METHOD ARTICLE

Revealing HIV viral load patterns using unsupervised machine learning and cluster summarization [version 1; peer review: 1 approved, 1 approved with reservations]

Samir A. Farooq¹, Samuel J. Weisenthal^{1,2}, Melissa Trayhan¹⁻³, Robert J. White^{1,2}, Kristen Bush^{1,2}, Peter R. Mariuz⁴, Martin S. Zand ¹⁻³

¹Rochester Center for Health Informatics, University of Rochester Medical Center, Rochester, NY, 14642, USA

²Clinical and Translational Science Institute, University of Rochester Medical Center, Rochester, NY, 14642, USA

³Department of Medicine - Division of Nephrology, University of Rochester Medical Center, Rochester, NY, 14642, USA

⁴Department of Medicine, Division of Infectious Diseases, Strong Memorial Hospital AIDS Center,, University of Rochester Medical Center, Rochester, NY, 14642, USA

V1 First published: 27 Jul 2018, 7:1144
<https://doi.org/10.12688/f1000research.15591.1>
 Latest published: 27 Jul 2018, 7:1144
<https://doi.org/10.12688/f1000research.15591.1>

Abstract

HIV RNA viral load (VL) is an important outcome variable in studies of HIV infected persons. There exists only a handful of methods which classify patients by VL patterns. Most methods place limits on the use of viral load measurements, are often specific to a particular study design, and do not account for complex, temporal variation. To address this issue, we propose a set of four unambiguous computable characteristics (features) of time-varying HIV viral load patterns, along with a novel centroid-based classification algorithm, which we use to classify a population of 1,576 HIV positive clinic patients into one of five different viral load patterns (clusters) often found in the literature: durably suppressed viral load (DSVL), sustained low viral load (SLVL), sustained high viral load (SHVL), high viral load suppression (HVLS), and rebounding viral load (RVL). The centroid algorithm summarizes these clusters in terms of their centroids and radii. We show that this allows new VL patterns to be assigned pattern membership based on the distance from the centroid relative to its radius, which we term radial normalization classification. This method has the benefit of providing an objective and quantitative method to assign VL pattern membership with a concise and interpretable model that aids clinical decision making. This method also facilitates meta-analyses by providing computably distinct HIV categories. Finally we propose that this novel centroid algorithm could also be useful in the areas of cluster comparison for outcomes research and data reduction in machine learning.

Open Peer Review

Approval Status

	1	2
version 1
27 Jul 2018	view	view

1. **Sally Blower**, University of California, Los Angeles (UCLA), Los Angeles, USA
2. **Amalio Telenti** , The Scripps Research Institute, La Jolla, USA

Any reports and responses or comments on the article can be found at the end of the article.

Keywords

Machine learning, HIV, viral load, feature extraction, HIV categories, centroid, cluster summarization, clinical interpretability

This article is included in the **Artificial Intelligence and Machine Learning** gateway.

Corresponding author: Martin S. Zand (martin_zand@urmc.rochester.edu)

Author roles: **Farooq SA:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Weisenthal SJ:** Investigation, Methodology, Software, Visualization, Writing – Review & Editing; **Trayhan M:** Conceptualization, Data Curation, Methodology, Resources, Software; **White RJ:** Conceptualization, Data Curation, Funding Acquisition, Project Administration, Resources, Supervision, Visualization, Writing – Review & Editing; **Bush K:** Investigation, Visualization, Writing – Review & Editing; **Mariuz PR:** Conceptualization, Supervision, Writing – Review & Editing; **Zand MS:** Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was partially funded by the University of Rochester Clinical and Translational Science Institute grants UL1 TR002001, and TL1 TR002000 from the National Center for Advancing Translational Sciences (NCATS), a component of the National Institutes of Health (NIH). This publication was also made possible through core services and support from the University of Rochester Center for AIDS Research (CFAR), an NIH-funded program (P30 AI078498).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Farooq SA *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Farooq SA, Weisenthal SJ, Trayhan M *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 <https://doi.org/10.12688/f1000research.15591.1>

First published: 27 Jul 2018, 7:1144 <https://doi.org/10.12688/f1000research.15591.1>

Introduction

The primary clinical goal of HIV treatment and patient engagement is suppression of the HIV viral load (VL), as measured by low or undetectable circulating HIV RNA levels. However, VL most often fluctuates over repeated measurements, with a range that spans 8 orders of magnitude from 0 (undetectable) - 10^7 copies/mL. VL is regularly monitored for signs of progression of HIV infection. Standard HIV treatment protocols are based on VL measurements¹, especially when monitoring responses to antiretroviral therapy (ART). Monitoring of VL helps to determine whether ART therapy was able to successfully suppress patient VL². Individuals with sustained high viral loads (SHVL) are at greater risk of secondary transmission, clinical progression to AIDS, and death³⁻⁶. In contrast, significant reduction in VL or high viral load suppression (HVLs) both lead to immune recovery, as measured by CD4 T cell levels⁷, and can reduce or eliminate the risks of SHVL. Furthermore, patients sustaining low-level viral load (SLVL), or with a rising VL after previous suppression, have a high incidence of treatment failure⁸. Thus, developing an objective measure of VL status, and categorization of patients by time varying patterns of VL, is critical for standardizing both therapy and comparing research protocol efficacy.

Reports in the current literature differ in the definition “high viral load”⁹⁻¹², and their findings of how long it takes a patient on highly active anti-retroviral therapy (HAART) to suppress their VL^{2,9,10,13}. We summarize some of the published approaches here (for greater detail see [Supplementary File 1](#)). With respect to VL levels, Terzian *et al.* defined SHVL as two consecutive viral load measurements (VLM) $\geq 100,000$ copies/mL⁹. Durable suppressed viral load (DSVL) was defined as all VLM < 400 copies/mL. In contrast, Greub *et al.* focused on detecting low level viral rebound (LLVR) by first considering patients with an initial consecutive VLM pair < 50 copies/mL, and classified LLVR as having subsequent maximum VLM between 51–500⁸. Alternatively, Rose *et al.* investigated the use of five different frameworks to categorize suppressed versus not-suppressed VL¹⁰. Their approach excluded patients with VLM < 200 at baseline, and stratified the remainder with regard to VL suppression using an 8 month window centered around 24 months after the start of VLM (18–30 months). Another approach was used by Phillips *et al.*, and characterized VLM responses to ART¹³, utilizing a 24–40 week window and a rule-based method to identify two populations of HIV patients (Viral Failure and Viral Rebound). Despite these studies, no formal standard has been adopted by the field to classify a patient as having DSVL, SHVL, HVLs, SLVL, or rebounding viral load (RVL) patterns.

Classifying patient VL states outside of research studies is further complicated in that real-world VL measurements are taken intermittently over time, and missing data is common due to a variety of factors (e.g. travel, social circumstances, non-adherence). This leads to incomplete and irregularly spaced data points. In addition, differences in the sensitivity of the multiple VL clinical assays available results in multiple cut off points

for “undetectable” viral loads analyzed at different facilities, further complicating analyses. Thus, there is a need for analytic techniques that can adjust for these details and classify VL states, both across research studies using different methodologies and to consistently classify patients in clinical practice.

Machine learning methods can provide objective, unsupervised classification of patient clinical status¹⁴. These methods begin by collecting a set of features from patient data (e.g. demographics, laboratory measurements, therapies) and then performing computational clustering to identify similar patient classes. Some groups have applied machine learning methods to HIV research studies¹⁵ to predict HIV VL responses¹⁶ or CD4 T cell counts¹⁷, to distinguish between suppressed and viremic patients¹⁸, and to select therapeutic regimens¹⁹. None, however, have used machine learning to create a standard classification for VL status with irregularly sampled VL measurements across a cohort of patients.

To address these issues, we propose a set of unambiguous features which, when combined as a feature vector, capture the distinct dynamic patterns present in VL measurements over time. In addition, we have developed a novel centroid algorithm to cluster HIV positive subjects based on these patterns. Here we present the derivation of this method, and demonstrate its application to clustering 1,576 HIV patients with repeated VL measurements over a 5 year period. We found that patient VL measurements can be clustered into five time-varying patterns that correspond well to clinically relevant states. We note that the method and resulting categories can be used to standardize definitions of VL patterns across research studies, and potentially for clinical classification.

Methods

Human subjects protection

This proposal was reviewed and approved by the University of Rochester Human Subjects Review Board (protocol number RSRB00068884). Consent was waived by the review board due to de-identification of the data set. The analysis in this paper is presented in compliance with Center for Medicare Services (CMS) current cell size suppression policy²⁰. Data were coded such that patients could not be identified directly in compliance with the Department of Health and Human Services Regulations for the Protection of Human Subjects (45 CFR 46.101(b)(4)).

Study data

We obtained medical encounter data from all patients with an HIV diagnosis in the University of Rochester Medical Center’s electronic medical record system (EMR) between 2011–2016, including age, gender, race, ethnicity, and VL measurements. There were a total of 1,892 patients with at least one VL measured, with 1,576 of these patients having at least three VL measurements.

Measurements ≤ 48 copies/mL, present as categorical values “NEG”, “POS < 20 ”, or “POS < 48 ” were transformed into numerical values of 0, 20, and 48 respectively. The deidentified study

data containing only viral load measurements and relative time are available at <https://doi.org/10.5281/zenodo.1313245>²¹.

Hardware and software specifications

Analyses were performed on a Windows 8 server with Intel(R) Xeon(R) CPUs E5-2620 v2 @ 2.10GHz and 256GB of RAM. Python 2.7 was used for most data mining and machine learning under Spyder v.3 installed from Anaconda2 (64-bit). The default packages available in Anaconda were used for analysis, including, but not limited to: NumPy, scikit-learn, SciPy, datetime, csv, math, Matplotlib, pip, operator, copy, random, and time. Using pip we installed the webcolors and pydotplus packages for rendering a decision tree. SQLite was used to store, query, and clean ~the data. Analytic code is available for download at https://github.com/SamirRCHI/Viral_Load_Data_Categorization.

Viral load analysis methods

Since VL data is asynchronous and noisy, with variable numbers of data points for each subject, we excluded patients with ≤ 2 VL measurements as too few to accurately assess VL patterns. Based on temporal patterns of VL described in the literature, the VL pair distribution of our data (Figure S1), and a further extensive investigation into the data, we hypothesized six potential temporal VL patterns, defined in Table 1 and illustrated in Figure 1.

It is important to note that these definitions are pattern based, and do not explicitly select absolute VL cutoff levels or a specific temporal window, as other reports have done^{8–10,13}. This has the advantage of allowing the absolute VL levels and critical time windows to emerge from the analysis. It also does not preclude incorporation of absolute levels (e.g. VLM>400) at a later stage into the pattern specification.

Feature vector definition

Mathematical notations for this work are described in Table 2. We next designed a feature vector to capture characteristics that would allow us to distinguish between VL patterns. VL values at the lower limit of detection are a function of the specific assay used, and appear in our data set as 0, 20, and 48 copies/mL (Figure S1). Thus, plots of the $\log_{10}$ transformed data have discretely spaced values at the lower level of detection, capturing the undetectable range of viral load. Additionally, we adjusted the data by $\log_{10}[VL + 10]$ to avoid $\log_{10}[0]$. The addition of 10 to VL (instead of 1) is used to minimize the distance between the undetectable values: 0, 20, and 48 (copies/mL). Thus, in our notation, all the values related to viral load are assumed to have been adjusted to this measure. For example, $\min_{VL} = \log_{10}[0 + 10] = 1$ and $\max_{VL} = \log_{10}[10^7 + 10] \approx 7$.

Using the transformed VL data, we next extract several relevant features of the VL measurements over time. These features are used for machine learning classification of individual patient VL time series, and designed to distinguish patterns in VL change while minimizing the effects of noise. We do not limit feature extraction based on the total elapsed time of viral load measurements because the optimal time-point for determining viral load class is not well established. The attributes for feature extraction are: relative area of viral exposure, weighted recency reliability, adjusted maximal difference, and interquartile range. The definitions include:

1. *Relative area of viral exposure (Area)* - the area under the viral load curve relative to the total viral load area possible, which has a range between [0,1]. We choose a normalized, relative score, as the total time span between the first and last viral load measurement, which differs between patients. This feature is similar to finding the

Table 1. Viral load state definitions.

Abbrv.	Name	Definition
<i>DSVL</i>	Durably Suppressed Viral Load	Having consistently suppressed their viral loads at or near the undetectable range
<i>SLVL</i>	Sustained Low Viral Load	Viral load counts which are constantly slightly higher than the undetectable range.
<i>SHVL</i>	Sustained High Viral Load	Viral load counts which are constantly in a range considered high risk for HIV complications (e.g. opportunistic infections, malignancy).
<i>HVLS</i>	High Viral Load Suppression	A viral load pattern in which the terminal portion of the curve has a negative slope and the terminal data point is in the low or suppressed range. This could have a few different speeds or styles of suppression - rapid, gradual, or slow.
<i>RVL</i>	Rebounding Viral Load	A viral load pattern in which viral loads are unstable, with the measurement at one time step seemingly being independent of the next measurement.
<i>EVL</i>	Emerging Viral Load	Having a steady, or rapid, emergence of high viral load while the first few measurements of viral load were suppressed. While we have found no mention of this type of pattern in the literature, and found that this pattern did not occur in our data set, VL data sets could contain this pattern.

*Colors are used throughout the manuscript to identify clusters

Figure 1. Possible HIV viral load patterns. Examples of each type of viral load pattern. Note that actual viral load patterns are noisier and may often be more difficult to distinguish. The magnitudes of viral load values reflect those found in the dataset.

Table 2. Mathematic notation.

Symbol	Description
N	The number of usable patients in the data. In our case 1576 patients.*
p	Refers to a single patient.
VL_{M_p}	The total number of viral load measurements patient p has taken.
$\overrightarrow{VL}_p$	All viral load counts of patient p in order of time.
$\overrightarrow{VL}_{p,i}$	Refers to the i^{th} viral load count of patient p in $\overrightarrow{VL}_p$, where $1 \leq i \leq VL_{M_p}$.
$\overrightarrow{t}_p$	All temporal instances corresponding to $\overrightarrow{VL}_p$.
$\overrightarrow{t}_{p,i}$	Temporal instance of viral load $\overrightarrow{VL}_{p,i}$, where $1 \leq i \leq VL_{M_p}$.
$\max_{VL}$	The maximum viral load for <i>all</i> patients, (10^7) .**
$\min_{VL}$	The minimum viral load for <i>all</i> patients, (0) .**
$\circ$	Hadamard Product - elemental-wise multiplication of arrays.

*This is after selecting for patients with ≥ 3 measurements.

**This value changes after transformation of the data.

mean and median, except it is sensitive to the dimension of time, hence yielding more information. The feature is calculated by summing the area of each trapezoid created by each pair of viral load values, followed by dividing by the total possible area (Equation 1).

$$\dot{A}_p = \frac{\sum_{i=2}^{VL_{M_p}} \frac{\overrightarrow{VL}_{p,i} + \overrightarrow{VL}_{p,i-1}}{2} (\overrightarrow{t}_{p,i} - \overrightarrow{t}_{p,i-1})}{(\max_{VL} - \min_{VL})(\overrightarrow{t}_{p,VL_{M_p}} - \overrightarrow{t}_{p,1})} \quad (1)$$

2. *Weighted recency reliability (wRR)* - Due to viral load noise, the last measurement may not be an accurate reflection of a patient's viral load trend. For example, a patient may have a VL whose average slope is negative, indicating high viral load suppression over time (HVLS). If, however, the last measurement is slightly higher than the trend, heavily weighting this last measurement could lead to misclassifying the patient as rebounding viral load (RVL). To account for this, we calculate a weighted mean where

the weight of the VL measurement increases with time. More specifically, the weight function follows an inverse square root function ($f(x) = \frac{1}{\sqrt{x}}$) rather than an inverse function ($g(x) = \frac{1}{x}$). This has the advantage of avoiding rapid convergence of $g(x)$ to zero when time is measured in units of days (Equation 2). Weighted recency is then calculated as the dot product of the viral loads and weights divided by the sum of the weights (Equation 3).

$$\overline{weight}_{p,i} = \frac{1}{\sqrt{\overline{t}_{p,VLMP} - \overline{t}_{p,i}} + 1} \quad (2)$$

$$wR_p = \frac{\overline{weight}_p \cdot \overline{VL}_p}{\sum_{i=1}^{VLMP} \overline{weight}_{p,i}} \quad (3)$$

We were also interested in how reliable wR is as a representation of the patient's viral load trend. To this end, we calculated the absolute deviations from the viral load measurements to wR (Equation 4). Rather than averaging the deviations, we take the median to reduce the effects of outliers and call this our weighted recency reliability measure (Equation 5). We take the inverse to force the range of the result to be between [0,1]; a property made to use in our next proposed feature, adjusted maximal difference.

$$\overline{dev}_{p,i} = |wR_p - \overline{VL}_{p,i}| \quad (4)$$

$$wRR_p = \frac{1}{\text{median}(\overline{dev}_p) + 1} \quad (5)$$

3. *Adjusted maximal difference (Adj MD)* - this is time-independent the difference between the “peak” and last VL measurements. To distinguish between viral load suppression or emergence, we calculate the “peak” as the maximum of the absolute deviations (Equation 4) and retain the sign of the result. We expected the positive scores to effectively isolate the EVL group, however, we instead found that retaining the positive (emergent) scores lead to mis-categorization of SHVL and RVL groups without clearly identifying EVL patterns. This, along with other investigation into the data, led us to conclude that the EVL pattern may not exist in our data, but we refrain to make generalizations to all healthcare facilities. With this consideration, we force (ground) the positive scores down to zero for proper labeling of SHVL and RVL (Equation 6).

Due to the varying nature in viral load measurements, we are hesitant to use the final viral load measurement as a means of judging suppression. Thus we propose to use wR instead. To reduce the effects of rebounding patients being falsely labeled as suppressed patients, we multiply our result by wRR - as rebounding patients are expected to have a low score in the range [0,1]. The maximal difference is necessary in order to ensure that the suppression type of viral load patterns are classified appropriately (Equation 7).

$$\text{grnd}(x) = \begin{cases} -1 & x < 0 \\ 0 & x \geq 0 \end{cases} \quad (6)$$

$$\check{D}_p = \text{grnd}(wR_p - \overline{VL}_{p, \arg\max(\overline{dev}_p)}) \cdot \max(\overline{dev}_p) \cdot wRR_p \quad (7)$$

4. *Interquartile range (IQR)* - This feature is added to further segregate the rebounding patients and follows the standard interquartile range calculation (Equation 8).

$$IQR_p = Q_3(\overline{VL}_p) - Q_1(\overline{VL}_p) \quad (8)$$

Statistical analysis

Machine learning methods for cluster classification were compared by calculating F_1 scores, the harmonic mean of precision and recall²², defined by Equation 9–Equation 11.

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (9)$$

$$\text{recall} = \frac{\text{True Positive}}{\text{Positive}} \quad (10)$$

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

Analytic terminology

Here we formally define keywords appearing in the analysis: Let *Feature extraction* be the process of determining the values $\check{A}$, wRR , $\check{D}$, and IQR from a set of patients (using their viral load patterns) with the formulations given above. Then a *feature vector* ($\check{F}_p$) contains the values $\check{A}_p$, wRR_p , $\check{D}_p$, and IQR_p extracted from patient p 's viral load pattern. The words *sample* or *point* are also used here RVL (black; n= 237) and HVLS (purple; n=316) clusters. interchangeably. The term *feature* (F) can be thought of as a column vector for all patients in the dataset consisting of the four attributes: F_A , F_{wRR} , F_D , and F_{IQR} . Finally, the terms *label assignment*, *VL pattern membership assignment*, *patient categorization*, and *prediction*, all refer to the same principle: To assign the most appropriate label which characterizes the viral load pattern of a patient. However, while the principle is the same, the method of assigning such an appropriate label differs depending on the categorization or the learning method used.

Results

Feature extraction and normalization

We began by transforming viral load data by min-max normalization²² to equally weight the temporal features of the VL series (Equation 12). That is, we normalize the features, F , to a range between [0, 1] using Equation 12 where $F^* = f(F)$.

$$F^* = f(F) = \frac{F - \min F}{\max F - \min F} \quad (12)$$

Next, we examined each of the four features for all patients with ≥ 3 viral load measurements ($N = 1,576$ patients), and did not find distinct bi-variate clustering (Figure S2). A feature correlation coefficient analysis (Supplementary Table S1) revealed that the *Adj MD* feature is linearly independent of *Area* and *wRR*. In contrast, there is modest linear dependence between *IQR* and *Adj MD*, and between *Area* and both *wRR* and *IQR*. As expected, the largest linear dependency is between *wRR* and *IQR*. These results suggest the separation between viral load patterns will be most noticeable between the *Area* and the *Adj MD* features - as we designed them to be. Also, although *Adj MD* is dependent upon *wRR*, we find that their correlation coefficient is very low (0.033).

Hierarchical clustering

We then performed hierarchical clustering of the individual subject VL patterns using a Euclidean distance metric and Ward's linkage criterion²³ to minimize the total within-cluster variance. Patients showed a clear separation into 5 distinct groups, which had clinical significance (Figure 2 and Figure 3). The cluster with the lowest viral loads and the highest weighted recency reliability ($n=442$) corresponds to the DSVL patient group. The patients corresponding to the SHVL group (orange; $n=46$) exhibited the highest relative *Area* and very low *IQR*. Compared to the DSVL cluster, the blue cluster ($n=535$) has slightly greater area and *IQR* with a significant difference in the weighted recency reliability. Using this information, along with the general patterns shown by Figure 3, we identify this as the SLVL group. The algorithm also identifies the RVL (black; $n=237$) and HVLS (purple; $n=316$) clusters. The RVL cluster has a low weighted recency reliability and high *IQR*. In contrast, the HVLS cluster has a lower area, higher weighted recency reliability, indicating little variation in the terminal portion of

the VL time series, and most importantly very low adjusted maximal differences (Figure 4).

VL patterns are similar within clusters, and dissimilar between clusters Figure 3. Interestingly, there are patients within each cluster whose last VL measurement occurs near 1,827 days. This is equivalent to the full span of five years of VL monitoring data set. This suggests that these clusters don't disappear after some elapsed time, but rather each type of pattern can be found at virtually any time point.

We found large VL spikes within the time series of the HVLS group. We hypothesize that this may be due to the asynchronous timing of measurements between subjects, the natural variation in biological responses, or patient variability in adherence to therapy. This observation also reflects one limitation of asynchronous outcomes data sampling, which lacks a "completion" endpoint characteristic of most prospective, randomized clinical trials. If measurements ended at a spike, the adjusted maximal difference feature may be weighted in the favor of the patient being classified as RVL. This may indicate that some patients classified as suppressing their viral loads should have been classified as having rebounding viral loads. Alternatively, may indicate that these features do not restrict a patient to forever to one category, but allow for dynamic classification as a function of biological or therapeutic responses.

Comparison of categorization methods

Using the same data set, we next compared our VL pattern categorization method to those previously published in the literature. Visually, we find that the SLVL group detected by our method is very similar to the LLVR group defined by Greub *et al.* (Figure 5). Furthermore, it appears that the methods

Figure 2. Dendrogram of hierarchically clustered patients. Clustered using the Euclidean distance along with Ward's method. Numbers on the bottom axis show number of patients in each cluster. The corresponding viral load pattern plots can be found in Figure 3. DSVL = Durably Suppressed Viral Load, SHVL = Sustained High Viral Load, SLVL = Sustained Low Viral Load, RVL = Rebounding Viral Load, HVLS = High Viral Load Suppression

Figure 3. Extracted patient viral load patterns. For each cluster categorization of the patient from Figure 2, the days since first viral load measurement are plotted against the viral load counts. The points on the plots indicate the last viral load measurement.

trying to capture SHVL, viral rebound, and viral failure patients did not succeed as well as the identification of SHVL and RVL patients in our method. RMVL repeat continuous visually appears to have performed very well in identifying patients who have suppressed their viral loads. However, the results suggest that our analysis performs slightly better in identifying the suppression group (HVLS), as we find that the last VL measurements (black dots in Figure 5) are consistently low using our method.

The other methods may not have performed as well as they rely on a window or a consecutive pair measure, which may be too subjective for assigning VL pattern membership. Furthermore, notice that patients with baseline VL < 200 (Figure 5) contain VL patterns which can reach as high as 10^6 copies/ml, which is in contrast to Rose *et al.*'s assumption that

these patients have consistently low viral variation. Lastly we wish to emphasize that while some of these categorization methods are successful in identifying a specific group of patients, our method is unique as it attempts to associate each VL pattern to a specific category, without using categories such as "Not Suppressed", "Unspecified", or "Omitted".

Supervised learning of VL patterns

We next used the classes identified by hierarchical clustering to compare several machine learning models, with the goal of identifying methods that could be trained to prospectively assign HIV patients to VL categories (i.e. SHVL, SVL, SLVL, DSVL, and RVL). Unsupervised learning (e.g. hierarchical clustering) is useful for establishing the data structure of VL categories and their locations in the feature vector space. Once the model is established (e.g. cluster boundaries), supervised

Figure 4. Feature segregation from hierarchical clustering. Each patient is colored corresponding to the results from the hierarchical clustering in Figure 2. The artificial line of points is a result of the grounding function used in *AdjMD*. Area = relative area of viral load exposure, *wRR* = weighted recency reliability, *IQR* = interquartile range, *AdjMD* = adjusted maximal viral load difference.

learning methods are better suited for prospective cluster assignment, given a robust "ground truth" for model training, as they do not depend on re-analysis of the entire population.

To this end, we compared the predictive power of several supervised learning methods for HIV cluster assignment, including: k-nearest neighbors (kNN), decision tree, support vector machine (SVM), Adaboost, and random forests. Models were trained on the original data set, and we then ranked their prediction power by their average F_1 score derived by leave-one-out cross-validation (LOOCV) on the clustered results (Table 3). We compared the ability of these methods to reconstruct the originally identified clusters, even when allowing for variability in cluster numbers (e.g. kNN with $k=\{7, 9\}$, or DT without a maximum depth specification). All methods performed comparably well, with the notable exception of Adaboost. This generally high performance was expected because the VL pattern categories are well-separated as a result of the

clustering. k-Nearest Neighbors and $k=5$, was computationally efficient and yielded the best results in Table 3.

We next considered the trade-off of predictive precision versus model interpretability. Critical clinical evaluation of machine learning results is important to protect against mis-categorization and clinical error. For this reason, many have advocated using models that are more clinically interpretable. kNN is dependent upon the entire training set for prediction, as it does not inherently "learn" patterns²⁴, hence it does not meet our interpretability criteria. In comparison, SVM offers a simpler model, but it's results could be non-intuitive for clinicians. And although Decision Trees offer the best interpretability, overly complex trees may be generated, as occurred in our study (Figure S3).

We found that pruned decision tree rules, with a maximum depth of 5 levels, met this interpretable criteria, however at

Figure 5. Comparison of patient categories with existing methods. A 2D binning of VLM counts for every patient category. Each row uses a different categorization method, and the method name is located to the right of the row, and the title of each subplot is the category assigned by the indicated method. The columns of each 2D bin are normalized based on the maximum number of logged viral load measurement (VLM) counts in the column: $\log_{10}[1 + VLM]$. Bin color for a count of 0 is copper, and other bin colors range from white to teal (the maximum of the $\log_{10}[VLM \text{ counts}]$ in the column of the bin). The black dots represent the last viral load measurement for the patient (opacity ≥ 0.3 ; 2D bins have variable opacity for the dots). The bottom row is our analysis is the same as Figure 3, but represented as a 2D bin. DSVL = durably suppressed viral load; LLVR = low level viral rebound; SLVL = sustained low viral load; SHVL = sustained high viral load; HVLS = high viral load suppression, RVL = rebounding viral load.

a slight cost to the predictive power (Table 3). The extracted decision rules are shown in Figure 6. Each category has a rule with a high proportion of true positive samples following the rule relative to all samples for the category (support). Similarly, a high proportion of the predicted class was found in the rule

(precision), indicating that the rules can be summarized into a majority rule. Note that the sum of the support does not necessarily add up to one for each class because some samples belonging to that class may have been otherwise placed into a different rule, making the precision of that rule weaker.

Table 3. F_1 prediction scores using LOOCV.

Group:	DSVL	SLVL	SHVL	HVLS	RVL	Average
Patients:	442	535	46	316	237	F_1 Score
kNN,k=5	0.9966	0.9925	0.9677	0.9889	0.9810	0.9853
kNN,k=9	0.9943	0.9907	0.9583	0.9841	0.9725	0.9800
kNN,k=7	0.9943	0.9897	0.9362	0.9873	0.9746	0.9764
Random Forest	0.9909	0.9841	0.9556	0.9685	0.9645	0.9727
Decision Tree (DT)	0.9898	0.9795	0.9670	0.9512	0.9432	0.9661
SVM	0.9955	0.9833	0.9111	0.9666	0.9387	0.9590
DT,max depth=5	0.9757	0.9659	0.9032	0.9375	0.9168	0.9398
<i>Polyhedron</i>	0.9727	0.9474	0.9011	0.8985	0.9109	0.9261
<i>Bounding Box</i>	0.9865	0.9630	0.8764	0.9038	0.8614	0.9182
<i>Push and Pull</i>	0.9737	0.9347	0.8842	0.9027	0.8767	0.9144
<i>Best Rep.</i>	0.9589	0.9280	0.9011	0.8598	0.8923	0.9080
<i>Mean</i>	0.9401	0.9004	0.9072	0.8212	0.8968	0.8931
<i>Smallest Disk</i>	0.9627	0.9017	0.8889	0.8246	0.8717	0.8899
<i>Median</i>	0.9271	0.8882	0.9167	0.7967	0.8953	0.8848
AdaBoost	0.9227	0.8248	0.5797	0.5033	0.6475	0.6956

LOOCV = leave-one-out cross validation, kNN = k nearest neighbor, DT = decision tree

SVM = support vector machine, DSVL = Durably Suppressed Viral Load

SHVL = Sustained High Viral Load, SLVL = Sustained Low Viral Load

RVL = Rebounding Viral Load, HVLS = High Viral Load Suppression

As an alternative interpretable model we explored the use of centroid cluster summarization, which is often used in clustering algorithms, and is flexible enough to accommodate different centroid determination methods^{22,25}. To determine the effects of different centroid determination algorithms, we compared seven different methods: multidimensional mean, multidimensional median, best representative center, bounding box method, smallest disk method, polyhedral center, and a novel “push and pull” (PnP) method inspired by force-directed graph drawing such as the Fruchterman-Reingold’s algorithm^{26,27} (see [Supplementary File 1](#)). Force directed clustering methods maximize inter-cluster center distances, while minimizing intra-cluster distance, and are the basis for modularity clustering in graph theory²⁸.

We then combined the centroid cluster summarization approach with a radius-based classification prediction algorithm. Let c_i be the i th cluster center with corresponding radius r_i , where r_i is calculated as the distance to the farthest intra-cluster sample from c_i , then for a new sample s choose its predicted cluster membership j such that $\frac{\|s - c_j\|}{r_j}$ is a minimum. We refer to this method as *radial normalization* classification.

Comparing the representative F_1 power of the centroid radial normalization methods (italicized in [Table 3](#)) to common machine learning algorithms, we find that the centroid interpretation loses some predictive power. However, the centroid

summary is highly interpretable because the entire model can be expressed concisely ([Supplementary Table S2](#)), and understood clearly. For example, a clinician classifying a patient by VL time series values would compare observed feature values with the ranges given in [Figure 6](#), and find which classification the patient’s data fits best within. In the case of the centroid method, if an observed value appears to fall in multiple categories, then they should be assigned to the one closest to the center (this allows a clinician to cross-check model predictions).

Temporal state variation

HIV patient viral load states are often fluid, with class changes (e.g. SHVL → HVLS) occurring due to therapy, viral genetics, social and other factors. To examine this aspect of classes, we use the k-Nearest Neighbors (k=5) model, fit to the original clusters, to predict the class state of each patient with ≥ 3 VL measurements using only partially retained VL data. For example if a patient has 6 viral load measurements, then we predict the class state at 3, 4, 5, and 6 VLM, which may yield SHVL → SHVL → RVL → HVLS as its prediction. We then constructed a state-transfer network using the trace-route method²⁹, revealing several interesting relationships:

1. Patients on therapy appear to suppress their viral loads at a positive linear rate throughout the entire 900 day span. This is quite different from the literature which suggests that if a patient is going to suppress their VL, it will be within 32 weeks, or 224 days¹³ ([Figure 7A](#)).

Figure 6. Extracted rules from pruned decision tree and polyhedral CM rule region. Support is the fraction of true positives satisfying the rule relative to all samples of the class. Precision is the proportion of true positives versus all positives in the rule. Rules are sorted in order of application, first by the level of the decision tree depth (Depth), and then by descending precision. The colored regions represent the values for which the rule holds (rule feature space). For the centroid method (CM; shaded gray) bounds were calculated by the polyhedron method, where the rectangular bar is the center and the radius is the area inside the parentheses. Area = relative area of viral load exposure, wRR = weighted recency reliability, IQR = interquartile range, AdjMD = adjusted maximal viral load difference.

Figure 7. Class state variation. **A**) Classification using kNN, with k=5, trained on the original five clusters to predict on partially retained viral load for patients ≥ 900 days of data. The number of patients in one class between 0–900 days are shown relative to the first state classification (i.e. third viral load measurement). **B**) A trace-route map of class state transfers ($class_1 \rightarrow class_2$) as a function of partially retained viral load derived from model. Nodes represent viral load classification and arrows reflect the volume of state transitions between successive VL measurements (e.g. SHVL $\rightarrow$ DSVL). Self-loops (e.g. RVL $\rightarrow$ RVL) indicate no change in state reflecting stable classification.

2. DSVL classification appears unstable for the first 400 days, suggesting that patients in this class should be monitored carefully during this initial period (Figure 7A).
3. The number of patients classified as SHVL drops considerably until ~500 days after first classification. After this point, those patients who have not yet left the SHVL category, may not do so (Figure 7A).
4. The two sets of classes {DSVL, SLVL} and {SHVL, RVL, HVLS} are well separated (i.e. without much transfer between sets; Figure 7B). This appears to suggest that patients whose viral load is consistently low or durably suppressed tend not to transfer into a high viral load state (i.e. RVL or SHVL), at least in this data cohort.
5. SHVL patients in this cohort tended to transfer out of the class at a much higher rate than the transfer in, suggesting positive patient care (Figure 7B). This observation is consistent with reports in the literature that entry into treatment, with adherence to a HAART regimen, generally results in viral load suppression.
6. The state transfer diagram illustrates that the most frequent state transition over time is remaining within the same cluster (Figure 7B) assignment.

Discussion

Researchers have previously performed HIV population case studies using differing schema to classify VL patterns^{2,9,10,13}. We have developed a unique method for standardizing the algorithmic classification of VL patterns using a set of optimally segregating features. These features have been specifically engineered to optimize unsupervised clustering of temporal sequences of VL data that are asynchronous and noisy. Our findings demonstrate their success in identifying five viral load patterns often reported in the literature⁷⁻¹². It is possible that additional viral load patterns may emerge in the future, for example due to new HIV variants that are resistant to current therapies. The method reported here is flexible enough to recognize such new temporal patterns of VL responses. It is also general enough that models could be trained on other viral infections that have patterns of natural or treatment related patient responses (e.g. hepatitis B and C, parvovirus B19), although this may require defining new features that capture disease specific pattern variants.

A common practice in data analytics is to calculate the centroid as the average of the points^{22,30}. However, Table 3 suggests that the mean is not necessarily the best centroid for HIV viral load data. We note two advantages of the centroid algorithm: First, we can choose the centroid best corresponding to the shape of the data, and second, we can use it to mathematically determine the amount of over-lap between n-dimensional cluster spheres (i.e. viral load categories). This method may facilitate cross-comparison of HIV research studies by providing a standard for VL pattern classification. Such standardization would be immensely useful in meta-analyses³¹⁻³⁵, potentially revealing the

influence of different patient care strategies or new relationships between different patient populations.

Our work also explored the trade-off with respect to predictive accuracy between model interpretability and more complex, “black box” approaches to classification. The interpretability versus predictability problem is well known in the deep learning literature³⁶⁻³⁸. Interpretability is a desirable attribute in clinical classification systems, allowing clinicians to integrate causal physiology and diagnostic information with data features in a way promotes clearer bedside clinical reasoning. Using an interpretable model for assigning viral load pattern membership may be advantageous when a clinician wishes to use the assigned pattern membership to aid in making a critical clinical decision (e.g. choosing between treatment options), or when examining features that may be linked to a mechanism (e.g. slope of VL decline and viral genotype). A “black box” or more complex model may make such decisions or interpretations more difficult³⁹, and can favor the use of simpler models at the expense of some predictive power.

Along these lines, we have also proposed a novel centroid-based algorithm for summarization of clustering results. This algorithm is not meant to supplant other well defined supervised learning algorithms, but rather to aid in interpretable assignment of VL patterns from other data sets into one of the five categories. The algorithm results are concise, allowing investigators to build the model in their preferred programming language. Hence this method may improve and standardize HIV population research by giving precise definitions to the varying temporal VL patterns, and potentially improving patient care.

Several caveats apply to our work. As noted, this is a single center study, and thus our method should be tested with a much larger data set to cross-validate the categories represented by the clusters. In addition, our feature vector was designed specifically based on observed VL patterns previously reported in the literature rather than objectively clustering the data using a standard time-series based clustering method⁴⁰⁻⁴². This may limit generalizability to other VL analyses. In addition, some of our features are slightly collinear - with the greatest correlation coefficient being between *IQR* and *wRR* (-0.717). However, while HVLS and RVL both have a varied range of *IQR*, it is clear that the HVLS class has greater *wRR* than the RVL class due to HVLS patients having a long consistent viral load tail. Furthermore *IQR* helps distinguish the HVLS and the SLVL or SHVL class, hence both *IQR* and *wRR* are necessary despite the slight correlation. Finally, because our method normalizes time into *number of days since first VL measurement*, we lose the ability to look for seasonal or yearly patterns in the data.

Our data set did not have patients in whom VL was initially suppressed, and then rebounded (EVL). We originally hypothesized the existence of six distinct VL patterns, we found that the *emergent VL* group was not a pattern identified in our data. Perhaps this is a consequence of a high rate of local patient engagement in therapy in this cohort study, access to care,

or the effectiveness of highly active anti-retroviral therapy regimens. We hypothesize that these conditions may not always exist (e.g. in areas where HAART is expensive, when people may lose the ability to pay for therapy), and that in such cases the EVL pattern may indeed be present and significant. Based on the formulation of the *adj MD* and *wRR* features, we hypothesize that a consequence of the grounding function is that any EVL pattern, if exists, will be grouped under RVL. This grouping may be appropriate as one can argue that going from a suppressed state to a high VL state is a form of rebounding. Clinical treatment of these patterns is likely to be similar. Further work with data sets that contain RVL patterns will need to be done to test these hypotheses. Unfortunately, we are not aware of any such data currently in the public domain.

Our method used hierarchical clustering to define groups, with a cutoff for group specification at a high level in the branching tree (i.e. level 5). Such thresholds or tuning parameters are characteristic of most unsupervised clustering algorithms^{22,43}. However, identification of important sub-clusters by using a lower threshold is also possible. Clustering results may change depending on the parameter chosen, revealing finer between-cluster differences as the number of clusters increase. The hierarchical clustering algorithm has the advantage that a proper cut-off can be easily visualized. For example, choosing a lower cut-off may reveal that the suppression group splits itself into categories with different rates of HIV viral load suppression during treatment. Researchers wishing to engineer a new feature vector for VL pattern segregation may find useful the [Supplementary material](#) on features we considered but subsequently removed due to poor performance.

Conclusions

We have proposed a set of four unambiguous features which have been successfully used in segregating five different types of temporal viral load patterns: durably suppressed viral load (DSVL), sustained low viral load (SLVL), sustained high viral load (SHVL), high viral load suppression (HVLS), and rebounding viral load (RVL). We have also proposed a novel centroid-based cluster summary algorithm. The use of this algorithm may improve meta-analyses or population studies of viral load patterns by standardizing the classification of HIV patient categories. Furthermore, the segregation process used in this paper (i.e. identifying domain specific features, performing unsupervised clustering, interpreting the results with a cluster summary) can be used to model other viral infections and the response of VL levels over time to treatment or natural disease progression. We also found that using a temporal state variation method is important when considering patient viral load classifications, as changes in patient response can continue to occur beyond previously estimated time frames.

Supplementary material

Supplementary Figures:

[Click here to access the data.](#)

- **Figure S1: Viral load distribution.** For each pair of viral load measurements, we calculate the change in days and the change in viral load counts for all patients and plot it as a scatter. The horizontal line of dots which appears between 0 and 2 are an artifact

Abbreviations

AdjMD = adjusted maximal viral load difference, *Area* = relative area of viral load exposure, ART = anti-retroviral therapy, DT = decision tree, EVL = **emerging viral load**, **DSVL** = Durably Suppressed Viral Load, HAART = highly active retroviral therapy, HIV = human immunodeficiency virus, *IQR* = interquartile range, kNN = k nearest neighbor, LLVL = low level viral load, LOOCV = leave-one-out cross validation, **SHVL** = Sustained High Viral Load, **SLVL** = Sustained Low Viral Load, RVL = Rebounding Viral Load, **HVLS** = High Viral Load Suppression, SVM = support vector machine, VL = viral load, VLM = viral load measurement, *wRR* = weighted recency reliability.

Data availability

Full access to the data is available on GitHub (Data S1): <https://doi.org/10.5281/zenodo.1313245>²¹

Data S1: Viral load data. The data set used for this study is provided in a completely deidentified format, CSV format where the first column represents a unique subject, with a random identifier. The subsequent values are as $t_{i,j}$, $V L_{i,j}$, where $t_{i,j}$ is the time from a universal T_0 for the VL measurement j for patient i , and $V L_{i,j}$ is the corresponding VL measurement. Each record (row) is of a unique length, depending on the number of VL measurements present for that subject. The study data, and code used for analysis, can be found at <https://doi.org/10.5281/zenodo.1313245>²¹.

Competing interests

No competing interests were disclosed.

Grant information

This work was partially funded by the University of Rochester Clinical and Translational Science Institute grants UL1 TR002001, and TL1 TR002000 from the National Center for Advancing Translational Sciences (NCATS), a component of the National Institutes of Health (NIH). This publication was also made possible through core services and support from the University of Rochester Center for AIDS Research (CFAR), an NIH-funded program (P30 AI078498).

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH).

Acknowledgements

We would like to thank Yusuf Bilgic (State University of New York at Geneseo) and James Java (University of Rochester), for discussions regarding the statistical analyses.

of using 20 and 48 in data to replace the “Pos <20” and “Pos <48” values which appeared in our data. The sequential range of viral load measurements shows that VL measurements taken within 10 days of each other may vary by $\pm 10^5$ copies/mL.

- **Figure S2: Patient feature extraction.** Feature extraction on 1576 patients displayed as 2D splicing of the 4 dimensional feature space. Each splice plots a dimension versus another in the form of a scatter plot.
- **Figure S3: Decision Tree.** While some useful rules may be pruned, the tree is otherwise complicated and difficult to draw useful conclusions from.
- **Figure S4: Seven centroid calculations on clustered viral load data.** For each cluster, the seven methods of calculating a globular cluster center are shown in comparison to each other (calculated on the normalized and clustered viral load data). Since the PnP method can have a center outside the range of [0,1], an indicator is shown for when the center goes beyond the range.
- **Figure S5: Centroid methods.** Gives a visual of how the seven methods work on an example point set. The green target signifies the exact center which is found according to the different methods in our algorithm.

Supplementary File 1: Review of existing viral load categorization methods and features and centroid detection methodologies that were considered but not used. A review of currently published viral load categorization methods.

[Click here to access the data.](#)

Supplementary Tables:

[Click here to access the data.](#)

- **Supplementary Table S1: Correlation coefficient matrix features.**
- **Supplementary Table S2: Centroids and radii from polyhedral CM.**

References

- Centers for Disease Control and Prevention (CDC): **Vital signs: HIV prevention through care and treatment—United States.** *MMWR Morb Mortal Wkly Rep.* 2011; **60**(47): 1618–23.
[PubMed Abstract](#)
- Yehia BR, Fleishman JA, Metlay JP, *et al.*: **Sustained viral suppression in HIV-infected patients receiving antiretroviral therapy.** *JAMA.* 2012; **308**(4): 339–42.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mellors JW, Muñoz A, Giorgi JV, *et al.*: **Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection.** *Ann Intern Med.* 1997; **126**(12): 946–54.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sterling TR, Vlahov D, Astemborski J, *et al.*: **Initial plasma HIV-1 RNA levels and progression to AIDS in women and men.** *N Engl J Med.* 2001; **344**(10): 720–725.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dybul M, Fauci AS, Bartlett JG, *et al.*: **Guidelines for using antiretroviral agents among HIV-infected adults and adolescents: recommendations of the Panel on Clinical Practices for Treatment of HIV.** *MMWR Recommendations and reports: Morbidity and mortality weekly report Recommendations and reports/Centers for Disease Control.* 2002; **51**(RR-7): 1–55.
[Reference Source](#)
- Attia S, Egger M, Müller M, *et al.*: **Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis.** *AIDS.* 2009; **23**(11): 1397–1404.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Viard JP, Burgard M, Hubert JB, *et al.*: **Impact of 5 years of maximally successful highly active antiretroviral therapy on CD4 cell count and HIV-1 DNA level.** *AIDS.* 2004; **18**(1): 45–49.
[PubMed Abstract](#)
- Greub G, Cozzi-Lepri A, Ledergerber B, *et al.*: **Intermittent and sustained low-level HIV viral rebound in patients receiving potent antiretroviral therapy.** *AIDS.* 2002; **16**(14): 1967–1969.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Terzian AS, Bodach SD, Wiewel EW, *et al.*: **Novel use of surveillance data to detect HIV-infected persons with sustained high viral load and durable virologic suppression in New York City.** *PLoS One.* 2012; **7**(1): e29679.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rose CE, Gardner L, Craw J, *et al.*: **A Comparison of Methods for Analyzing Viral Load Data in Studies of HIV Patients.** *PLoS One.* 2015; **10**(6): e0130090.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- de Jong MD, Simmons CP, Thanh TT, *et al.*: **Fatal outcome of human influenza A (H5N1) is associated with high viral load and hypercytokinemia.** *Nat Med.* 2006; **12**(10): 1203–1207.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yiitolo N, Sørensen P, Josefsson AM, *et al.*: **Consistent high viral load of human papillomavirus 16 and risk of cervical carcinoma in situ: a nested case-control study.** *Lancet.* 2000; **355**(9222): 2194–2198.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Phillips AN, Staszewski S, Weber R, *et al.*: **HIV viral load response to antiretroviral therapy according to the baseline CD4 cell count and viral load.** *JAMA.* 2001; **286**(20): 2560–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kononenko I: **Machine learning for medical diagnosis: history, state of the art and perspective.** *Artif Intell Med.* 2001; **23**(1): 89–109.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dubey A: **Applications of Machine Learning: Cutting Edge Technology in HIV Diagnosis, Treatment and Further Research.** *Computational Molecular Biology.* 2016; **6**(3): 1–6.
[Publisher Full Text](#)
- Rosa RS, Santos RH, Brito AY, *et al.*: **Insights on prediction of patients' response to anti-HIV therapies through machine learning.** In: *Neural Networks*

- (IJCNN), 2014 International Joint Conference on. *IEEE*; 2014; 3697–3704.
[Publisher Full Text](#)
17. Rodríguez JO, Prieto SE, Correa C, *et al.*: Predictions of CD4 lymphocytes' count in HIV patients from complete blood count. *BMC Med Phys*. 2013; 13(1): 3.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 18. Ramirez CM, Sinclair E, Epling L, *et al.*: Immunologic profiles distinguish aviremic HIV-infected adults. *AIDS*. 2016; 30(10): 1553–1562.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 19. Parbhoo S, Bogojeska J, Zazzi M, *et al.*: Combining Kernel and Model Based Learning for HIV Therapy Selection. *AAMIA Jt Summits Transl Sci Proc*. 2017; 2017: 239–248.
[PubMed Abstract](#) | [Free Full Text](#)
 20. Center for Medicare Services: CMS Cell Size Suppression Policy. 2015. [Online; accessed 29-November-2017].
[Reference Source](#)
 21. SamirRCHI: Samir-RCHI/Viral_Load_Data_Categorization: HIV Viral Load Categorization Release (Version v0.1-alpha). *Zenodo*. 2018.
<http://www.doi.org/10.5281/zenodo.1313245>
 22. Han J, Pei J, Kamber M: Data mining: concepts and techniques. Elsevier; 2011.
[Reference Source](#)
 23. Punj G, Stewart DW: Cluster Analysis in Marketing Research: Review and Suggestions for Application. *J Mark Res*. 1983; 20(2): 134–148.
[Publisher Full Text](#)
 24. Keller JM, Gray MR, Givens JA: A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*. 1985; SMC-15(4): 580–585.
[Publisher Full Text](#)
 25. Maire F: An algorithm for the exact computation of the centroid of higher dimensional polyhedra and its application to kernel machines. In: *Third IEEE International Conference on Data Mining. IEEE Comput Soc*. 2003.
[Publisher Full Text](#)
 26. Kobourov SG: Spring Embedders and Force Directed Graph Drawing Algorithms. *arXiv preprint arXiv 12013011*. 2012.
[Reference Source](#)
 27. Fruchterman TMJ, Reingold EM: Graph drawing by force-directed placement. *Software: Practice and Experience*. 1991; 21(11): 1129–1164.
[Publisher Full Text](#)
 28. Noack A: Modularity clustering is force-directed layout. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2009; 79(2 Pt 2): 026102.
[PubMed Abstract](#) | [Publisher Full Text](#)
 29. Zand MS, Trayhan M, Farooq SA, *et al.*: Properties of healthcare teaming networks as a function of network construction algorithms. *PLoS One*. 2017; 12(4): e0175876.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 30. Abdi H: Centroids. *Wiley Interdiscip Rev Comput Stat*. 2009; 1(2): 259–260.
[Publisher Full Text](#)
 31. Etter P, Landovitz R, Sibeko S, *et al.*: Recommendations for the follow-up of study participants with breakthrough HIV infections during HIV/AIDS biomedical prevention studies. *AIDS*. 2013; 27(7): 1119–1128.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 32. Olsen CM, Knight LL, Green AC: Risk of melanoma in people with HIV/AIDS in the pre- and post-HAART eras: a systematic review and meta-analysis of cohort studies. *PLoS One*. 2014; 9(4): e95096.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 33. Blaser N, Wettstein C, Estill J, *et al.*: Impact of viral load and the duration of primary infection on HIV transmission: systematic review and meta-analysis. *AIDS*. 2014; 28(7): 1021–1029.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 34. Boender TS, Sigaloff KC, McMahon JH, *et al.*: Long-term Virological Outcomes of First-Line Antiretroviral Therapy for HIV-1 in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis. *Clin Infect Dis*. 2015; 61(9): 1453–1461.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 35. Boerma RS, Boender TS, Bussink AP, *et al.*: Suboptimal Viral Suppression Rates Among HIV-Infected Children in Low- and Middle-Income Countries: A Meta-analysis. *Clin Infect Dis*. 2016; 63(12): 1645–1654.
[PubMed Abstract](#) | [Publisher Full Text](#)
 36. Bologna G: Symbolic Rule Extraction from the DIMLP Neural Network. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg; 2000; 240–254.
[Publisher Full Text](#)
 37. Bologna G: A model for single and multiple knowledge based networks. *Artif Intell Med*. 2003; 28(2): 141–163.
[PubMed Abstract](#) | [Publisher Full Text](#)
 38. Intrator O, Intrator N: Interpreting neural-network results: a simulation study. *Comput Stat Data Anal*. 2001; 37(3): 373–393.
[Publisher Full Text](#)
 39. Shickel B, Tighe PJ, Bihorac A, *et al.*: Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform*. 2017; 1–1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 40. Klapper-Rybicka M, Schraudolph NN, Schmidhuber J: Unsupervised Learning in LSTM Recurrent Neural Networks. In: *Artificial Neural Networks — ICANN 2001*. Springer Berlin Heidelberg; 2001; 684–691.
[Publisher Full Text](#)
 41. Bahadori MT, Kale D, Fan Y, *et al.*: Functional subspace clustering with application to time series. In: *International Conference on Machine Learning*. 2015; 228–237.
[Reference Source](#)
 42. Kontaki M, Papadopoulos AN, Manolopoulos Y: Continuous subspace clustering in streaming time series. *Inf Syst*. 2008; 33(2): 240–260.
[Publisher Full Text](#)
 43. Karypis G, Han EH, Kumar V: Chameleon: hierarchical clustering using dynamic modeling. *Computer*. 1999; 32(8): 68–75.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 29 October 2018

<https://doi.org/10.5256/f1000research.17007.r39645>

© 2018 Telenti A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Amalio Telenti

The Scripps Research Institute, La Jolla, CA, USA

This report brings machine learning approaches to the classification of patterns of viral control in HIV infected individuals. This is welcome because, although this is a mature field in HIV, it signals the opportunity for new models in this and in other infections.

Strengths: very well developed models and excellent reporting of the results through figures and documentation. The code and datasets are available in Github.

Weakness: the model is trained and implemented on a suboptimal dataset. Treatment response in HIV infection (and thus the modeling of viral response) is well understood and best modeled with the knowledge of the time of treatment initiation, and a full understanding of variable influencing treatment response. Having a cohort that is described solely by "time from first measured viral load" is to all purposes, suboptimal. An additional issue is the reliance of a limited number of viral load determinations for an unclear number of individuals. Depending on the circumstances of sampling, having three viral load over an undisclosed time period is not devoid of many uncontrolled biases. Lastly, the text is equivocal in the utilization of the last time point – the reviewer understands that the information contained in the last point may be weighted because of the possibility that it is noisy. Unfortunately, in real life, that is the moment where strong predictive models are needed. It is possible that this was actually the goal of the authors.

Summary: this work is a valuable contribution to the field, and the basic concepts and models will hopefully be deployed in the study of datasets that are more appropriate for this exercise. It is desirable that future modeling includes a more ambitious plan to move from the current train-test approach to one that establishes the generalization of the model. It will also be critical to observe the predictive value of the model on longer term outcomes.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Host and pathogen genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 28 September 2018

<https://doi.org/10.5256/f1000research.17007.r37937>

© 2018 Blower S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sally Blower

Center for Biomedical Modeling, Semel Institute of Neuroscience and Human Behavior, David Geffen School of Medicine, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

This is an extremely interesting study that proposes a novel quantitative methodology for classifying HIV patients by viral load patterns. The authors propose four computable characteristics of time-varying viral load patterns and a novel classification algorithm. They demonstrate their approach by classifying a group of 1,576 HIV positive patients into five categories based on viral load patterns.

This is an extremely well written interesting paper with excellent figures. The proposed methodology has great importance and utility for both research studies and clinical programs.

My only very minor comments are:

1. For the descriptions given in Table 2 for mathematical notation. I suggest changing the description of "refers to a single patient" to "refers to a specific patient".

2. Equation 10, please clarify what the denominator means; i.e., how does "positive" differ from "true positive" or "false positive".
3. In the paragraph on page 6, under the heading Analytic terminology, editing is needed on line 8, where it says: clusters. interchangeably.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: mathematical modeling of HIV

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research