HE²RO Policy Brief

Number 32 · December 2019

APPLYING MACHINE LEARNING TO LABORATORY DATA: PREDICTING SUPPRESSION OF NEXT HIV VIRAL LOAD IN SOUTH AFRICA

Background

With more than seven million HIV-infected people [1], South Africa is home to more people living with HIV than any other country in the world,[2] and their national ART program is the world's largest.[3] South Africa adopted a "treat all" policy to provide ART to all HIV-positive people, regardless of CD4 cell count, in September 2016.[3] Expanded ART availability has dramatically altered the health and quality of life of people living with HIV/AIDS, with an estimated life expectancy gain of 11.3 years between 2003 and 2011 due to ART[4] and a 77% decrease in HIV transmission in stable serodiscordant couples.[5] However, the rapid scale-up of the national ART program has put tremendous pressure on the limited resources of a public health sector. By expanding eligibility, UTT has eliminated "pre-ART" care for most patients.[6] It has also led to faster ART initiation with fewer required clinic visits prior to dispensing drugs. These changes have compressed the care cascade, likely leading to greater ART uptake. At the same time, retention of new patients may have suffered. The problem of patient loss to follow-up (LTFU) within the public sector in South Africa has been well documented.[7] The potential for improved health through expanded ART availability will only be realized if individuals sustain engagement in HIV care.

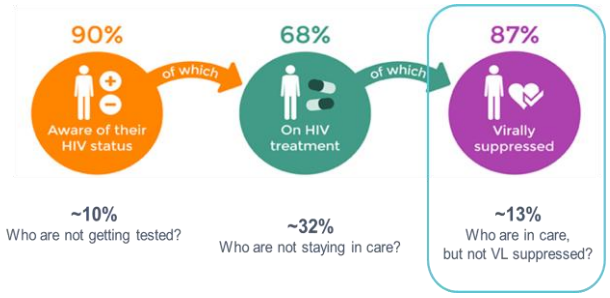

Figure 1: SA HIV care cascade, 2018 (avert.org)

To optimize South Africa's HIV response and reach targets of 95% tested, 95% treated, and 95% virally suppressed, numbers of patients initiating and successfully maintaining viral suppression on antiretroviral therapy must increase. In 2018, just 53% of people living with HIV (PLWH) in South Africa were virally suppressed.[1] While much effort and resources have been focused on tracing those LTFU and returning them to care, very little prior work has successfully addressed identifying those most at risk of poor treatment outcomes while still engaged in care.

Methods

We applied machine learning and modelling algorithms developed by Palindrome, data science implementers, (https://www.palindromedata.com) to de-identified HIV programmatic data collected from public sector treatment facilities based in two districts supported by Right to Care between 2015 and 2019. We included data for patients all patients who had accessed HIV care, initiated treatment and were retained through to virologic monitoring. HIV viral load (VL) suppression at next VL test was selected as primary outcome as it is an established clinical treatment outcome and objectively defined (diagnostically measurable reading) and thus made for a good target outcome to build confidence around the approach. High (>1000) Viral Loads that followed shortly (<6 months) after a previously high Viral Load were excluded from analyses due to the high probability of also being high.

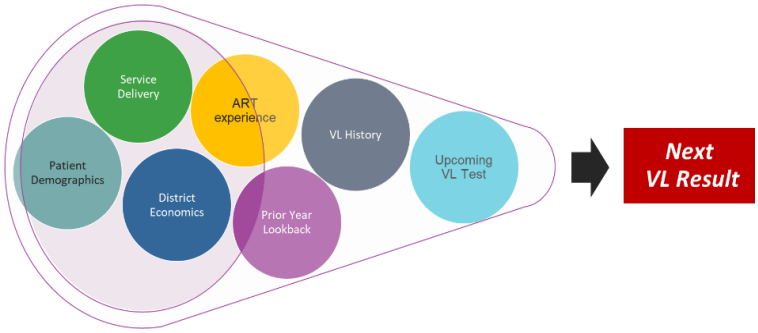

Figure 2: Framework for modelling compound effects

As the likelihood of viral load suppression is impacted by multiple components of the patient's treatment journey (Figure 2), demographic, clinical, behavioral (e.g. visit patterns) and laboratory data were investigated as potential predictor variables of VL result at next visit. Multiple models were created using various combinations of predictor variables and classification algorithms. These were then tested against unseen data to identify the optimal balance between predictive power and implementation feasibility. The final models were built using a random forest classifier and combined features. Models were evaluated using receiver operating characteristic (ROC) curves which assess the performance of each model's predictions against a test set of unseen data with known outcomes. The area under the curve (AUC) measures how well a variable can classify into two groups – in this case VL suppressed or unsuppressed. AUC values range between 0.5 (poor classifier) and 1.0 (excellent classifier).

Results

We included data on 688,614 VL results, during the study period 1 January 2016 – July 2019. We tested >50 potential input features per patient in 7 different models using multiple combinations of input features and classification algorithms. Each model was tested against unseen data to identify optimal predictive performance. Model results ranged from AUC of 0.57 for the poorest performing model (included gender and age at ART start only) to an AUC of 0.739 for the best performing model (Figure 3). Practically, this means the model correctly anticipated whether the VL result at next test would be suppressed or unsuppressed in approximately 3 out of 4 patients. The model consistently achieved an accuracy of 75% per month over the most recent months of patient data in 2019 suggesting it is both historically accurate but also relevant to patients currently accessing care.

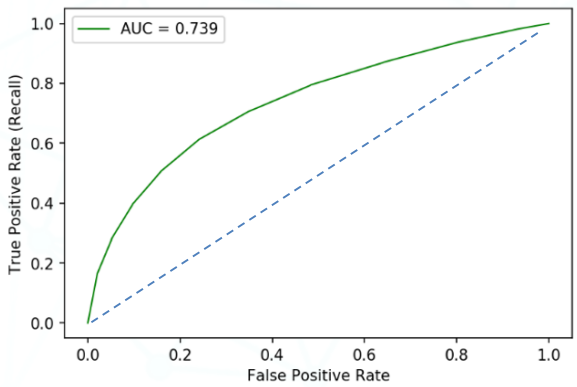

Figure 3. ROC curve for final model

Several patient characteristics were found to contribute a larger importance in the Random Forest model in terms of predicting their risk of having an unsuppressed VL at next visit (Figure 4). These included: age at ART initiation, most recent VL result value, time on ART, pattern of previous missed visits and month they accessed treatment. As a reminder, the Random Forest method observes the correlation of these features *in combination* – as such the figure should be read as a group, rather than an ordered list of priority or individual causation.

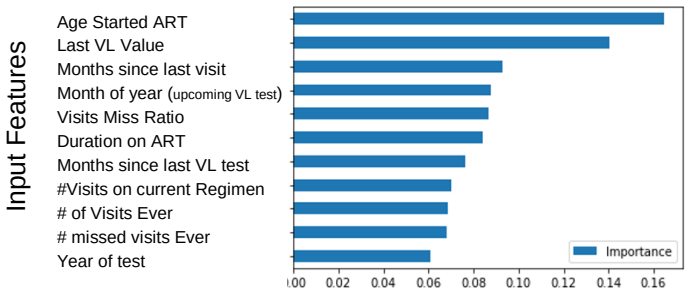

Figure 4. The top 11 features with largest predictive importance in final model, from the original 50

Policy relevance

As South Africa increases efforts towards 95-95-95 goals, knowing which patients require additional services and interventions in order to achieve successful treatment outcomes at each step of the cascade is critical. Our model was able to effectively separate high-risk from low-risk patients using a combination of clinical, laboratory and behavioral (visit patterns) data. Early detection of patients at high risk of becoming virologically unsuppressed has implications not only for the individual patient's health, but also for the risk of onward transmission of the virus and impact on breaking transmission chains. Potential operational application of these results could include the ability to score patients into risk categories at each visit and triage their care accordingly – high risk patients get prioritized to receive intensive intervention at point of care while low risk patients are expedited through the visit. On-going work will also continue to develop the model and explore other predicted outcomes such as risk of disengaging from care at next scheduled visit.

Leveraging predictive models to better understand the risk of individuals will allow for health care services to better triage patients, improving efficiency and resource utilization. By prioritizing those most at-risk, clinics can realize better health outcomes without additional investments in data collection and staff. Moreover, by anticipating future issues before any visible signs are present (e.g. an unsuppressed VL), clinics can intervene pro-actively while patients are still accessible, engaged in health services and provide targeted services earlier.

References

1. Statistics South Africa. Mid-year population estimates 2018. Pretoria, 2018 https://www.statssa.gov.za/publications/P0302/P03022018.pdf.
2. Joint United Nations Programme on HIV/AIDS (UNAIDS). Global report: UNAIDS report on the global AIDS epidemic 2013. Geneva, 2013 http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf.
3. Joint United Nations Programme on HIV/AIDS (UNAIDS). South Africa takes bold step to provide HIV treatment for all. 2016; published online May 13. http://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2016/may/20160513_UTT.
4. Bor J, Herbst AJ, Newell M-L, Bärnighausen T. Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. *Science (80-)* 2013; **339**: 961–5.
5. Oldenburg CE, Bärnighausen T, Tanser F, *et al.* Antiretroviral therapy to prevent HIV acquisition in serodiscordant couples in a hyperendemic community in rural South Africa. *Clin Infect Dis* 2016; **63**: 548–54.
6. Fox MP, Rosen S. A new cascade of HIV care for the era of "treat all". *PLOS Med* 2017; **14**: e1002268.
7. Fox MP, Rosen S. Patient retention in antiretroviral therapy programs up to three years on treatment in sub-Saharan Africa, 2007-2009: systematic review. *Trop Med Int Heal* 2010; **15**: 1–15.

This study has been made possible by the generous support of the American People and the President's Emergency Plan for AIDS Relief (PEPFAR) through USAID under the terms of Cooperative Agreement 72067419CA00004 to HE2RO. The contents are the responsibility of the authors and do not necessarily reflect the views of PEPFAR, USAID or the United States Government.

Suggested citation: Maskew M, Crompton T, Sharpey-Schafer K, De Voux L, Bor J, Rennick M, Pisa P, Miot J. Applying machine learning to laboratory data: predicting suppression of next HIV viral load in South Africa. Johannesburg: HE²RO Policy Brief Number 32, Health Economics and Epidemiology Research Office, December 2019.