Machine learning to predict persons with HIV at risk of falling out of care

1307

Ravi Goyal[1], Gordon Honerkamp-Smith[1], Susan J Little[1], Alan Wells[1], **Thomas CS Martin**[1]
[1]Department of Infectious Diseases and Global Public Health, University of California San Diego, San Diego, USA

BACKGROUND

Effective antiretroviral therapy to maintain durable viral suppression is key to ending the HIV epidemic in the United States. Current public health strategies to identify persons with HIV (PWH) who are out of care are often retrospective in nature and inaccurate as many individuals identified by standard algorithms are actually in care, have moved, are incarcerated, or have died. We hypothesized that:

- Machine learning algorithms could improve the prediction of HIV outcomes (e.g., unsuppressed viral load (VL) within 12 months) among PWH
- Incorporating viral testing variables (eg testing interval/frequency/variance of testing interval) could improve prediction of developing detectable VL.

METHODS

- This study used de-identified data from all PWH in San Diego County, USA using mandatorily reported public health data between 2017 and 2022
- Outcome measure: VL>200 copies/ml in a year-long prediction window. Where no VL was available, next observation carry back was used to ascertain the outcome
- Predictive variables included age at diagnosis, sex, gender, race/ethnicity, HIV stage at diagnosis, transmission risk, baseline CD4 and longitudinal HIV VL and CD4
- VL testing patterns prior to prediction window were incorporated: proportion unsuppressed, mean and variance of testing intervals, duration since last VL, linear trend VL.
- The data was divided into three datasets: training (60%), refinement (20%) and validation (20%)
- A Random Forest model was trained and refined on the training and refinement datasets sequentially before testing its predictive capacity on the validation dataset
- Predictive ability of the algorithm was reported by sensitivity, specificity, AUC, and positive and negative predictive values (PPV/NPV).
- Algorithm bias assessment: Performance was evaluated on racial/ethnic subgroups
- Relative variable importance was assessed using permutation analysis and mediation via mediation analysis

Over 50% of PWH identified by a machine learning algorithm of being at risk of unsuppressed viral load had a viral load >200 copies/ml within the next 12 months

RESULTS

- 18,916 PWH were included shown in **Table 1**
- Median of 0.92 years VL data and median of 3 VL measurements prior to prediction window

VALIDATION OUTCOMES

- Out of set prediction had sensitivity of 33.8%, specificity of 96.5% and PPV of 55.7%
- Prediction performance did not differ by race or ethnicity (**Figure 1**)
- VL testing variables ranked higher than 'traditional' risk factors for developing unsuppressed viral load (**Figure 2**)
- Mediation analysis indicated some, but not all, effects of demographic/transmission risk were mediated through VL testing variables (eg. Race mediated by proportion unsuppressed VL [~39.9%])

Table 1: Characteristics of the overall population and breakdown by datasets used to train and evaluate the machine learning algorithm

Characteristic	Level	Overall N = 18,916	Training N = 11,350	Refinement N = 3,783	Validation N = 3,783	p-value
Current Age	<40	3,821 (20%)	1,990 (18%)	886 (23%)	945 (25%)	<0.001***
	(40, 60)	9,062 (48%)	5,520 (49%)	1,741 (46%)	1,801 (48%)	
	60+	6,016 (32%)	3,839 (34%)	1,148 (30%)	1,029 (27%)	
Age at Diagnosis	<20	732 (3.9%)	423 (3.7%)	154 (4.1%)	155 (4.1%)	0.143
	(20, 30)	6,653 (35%)	3,920 (35%)	1,365 (36%)	1,368 (36%)	
	(30, 40)	6,548 (35%)	4,014 (35%)	1,262 (33%)	1,272 (34%)	
	40+	4,966 (26%)	2,992 (26%)	994 (26%)	980 (26%)	
Birth Sex	Male	16,952 (90%)	10,183 (90%)	3,399 (90%)	3,370 (89%)	0.545
	Female	1,942 (10%)	1,165 (10%)	374 (9.9%)	403 (11%)	
Race/Ethnicity	Hispanic, any race	6,474 (34%)	3,866 (34%)	1,279 (34%)	1,329 (36%)	0.05
	Other	1,429 (7.6%)	861 (7.6%)	266 (7.1%)	302 (8.1%)	
	Black	2,567 (14%)	1,508 (13%)	557 (15%)	502 (13%)	
	White	8,366 (44%)	5,111 (45%)	1,652 (44%)	1,603 (43%)	
	Missing	80	4	29	47	
Exposure Category	MSM	13,298 (74%)	8,041 (73%)	2,655 (75%)	2,602 (74%)	0.051
	IDU	1,087 (6.0%)	672 (6.1%)	214 (6.1%)	201 (5.8%)	
	MSM & IDU	1,631 (9.1%)	1,054 (9.6%)	294 (8.3%)	283 (8.1%)	
	Other	171 (0.9%)	104 (0.9%)	35 (1.0%)	32 (0.9%)	
	Heterosexual contact	1,829 (10%)	1,125 (10%)	328 (9.3%)	376 (11%)	
	Missing	900	354	257	289	
Disease Category	HIV only (HIV, stage 1 or 2)	8,748 (48%)	4,999 (46%)	1,837 (50%)	1,912 (52%)	<0.001***
	HIV and later AIDS	6,438 (35%)	3,993 (36%)	1,281 (35%)	1,164 (32%)	
	HIV/AIDS simultaneously	3,133 (17%)	1,964 (18%)	561 (15%)	608 (17%)	
	Missing	597	394	104	99	
Testing History Length (years)		0.92 (0.21, 1.35)	0.84 (0.28, 1.10)	0.67 (0.00, 1.94)	2.53 (0.45, 3.01)	
Number of Viral Loads		3.00 (2.00, 5.00)	3.00 (2.00, 4.00)	3.00 (1.00, 6.00)	5.00 (2.00, 8.00)	<0.001***

Figure 1: Model performance on out of set prediction subdivided by ethnicity and race

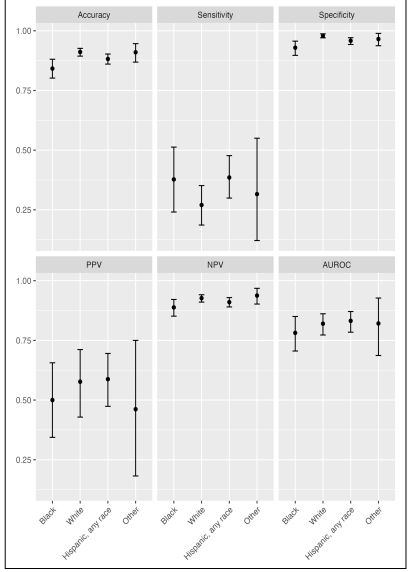


Figure 2: Relative importance of variables included in predicting the outcome of viral load >200 performed by permutation analysis

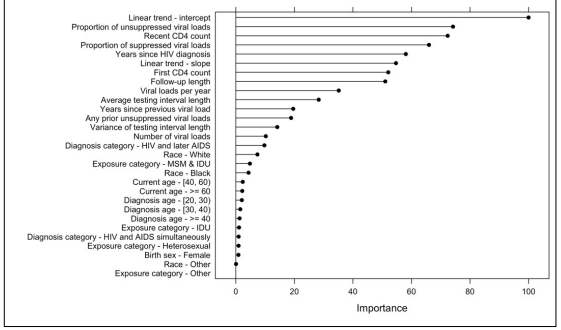


Table 2: Prediction performance of the model on training, refinement, and validation datasets

Dataset	Sensitivity	Specificity	PPV	NPV	AUROC
Training	33.9 (26.2, 41.5)	97.5 (96.3, 98.7)	66.8 (55.5, 78.0)	91.0 (90.0, 91.9)	83.3 (79.2, 87.5)
Refinement	43.8 (38.5, 49.5)	94.5 (93.5, 95.4)	51.5 (45.5, 57.5)	92.7 (91.7, 93.7)	83.8 (81.3, 86.3)
Validation	33.8 (28.6, 39.0)	96.5 (95.7, 97.2)	55.7 (48.7, 62.8)	91.7 (90.6, 92.8)	82.2 (79.3, 85.0)

CONCLUSIONS

- **Machine learning algorithms may improve prediction of PWH who are at risk of developing detectable VL or falling out of care**
- **In this study, >50% of persons identified by the algorithm had VL >200 copies/ml in the subsequent year**
- **VL testing characteristics were found to rank as more important than 'traditional' risk factors such as race, sex, age or transmission risk and should be included in predictive models**
- **Improved data linkage and use of novel machine learning techniques may be able to help allocate intensive multi-modal interventions to prevent falling out of care or enable the early re-engagement for individuals that have fallen out of care**

ADDITIONAL KEY INFORMATION

Contact: thm001@health.ucsd.edu

We would like to acknowledge San Diego County Public Health