct classes of individuals VL trajectories. Unlike more conventional growth model approaches such as the mixed effects model that assume a single estimate of growth parameters can adequately describe the VL growth trajectories of all individuals in the data, it allows for different average growth curves to accommodate different classes of individuals.[25,26]

3.5 Number of clusters

Determining an ideal number of clusters is a major problem yet to be fully addressed in cluster analysis. Most nonparametric clustering procedures offer little insight as to the number of clusters that exist in the data. Nonparametric approaches for selecting an optimal number of clusters vary. Direct approaches such as using the average silhouette[27], or maximizing between cluster distance while minimizing within cluster distance[28] can be attempted. There are also testing procedures with the null hypothesis defined as $k = 1$.[29,30] Despite the numerous techniques, there is no unanimously accepted method. Like nonparametric approaches, the problem of choosing an ideal number of clusters persists when applying model-based techniques. Information theory techniques such as Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can be used to determine cluster numbers by selecting models with lowest loss of information.

For this research, the number of clusters will be selected based on domain knowledge and computational feasibility. For model-based clustering approach, the optimal number of clusters will be determined using both the AIC and BIC of the models and domain knowledge on the types of trajectories expected in the data set.

3.6 Determining cluster membership

Descriptive statistics of all variables that will be included in the clustering process shall be reported using proportions, means with standard deviations and medians with interquartile ranges as appropriate. The cluster sizes - i.e. the proportion of the data that each cluster accounts for - will be presented. Clustering results will be described graphically by plotting the mean trajectory for each cluster identified for each clustering approach used.

After performing the cluster analysis and detecting the latent longitudinal VL trajectory sub-populations, the next step will be to identify which risk factors associated with cluster membership. Descriptive statistics will be stratified by clusters obtained to determine which variables may influence cluster membership for individuals. The dataset is limited with regard to the variables available for analysis, hence there will be few variables used. The variables that will be used for analysis include:

- Age (in years)
- Gender
- Facility type

- HIV viral load
- Time in months since first VL was measured

4. Potential limitations

There are many ways to cluster data. Although the body of literature for clustering longitudinal data is relatively smaller compared to other forms of multivariate data, it is still substantial. As a result, the clustering methods identified are unlikely to be exhaustive. Additionally, the use of routine data means that the standards with which they are collected are usually not as high as data for research specific purposes which may compromises data accuracy. This also makes it difficult to improve data quality. The data obtained may be incomplete for various reasons. For example, there is the possibility of high mobility of individuals between services across geographical regions – i.e. moving to other provinces may influence results obtained.

5. Ethical considerations

All the necessary ethical considerations for undertaking this study have been accounted for. There is no conflict of interest to report. There are no incentives to be received and no proprietary interests involving any agent, device or software being evaluated by conducting this research.

5.1 Consent

This is an analysis of routinely collected retrospective secondary data and individual consent for use was not sought. The objectives of the study related to the identification of areas in the public sector provision of ART services that could be strengthened, improving care for all individuals living with HIV.

5.2 Risks

There are no direct risks to participants as this is an analysis of secondary data and no participants will be recruited or enrolled. There is a small indirect risk relating to the potential for loss of privacy and anonymity. This risk will be mitigated by ensuring that no personal data will be made available to the researchers. Individual names, ID numbers and other identifying information will be removed by the parent study data team. Individuals will be identified by a unique, study specific, random identifier. These identifiers cannot be linked to a participant's name or original identity.

5.3 Benefits

There are no direct benefits to participants contributing data as this is an analysis of secondary data and no participants are recruited or enrolled. There may be indirect benefits to people living with HIV as results from this research may inform policy on better management of people living with HIV and attending public sector services in the Western Cape.

6. Budget

This study constitutes the research component of the Master of Public Health (MPH) degree and therefore a budget is not necessary.

7. Timetable

Table 1. Time frame from start of study to completion

	October 2019	November 2019	December 2019	January 2020	February 2020
Literature review	✓	✓	✓		
Data management		✓	✓	✓	
Data analysis			✓	✓	
Results				✓	✓
Write up					✓

8. Stakeholders and dissemination

The findings from this research may prove important for current efforts towards increased VL suppression rates in South Africa. Stakeholders who may be interested in the outcome of the research include researchers and clinicians in the field of HIV, relevant policy makers such as the South African National Department of Health , the Western Cape provincial health authority, and individuals who have initiated or about to initiate ART. Results from this research will be made available through publication in a peer-reviewed journal and/or by presentation of the findings at local or international conferences.

9. References

1. Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinformatics. 2014;15 Suppl 11(Suppl 11):S11-S.

2. Heggeseth B. Longitudinal Cluster Analysis with Applications to Growth Trajectories: UC Berkeley; 2013.

3. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis: John Wiley & Sons; 2009.

4. Windgassen S, Moss-Morris R, Goldsmith K, Chalder T. The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome. Taylor & Francis; 2018.

5. Chan PA, Rose J, Maher J, Benben S, Pfeiffer K, Almonte A, et al. A latent class analysis of risk factors for acquiring HIV among men who have sex with men: implications for implementing pre-exposure prophylaxis programs. AIDS patient care and STDs. 2015;29(11):597-605.

6. van Gorp WG, Hinkin C, Satz P, Miller EN, Weisman J, Holston S, et al. Subtypes of HIV-related neuropsychological functioning: A cluster analysis approach. Neuropsychology. 1993;7(1):62.

7. Grabowski MK, Herbeck JT, Poon AF. Genetic cluster analysis for HIV prevention. Current HIV/AIDS Reports. 2018;15(2):182-9.

8. Human Sciences Research Council. South African Nationa HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017 2018 [Available from: http://www.hsrc.ac.za/uploads/pageContent/9234/FINAL%20Presentation%20for%2017%20July%20launch.pdf.

9. Mayosi BM, Lawn JE, Van Niekerk A, Bradshaw D, Karim SSA, Coovadia HM, et al. Health in South Africa: changes and challenges since 2009. The Lancet. 2012;380(9858):2029-43.

10. UNAIDS. South Africa: UNAIDS; [Available from: https://www.unaids.org/en/regionscountries/countries/southafrica.

11. Boulle A, Schomaker M, May MT, Hogg RS, Shepherd BE, Monge S, et al. Mortality in patients with HIV-1 infection starting antiretroviral therapy in South Africa, Europe, or North America: a collaborative analysis of prospective studies. PLoS medicine. 2014;11(9):e1001718.

12. UNAIDS. UNAIDS DATA 2018 2018 [cited 2019 06/06/2019]. Available from: https://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf.

13. Johnson LF, Mossong J, Dorrington RE, Schomaker M, Hoffmann CJ, Keiser O, et al. Life expectancies of South African adults starting antiretroviral treatment: collaborative analysis of cohort studies. PLoS medicine. 2013;10(4):e1001418.

14. Joint United Nations Programme on HIV/AIDS. 90-90-90—an ambitious treatment target to help end the AIDS epidemic. Geneva: UNAIDS; 2014 [Available from: https://www.unaids.org/sites/default/files/media_asset/90-90-90_en_0.pdf.

15. Johnson LF, Dorrington RE, Moolla H. Progress towards the 2020 targets for HIV diagnosis and antiretroviral treatment in South Africa. Southern African journal of HIV medicine. 2017;18(1).

16. UNAIDS. The need for routine viral load testing 2016 [Available from: https://www.unaids.org/sites/default/files/media_asset/JC2845_en.pdf.

17. National Health Laboratory Service. Who are we? : National Health Laboratory Service; 2019 [Available from: http://www.nhls.ac.za/?page=who_are_we&id=17.

18. Gong H, Xun X, Zhou Y. Profile clustering in clinical trials with longitudinal and functional data methods. Journal of biopharmaceutical statistics. 2019:1-17.

19. Hartigan JA. Statistical Clustering. In: Wright JD, editor. International Encyclopedia of the Social & Behavioral Sciences (Second Edition). Oxford: Elsevier; 2015. p. 392-6.

20. Genolini C, Falissard B. KmL: k-means for longitudinal data. Computational Statistics. 2010;25(2):317-28.

21. Genolini C, Ecochard R, Benghezal M, Driss T, Andrieu S, Subtil F. kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes. Plos one. 2016;11(6):e0150738.

22. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1994;43(3):429-53.

23. Jung T, Wickrama K. An introduction to latent class growth analysis and growth mixture modeling. Social and personality psychology compass. 2008;2(1):302-17.

24. The SAGE Handbook of Quantitative Methodology for the Social Sciences. 2004 2019/06/30. SAGE Publications, Inc. Available from: https://methods.sagepub.com/book/the-sage-handbook-of-quantitative-methodology-for-the-social-sciences.

25. Thinsungnoena T, Kaoungkub N, Durongdumronchaib P, Kerdprasopb K, Kerdprasopb N. The clustering validity with silhouette and sum of squared errors. learning. 2015;3:7.

26. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;50(2):159-79.

27. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome biology. 2002;3(7):research0036. 1.

28. Yan M. Methods of determining the number of clusters in a data set and a new clustering criterion: Virginia Tech; 2005.

B. LITERATURE REVIEW

1. Introduction

South Africa is a middle-income country in Southern Africa that has the largest burden of human immunodeficiency virus (HIV) globally.[31] It is estimated that there are more than 7 million persons living with HIV (PLHIV) in the country.[32] As of 2017, the estimated HIV prevalence amongst all South Africans was 14%.[10] Between 1997 and 2010, it is estimated that more than 2.8 million people lost their lives to the disease – a figure higher than any other country in Sub-Saharan Africa.[33] There are several interventions in place to decrease the number of new HIV infections annually. These interventions include voluntary medical male circumcision, promotion of condom use, increased HIV education and awareness campaigns.[34] The most widespread and effective method for prevention of HIV is enrollment on antiretroviral therapy (ART) for PLHIV.[17,35,36]

1.1 The success of ART in South Africa

The use of ART has transformed the lives of PLHIV and contributed to reducing HIV incidence in the country.[15] Since the rapid scale up of the South African national ART program, started in 2004, there has been a great improvement in life expectancy over the years.[15] In 2018, the overall life expectancy for men and women had increased to 61.1 years and 67.3 years respectively.[37] The scale up has been so extensive that South Africa accounts for 20% of all PLHIV on ART globally, making it the largest HIV treatment program in the world.[38] The South African national ART program has had a tremendous impact on HIV prevention efforts. The proportion of new HIV infections has decreased by 40% from 2010-2018.[39] There has been progress in HIV-related mortality with a decrease from an estimated 200,000 in 2004 to 71,000 in 2018.[32] South Africa still contributes the largest of proportion of all HIV-related deaths at 29% in Southern Africa, which when given the size of its HIV burden, is in line with expectations; however, HIV incidence remains unacceptably high.[39]

In 2014, the Joint United Nations Programme on HIV/AIDS (UNAIDS) launched an ambitious plan targeted at the elimination of new HIV infections globally by 2020 (known as 90-90-90 targets).[16] This plan requires that 90% of all infected persons know their HIV status, 90% of those that know their status to be on treatment, and 90% of those on treatment to be virally suppressed. Success in achieving these targets would lead to at least 73% of PLHIV achieving viral suppression worldwide. However, for this to successfully occur, every country worldwide would have needed to achieve 73% viral suppression of PLHIV.

1.2 South Africa's progress towards UNAIDS 2020 targets

Currently, South Africa has failed to achieve 90-90-90 targets and is unlikely to do so before the target deadlines in 2020; having met the first 90% target, it has fallen short on the remaining two.[17] As a result, it is estimated that only between 38-54% of PLHIV have achieved viral suppression, well behind desired 2020 targets.[17,39] In Southern Africa, South Africa's neighboring countries have shown that these targets are achievable.[39] Eswatini, Namibia and Botswana have reached all three 90% targets while Rwanda has achieved the first two and is close to achieving the third, however, the magnitude and history of the epidemic varies by country.[39] South Africa's failure to meet the 90-90-90 targets while having similar resources available may be a reflection of the difference in magnitude of epidemics or in the health system response between these countries.

Achieving a 73% population viral suppression rate nationally is highly dependent on the strength of ART adherence – which is a potent predictor of the success HIV treatment programs.[40] There are numerous factors that influence ART adherence amongst PLHIV; and ability to adhere to lifelong ART often varies on an individual, social and demographic level. Individual level factors include adherence self-efficacy, current substance abuse, concerns/doubts about ART and trust in an HIV care provider.[41] Social factors play as big a role in treatment adherence as individual factors. Strong social support has been linked with improved adherence whilst high levels of social stigma have been shown to be detrimental.[41,42] Demographic factors such as age and female gender have been associated with a positive impact on ART adherence.[43]

1.3 Objectives

Understanding longitudinal patterns of viral suppression among PLHIV on ART may support targeted programs to help the health system achieve 90-90-90 targets. To inform this research, the objectives of this literature review are as follows:

- To provide an overview of VL monitoring in South Africa
- To explore the use of cluster analysis and its applications to longitudinal health data
- To briefly describe methods to cluster longitudinal health data

2. Summary of literature

2.1 Viral load monitoring in South Africa

Routine VL monitoring is considered the gold standard in HIV treatment monitoring.[18] According to the World Health Organization (WHO) guidelines, it is recommended that VL monitoring should be used as the primary method for treatment failure or success confirmation for PLHIV enrolled on ART.[44] It is the most accurate measure of ART's effect on an individual's viral suppression. Routine monitoring may have the potential to improve the quality of treatment and health outcomes for PLHIV by reducing the time spent on a first-line ART regimen that may be failing and aid in identifying cases of poor adherence.[45]

In South Africa, routine VL monitoring is used preferentially to CD4 cell count monitoring as a gauge for program performance.[46] A reason for this is because VL monitoring has been shown to be better at predicting mortality of PLHIV on ART.[47] All public sector VL testing services fall under the jurisdiction of the National Health Laboratory Services (NHLS); the largest diagnostic pathology service in the country with availability in every province and an estimated 80% population coverage.[19] There were initial challenges in implementing VL monitoring as there was a higher level of skill required to conduct testing and a significant lack of skilled labor available, commercial assays were costly and cold chain required for sample transportation was not always readily available.[48] Despite these initial challenges, the volume of VL testing services has grown tremendously from 1,625,236 viral assays in 2008 (January to September) to 4,949,891 between 2017-2018.[48,49] This growth is further expected to rise to meet UNAIDS 2020 targets.[50]

South African national guidelines recommend that routine VL testing should be carried out 6 months after patients initiate ART and every 12 months for those who remain virally suppressed.[44,46] The definition of viral suppression has varied in different settings but is defined as <400 copies/mL according to South African national guidelines.[46] UNAIDS and the WHO do not explicitly define a VL suppression threshold but it is recommended that <1000 copies/mL should be used as there are well described clinical benefits for individuals that can maintain VL <1000 copies/mL.[44,51] A stricter and less used definition of viral suppression is <50 copies/mL which may also be referred to as below the detectable limit (BDL), and is largely regarded as the threshold below which transmission will not occur.[44,46]

In South Africa, there are several steps in the process leading up to a VL being "done".[52] Firstly, a blood sample is taken from an HIV infected individual who is aware of their status by a health care provider. The blood sample is then transported to and processed at a NHLS facility.[53] Results of the blood sample are

then conveyed back to the health care provider electronically or by delivery. The health care provider then explains the results of the VL test to the individual and suggests options for follow up.[53] In cases where results do not come back with a suppressed VL and regular ART uptake is assumed to be an underlying reason for this, enhanced adherence counselling (EAC) is provided and a follow up VL measurement is taken.[53] First-line ART resistance is assumed where subsequent VL results remain unsuppressed and a switch to second-line ART is initiated.[46]

Individuals undergoing routine VL monitoring will typically have several repeat measurements of their VL over time. This leads to longitudinal VL trajectories which may vary according to pattern and level between individuals in the monitoring program.

2.2 Cluster analysis

Routine VL monitoring program generates multivariate longitudinal data – which may be defined as repeat measurements of one or more variables over time within a group or population of individuals. This data is often heterogenous in nature with underlying correlation structures. Cluster analysis is an increasingly used data exploration tool used for finding groups in large multivariate longitudinal datasets. It has applications in different fields of research including marketing, computer science and machine learning theory.[2] Despite having various applications in different fields, the underlying principal is the same; grouping together objects that are similar and keeping these groups of objects distinct from each other.[3]

The significance of cluster analysis has been on the rise in recent times – in part, because of the rise of big data and the need for practical and efficient instruments to explore and reveal vital information concealed within it.[54] Clustering of big data has become more frequent as powerful computers capable of handling growingly complex algorithms needed for analysis are more widely available than ever.[3] There has been a push towards increasing the volume of repository data kept in patient registries and electronic health records (EHR) databases in both developed and developing countries.[55,56] Schulam et al.[57] note the exciting opportunities this affords us for identifying disease subgroups in an objective manner. These large datasets containing patient level data with repeated measurements are often complex and have a substantial amount of diverse information within them. This diversity may be found in – but is not limited to – patient attributes, severity of illness or symptoms, and treatment responses.[4] Using cluster analysis to gain a deeper understanding of large, longitudinal multidimensional datasets has a few advantages. Firstly, it allows for the identification of clinically relevant patient groups with key attributes that could lead to more tailored and enhanced care.[4] Additionally, in situations where the diagnostic criteria for

certain mental and physical illnesses are either too limited or too vague, cluster analysis can provide clearer and more valid criteria to aid health professionals.[4]

2.3 Cluster analysis and Longitudinal studies

Longitudinal studies are an important, intrinsic feature in epidemiological research.[2] They allow researchers to observe changes in the outcome of interest in individuals, families or populations over time and allow us to gain important insights in growth, development, and life span related health topics.[58] A key advantage of longitudinal studies is their ability to relate outcomes to specific exposures in a more flexible way than cross-sectional studies because of their temporal nature.[58] Longitudinal data require non-conventional, specific clustering techniques.[2] This is because longitudinal data is often sparsely and irregularly measured and commonly has a significant amount of missing data.[2] Furthermore, objects are rarely independent from each other due to the time ordering of their measurements.[2]

2.3.1 Argument for cluster analysis for VL monitoring

A systematic review by Pinaire et al.[59] highlighted the growing interest in the study of individuals' longitudinal health trajectories using health indicators like VL as monitoring tools. The review noted that despite the increase in these types of studies in recent years, cluster analysis is not a method commonly used for their analysis.[59] This finding may be extended to South Africa, which has little literature regarding clustering of health related trajectories. The large volume of longitudinal data from South Africa's routine VL monitoring program requires objective analytic approaches such as cluster analysis to account for the heterogeneity that may be found within it. Conventional analyses performed on the data are often limited to the overall population suppression rates – commonly stratified by arbitrary categories such as age groups, sex and province.[12,17] This approach yields important information on program performance but has its limitations. Firstly, the repeated use of subjective categories based on previous literature hinders the discovery of other clinically relevant population subgroups. De Keulenaer et al.[60] argue for the use of more objective, data-driven methods for identifying potentially important population subgroups when studying chronic diseases such as heart failure. Another drawback is that these analyses frequently focus solely on viral suppression as a proportion of PLHIV and ignore the pattern of individuals' viral load trajectories – which may help us understand disease progression. Genolini et al.[23] posit that the clustering individuals' trajectories according to shape could help reveal important insights that classical analyses cannot.

A better comprehension of trajectories permits us to anticipate the people that have the greatest risk for adverse trajectories and outcomes, better understand factors that influence variation in health over time, grant the ability to investigate effects of interventions on the trajectories – including identifying for whom and at what point in the trajectory interventions may be most effective.[61]

2.3.2 Applications of longitudinal data clustering in general health research

There are many applications of longitudinal data clustering in health research that cover a vast and diverse range of research fields such as clinical trials[20], genetics[62], and physiology[63]. A few examples will be provided to demonstrate the flexibility of cluster analysis in varying health research fields.

Pate et al.[64] identify three distinct physical activity trajectories amongst youth progressing from grade five through to grade eleven. Of the trajectories they discovered, there was one that had youth's physical activity decreasing to very low levels as they aged from 10-16 years of age.[64] Gender and their fifth grade body mass index were factors associated with this group.[64] Shoji et al.[65] presented a novel approach to clustering hospital laboratory examinations. They showed that their approach could be used to analyse relationships between platelet count and albumin trajectories and how they influence fibrotic stages.[65] Their findings showed varying distribution of fibrotic stages in discovered clusters suggesting an underlying influence between the trajectories and fibrotic stages.[65]

Jensen et al.[66] carried out a massive, population wide, disease trajectory clustering using the Danish national patient registry that consisted of 6.2 million patients. They transformed this vast amount of data – which used the entire range of diseases found in 14.9 years worth of registry data - into 1,171 disease trajectories. Although their research was focused more on grouping of disease trajectories rather than patient trajectories, they were nevertheless able to show that this approach can reveal diagnosis linkages that may have previously had uncertain or contradictory associations through conventional epidemiological studies.[66]

2.3.3 Applications of longitudinal data clustering in HIV research

Cluster analysis has previously been applied to HIV research in South Africa. Gosset et al.[67] investigated retention in care (RIC) trajectories for 777 individuals eligible for ART using group-based cluster analysis in rural South Africa. They found four clustered trajectories: A "remained in care" cluster that accounted for 71.3% of the patients, an "exited care then returned" cluster that accounted for 5.2% of patients, an "exited care rapidly" cluster that accounted for 12.6% of patients, and finally, an "exited care later" that

accounted for 10.9% of patients.[67] They concluded that in order to maximize RIC, swift ART initiation and focused support for young and newly diagnosed PLHIV is necessary.[67]

Likewise, cluster analysis has been used for understanding VL suppression trajectories. Kassaye et al.[68] used a cohort study of 1989 HIV infected women in the United States to find three distinct trajectory groups representing varying probability of viremia. The groups were categorized as low probability (28.6% of participants), intermediate probability (39.4% of participants) and high probability (32% of participants).[68] Factors such as drug use, depression symptoms and younger age were associated with the high probability group which highlighted the need to focus efforts towards mental health and other behavioral factors that are associated with a higher viremia probability.[68]

2.4 Clustering methods for longitudinal data

There are numerous ways to cluster trajectories, but they can be broadly categorized into two methods; parametric and non-parametric clustering.[2,20] The key difference between the two methods is that non-parametric clustering makes few assumptions of the distribution of the data, instead focusing on explicitly defining similarity.[2]

2.4.1 Nonparametric clustering

The most important components of this approach involve the choice of measure for distinctiveness between individuals, the clustering algorithm and the number of clusters in the data. This clustering technique is often described as "hard clustering" because an object can belong to only one cluster.[2]

Measures of distinctiveness

There are several terms used for a measure to quantify the distinctiveness between two individuals. These terms are sometimes called a metric, distance function, dissimilarity function or similarity function. Although closely related to each other, these terms do not necessarily measure distinctiveness the same way. For example, all metrics can be said to be distance functions, but the reverse is not always the case.[2] The choice of distance measure is crucial and can often have a large impact on clustering results, therefore, it must be carefully considered before being used.[2]

Clustering algorithm

These are generally classified as Hierarchical and partitioning algorithms.[3] Partitioning algorithms cluster the data into two or more groups. These groups must jointly satisfy the following conditions of a partition:

- At least one object must be in each group
- An object cannot belong to more than one group

A "good" partition is one where the objects in a group are very similar, while objects in different clusters are very different.[2,3]

Hierarchical algorithms work differently from partitioning algorithms. Rather than partition the data into *k* clusters at an instance, they handle all values of *k* in a single run.[3] This means that a partition with a single cluster (*k* = 1) and a partition with each object forming a separate cluster (*k* = *n*) are presented together, with all values of *k* = 2, 3, . . . , *n*-1 shown in between as a gradual transition. This method can either be agglomerative (starting from *n* clusters and gradually combining them into one cluster) or divisive (starting from one cluster and gradually dividing them into *n* clusters). Hierarchical algorithms often require a priori knowledge on the number of clusters to be selected.[2]

Number of clusters

Determining the optimal number of clusters in a dataset has historically been a challenge.[2] There is no gold standard in approach for selecting an optimal number of clusters. Non testing procedures such as minimizing the within cluster dissimilarity while maximizing the between cluster dissimilarity[28] or using the average silhouette[27] are available options. More formal testing approaches are also available that use a null hypothesis that not more than one cluster exists in the dataset.[29,30]

2.4.2 Model-based clustering

Unlike nonparametric approaches, which require each object to explicitly belong to only one cluster, model-based clustering and related methods allows objects may belong to several clusters with varying probability of membership. This is known as fuzzy or soft clustering and allows for uncertainty in the data.[3] This is useful for objects between the boundary of two clusters. Assumptions on the data distribution are made a priori and have an advantage of providing more comprehensive information on the data structure.[3] The most used model for clustering longitudinal data is the finite mixture model.[20]

Just as with nonparametric approaches, determining the optimal number of clusters poses a challenge for researchers.[2] A common approach to selecting a model that uses the ideal number of clusters is the use of information theory methods. Akaike information criterion (AIC) and Bayesian information criterion (BIC) are two of the most popular information criterions used for this. Models are compared based on these criterion and the model with the smallest loss of information is chosen.[2]

3. Conclusion

The size of South Africa's VL monitoring program has had a tremendous growth in capacity in recent years despite initial struggles in its launch. The program is crucial to measuring the programmatic success of ART in the country. Cluster analysis may be an approach to uncover latent trajectories of individuals' VL patterns that may help in the development of targeted HIV prevention programs. This is especially true for large multidimensional datasets like that from NHLS data on routine VL monitoring. Cluster analysis has been demonstrated to have several advantages beyond conventional, subjective analysis of longitudinal data.

4. Areas for further research

The review has brought focus to the considerable body of evidence for cluster analysis in health research. Despite being applicable to numerous scenarios and settings, there are few studies that use cluster analysis in for HIV research in South Africa. There is an application for further HIV-related studies that use cluster analysis – not limited to analyzing trajectories – that may help identify HIV disease and patient subgroups that may have clinical relevance, if any exist. This information may help define better treatment approaches for PLHIV.

5. References

1. UNAIDS. UNAIDS Programme Coordinating Board sees South Africa's AIDS response first-hand. 2018 [updated November 7, 2018. Available from: https://www.unaids.org/en/resources/presscentre/featurestories/2018/november/pcb-field-visit-south-africa.

2. UNAIDS. Country factsheets: South Africa 2018 [Available from: https://www.unaids.org/en/regionscountries/countries/southafrica.

3. Human Sciences Research Council. South African Nationa HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017 2018 [Available from: http://www.hsrc.ac.za/uploads/pageContent/9234/FINAL%20Presentation%20for%2017%20July%20launch.pdf.

4. Bradshaw D, Msemburi W, Dorrington R, Pillay-van Wyk V, Laubscher R, Groenewald P. HIV/AIDS in South Africa: how many people died from the disease between 1997 and 2010? Aids. 2016;30(5):771-8.

5. Krishnaratne S, Hensen B, Cordes J, Enstone J, Hargreaves JR. Interventions to strengthen the HIV prevention cascade: a systematic review of reviews. The Lancet HIV. 2016;3(7):e307-e17.

6. Williams BG, Lima V, Gouws E. Modelling the impact of antiretroviral therapy on the epidemic of HIV. Curr HIV Res. 2011;9(6):367-82.

7. Chibawara T, Mbuagbaw L, Kitenge M, Nyasulu P. Effects of antiretroviral therapy in HIV-positive adults on new HIV infections among young women: a systematic review protocol. Systematic reviews. 2019;8(1):68.

8. Johnson LF, Dorrington RE, Moolla H. Progress towards the 2020 targets for HIV diagnosis and antiretroviral treatment in South Africa. Southern African journal of HIV medicine. 2017;18(1).

9. Johnson LF, Mossong J, Dorrington RE, Schomaker M, Hoffmann CJ, Keiser O, et al. Life expectancies of South African adults starting antiretroviral treatment: collaborative analysis of cohort studies. PLoS medicine. 2013;10(4):e1001418.

10. Statistics South Africa. Statistical release P0302: Mid-year population estimates 2018 [Available from: https://www.statssa.gov.za/publications/P0302/P03022018.pdf.

11. Avert. HIV and AIDS in South Africa [Available from: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa.

12. UNAIDS. UNAIDS DATA 2019 2019 [Available from: https://www.unaids.org/sites/default/files/media_asset/2019-UNAIDS-data_en.pdf.

13. Joint United Nations Programme on HIV/AIDS. 90-90-90—an ambitious treatment target to help end the AIDS epidemic. Geneva: UNAIDS; 2014 [Available from: https://www.unaids.org/sites/default/files/media_asset/90-90-90_en_0.pdf.

14. Bezabhe WM, Chalmers L, Bereznicki LR, Peterson GM. Adherence to Antiretroviral Therapy and Virologic Failure: A Meta-Analysis. Medicine. 2016;95(15):e3361-e.

15. Langebeek N, Gisolf EH, Reiss P, Vervoort SC, Hafsteinsdóttir TB, Richter C, et al. Predictors and correlates of adherence to combination antiretroviral therapy (ART) for chronic HIV infection: a meta-analysis. BMC medicine. 2014;12(1):142.

16. Katz IT, Ryu AE, Onuegbu AG, Psaros C, Weiser SD, Bangsberg DR, et al. Impact of HIV-related stigma on treatment adherence: systematic review and meta-synthesis. Journal of the International AIDS Society. 2013;16:18640.

17. Bernaards CM, Hildebrandt VH, Hendriksen IJM. Correlates of sedentary time in different age groups: results from a large cross sectional Dutch survey. BMC Public Health. 2016;16:1121.

18. UNAIDS. The need for routine viral load testing 2016 [Available from: https://www.unaids.org/sites/default/files/media_asset/JC2845_en.pdf.

19. World Health Organization. What's New in Treatment Monitoring: Viral Load and CD4 Testing. 2017.

20. Sigaloff KC, de Wit TFR. ART in sub-Saharan Africa: the value of viral load monitoring. The Lancet HIV. 2015;2(7):e261-e2.

21. National Department of Health. National consolidated guidelines for the prevention of mother-to-child transmission of HIV (PMTCT) and the management of HIV in children, adolescents and adults. National Department of Health Pretoria; 2015.

22. Shoko C, Chikobvu D. A superiority of viral load over CD4 cell count when predicting mortality in HIV patients on therapy. BMC Infectious Diseases. 2019;19(1):169.

23. National Health Laboratory Service. Who are we? : National Health Laboratory Service; 2019 [Available from: http://www.nhls.ac.za/?page=who_are_we&id=17.

24. Stevens WS, Marshall TM. Challenges in implenting HIV load testing in South Africa. The Journal of Infectious Diseases. 2010;201(Supplement_1):S78-S84.

25. National Health Laboratory Service. Annual Report 2018 [Available from: https://nationalgovernment.co.za/entity_annual/1714/2018-national-health-laboratory-service-(nhls)-annual-report.pdf.

26. Moyo F, Haeri Mazanderani A, Feucht UD, Ngoma K, Bhardwaj S, Goosen M, et al. Monitoring diagnosis, retention in care and viral load suppression in children testing HIV polymerase chain reaction-positive in two districts in South Africa2019.

27. Ellman TM, Alemayehu B, Abrams EJ, Arpadi S, Howard AA, El-Sadr WM. Selecting a viral load threshold for routine monitoring in resource-limited settings: optimizing individual health and population impact. Journal of the International AIDS Society. 2017;20 Suppl 7(Suppl 7):e25007.

28. El-Sadr WM, Rabkin M, Nkengasong J, Birx DL. Realizing the potential of routine viral load testing in sub-Saharan Africa. Journal of the International AIDS Society. 2017;20(S7):e25010.

29. Euvrard J, Schulz T, Hilderbrand K, Bosland M, Osler M, Boulle A, et al. How accurately do routinely reported HIV viral load suppression proportions reflect progress towards the 90-90-90 target in the population on antiretroviral treatment in Khayelitsha, South Africa?2019.

30. Heggeseth B. Longitudinal Cluster Analysis with Applications to Growth Trajectories: UC Berkeley; 2013.

31. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis: John Wiley & Sons; 2009.

32. Chen M, Ludwig SA, Keqin L. Clustering in Big Data. In: Li K-C, Jiang H, Zomaya A, editors. Big Data Management and Processing. New York. 2017.

33. Bukachi F, Pakenham-Walsh N. Information technology for health in developing countries. Chest. 2007;132(5):1624-30.

34. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. New England Journal of Medicine. 2010;362(3):192-5.

35. Schulam P, Wigley F, Saria S, editors. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015.

36. Windgassen S, Moss-Morris R, Goldsmith K, Chalder T. The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome. Taylor & Francis; 2018.

37. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. J Thorac Dis. 2015;7(11):E537-E40.

38. Pinaire J, Azé J, Bringay S, Landais P. Patient healthcare trajectory. An essential monitoring tool: a systematic review. Health Information Science and Systems. 2017;5(1):1.

39. UNAIDS. South Africa: UNAIDS; [Available from: https://www.unaids.org/en/regionscountries/countries/southafrica.

40. Keulenaer GWD, Brutsaert DL. The Heart Failure Spectrum. Circulation. 2009;119(24):3044-6.
41. Genolini C, Ecochard R, Benghezal M, Driss T, Andrieu S, Subtil F. kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes. Plos one. 2016;11(6):e0150738.
42. Henly SJ, Wyman JF, Findorff MJ. Health and illness over time: the trajectory perspective in nursing science. Nurs Res. 2011;60(3 Suppl):S5-S14.
43. Gong H, Xun X, Zhou Y. Profile clustering in clinical trials with longitudinal and functional data methods. Journal of biopharmaceutical statistics. 2019:1-17.
44. Coffey N, Hinde J, Holian E. Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. Computational Statistics & Data Analysis. 2014;71:14-29.
45. Ilmarinen P, Tuomisto LE, Niemelä O, Tommola M, Haanpää J, Kankaanranta H. Cluster analysis on longitudinal data of patients with adult-onset asthma. The Journal of Allergy and Clinical Immunology: In Practice. 2017;5(4):967-78. e3.
46. Pate RR, Schenkelberg MA, Dowda M, McIver KL. Group-based physical activity trajectories in children transitioning from elementary to high school. BMC Public Health. 2019;19(1):323.
47. Hirano S, Tsumoto S, editors. Cluster analysis of trajectory data on hospital laboratory examinations. AMIA Annual Symposium Proceedings; 2007: American Medical Informatics Association.
48. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature Communications. 2014;5(1):4022.
49. Gosset A, Protopopescu C, Larmarange J, Orne-Gliemann J, McGrath N, Pillay D, et al. Retention in Care Trajectories of HIV-Positive Individuals Participating in a Universal Test-and-Treat Program in Rural South Africa (ANRS 12249 TasP Trial). J Acquir Immune Defic Syndr. 2019;80(4):375-85.
50. Kassaye SG, Wang C, Ocampo JMF, Wilson TE, Anastos K, Cohen M, et al. Viremia Trajectories of HIV in HIV-Positive Women in the United States, 1994-2017. JAMA network open. 2019;2(5):e193822-e.
51. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;50(2):159-79.
52. Thinsungnoena T, Kaoungkub N, Durongdumronchaib P, Kerdprasopb K, Kerdprasopb N. The clustering validity with silhouette and sum of squared errors. learning. 2015;3:7.
53. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome biology. 2002;3(7):research0036. 1.

54. Yan M. Methods of determining the number of clusters in a data set and a new clustering criterion: Virginia Tech; 2005.

C. MANUSCRIPT

Title: Clustering of Longitudinal Viral Load Trajectories in the Western Cape

Author: Eke Nnanna Arua[1§]

1. Division of Epidemiology & Biostatistics, School of Public Health & Family Medicine, University of Cape Town

[§]Corresponding Author

Address: Division of Epidemiology & Biostatistics
School of Public Health & Family Medicine
University of Cape Town
Cape Town, Western Cape, South Africa, 7935

Email: eke2892@yahoo.com

Mobile: +27 21 650 4532

Keywords: HIV, Routine VL monitoring, VL trajectories, cluster analysis, South Africa

Journal: The Journal of the International AIDS Society (JIAS)

Manuscript word count (excluding Tables) = 2391

Abstract = 244

Tables: 4

Figures: 3

The article meets the requirements set out in the Instructions for Authors for the Journal of the International AIDS Society (JIAS) excluding competing interest, acknowledgement statement, and tables listed sequentially (for ease of review). As per the MPH mini-dissertation guidelines, sole authorship is reported but the contributions of co-authors are noted in the acknowledgments section of this dissertation. The JIAS Instructions for Authors are included in Appendix B of the dissertation.

Abstract

Introduction: Data from routine viral load monitoring is important for assessing the programmatic effectiveness of antiretroviral treatment (ART) in South Africa, but this is usually analysed in a cross-sectional manner and there are few analyses of longitudinal viral load trajectories over time. Cluster analysis was used to identify latent viral load trajectory patterns in a large dataset of routinely collected viral load measurements in the Western Cape, South Africa.

Methods: We analysed available VL measurements collected during routine care from the Western Cape public sector antiretroviral treatment programs, including all people living with HIV who were enrolled on ART and had experienced at least one elevated (>1000 copies/mL) viral load test result between 2008 and 2018. Empirical rules-based classification, nonparametric clustering using the KML-Shape algorithm and model-based clustering using latent class mixture modelling were used to cluster viral load trajectories.

Results: Both the nonparametric and model-based clustering techniques identified five latent viral load trajectory subgroups. The shapes and magnitudes of these subgroups differed according to method. Majority of individuals' trajectories belonged to clusters that had a decreasing VL trend. Most of the trajectory subgroups identified had prolonged periods of low-level viremia. Both methods identified viral load trajectory clusters that increased over time.

Conclusion: Cluster analysis is a useful tool for identifying latent VL trajectory clusters in the Western Cape and the dynamic VL cluster patterns that exist among the large VL dataset can offer important insights. Further research is needed to understand factors associated with belonging to these clusters to improve population viral load suppression rates and aid HIV prevention.

1. Introduction

South Africa has the largest ART program in the world.[12] The program has played a significant role in reducing HIV related mortality and improving HIV prevention efforts in the country.[13,14] It is estimated that over 6 million adult life years have been saved and the cumulative number of HIV related deaths in adults was lessened by 1.72 million because of the national ART program.[69] Routine viral load (VL) monitoring is important for evaluating the ART program effectiveness over times. Routine viral load (VL) monitoring is the World Health Organization (WHO) recommended strategy for measuring treatment response and identifying treatment failure among people living with HIV (PLHIV)[5,6]. Routinely collected data from treatment programs can be used to understand population viral suppression rates cross sectionally as well as longitudinally and to identifying groups of individuals who may have difficulty with ART adherence or be experiencing treatment failure.[18]

The conventional approach to analysing routine VL data is often focused on cross sectional viral suppression rates.[50,53,70,71] While cross-sectional approaches yield important information, they ignore individuals' longitudinal VL trajectory or patterns of viremia. Most work evaluating viral suppression has also used pre-specified population subgroups based on existing literature such as sex, race, geographic location, or age group. There has been a substantial increase in the volume of the country's VL testing services; reaching approximately 5 million viral load tests in 2017/18.[49] This increase in volume provides an exciting opportunity to use data-driven approaches to identify latent subgroups of VL trajectories, clusters not defined using a-priori categories. Cluster analysis is used in many disciplines including marketing, politics and economics and can be a powerful data exploration tool for finding unobserved patterns in multivariate data.[3]

There is a dearth of research on how longitudinal VL patterns for individuals who experience at least one elevated VL (defined as > 1000 copies/mL) change over time in South Africa. It is likely that there are several latent subgroups of VL trajectories for this population and an improved understanding of different patterns and frequency of pattern occurrence may be useful in targeting support programs. Discovering these subgroups of VL trajectories may help identify characteristics of individuals in care who are more likely to develop VL trajectories associated with negative outcomes before they occur and may provide deeper a deeper understanding of the underlying mechanism for their trajectories. This study aims to bridge the knowledge gap between routine VL monitoring and the type of longitudinal VL patterns that exist for PLHIV on ART in South Africa.

2. Methods

2.1 Data

We reviewed routinely collected HIV VL measurements from the Western Cape province in South Africa collected by the public sector ART program and processed by the National Health Laboratory Service (NHLS) over a 10 year period from 2008 to 2018. Data was provided as linked records giving a unique longitudinal time series for individuals. We excluded individuals from all analyses that either had less than two VL measurements or only had VL results of <1000 copies/mL throughout the 10 year observation period. There was minimal clinical and demographic characteristics available from the dataset, including age, date of VL test, sex and facility type.

Individual characteristics were summarized by frequency (percentage) or median (interquartile range) as appropriate. We used three thresholds for viral suppression to compare clustering results; the first was the World Health Organization (WHO) threshold defined as any VL <1000 copies/mL, the second was the South African national guidelines threshold defined as any VL <400 copies/mL, and the third was a very conservative threshold of any VL <50 copies/mL.

2.2 Clustering

Raw VL measurements (copies/mL) were $\log_{10}$ transformed prior to all analyses. An empirical rules-based approach to classification was used to describe the VL trajectories. Individuals were classified into one of 4 groups, defined by a single elevated (>1000 copies/mL) VL at or not at the first visit, multiple elevated VL with at least one VL <1000 copies/mL, and all VL elevated.

Both model-based and nonparametric approaches to cluster the VL trajectories were applied. A k-means approach for longitudinal data based on shape (KML-Shape) was applied to the data, first applying a data size reduction technique recommend by Genolini et al.[22] to make the algorithm less computationally expensive. For the model-based approach we first modelled individuals' repeated VL measurements using a mixed effects model with fractional polynomial terms to account for non-linear VL trajectories. The fractional polynomial terms that provide the best fit to the mixed effect model were then used in a latent class mixture model to identify distinct clusters of individuals' VL trajectories.

We compared the available demographic characteristics of individuals amongst the clusters obtained from both clustering methods to identify which factors may influence cluster membership. Longitudinal profiles

and mean cluster profiles were plotted for visual inspection and to assist with description of the resulting clusters. All analysis was carried out in R (R Foundation, Vienna, Austria).

The University of Cape Town Human Research Ethics Committee reviewed and approved the research protocol.

Table 1. Characteristics of individuals used for clustering.

Variables	Value
Number of individuals (N)	109092
Number of tests (N)	680751
Number of tests per person [median (IQR)]	5 (3, 9)
Duration between tests (months) [median (IQR)]	7.2 (5.1, 10.7)
Number of virally suppressed tests [N (%)]	
< 1000 copies/mL	397889 (58.4)
< 400 copies/mL	364399 (53.5)
< 50 copies/mL	282278 (41.5)
First test results [N (%)]	
< 1000 copies/mL	50580 (46.4)
< 400 copies/mL	46909 (43)
< 50 copies/mL	36985 (33.9)
Number of individuals always elevated [N (%)]	
> 1000 copies/mL	14297 (13.1)
> 400 copies/mL	17277 (15.8)
> 50 copies/mL	28944 (26.5)
Age in years [median (IQR)]	32 (26-38)
Sex [N (%)]	
Female	71080 (65.2)
Male	38012 (34.8)
Facility type [N (%)]	
Clinic	53055 (48.6)
Hospital	53928 (49.4)
Military	155 (0.1)
Prison	1954 (1.8)
VL categories [N (%)]	
Single elevated VL at first visit	22556 (20.7)
Single elevated VL not at first visit	24234 (22.2)
Multiple elevated VL with at least one suppressed VL	48005 (44)
Elevated throughout	14297 (13.1)

3. Results

3.1 Background characteristics of patients

Of the 469,161 individuals in the NHLS dataset, we removed 122,313 (26%) that had only one VL measurement, we then removed 237,561 (50.7%) that did not have at least one elevated VL (≥ 1000 copies/mL). A total of109,092 (23.3%) individuals and 680,751 VL tests remained eligible for analysis. The demographic characteristics of the individuals are displayed in Table 1. The majority of individuals were female (65.2%) and the median (IQR) age in years for participants was 32 (26-38). The median (IQR) number of months between tests for individuals was 7.2 (5.1-10.7). Most VL tests were performed at either a clinic (48.6%) or a hospital (49.4%). The proportion of VL test results that were virally suppressed at first visit was less than half using <1000 copies/mL threshold (46.4%), <400 copies/mL (43.0%) and <50 copies/mL (33.9%)

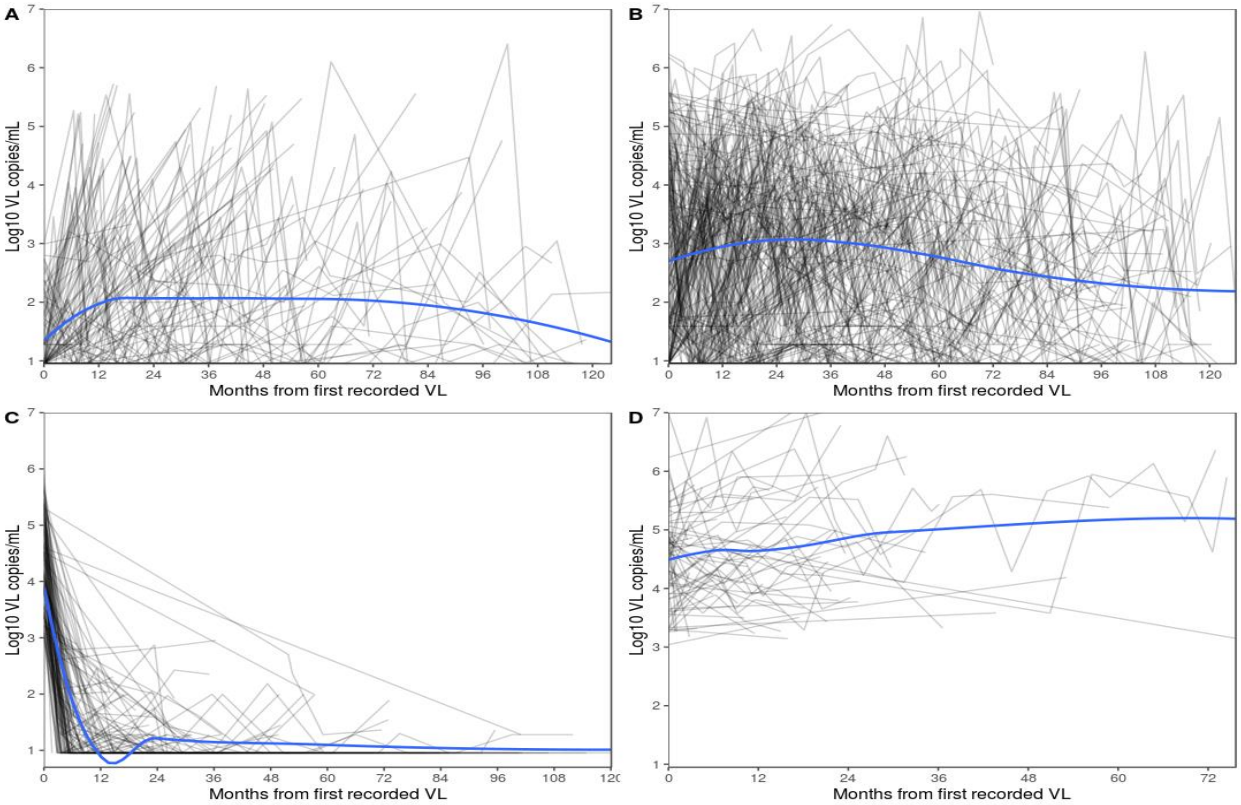

Figure 1. A sample of 500 individual VL trajectories with a loess smoother (blue line) illustrating pattern of change over time categorized into 4 groups. **A)** Single elevated not at first visit: Individuals with a single elevated VL that does not occur on their first visit. **B)** Multiple elevated: Individuals who have multiple elevated VLs with at least one suppressed VL. **C)** Single elevated at first: individuals who have a single elevated VL only at their first visit. **D)** Elevated throughout: Individuals who never have a suppressed VL

Because all participants had at least one elevated VL in the analysis set, we classified individuals based on the number and timing of their elevated VL test results into four categories. The largest single group of

individuals had multiple elevated VL with at least one VL < 1000 copies/mL (44%), 22% had a single elevated VL not occurring at the first visit, 21% a single elevated viral load occurring at the first visit and the remaining 13% had elevated VL (>1000 copies/mL) throughout (Table 1). A sample of 500 VL trajectories selected randomly from these groups is plotted with a loess smoother superimposed to show the nature of the longitudinal VL patterns (Figure 1).

3.2 Clusters identified

The KML-Shape approach yielded five VL trajectory clusters that varied in pattern and direction (Figure 2). Three clusters contained the majority of individuals and are described as "Elevated with steady decrease" (41%), "Steady low-level viremia (<1000 copies/mL)" (30%) and "Elevated with steady increase" (23%). The remaining two clusters contained a much smaller proportion of individuals and are described as "Elevated with rapid decrease" (5%) and "Elevated throughout" (<1%). Individuals in the "Elevated throughout" cluster did not have records longer than 72 months, which may be an indicator of late entry to the data set or loss to follow up from the VL monitoring program.

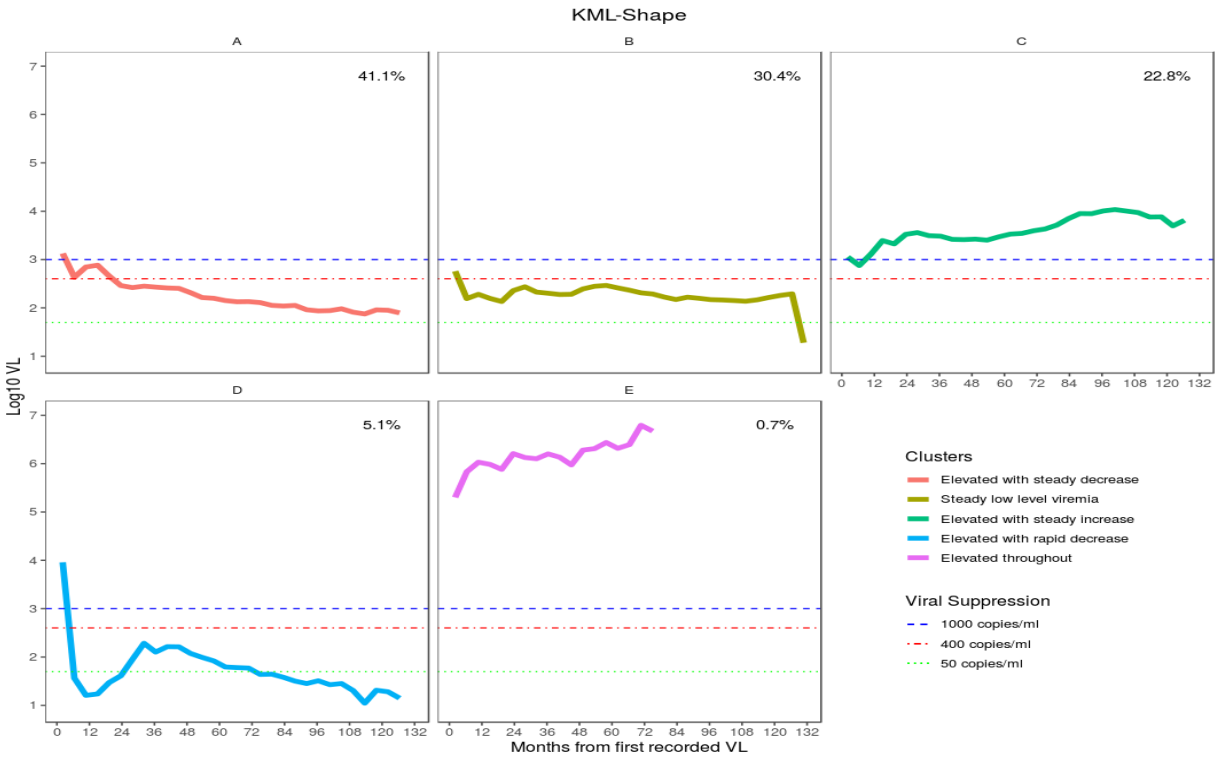

Figure 2. Five VL trajectory clusters identified using the nonparametric approach KML-Shape. **A)** Initially elevated, then steady decline VL trajectory cluster. **B)** < 1000 copies/mL throughout. **C)** Initially elevated, then steady increase VL trajectory cluster. **D)** Initially elevated, then rapid decline. **E)** Very high elevated VL. Trajectories are plotted by averaging over a bin of 25 time points. The proportion of individuals classified to each cluster is indicated at the top right.

This observation was similar to the sample VL trajectories identified through the rules-based approach as "elevated throughout" and shown in Figure 1D that have shorter than average follow up periods (<60 months).

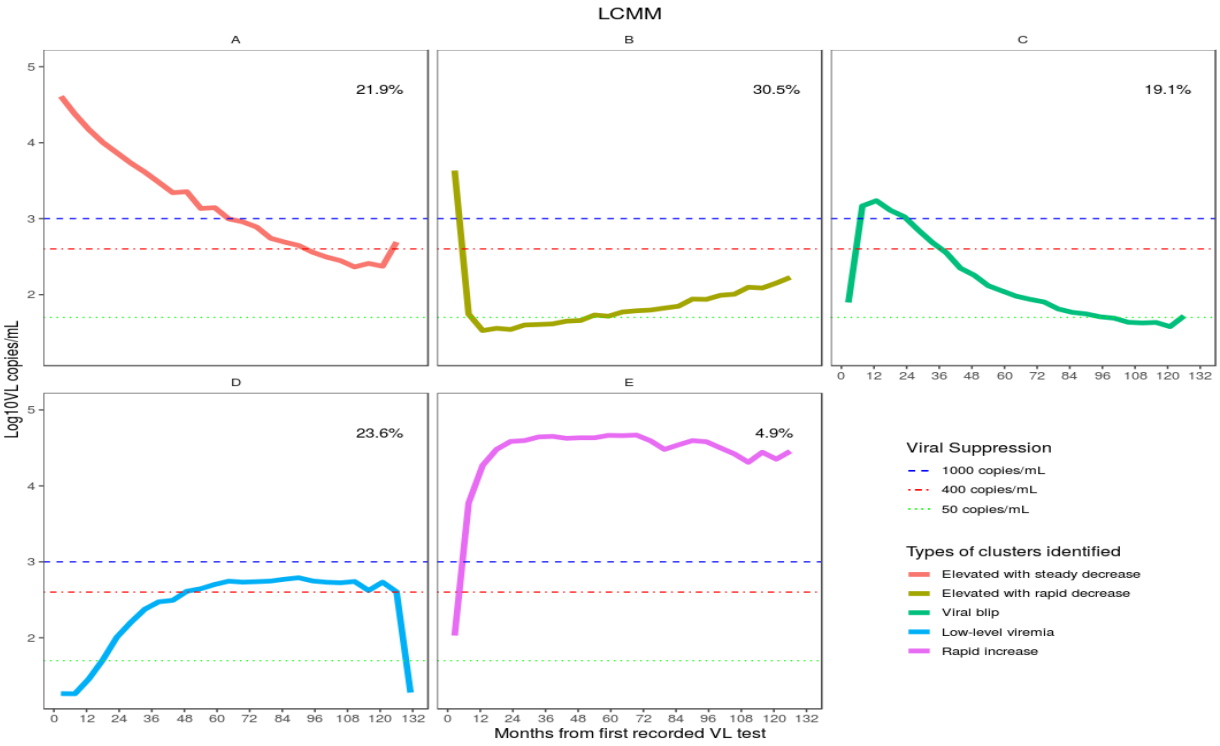

Figure 3. Five VL trajectory clusters identified using model-based approach LCMM with fractional polynomials. Proportion of individuals are displayed in the top right of each cluster. **A)** Initially elevated, then steady decline. **B)** Initially elevated, then rapid decline. **C)** Elevated blip. **D)** Undetected, then low-level viremia. **E)** Rapid elevation. Trajectories are plotted by averaging over a bin of 25 time points. The proportion of individuals classified to each cluster is indicated at the top right.

The LCMM approach also identified five VL trajectory clusters which generally differed in shape compared to those of the KML-Shape approach (Figure 3). Somewhat unlike the KML-Shape approach, individuals were more broadly dispersed, with four clusters containing 19% or more of the individuals. These four main clusters can be described as "Elevated with steady decrease" (22%), "Elevated with rapid decrease" (31%), "Viral blip" (19%), "Low-level viremia" (24%). The fifth cluster, containing only 5% of individuals can be described as one of "Rapid increase". The mean VL trajectories resulting from this approach were more dynamic in shape compared to the KLM-Shape approach and although have been described using similar terms, the range of VL values covered by each cluster also differed (Figure 2 and Figure 3). The mean LCMM VL trajectories captured more rapid change, which can be observed by the relatively steeper slopes when compared to those of KML-Shape.

Table 2. A comparison of cluster types between both KML-Shape and LCMM to Empirical clusters in the dataset

	Empirical	KML-Shape	LCMM
Type of cluster obtained*			
Elevated with rapid decrease	22556 (20.7)	5513 (5.1)	33318 (30.5)
Elevated with steady decrease		44785 (41.1)	23851 (21.9)
Viral Blip	48005 (44)		20878 (19.1)
Low-level viremia	24234 (22.2)	33203 (30.4)	25737 (23.6)
Suppressed then sharp increase			5308 (4.9)
Elevated with steady increase		24869 (22.8)	
Elevated VL throughout	14297 (13.1)	722 (0.7)	

*Blanks indicate that the cluster type was not found under the method.

Table 3. Characteristics of individuals stratified by cluster types obtained from KML-Shape

	Clusters from KML-Shape				
	Elevated with steady decrease	**Steady low-level viremia**	**Elevated with steady increase**	**Elevated with rapid decrease**	**Elevated VL throughout**
Number [N (%)]	44785 (41.1)	33203 (30.4)	24869 (22.8)	5513 (5.1)	722 (0.7)
Gender [N (%)]					
Female	29557 (66)	21890 (65.9)	15807 (63.6)	3400 (61.7)	426 (59)
Male	15228 (34)	11313 (34.1)	9062 (36.4)	2113 (38.3)	296 (41)
Facility Type [N (%)]					
Clinic	22095 (49.3)	15570 (46.9)	12512 (50.3)	2513 (45.6)	365 (50.6)
Hospital	21731 (48.5)	17125 (51.6)	11902 (47.9)	2824 (51.2)	346 (47.9)
Military	74 (0.2)	39 (0.1)	32 (0.1)	10 (0.2)	N/A
Prison	885 (2)	469 (1.4)	423 (1.7)	166 (3)	11 (1.5)
Age [median (IQR)]	31 (26-38)	32 (25-38)	31 (26-38)	32 (26-39)	32 (26-40)
Duration between tests (months) [median (IQR)]	6.1 (3.8-10.8)	6.3 (4.3-9.5)	6.7 (4.2-12.2)	6.5 (4.4-9.1)	5.2 (2.9-9.7)
Number of tests [median (IQR)]	5 (3-8)	7 (4-10)	4 (3-7)	5 (3-7)	2 (2-3)
Number of individuals suppressed at first test [N (%)]					
VL < 1000	18095 (40.4)	18050 (54.4)	13264 (53.3)	1103 (20)	68 (9.4)
VL < 400	16461 (36.8)	17018 (51.3)	12358 (49.7)	1012 (18.4)	60 (8.3)
VL < 50	12334 (27.5)	14264 (43)	9590 (38.6)	765 (13.9)	32 (4.4)
Number of individuals always elevated [N (%)]					
VL always > 1000	5415 (12.1)	425 (1.3)	7473 (30)	1 (0)	608 (84.2)
VL always > 400	6490 (14.5)	1630 (4.9)	8469 (34.1)	19 (0.3)	619 (85.7)
VL always >50	11288 (25.2)	5724 (17.2)	11131 (44.8)	152 (2.8)	649 (89.9)
VL categories [N (%)]					
Single elevated VL at first visit	11763 (26.3)	7285 (21.9)	101 (0.4)	3407 (61.8)	N/A
Single elevated VL not at first visit	9261 (20.7)	8722 (26.3)	5598 (22.5)	609 (11)	44 (6.1)
Multiple elevated VL with at least one suppressed VL	18201 (40.6)	16754 (50.5)	11521 (46.3)	1496 (27.1)	33 (4.6)
Elevated throughout	5560 (12.4)	442 (1.3)	7649 (30.8)	1 (0)	645 (89.3)

Table 2 compares the methods of classification using the descriptive terminology along with the number (percent) of individuals classified into each group. By inspection, only the "Elevated with rapid decrease" VL cluster type was present under all three approaches, aligning very much to a first elevated VL followed by viral suppression. However, the sizes of these clusters varied with KML-Shape being the smallest. KML-Shape identified more cluster types (3 in total) that were expected a priori compared to LCMM (2 in total); however, the difference in proportion of individuals in each cluster type for all three varied greatly suggesting possible misclassifications of VL trajectory clusters.

Table 4. Characteristics of individuals stratified by cluster type obtained from LCMM

	Clusters from LCMM				
	Elevated with steady decrease	**Elevated with rapid decrease**	**Viral blip**	**Low-level viremia**	**Rapid increase**
Number of individuals [N (%)]	23851 (21.9)	33318 (30.5)	20878 (19.1)	25737 (23.6)	5308 (4.9)
Gender [N (%)]					
Female	14515 (60.9)	22081 (66.3)	13992 (67)	17344 (67.4)	3148 (59.3)
Male	9336 (39.1)	11237 (33.7)	6886 (33)	8393 (32.6)	2160 (40.7)
Facility type [N (%)]					
Clinic	12082 (50.7)	15635 (46.9)	10620 (50.9)	11997 (46.6)	2721 (51.3)
Hospital	11145 (46.7)	16591 (49.8)	10060 (48.2)	13596 (52.8)	2536 (47.8)
Military	55 (0.2)	85 (0.3)	2 (0)	11 (0)	2 (0)
Prison	569 (2.4)	1007 (3)	196 (0.9)	133 (0.5)	49 (0.9)
Age in years [median (IQR)]	31 (25-38)	31 (25-38)	32 (26-38)	32 (26-39)	32 (27-39)
Duration between tests (months) [median (IQR)]	5.5 (3.3-10.5)	5 (3.4-8.3)	7.5 (5.5-12)	6.9 (5.5-10)	9.2 (5.9-19.4)
Number of tests [median (IQR)]	4 (2-7)	4 (3-7)	6 (3-10)	7 (5-11)	5 (3-8)
Number of first test results [N (%)]					
VL < 1000 copies/mL	588 (2.5)	579 (1.7)	19339 (92.6)	25350 (98.5)	4724 (89)
VL < 400 copies/mL	146 (0.6)	283 (0.8)	17439 (83.5)	24798 (96.4)	4243 (79.9)
VL < 50 copies/mL	14 (0.1)	182 (0.5)	12611 (60.4)	21553 (83.7)	2625 (49.5)
Number of individuals always elevated [N (%)]					
VL always >1000 copies/mL	13217 (55.4)	668 (2)	186 (0.9)	9 (0)	217 (4.1)
VL always > 400 copies/mL	14687 (61.6)	1427 (4.3)	906 (4.3)	26 (0.1)	611 (11.5)
VL always > 50 copies/mL	17950 (75.3)	5117 (15.4)	3813 (18.3)	436 (1.7)	2251 (42.4)
VL categories [N (%)]					
Single elevated VL at first visit	10396 (43.6)	9983 (30)	11565 (55.4)	11940 (46.4)	4121 (77.6)
Single elevated VL not at first visit	152 (0.6)	417 (1.3)	8989 (43.1)	13706 (53.3)	970 (18.3)
Multiple elevated VL with at least one suppressed VL	86 (0.4)	22250 (66.8)	138 (0.7)	82 (0.3)	N/A
Elevated throughout	13217 (55.4)	668 (3)	186 (0.9)	9 (0)	217 (2.1)

Individuals and test characteristics were compared across the different clusters obtained from both LCMM and KML-Shape approaches (Tables 3 & 4). In the LCMM method, median duration between tests in months was shorter for the VL trajectory clusters that showed a downward trend compared to those that

had a rising trend. The median number of tests was also lower for downward trend clusters compared to the other clusters, suggesting that LCMM is in part classifying individuals based on the number of available tests. There was a negligible difference in the median duration between tests in months for all clusters in the KML-Shape, with the exception of the "Elevated VL throughout" cluster. The "Elevated throughout" cluster had a higher proportion of males when compared to other clusters (KML-Shape).

4. Discussion

This study identified five latent VL trajectory clusters using a nonparametric and model-based approach for individuals in the Western Cape who have had at least one elevated VL. To the best of our knowledge, a study of this nature and has not been carried out using routine VL data in South Africa. The findings of the study provided a unique insight on the VL trajectory patterns of individuals who have had at least one elevated VL over a 10-year period. It builds upon the base of knowledge for VL monitoring and ART performance in South Africa. Study findings showed that beyond the trajectories expected to be in the data, there are several other latent trajectories that require further investigation. The likelihood of being classified in groups with undesirable VL trajectories may be due to factors such as poor adherence or the distance to the nearest health care facility; targeted program interventions can be directed towards high risk patients to improve their long-term outcomes. The analysis presented also demonstrates the importance of method selection, as the three approaches did not align well in terms of classification to common clusters.

Most trajectories identified in this analysis, and the largest proportion of individuals, were assigned to clusters that had an overall trend towards viral suppression. This is a positive finding and suggests that overall, individuals with elevated viral loads achieve re-suppression. Under all approaches there remained a small cluster of individuals with consistently elevated VL results. This observation is concerning as it suggests there is a small group of individuals in care who remain viremic. This group differed from other clusters only in the average number of tests, and so may reflect a shortcoming of the data rather than a clinical issue of concern.

All VL trajectory clusters identified that showed an initially elevated VL usually declined – either rapidly or steadily – towards viral suppression but still had prolonged periods of low-level viremia (LLV). LLV can be defined using different thresholds depending on the level of development of the country, but in the case of South Africa is generally considered as either between 50 to 400 copies/mL (very low) or 400 to 1000 copies/mL. LLV is associated with an increased risk of virologic failure, switch to second-line ART and HIV

related mortality.[72] The finding that there may be reason to consider the rate of viral load change over time is an important one that should be further investigated.

This study has several strengths. Firstly, the discovery of these latent VL trajectories was done using objective, data-driven approaches. Secondly, the large number of records in the NHLS database covers the entirety of the region which greatly reduces the occurrence of selection bias. Furthermore, the high rates of public sector care in the Western Cape over a considerable time period means there is good population representation of individuals in care.

There are also several limitations to the study. Because the data was collected from routine data, the number of covariates, and hence possible risk factors, was limited, meaning further investigation to ascertain factors associated with membership of clusters of interest was not possible. The time between VL measurements are unevenly spaced out. As a result, there may be long periods between measurements were the VL may have varied significantly. The lack of data between these periods may be considered missing in this context as it would have been useful for clustering. Another limitation was that there was little agreement in the level, slope and size of the clusters obtained between the two approaches. Although, this outcome was expected because the KML-Shape algorithm focuses more on grouping trajectories of similar shapes together rather than trajectories of similar level, it leads to difficulty in interpretation of the resulting clusters. The demonstrated difference in cluster size between the two approaches and what was shown empirically is noteworthy but also understandable within the context of the clustering approaches not using explicit elevated VL thresholds.

5. Conclusion

Data from routine VL monitoring has a greater potential for offering more insights on the patterns of VL progression over time. Cluster analysis provided a useful tool for unearthing previously unobserved patterns in large heterogenous data like that of NHLS VL data. Further research is needed to understand how these patterns affect population suppression rates which may be able to contribute to ongoing efforts in trying to HIV eradication.

6. References

1. Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinformatics. 2014;15 Suppl 11(Suppl 11):S11-S.
2. Heggeseth B. Longitudinal Cluster Analysis with Applications to Growth Trajectories: UC Berkeley; 2013.
3. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis: John Wiley & Sons; 2009.
4. Windgassen S, Moss-Morris R, Goldsmith K, Chalder T. The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome. Taylor & Francis; 2018.
5. Chan PA, Rose J, Maher J, Benben S, Pfeiffer K, Almonte A, et al. A latent class analysis of risk factors for acquiring HIV among men who have sex with men: implications for implementing pre-exposure prophylaxis programs. AIDS patient care and STDs. 2015;29(11):597-605.
6. van Gorp WG, Hinkin C, Satz P, Miller EN, Weisman J, Holston S, et al. Subtypes of HIV-related neuropsychological functioning: A cluster analysis approach. Neuropsychology. 1993;7(1):62.
7. Grabowski MK, Herbeck JT, Poon AF. Genetic cluster analysis for HIV prevention. Current HIV/AIDS Reports. 2018;15(2):182-9.
8. Mugavero MJ, Amico KR, Westfall AO, Crane HM, Zinski A, Willig JH, et al. Early retention in HIV care and viral load suppression: implications for a test and treat approach to HIV prevention. J Acquir Immune Defic Syndr. 2012;59(1):86-93.
9. Farooq S, Weisenthal S, Trayhan M, White R, Bush K, Mariuz P, et al. Revealing HIV viral load patterns using unsupervised machine learning and cluster summarization [version 1; peer review: 1 approved, 1 approved with reservations]. F1000Research. 2018;7(1144).
10. Human Sciences Research Council. South African Nationa HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017 2018 [Available from: http://www.hsrc.ac.za/uploads/pageContent/9234/FINAL%20Presentation%20for%2017%20July%20launch.pdf.
11. Mayosi BM, Lawn JE, Van Niekerk A, Bradshaw D, Karim SSA, Coovadia HM, et al. Health in South Africa: changes and challenges since 2009. The Lancet. 2012;380(9858):2029-43.
12. UNAIDS. South Africa: UNAIDS; [Available from: https://www.unaids.org/en/regionscountries/countries/southafrica.
13. Boulle A, Schomaker M, May MT, Hogg RS, Shepherd BE, Monge S, et al. Mortality in patients with HIV-1 infection starting antiretroviral therapy in South Africa, Europe, or North America: a collaborative analysis of prospective studies. PLoS medicine. 2014;11(9):e1001718.
14. UNAIDS. UNAIDS DATA 2018 2018 [cited 2019 06/06/2019]. Available from: https://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf.
15. Johnson LF, Mossong J, Dorrington RE, Schomaker M, Hoffmann CJ, Keiser O, et al. Life expectancies of South African adults starting antiretroviral treatment: collaborative analysis of cohort studies. PLoS medicine. 2013;10(4):e1001418.
16. Joint United Nations Programme on HIV/AIDS. 90-90-90—an ambitious treatment target to help end the AIDS epidemic. Geneva: UNAIDS; 2014 [Available from: https://www.unaids.org/sites/default/files/media_asset/90-90-90_en_0.pdf.
17. Johnson LF, Dorrington RE, Moolla H. Progress towards the 2020 targets for HIV diagnosis and antiretroviral treatment in South Africa. Southern African journal of HIV medicine. 2017;18(1).
18. UNAIDS. The need for routine viral load testing 2016 [Available from: https://www.unaids.org/sites/default/files/media_asset/JC2845_en.pdf.
19. National Health Laboratory Service. Who are we? : National Health Laboratory Service; 2019 [Available from: http://www.nhls.ac.za/?page=who_are_we&id=17.

20. Gong H, Xun X, Zhou Y. Profile clustering in clinical trials with longitudinal and functional data methods. Journal of biopharmaceutical statistics. 2019:1-17.
21. Hartigan JA. Statistical Clustering. In: Wright JD, editor. International Encyclopedia of the Social & Behavioral Sciences (Second Edition). Oxford: Elsevier; 2015. p. 392-6.
22. Genolini C, Falissard B. KmL: k-means for longitudinal data. Computational Statistics. 2010;25(2):317-28.
23. Genolini C, Ecochard R, Benghezal M, Driss T, Andrieu S, Subtil F. kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes. Plos one. 2016;11(6):e0150738.
24. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1994;43(3):429-53.
25. Jung T, Wickrama K. An introduction to latent class growth analysis and growth mixture modeling. Social and personality psychology compass. 2008;2(1):302-17.
26. The SAGE Handbook of Quantitative Methodology for the Social Sciences. 2004 2019/06/30. SAGE Publications, Inc. Available from: https://methods.sagepub.com/book/the-sage-handbook-of-quantitative-methodology-for-the-social-sciences.
27. Thinsungnoena T, Kaoungkub N, Durongdumronchaib P, Kerdprasopb K, Kerdprasopb N. The clustering validity with silhouette and sum of squared errors. learning. 2015;3:7.
28. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;50(2):159-79.
29. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome biology. 2002;3(7):research0036. 1.
30. Yan M. Methods of determining the number of clusters in a data set and a new clustering criterion: Virginia Tech; 2005.
31. UNAIDS. UNAIDS Programme Coordinating Board sees South Africa's AIDS response first-hand. 2018 [updated November 7, 2018. Available from: https://www.unaids.org/en/resources/presscentre/featurestories/2018/november/pcb-field-visit-south-africa.
32. UNAIDS. Country factsheets: South Africa 2018 [Available from: https://www.unaids.org/en/regionscountries/countries/southafrica.
33. Bradshaw D, Msemburi W, Dorrington R, Pillay-van Wyk V, Laubscher R, Groenewald P. HIV/AIDS in South Africa: how many people died from the disease between 1997 and 2010? Aids. 2016;30(5):771-8.
34. Krishnaratne S, Hensen B, Cordes J, Enstone J, Hargreaves JR. Interventions to strengthen the HIV prevention cascade: a systematic review of reviews. The Lancet HIV. 2016;3(7):e307-e17.
35. Williams BG, Lima V, Gouws E. Modelling the impact of antiretroviral therapy on the epidemic of HIV. Curr HIV Res. 2011;9(6):367-82.
36. Chibawara T, Mbuagbaw L, Kitenge M, Nyasulu P. Effects of antiretroviral therapy in HIV-positive adults on new HIV infections among young women: a systematic review protocol. Systematic reviews. 2019;8(1):68.
37. Statistics South Africa. Statistical release P0302: Mid-year population estimates 2018 [Available from: https://www.statssa.gov.za/publications/P0302/P03022018.pdf.
38. Avert. HIV and AIDS in South Africa [Available from: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa.
39. UNAIDS. UNAIDS DATA 2019 2019 [Available from: https://www.unaids.org/sites/default/files/media_asset/2019-UNAIDS-data_en.pdf.
40. Bezabhe WM, Chalmers L, Bereznicki LR, Peterson GM. Adherence to Antiretroviral Therapy and Virologic Failure: A Meta-Analysis. Medicine. 2016;95(15):e3361-e.

41. Langebeek N, Gisolf EH, Reiss P, Vervoort SC, Hafsteinsdóttir TB, Richter C, et al. Predictors and correlates of adherence to combination antiretroviral therapy (ART) for chronic HIV infection: a meta-analysis. BMC medicine. 2014;12(1):142.
42. Katz IT, Ryu AE, Onuegbu AG, Psaros C, Weiser SD, Bangsberg DR, et al. Impact of HIV-related stigma on treatment adherence: systematic review and meta-synthesis. Journal of the International AIDS Society. 2013;16:18640.
43. Bernaards CM, Hildebrandt VH, Hendriksen IJM. Correlates of sedentary time in different age groups: results from a large cross sectional Dutch survey. BMC Public Health. 2016;16:1121.
44. World Health Organization. What's New in Treatment Monitoring: Viral Load and CD4 Testing. 2017.
45. Sigaloff KC, de Wit TFR. ART in sub-Saharan Africa: the value of viral load monitoring. The Lancet HIV. 2015;2(7):e261-e2.
46. National Department of Health. National consolidated guidelines for the prevention of mother-to-child transmission of HIV (PMTCT) and the management of HIV in children, adolescents and adults. National Department of Health Pretoria; 2015.
47. Shoko C, Chikobvu D. A superiority of viral load over CD4 cell count when predicting mortality in HIV patients on therapy. BMC Infectious Diseases. 2019;19(1):169.
48. Stevens WS, Marshall TM. Challenges in implenting HIV load testing in South Africa. The Journal of Infectious Diseases. 2010;201(Supplement_1):S78-S84.
49. National Health Laboratory Service. Annual Report 2018 [Available from: https://nationalgovernment.co.za/entity_annual/1714/2018-national-health-laboratory-service-(nhls)-annual-report.pdf.
50. Moyo F, Haeri Mazanderani A, Feucht UD, Ngoma K, Bhardwaj S, Goosen M, et al. Monitoring diagnosis, retention in care and viral load suppression in children testing HIV polymerase chain reaction-positive in two districts in South Africa2019.
51. Ellman TM, Alemayehu B, Abrams EJ, Arpadi S, Howard AA, El-Sadr WM. Selecting a viral load threshold for routine monitoring in resource-limited settings: optimizing individual health and population impact. Journal of the International AIDS Society. 2017;20 Suppl 7(Suppl 7):e25007.
52. El-Sadr WM, Rabkin M, Nkengasong J, Birx DL. Realizing the potential of routine viral load testing in sub-Saharan Africa. Journal of the International AIDS Society. 2017;20(S7):e25010.
53. Euvrard J, Schulz T, Hilderbrand K, Bosland M, Osler M, Boulle A, et al. How accurately do routinely reported HIV viral load suppression proportions reflect progress towards the 90-90-90 target in the population on antiretroviral treatment in Khayelitsha, South Africa?2019.
54. Chen M, Ludwig SA, Keqin L. Clustering in Big Data. In: Li K-C, Jiang H, Zomaya A, editors. Big Data Management and Processing. New York. 2017.
55. Bukachi F, Pakenham-Walsh N. Information technology for health in developing countries. Chest. 2007;132(5):1624-30.
56. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. New England Journal of Medicine. 2010;362(3):192-5.
57. Schulam P, Wigley F, Saria S, editors. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015.
58. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. J Thorac Dis. 2015;7(11):E537-E40.
59. Pinaire J, Azé J, Bringay S, Landais P. Patient healthcare trajectory. An essential monitoring tool: a systematic review. Health Information Science and Systems. 2017;5(1):1.
60. Keulenaer GWD, Brutsaert DL. The Heart Failure Spectrum. Circulation. 2009;119(24):3044-6.

61. Henly SJ, Wyman JF, Findorff MJ. Health and illness over time: the trajectory perspective in nursing science. Nurs Res. 2011;60(3 Suppl):S5-S14.
62. Coffey N, Hinde J, Holian E. Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. Computational Statistics & Data Analysis. 2014;71:14-29.
63. Ilmarinen P, Tuomisto LE, Niemelä O, Tommola M, Haanpää J, Kankaanranta H. Cluster analysis on longitudinal data of patients with adult-onset asthma. The Journal of Allergy and Clinical Immunology: In Practice. 2017;5(4):967-78. e3.
64. Pate RR, Schenkelberg MA, Dowda M, McIver KL. Group-based physical activity trajectories in children transitioning from elementary to high school. BMC Public Health. 2019;19(1):323.
65. Hirano S, Tsumoto S, editors. Cluster analysis of trajectory data on hospital laboratory examinations. AMIA Annual Symposium Proceedings; 2007: American Medical Informatics Association.
66. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature Communications. 2014;5(1):4022.
67. Gosset A, Protopopescu C, Larmarange J, Orne-Gliemann J, McGrath N, Pillay D, et al. Retention in Care Trajectories of HIV-Positive Individuals Participating in a Universal Test-and-Treat Program in Rural South Africa (ANRS 12249 TasP Trial). J Acquir Immune Defic Syndr. 2019;80(4):375-85.
68. Kassaye SG, Wang C, Ocampo JMF, Wilson TE, Anastos K, Cohen M, et al. Viremia Trajectories of HIV in HIV-Positive Women in the United States, 1994-2017. JAMA network open. 2019;2(5):e193822-e.
69. Johnson LF, May MT, Dorrington RE, Cornell M, Boulle A, Egger M, et al. Estimating the impact of antiretroviral treatment on adult mortality trends in South Africa: A mathematical modelling study. PLoS medicine. 2017;14(12):e1002468.
70. Woldesenbet SA, Kufa T, Barron P, Chirombo BC, Cheyip M, Ayalew K, et al. Viral suppression and factors associated with failure to achieve viral suppression among pregnant women in South Africa: a national cross-sectional survey. AIDS. 2019.
71. Hunt GM, Dokubo EK, Takuva S, de Oliveira T, Ledwaba J, Dube N, et al. Rates of virological suppression and drug resistance in adult HIV-1-positive patients attending primary healthcare facilities in KwaZulu-Natal, South Africa. Journal of Antimicrobial Chemotherapy. 2017;72(11):3141-8.
72. Bernal E, Gómez JM, Jarrín I, Cano A, Muñoz A, Alcaraz A, et al. Low-level viremia is associated with clinical progression in HIV-infected patients receiving antiretroviral treatment. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2018;78(3):329-37.

D. APPENDICES

Appendix A. Ethics approval for study

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee

Room E53-46 Old Main Building
Groote Schuur Hospital
Observatory 7925
Telephone [021] 406 6492
Email: sumayah.arlefdien@uct.ac.za
Website: www.health.uct.ac.za/fhs/research/humanethics/forms

28 October 2019

HREC REF: 728/2019

A/Prof M Lesosky
Division of Epidemiology & Biostatistics
5.39 Falmouth Building
FHS

Dear A/Prof Lesosky

PROJECT TITLE: CLUSTERING OF LONGITUDINAL VIRAL LOAD TRAJECTORIES IN THE WESTERN CAPE (SUB-STUDY _ 329/2014) (MASTER'S DEGREE - MR EKE ARUA)

Thank you for submitting your study to the Faculty of Health Sciences Human Research Ethics Committee (HREC) for review.

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

Approval is granted for one year until the 30 October 2020.

Please submit a progress form, using the standardised Annual Report Form if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.
(Forms can be found on our website: www.health.uct.ac.za/fhs/research/humanethics/forms)

The HREC acknowledge that the student: Mr Eke Arua will also be involved in this study.

Please quote the HREC REF in all your correspondence.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please note that for all studies approved by the HREC, the principal investigator **must** obtain appropriate institutional approval, where necessary, before the research may occur.

Yours sincerely

Signature Removed

PROFESSOR M BLOCKMAN
CHAIRPERSON, FHS HUMAN RESEARCH ETHICS COMMITTEE

Federal Wide Assurance Number: FWA00001637.
Institutional Review Board (IRB) number: IRB00001938
NHREC-registration number: REC-210208-007

HREC 728/2019

Appendix B. Journal submission guidelines

Sections

1. SUBMISSION

Please carefully read through the Instructions for Authors and prepare your manuscript according to the guidelines, including structuring it manuscript based on the chosen article category. Manuscripts that do not follow the instructions may be returned to the authors for corrections.

Authors should kindly note that submission implies that the content has not been published or submitted for publication elsewhere except as a brief abstract in the proceedings of a scientific meeting or symposium.

Once the submission materials have been prepared in accordance with the Author Guidelines, manuscripts should be submitted online at https://mc.manuscriptcentral.com/jias. The submission system will prompt authors to use an ORCID iD (a unique author identifier) to help distinguish their work from that of other researchers. Click here to find out more.

You will be asked to suggest potential peer reviewers for your manuscript: they should be experts in the field and be able to provide an objective assessment of the manuscript. Any suggested peer reviewers should not have published with any of the authors of the manuscript within the past five years, should not be current collaborators, and should not be members of the same institution. Suggested reviewers will be considered alongside potential reviewers identified by the Editorial team.

Click here for more details on how to use ScholarOne.

2. AIMS AND SCOPE

The *JIAS* welcomes submissions on HIV-related topics from across all scientific disciplines, including but not limited to:

- Basic and biomedical sciences
- Behavioural sciences
- Epidemiology
- Clinical sciences
- Health economics and health policy
- Operations research and implementation sciences
- Social sciences and humanities, including political sciences and media

The *JIAS* prioritizes submissions from operational research and implementation science as publication of such material can provide valuable information on various algorithms for monitoring

and providing support for comprehensive, yet affordable and sustainable treatment, prevention and care programmes in different contexts.

Submission of HIV research carried out in low- and middle-income countries is strongly encouraged.

3. MANUSCRIPT CATEGORIES AND REQUIREMENTS

The *JIAS* accepts submissions in the following categories:

- **Research**
- **Short report**
- **Review**
- **Debate**
- **Commentary**
- **Letter to the Editor**
- **Viewpoint**

Research - full reports of data from original research studies

Abstract:
Headings: Introduction, Methods, Results, Conclusions
Word limit: 350 words

Main text:
Headings: Introduction, Methods, Results, Discussion, Conclusions
Word limit: 3500 words
Numbers of figures and tables: Unlimited
Additional files: Yes

Download the manuscript template

Short report - brief reports of data from original research, such as follow-up or confirmatory studies, case series and negative results

Abstract:
Headings: Introduction, Methods, Results and discussion, Conclusions
Word limit: 350 words

Main text:
Headings: Introduction, Methods, Results and discussion, Conclusions
Word limit: 2000 words
Numbers of figures and tables: 3
Additional files: No

Download the manuscript template

Review - comprehensive, authoritative descriptions and summaries of a specific subject area providing a systematic and substantial overview of the field

Abstract:
Headings: Introduction, Methods, Results and discussion (if applicable, otherwise Discussion only), Conclusions
Word limit: 350 words

Main text:
Headings: Introduction, Methods, Results and discussion (if applicable, otherwise Discussion only), Conclusions
Word limit: 5000 words
Numbers of figures and tables: Unlimited
Additional files: Yes

Download the manuscript template

Debate - presentation of an evidence-based argument

Abstract:
Headings: Introduction, Discussion, Conclusions
Word limit: 350 words

Main text:
Headings: Introduction, Discussion, Conclusions
Word limit: 3500 words
Numbers of figures and tables: 4
Additional files: No

Download the manuscript template

Commentary - focused and opinionated articles on important and timely issues, no original data

Abstract:
Headings: Introduction, Discussion, Conclusions
Word limit: 350 words

Main text:
Headings: Introduction, Discussion, Conclusions
Word limit: 2500 words
Numbers of figures and tables: 1
Additional files: No

Download the manuscript template

Letter to the Editor - comments on and responses to published articles

Abstract:
None

Main text:
Headings: None
Word limit: 500 words
Numbers of figures and tables: None
Additional files: No

Download the manuscript template

Viewpoint - constructive, stand-alone views on current topics

Abstract:
None

Main text:
Headings: None
Word limit: 1000 words
Numbers of figures and tables: 1
Additional files: No

Download the manuscript template

4. PREPARING THE SUBMISSION

Cover letter
In the cover letter, please explain why your manuscript should be published in the journal. If necessary, address any issues relating to our editorial policies .and declare any competing interests (see **Editorial Policies and Ethical Considerations**)

Parts of the Manuscript
The manuscript should be submitted as a main text file including figures and appendices and supporting information should be supplied as separate files.

Main Text File
The text file should be presented in the following order:

1. Title page;
2. Keywords;
3. Abstract;
4. Main text;
5. Conflict of Interest Statement;
6. Authorship;
7. Acknowledgments;
8. References;
9. Tables;
10. Figures;

Title page
The title should not contain abbreviations, except commonly used abbreviations such as HIV or AIDS (see **Wiley's best practice SEO tips**).

On the title page, you should mention the title of the manuscript, list all authors' names in full, and list any study groups if applicable. Each authors' affiliation should be numbered in superscript consecutively and listed underneath, including department, institution, city and country.

The corresponding author should be marked with the symbol § in superscript and full contact details should be provided, including a telephone number with country code. Authors who have contributed equally to the work should be marked with the symbol * in superscript. Deceased authors should be marked with the symbol ^ in superscript. The email addresses of all authors should be listed by their initials.

Keywords
Pease provide six keywords. Keywords should be taken from those recommended by the US National Library of Medicine's Medical Subject Headings (MeSH) browser list at **https://www.nlm.nih.gov/mesh/**. Preferably alternate words to those found in the abstract in order to improve search hits for the article in repositories.

Abstract
The Abstract should not exceed 350 words and should be structured according to the headings of the selected article category (see above), excluding the heading “Discussion” for Research articles. Avoid using abbreviations and do not cite references in the Abstract. If you are reporting results from a controlled health care intervention, please include your trial registry, together with your unique identifying number at the end of the Abstract. For randomized controlled trials, follow the **CONSORT extension for abstracts** .

Main Text

Article sections
Introduction
The Introduction section should introduce the topic to readers without specialist knowledge in that area and must clearly outline the current state of knowledge in this field, the motivation and the aim of the study or the article.

Methods
The Methods section should include all information necessary to repeat the study, in particular, the study design, how data was collected and analyzed, clarifying the choice of methods that were made. If applicable, you should describe the setting of the study, the dates the study were conducted, and the sample or participants, as well as necessary power calculations and materials, including statistical packages, used. Interventions and programmes should be described in detail. Generic names for drugs or any molecules should be used.

All studies involving humans or animals require a statement on ethical approval, and for the former, the consent procedure that was followed. Please include the names of the ethics review board(s) that approved the study. If the research study was specific to one sex/gender, the reasons for this should be clearly stated.

Results
This section should include only data and findings from the authors' study. Presentation of statistical results should mention confidence intervals and levels of significance where appropriate. Quotes from qualitative study participants of less than three lines should be quoted in the text using quotation marks. For quotes longer than three lines, place the quote in a separate, indented paragraph and introduce it with a colon. No quotation marks are needed in this case. Details of the participant can be added in round brackets following the quote, but should not contain identifiable information to ensure confidentiality. Clarifications within the quotation should be placed in square brackets.

Submitting authors are strongly encouraged to include data disaggregated by sex (and, whenever possible, by race) and provide a comprehensive analysis of gender and racial differences. The authors should include the number and percentage of men, women and, if appropriate, transgender persons who participated in the research study. Anatomical and physiological differences between men and women (height, weight, body fat-to-muscle ratios, cell counts, hormonal cycles, etc.), as well as social and cultural variables (socio-economic, education, access to care, etc.), should be taken into consideration in the presentation of data and/or analysis of the results.

Discussion
In the Discussion section, you should discuss your main findings and place these within the context of the current body of knowledge in the field. Limitations of the study, for example, selection bias, can also be discussed, and should address how these influence the results and conclusions. If statistically significant differences were found between men and women or between different racial or cultural groups in the effects of the studied intervention, the implications, if any, for clinical and/or public health should be adequately discussed.

Conclusions
In your Conclusions section, state your key messages from the study and explain their importance and relevance, as well as implications. Future studies and recommendations can be included in this section. The conclusions drawn must be strictly based on the data provided.

Conflict of Interest Statement
Authors will be asked to provide a conflict of interest statement during the submission process. For details on what to include in this section, see the 'Conflict of Interest' section in the **Editorial Policies and Ethical Considerations** section below. Submitting authors should ensure they liaise with all co-authors to confirm agreement with the final statement.

Authorship
Please refer to the journal's Authorship policy in the **Editorial Policies and Ethical Considerations** section for details on author listing eligibility. The individual contributions of each author must be specified in the Authors' Contributions section. Please use authors' initials and state

that all authors have read and approved the final manuscript. An example of a suitable statement is: “S.W., N.J., D.W. and S.S. performed the research. S.W., N.J., H.H. and T.L. designed the research study. H.H. and S.S. contributed essential reagents or tools. S.W., N.J. and D.W. analysed the data. S.W. and N.J. wrote the paper.” Please see the ‘Authorship’ section in the Editorial Policies and Ethical Considerations section below for what constitutes authorship.

Acknowledgments

Contributions from anyone who does not meet the criteria for authorship should be listed, with permission from the contributor, in an Acknowledgments section. Financial and material support should also be mentioned. Thanks to anonymous reviewers are not appropriate.

References

All external sources of information should be referenced within the text, the tables and figures, using consecutive numbering in square brackets, e.g. [1], [3-5], [3,4]. The references should be up to date and adequately reflect the current state of knowledge in the field. Citation bias, for example, by country or point of view must be avoided. Numbers of references are unlimited for all article categories and should be formatted in standard Vancouver style; **see Sample references from ICMJE** . Unpublished observations, personal communications and manuscripts currently under consideration should be cited in the text in round brackets and not in the reference list.

Tables

They should be supplied as editable files, not pasted as images. Tables should be inserted into the text. They should have the header: "Table 1. Title of table". All tables should be cited in the text in consecutive order. The tables should not contain colour or shading, and no vertical, visible lines. If tables are copied or adapted from another source, permission must be sought by the authors prior to publication and these should be clearly cited as such. If a table spans more than one page, authors may want to consider uploading the table as an additional file instead. Tables should be self-contained and complement, not duplicate, information contained in the text. A legend can be provided underneath the title, listing any abbreviations or meanings of symbols used. If several tables are included, please ensure that symbols are used consistently. Legends should be concise but comprehensive – the table, legend, and footnotes must be understandable without reference to the text. All abbreviations must be defined in footnotes. Footnote symbols: †, ‡, §, ¶, should be used (in that order) and *, **, *** should be reserved for P-values. Statistical measures such as SD or SEM should be identified in the headings.

Figures

Figures should be cropped as closely as possible and have the header: "Figure 1. Title of figure". All figures need to be cited in the text in consecutive order.

Although authors are encouraged to send the highest-quality figures possible, for peer-review purposes, a wide variety of formats, sizes, and resolutions are accepted. **Click here** for the basic figure requirements for figures submitted with manuscripts for initial peer review, as well as the more detailed post-acceptance figure requirements.

Figure legends should be concise but comprehensive – the figure and its legend must be understandable without reference to the text. Include definitions of any symbols used and define/explain all abbreviations and units of measurement. If several figures are included, please ensure that symbols are used consistently.

Additional Files

Appendices

Appendices will be published after the references. For submission, they should be supplied as separate files but referred to in the text.

Supporting Information

Supporting information is information that is not essential to the article, but provides greater depth and background. It is hosted online and appears without editing or typesetting. It may include tables, figures, videos, datasets, etc. **Click here** for Wiley’s FAQs on supporting information.

Note : if data, scripts, or other artefacts used to generate the analyses presented in the paper are available via a publicly available data repository, authors should include a reference to the location of the material within their paper.

General Style Points

The following points provide general advice on formatting and style:

• **Abbreviations:** In general, terms should not be abbreviated unless they are used repeatedly and the abbreviation is helpful to the reader. Initially, use the word in full, followed by the abbreviation in parentheses. Thereafter use the abbreviation only.
• **Acronyms:** Acronyms should be used sparingly, and not in headings or in the Abstract. Only commonly known acronyms may be used, and they should be spelt out at first use followed by the abbreviation in brackets. SI units should be used, with litre and molar being permitted.
• **Units of measurement:** Measurements should be given in SI or SI-derived units. Visit the Bureau International des Poids et Mesures (BIPM) website **here** for more information about SI units.
• **Numbers:** Numbers under 10 are spelt out, except for: measurements with a unit (8mmol/l); age (6 weeks old), or lists with other numbers (11 dogs, 9 cats, 4 gerbils).
• **Trade Names:** Chemical substances should be referred to by the generic name only. Trade names should not be used. Drugs should be referred to by their generic names. If proprietary drugs have been used in the study, refer to these by their generic name, mentioning the proprietary name and the name and location of the manufacturer in parentheses.
• **Footnotes:** Footnotes are not allowed in the text, the information shall be included directly into the text, where it fits best, and if these are references, to include in the reference section at the end.
• **Language**: All submissions must be in UK English (International) and UN-accepted terminology should be followed. No capitalization should be used except for grammatically correct use, official names and titles, and abbreviations.
• **General recommendation:** Use line spacing of 1.5 and an easily readable font, for example, Times New Roman, size 12. Your manuscript should contain line numbers to facilitate editors' and reviewers' comments

Wiley Author Resources

Manuscript Preparation Tips: Wiley has a range of resources for authors preparing manuscripts for submission available **here** . In particular, authors may benefit from referring to Wiley's best practice tips on **Writing for Search Engine Optimization** .

Editing, Translation, and Formatting Support: **Wiley Editing Services** can greatly improve the chances of a manuscript being accepted. Offering expert help in English language editing, translation, manuscript formatting, and figure preparation, Wiley Editing Services ensures that the manuscript is ready for submission.

5. EDITORIAL POLICIES AND ETHICAL CONSIDERATIONS

Editorial Review and Acceptance

The acceptance criteria for all papers are the quality and originality of the research and its significance to journal readership. Except where otherwise stated, manuscripts are single-blind peer reviewed, meaning that reviewers remain anonymous to the authors, although the authors' identity is known to the reviewers. Papers will only be sent to review if the Editors-in-Chief determine that the paper meets the appropriate quality and relevance requirements.

All manuscripts are reviewed by at least two independent experts with experience in the subject area and selected by the Editors. Dedicated statistical reviewers may be used if needed. Reviewers have to declare any competing interests to the Editors. Authors can suggest peer reviewers during the submission step. Suggested peer reviewers should not have co-authored publications with any of the authors during the past five years, should not be current collaborators, and should not be members of the same institution. Suggested reviewers will be considered alongside potential reviewers identified by the Editorial team. Authors may also request exclusion of individuals as potential reviewers: those who have clear competing interests, are close collaborators, or have given input into the manuscript previously.

The Editors assess revised manuscripts based on whether the authors have adequately addressed all comments. Re-reviews are only requested when revisions fall out of the technical expertise of the Editors. Further rounds of major revisions are usually not allowed, and manuscripts that have not been satisfactorily revised will be rejected. Minor revisions though may be requested as needed.

Wiley's policy on the confidentiality of the review process is available **here**.

Data Storage and Documentation
The *Journal of the International AIDS Society* expects that data supporting the results in the paper will be archived in an appropriate public repository. Whenever possible the scripts and other artefacts used to generate the analyses presented in the paper should also be publicly archived. Exceptions may be granted at the discretion of the editor for sensitive information such as human subject data or the location of endangered species. Authors are expected to provide a data accessibility statement, including a link to the repository they have used, to accompany their paper.

Protein and nucleotide sequences
For nucleic acid sequences, protein sequences or atomic coordinates, which are cited in the manuscript, and the accession number, together with the database where the information was deposited, should be cited in square brackets in the text, for example, [EMBL:AB026295, EMBL:AC137000, DDBJ:AE000812, GenBank:U49845, PDB:1BFM, Swiss-Prot:Q96KQ7, PIR:S66116]. Relevant databases are: EMBL Nucleotide Sequence Database (**EMBL**), DNA Data Bank of Japan (**DDBJ**), GenBank at the NCBI (GenBank), Protein Data Bank (**PDB**), Protein Information Resource (**PIR**) and the Swiss-Prot Protein Database (**Swiss-Prot**).

Mass spectrometry
Mass spectrometry data should be provided in the mzML format according to the **HUPO Protein Standards Initiative Mass Spectrometry Standards Working Group guidelines**. The data should also be deposited in the **ProteomeExchange** through the **PRIDE** website, and protein interaction data can be deposited through members of the IMEx consortium.

Structures
Protein structures can be submitted with one of the members of the **Worldwide Protein Data Bank**. Nucleic acid structures can be deposited with the **Nucleic Acid Database** at Rutgers. Crystal structures of organic compounds can be deposited with the **Cambridge Crystallographic Data Centre**.

Chemical structures and assays
Structures of chemical substances can be deposited with **PubChem Substance**. Bioactivity screens of chemical substances can be deposited with **PubChem BioAssay**.

Functional genomics data (such as microarray or CHIP-Seq data)
Please refer to standards proposed by the **Functional Genomics Data Society** and deposit your microarray data in MIAME-compliant format in one of the public repositories, for example, **ArrayExpress** or **Gene Expression Omnibus** (GEO). Deposition of high-throughput functional genomics sequencing data (such as RNA-Seq or ChIP-Seq data) with ArrayExpress or GEO in compliance with MINSEQE is also needed.

Computational modelling
Please prepare models of biochemical reaction networks using the **Systems Biology Markup Language** and submit your model to the **BioModels database**, as well as providing it as an additional file with your submission.

Plasmids
Please submit copies of your plasmids as DNA or bacterial stocks with **Addgene**, a non-profit repository, or **PlasmID**, the Plasmid Information Database at Harvard.

Ethical approval – Human and animal studies

Human Studies and Subjects
For manuscripts reporting medical studies that involve human participants, a statement identifying the ethics committee that approved the study and confirmation that the study conforms to recognized standards is required, for example: **Declaration of Helsinki**; **US Federal Policy for the Protection of Human Subjects**; or **European Medicines Agency Guidelines for Good Clinical Practice**.

A statement on the ethical aspects, including the consent procedure followed, must be included in the Methods section of the manuscript. The Editors may reject manuscripts where the research has not been carried out within an ethical framework. Images and information from individual participants will only be published where the authors have obtained the individual's free prior informed consent. Confidentiality of study participants must be ensured at all stages of research and reporting. Authors do not need to provide a copy of the consent form to the publisher; however, in signing the author license to publish, authors are required to confirm that consent has been obtained. Wiley has a **standard patient consent form available for use**.

Animal Studies
A statement indicating that the protocol and procedures employed were ethically reviewed and approved, as well as the name of the body giving approval, must be included in the Methods section of the manuscript. Authors are encouraged to adhere to animal research reporting standards, for example the **The Gold Standard Publication Checklist from Hooijmans and colleagues** or the **ARRIVE reporting guidelines** for reporting study design and statistical analysis; experimental procedures; experimental animals and housing and husbandry. Authors should also state whether experiments were performed in accordance with relevant institutional and national guidelines for the care and use of laboratory animals:

• US authors should cite compliance with the US National Research Council's **Guide for the Care and Use of Laboratory Animals**, the US Public Health Service's **Policy on Humane Care and Use of Laboratory Animals**, and **Guide for the Care and Use of Laboratory Animals**.
• UK authors should conform to UK legislation under the **Animals (Scientific Procedures) Act 1986 Amendment Regulations (SI 2012/3039)**.
• European authors outside the UK should conform to **Directive 2010/63/EU**.

Clinical Trial Registration
The journal requires that clinical trials are prospectively registered in a publicly accessible database and clinical trial registration numbers should be included in all papers that report their results. Authors are asked to include the name of the trial register and the clinical trial registration number at the end of the abstract. If the trial is not registered, or was registered retrospectively, the reasons for this should be explained.

Research Reporting Guidelines

Standard of reporting
The *JIAS* endorses international standards of reporting. Please see the **Uniform Requirements for Manuscripts Submitted to Biomedical Journals** guidelines produced by ICMJE as a reference standard of reporting. Authors are also referred to the **EQUATOR** network website for further information on the available reporting guidelines for health research, and the **MIBBI** Portal for prescriptive checklists for reporting biological and biomedical research where applicable. A number of checklists are available for various study designs, including randomized controlled trials (**CONSORT**), interventional trials (**SPIRIT**), qualitative research (**COREQ**), systematic reviews (**PRISMA**), observational studies (**STROBE**), economic evaluations of health interventions (**CHEERS**), meta-analyses of observational studies (**MOOSE**) and diagnostic / prognostic studies (**STARD** and **TRIPOD**). For systematic reviews, an additional file should be provided by the authors listing all details concerning the search strategy. Please refer to the **Cochrane Reviewers' Handbook** for an example of how a search strategy should be presented.

Guidelines on mutation nomenclature are provided by the **Human Genome Variation Society** , and authors should use the recommended gene name by referring to the appropriate genetic nomenclature database, for example, HUGO for human genes, and the International Committee on Standardized Genetic Nomenclature for Mice. When describing human phenotypes, please use standardized terms, such as those proposed by the Elements of Morphology working group (see **http://research.nhgri.nih.gov/morphology/index.cgi**).

Contributions from pharmaceutical companies or other commercial organizations should follow the **Good Publication Practice guidelinEs for pharmaceutical companies**, which also apply to any companies or individuals that work on industry-sponsored publications, such as freelance writers, contract research organizations and communications companies.

The *JIAS* supports international standards of reporting of trials, in particular, prospective registering and numbering of clinical trials. Clinical trials are defined by the World Health Organization as all phase I to IV trials, which are research studies that prospectively assign human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes. Trials need to be registered prior to submission in a suitable, publicly available registry. Links to existing registries can be found through ICMJE **here** or through the primary registers that participate in the **WHO International Clinical Trials Registry Platform.** The trial registration number should be included as the last line of the manuscript Abstract.

Reporting by gender and race
Submitting authors shall include data disaggregated by sex (and, whenever possible, by race) and provide an analysis of gender and racial differences. The authors should include the number and percentage of men, women and, if appropriate, transgender persons, who participated in the research study. Anatomical and physiological differences between men and women (height, weight, body fat-to-muscle ratios, cell counts, hormonal cycles, etc.), as well as social and cultural variables (socio-economic, education, access to care, etc.), should be taken into consideration in the presentation of data and/or analysis of the results. If statistically significant differences were found between men and women or between different racial or cultural groups in the effects of the studied intervention, the implications, if any, for clinical and/or public health should be adequately discussed. If the research study was specific to one sex/gender, the reasons for this should be clearly stated. Please refer to the **SAGER guidelines** for more information

Species Names
Upon its first use in the title, abstract, and text, the common name of a species should be followed by the scientific name (genus, species, and authority) in parentheses. For well-known species, however, scientific names may be omitted from article titles. If no common name exists in English, only the scientific name should be used.

Genetic Nomenclature
Sequence variants should be described in the text and tables using both DNA and protein designations whenever appropriate. Sequence variant nomenclature must follow the current HGVS guidelines; see **varnomen.hgvs.org**, where examples of acceptable nomenclature are provided.

Conflict of Interest
Authors are required to submit a statement on competing interests, which exist when personal or financial relationships with persons or organizations may influence the interpretation of the data or how the author's work is presented. For details, see ICMJE's policy on competing interests **here**. In brief, all financial competing interests must be disclosed in this statement (reimbursements, fees, funding, salary payments from or ownership of any stocks or shares in an organization that may in any way gain or lose financially from the publication of the manuscript, either now or in the future, or applications for patents relating to the content of the manuscript), as well as non-financial competing interests (such as political, personal, religious, ideological, academic and/or intellectual interests) that are related to the work submitted. The competing interest statement should be included in the manuscript and will be published in the final article. If no competing interests exist, please state in

this section, "The authors declare that they have (or The author declares that he/she has) no competing interests."

Copyright and libel
Legal responsibility to ensure that no material is published that infringes copyright or that includes libellous or defamatory content lies with the Journal of the International AIDS Society's publisher, the International AIDS Society. If a manuscript is judged by the journal Editors to include potentially libellous content, authors will be requested to adjust wording as necessary.

Commercial writers and editors
The involvement of scientific (medical) writers or anyone else who assisted with the preparation of the manuscript content should be acknowledged, along with their source of funding, as described in the European Medical Writers Association (EMWA) **guidelines** on the role of medical writers in developing peer-reviewed publications.

Funding
Authors should list all funding sources in the Acknowledgments section. Authors are responsible for the accuracy of their funder designation. If in doubt, please check the Open Funder Registry for the correct nomenclature: **https://www.crossref.org/services/funder-registry/**

Authorship
It is understood that all authors listed on submitted manuscripts have read and agreed to its content, and meet the authorship requirements as detailed by ICMJE **here**. The list of authors should accurately illustrate who contributed to the work and how. All those listed as authors should qualify for authorship according to the following criteria:

1. Have made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data;
2. Have been involved in drafting the manuscript or revising it critically for important intellectual content;
3.Have given final approval of the version to be published. Each author should have participated sufficiently in the work to take public responsibility for appropriate portions of the content; and
4. Have agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Each author should have participated sufficiently in the work to take public responsibility for appropriate portions of the content. Acquisition of funding, collection of data, or general supervision of the research group, alone, does not justify authorship.

All contributors who do not meet the criteria for authorship should be listed in the Acknowledgements section. Examples of those who might be acknowledged include a person who provided purely technical help or writing assistance, or a head of department, who provided only general support. Prior to submitting the article all authors should agree on the order in which their names will be listed in the manuscript.

Additional Authorship Options: Joint first or senior authorship: In the case of joint first authorship, a footnote should be added to the author listing, e.g. ‘X and Y should be considered joint first author’ or ‘X and Y should be considered joint senior author.’

ORCID
As part of the journal’s commitment to supporting authors at every step of the publishing process, the journal requires the submitting author (only) to provide an ORCID iD when submitting a manuscript. This takes around 2 minutes to complete. **Find more information here**.

Publication Ethics
This journal is a member of the Committee on Publication Ethics (**COPE**) and endorses the World Association of Medical Editors' (WAME's) Policy Statement on Geopolitical Intrusion on Editorial Decisions. Note this journal uses iThenticate’s CrossCheck software to detect instances of overlapping and similar text in submitted manuscripts. Any misconduct by authors in reporting their data, for example, falsification, will lead to rejection of their manuscript and other consequences

decided on by the Editors. Please see COPE and International Committee of Medical Journal Editors (ICMJE) for further information on ethical issues in publishing

Read Wiley's Top 10 Publishing Ethics Tips for Authors **here**. Wiley's Publication Ethics Guidelines can be found **here** .

6. AUTHOR LICENSING

Journal of the International AIDS Society is an Open Access journal: authors of accepted papers pay an Article Publication Charge and their papers are published under a Creative Commons license. With Creative Commons licenses, the author retains copyright and the public is allowed to reuse the content. The author grants Wiley a license to publish the article and identify as the original publisher.

Open Access Fees: Information on the Article Publication Charge for publishing in the journal is available **here** .

If a paper is accepted for publication, the author identified as the formal corresponding author will receive an email prompting them to login to Author Services, where via the Wiley Author Licensing Service (WALS), they will be able to complete the license agreement on behalf of all authors on the paper.

To find out which Created Commons Licenses are available for the journal, click **here** . To learn more about Creative Commons Licenses and to preview terms and conditions of the agreements, please **click here**. Note that certain funders mandate a particular type of CC license be used; to check this, please click **here**.

7. PUBLICATION PROCESS AFTER ACCEPTANCE

Accepted Article Received in Production
When an accepted article is received by Wiley's production team, the corresponding author will receive an email asking them to login or register with **Wiley Author Services**. The author will be asked to sign a publication license at this point.

Proofs
Once the paper is typeset, the author will receive an email notification with the URL to download a PDF typeset page proof, as well as associated forms and full instructions on how to correct and return the file.

Please note that the author is responsible for all statements made in their work, including changes made during the editorial process – authors should check proofs carefully. Note that proofs should be returned within 48 hours from receipt of first proof.

8. POST PUBLICATION

Access and Sharing
When the article is published online:

• The author receives an email alert (if requested).
• The link to the published article can be shared through social media.

Promoting the Article
To find out how to best promote an article, click **here**.

Measuring the Impact of an Article
Wiley also helps authors measure the impact of their research through specialist partnerships with **Kudos** and **Altmetric**.

9. EDITORIAL OFFICE CONTACT DETAILS

Journal of the International AIDS Society
Avenue de France 23
CH - 1202 Geneva
Switzerland
Tel: 41 (0)22 7 100 800

Principal Contact

Editorial Team
Journal of the International AIDS Society
Avenue de France 23
CH - 1202 Geneva
Switzerland
Phone: 41 (0)22 7 100 800
Fax: 41 (0)22 7 100 899
Email: **editorial@jiasociety.org**

Published on behalf of the International AIDS Society