01137

Improving HIV case-finding with machine learning

Pavlo Smyrnov[1], Yuliia Sereda[2], Artem Lytvyn[2], Olga Denisiuk[1]
[1]Alliance for Public Health, Kyiv, Ukraine, [2]Independent consultant, Kyiv, Ukraine

BACKGROUND

Alliance for Public Health implements social network strategy for HIV case-finding among key populations since 2016. We developed a model to improve recruitment of undiagnosed HIV-positives in the social network using machine learning (ML) algorithm.

METHODS

Since 2016 more than 130,000 people who inject drugs and their extended risk network peers were recruited in 12 regions of Ukraine. Recruitment starts from each HIV positive case with following criteria: 14+ years, inject drugs.

Participants are provided with 3 coupons (**basic rule**) to invite their peers defined as an injecting or sexual partner or other who can be at risk of HIV. Recruitment stops after two consecutive HIV-negative cases in a chain. Data on recruitment chains and participants' characteristics is collected in mobile application and cloud database.

In addition we conducted logistic regression to find association of probable seroconversion and characteristics of peers immediately adjacent in recruitment chain from IBBS 2015 data (PWID, n=9,405). From 2017 additional coupons were provided to participants with characteristics associated with seroconversion of peer, such as “over 10 years of injecting drugs” (+1 coupon), “history of incarceration” (+1 coupon), “recent HIV infection” (+3 coupons) and combinations (max +5 coupons) (**rule based algorithm**).

Later we developed Random Forest Based ML algorithm to predict the probability of recruiting of an undiagnosed HIV-positive person within one step from the participant. Considering the estimated probability, recommendations on a number of additional coupons are provided: probability=0.05-0.1 (+1 coupon), 0.1-0.15 (+2 coupons), 0.15-0.2 (+3 coupons), 0.2-0.25 (+4 coupons), >0.25 (+5 coupons) (**ML algorithm**).

First, we analyzed theoretical performance of ML and actual recruitment results (Figure 1). From November 2019 we randomly assigned ML algorithm and rule based algorithm to all cases by 50/50 AB (split) randomization. We analyzed actual performance of model on data for period from November 1, 2019 till January 24, 2020. We update ML model once in a months to account for new data for model training. We exclude last 20 days from the training dataset to account for outcome period for coupons return.

ML algorithm has performed better or equally well in comparison with rule based algorithm on making decision who should receive additional recruitment coupons due to higher probability of undiagnosed HIV cases among recruited peers. ML algorithm was significantly better than rule based algorithm on making “big” decisions on who should get five or more recruitment coupons.

RESULTS

In theoretical comparison with rule based recruitment algorithm ML model shown to be 1.5-2.5 times more efficient (based on GINI index) in predicting chances of undiagnosed positive case (Figure 1). ML model with 42 predictors yielded a GINI index of 34% for classification of HIV-positives and negatives. The most informative predictors of recruitment of HIV-positives included “HIV test result”, “Region”, “Years of injecting drugs”, while “Age” and “Marital Status” had the lowest contribution to prediction. Comparison of actual application of rule based algorithm and ML algorithm shown that ML algorithm is performing better or similar to rule based approach (Table 1 and Table 2).

Table 1

Each coupon given away with ML is as effective as coupon of rulebased aproach in terms of event per coupon, the yield is not decreasing with increased number of distributed coupons

	target	missed	total	Event per coupon
ML algorithm	**218**	**7627**	**7845**	**0,0278**
RB algorithm	152	5640	5792	0,0262
Grand Total	370	13267	13637	0,0271
p-value (chi-squared test of homogeneity)	0,5830			

Number of coupons given to those who eventually participate in case management = target; Number of coupons not given to positives who eventually participate in case management = missed; Total number of coupons = total

Figure 1



AUC – area under the receiver operating characteristic curve; AUC_current – AUC for the Standard of Care (Coupon Issue Strategy); AUC_logit – AUC for logistic regression classification; AUC_RF – AUC for random forest classification. GINI=AUC*2-1

CONCLUSIONS

Higher level of discriminatory power (an ability to distinguish between successful and not successful recruitment) of ML model suggests that application of ML algorithm during recruitment could improve HIV-positive yield among recruited participants and guide HIV testing to address gap in locating undiagnosed cases. Piloting of ML algorithm with randomizing recruiters confirmed the effectiveness of ML in improving case-finding in groups connected in risk networks with high prevalence of HIV.

Table 2

ML algorithm is performing significantly better than rulebased aproach in terms of bringing at least one positive to the case management program per visit

Event = "client led to at least one positive who participated in case management"; 1 -- event, 0 -- no event

	# of visits				
	0	1	Share of "0"	Share of "1"	**Total**
ML algorithm	**1403**	**176**	**88,85%**	**11,15%**	**1579**
RB algorithm	1520	136	91,79%	8,21%	1656
Grand Total	2923	312	90,36%	9,64%	3235
p-value (chi-squared test of homogeneity)				0,0047	

While comparing performance of ML aproach and rulebased aproach it is clear that ML shows better results when issuing high number of coupons (5+)

Coupons Issued	0	1	Share of "0"	Share of "1"	**Total**
3	776	68	91,94%	8,06%	844
ML algorithm	50	9	84,75%	15,25%	59
RB algorithm	726	59	92,48%	7,52%	785
4	423	50	89,43%	10,57%	473
ML algorithm	31	4	88,57%	11,43%	35
RB algorithm	392	46	89,50%	10,50%	438
5	1015	139	87,95%	12,05%	1154
ML algorithm*	717	108	86,91%	**13,09%**	825
RB algorithm	298	31	90,58%	9,42%	329
6	478	53	90,02%	9,98%	531
ML algorithm	474	53	89,94%	10,06%	527
RB algorithm	4		100,00%	0,00%	4
Grand Total	**2692**	**310**	**89,67%**	**10,33%**	**3002**
5&6 merged	**1493**	**192**	**88,61%**	**11,39%**	**1685**
ML algorithm**	1191	161	88,09%	**11,91%**	1352
RB algorithm	302	31	90,69%	9,31%	333
3&4 merged	**1199**	**118**	**91,04%**	**8,96%**	**1317**
ML algorithm	81	13	86,17%	13,83%	94
RB algorithm	1118	105	91,41%	8,59%	1223
* p-value (Fisher's exact test)				0,0180	
** p-value (Fisher's exact test)				0,0323	

Author contact information: smyrnov@aph.org.ua