Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210265

How to Identify Potential Candidates for HIV Pre-Exposure Prophylaxis: An AI Algorithm Reusing Real-World Hospital Data

Jean-Charles DUTHE[a,b,1], Guillaume BOUZILLE[a], Emmanuelle SYLVESTRE[a,c], Emmanuel CHAZARD[d], Cedric ARVIEUX[b,e] and Marc CUGGIA[a]
[a] *Univ. Rennes, CHU Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France*
[b] *COREVIH Bretagne, France*
[c] *CHU Martinique, Centre de données cliniques, F-97200 Martinique, France*
[d] *Univ. Lille, CHU Lille, ULR 2964 -METRICS, CERIM, Public health dept., F-59000 Lille, France*
[e] *Infectious Diseases and Emergency Unit , CHU Rennes, F-35000 Rennes, France*

Abstract. HIV Pre-Exposure Prophylaxis (PrEP) is effective in Men who have Sex with Men (MSM), and is reimbursed by the social security in France. Yet, PrEP is underused due to the difficulty to identify people at risk of HIV infection outside the "sexual health" care path. We developed and validated an automated algorithm that re-uses Electronic Health Record (EHR) data available in eHOP, the Clinical Data Warehouse of Rennes University Hospital (France). Using machine learning methods, we developed five models to predict incident HIV infections with 162 variables that might be exploited to predict HIV risk using EHR data. We divided patients aged 18 or more having at least one hospital admission between 2013 and 2019 in two groups: cases (patients with known HIV infection in the study period) and controls (patients without known HIV infection and no PrEP in the study period, but with at least one HIV risk factor). Among the 624,708 admissions, we selected 156 cases (incident HIV infection) and 761 controls. The best performing model for identifying incident HIV infections was the combined model (LASSO, Random Forest, and Generalized Linear Model): AUC = 0.88 (95% CI: 0.8143-0.9619), specificity = 0.887, and sensitivity = 0.733 using the test dataset. The algorithm seems to efficiently identify patients at risk of HIV infection.

Keywords. Pre-exposure prophylaxis (PrEP), HIV prevention, clinical informatics, predictive analytics, sexual health, risk reduction practices, machine learning

1. Introduction

Each year, approximately 6,500 new Human Immunodeficiency Virus (HIV) infections are recorded in France. In 2018, the number of new HIV infections slightly decreased (-7%) compared with the previous years due to the effect of prevention interventions, including Pre-Exposure Prophylaxis (PrEP) [1]. PrEP is a strategy to reduce the risk of HIV infection based on the administration of an antiretroviral treatment during the

[1] Corresponding Author, E-mail: jeancharles.duthe@chu-rennes.fr

exposure period. PrEP (the combination of tenofovir-disoproxil-fumarate and emtricitabine) is authorized in France since 2016 and is reimbursed by the French Social Security. PrEP role in HIV prevention is now well established (86% of effectiveness, especially among Men having Sex with Men (MSM)) [2]. In France, the number of people taking PrEP has increased rapidly, from 1,166 in 2016 to 15,501 individuals in 2019, mainly among MSM [3]. However, the number of MSM at high risk of HIV ranges between 32,000 [4] and 50,000 [5], indicating a low PrEP coverage rate. Therefore, our objective was to develop a model to identify candidates for PrEP exploiting clinical data. Several studies on HIV [6, 7] used this strategy to identify people at risk of HIV infection with the aim of improving and better targeting prevention in this population. Here, we used machine learning methods and clinical data from Rennes academic hospital (Rennes CHU).

2. Materials and Methods

Clinical data were extracted from eHOP [8], a centralized Clinical Data Warehouse (CDW) that contains the data of 1,600,000 patients hospitalized at Rennes CHU.

Among all patients aged 18 or more, having at least one admission at Rennes CHU between January 1, 2013 and December 31, 2019, two groups of patients were considered: the “case” group and the “control” group.

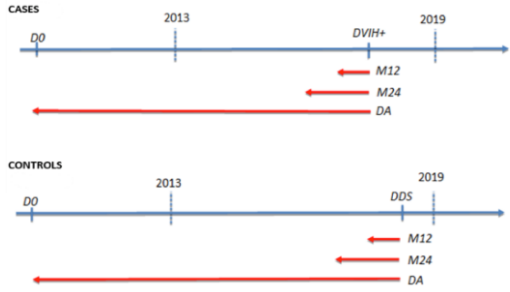

Figure 1. Selection of the "CASE" and "CONTROL" groups (D0 : first admission at Rennes CHU. DHIV+ : date of the event of interest. DDS : date of last admission. M12 : data for the last 12 months. M24 : data for the last 24 months. DA : all available data)

The event of interest was the diagnosis of HIV infection during the study period. The case group included all patients with incident HIV infection and with at least one admission (for any reason) at Rennes CHU at least 7 days before the date of HIV diagnosis (DHIV+). The control group included patients without diagnosis of HIV infection and with at least one admission before the date of the last visit within the study period (DDS) and whose EHR included at least one HIV risk factor. Patients with at least one consultation for PrEP were excluded. All available data (DA) before the event of interest in each group (DHIV+ or DDS) were used for the analysis (Figure 1).

From the demographic, clinical, laboratory and treatment data in eHOP, 162 constructed variables were identified as potential predictors of HIV risk by clinicians from the Rennes CHU infectious diseases department. These variables were constructed from known factors of HIV vulnerability, previous screening and recommendations (history of Sexually Transmitted Infections (STIs), sexual risk-taking behavior), lifestyle

habits that can be closely linked to taking sexual risks (drug, alcohol use), healthcare consumption that can reflect the quality of care (practices, number of visits), and admission to medical departments that are linked to sexual health (dermatology, proctology).

Three periods were identified for each variable (Figure 1): DA (79 variables); last 12 months before the event of interest (M12) (41 variables); M24 (42 variables). Some variables were constructed from structured data alone (diagnosis from Diagnosis Related Group (DRG) data, laboratory and EHR forms), from textual data only (using rule-based Natural Language Processing (NLP) extraction methods), and from a combination of these data types. Each patient was phenotyped by automatically extracting his/her characteristics. It was assumed that if information on a feature was not present in the CDW, that characteristic was absent (paradigm of the closed world).

This project was carried out following the rules and regulations for re-using data from Rennes CHU (patient information, respect for the right to object).

For the identification of patients at risk of HIV, four different simple models (Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Support Vector Machine (SVM), and Gradient boosting) and a combined model (LASSO, RF and General Linear Model (GLM)) were tested. For each model, a workflow was performed that included a data standardization step (V1), and a normalization step associated with a class-rebalancing step (V2). Cases and controls were matched based on sex, age and number of admissions present in eHOP, with a proportion of 1 case for 5 controls. Each model was cross-validated using the training sample (126 cases; 584 controls among all included patients) to select the best set of hyper-parameters for each model. The Receiver Operating Characteristic (ROC) curve and the test sample (30 cases; 177 controls) was used to compare the performance of the different models for discriminating between cases and controls. For each model, sensitivity and specificity were computed at the threshold that gave the minimum Euclidean distance between the (specificity, sensitivity) coordinates and the optimal point (1,1). The R software (v 3.5.2) and the 'caret' package were used to develop and evaluate the machine learning-based models.

3. Results

Among the 624,708 patients with at least one admission during the study period, 422 patients had incident HIV (0.07%) among whom 156 patients were selected for the study because of the availability in eHOP of data before the HIV diagnosis date. The case group included 66.1% of men (9.7% of MSM) and 33.9% of women. Among the 621,370 patients without known HIV diagnosis and without any PrEP prescription, 247,494 (39.8%) patients had at least one variable suggesting HIV risk. Among them, 761 controls were selected, including 1.8% of MSM in men population.

Among the 79 referenced variables in the DA period (Figure 1), 3 variables were generated from textual data alone, 51 variables were extracted from structured data alone, and 22 of the other 25 variables combined textual and structured data. More than 50% of patients were identified thanks to the textual search; and more than 75% of patients had 19 variables. Table 1 describes the performance of the different models.

Table 1. AUC, sensitivity, and specificity of the tested models

		LASSO		SVM		RF		Boosting		LASSO/ RF/GLM
		V1	V2	V1	V2	V1	V2	V1	V2	V2
TRAINING	cv-AUC	0.939	0.947	0.985	0.988	0.999	0.992	0.997	0.992	0.991
	95% CI	0.912-0.965	0.925-0.969	0.976-0.993	0.981-0.995	0.999-1	0.988-0.996	0.994-0.999	0.988-0.996	0.986-0.995
	Optimal cut-off	0.186	0.395	0.092	0.485	0.292	0.313	0.200	0.237	0.454
	Sensitivity	0.913	0.889	0.976	0.952	1.000	0.976	0.984	0.976	0.944
	Specificity	0.861	0.918	0.935	0.957	0.991	0.954	0.966	0.945	0.949
TEST	AUC	0.751	0.768	0.837	0.815	0.877	0.866	0.854	0.880	0.888
	95% CI	0.620-0.882	0.638-0.897	0.746-0.928	0.715-0.915	0.801-0.953	0.787-0.945	0.776-0.932	0.805-0.955	0.814-0.962
	Sensitivity	**0.667**	**0.700**	**0.867**	**0.500**	**0.667**	**0.667**	**0.600**	**0.633**	**0.733**
	Specificity	**0.831**	**0.893**	**0.678**	**0.960**	**0.921**	**0.927**	**0.881**	**0.876**	**0.887**

The LASSO V1 model retained 40 predictive variables. The strongest predictor variables for incident HIV cases were sexual orientation (MSM) (OR 1.285), known history of syphilis (OR 1.206), and history of schizophrenia (OR 1.176). Conversely, other variables better identified controls, such as hepatitis C virus testing in the last 12 months (OR 0.460), previous STI testing (OR 0.560), and previous sexual disorders (OR 0.852).

4. Discussion and Conclusion

The combined model (LASSO, RF, and GLM) showed the best results in terms of discrimination between HIV cases and controls (AUC = 0.88 with 95% CI= 0.8143-0.9619) using the test sample. This model was trained on data extracted from an academic hospital database. This combined model identified 73% of patients with HIV infection, which is the second highest sensitivity among the models tested in this study. Very few studies, mostly from the USA, have tried to predict HIV risk using real-life data. Our results are very similar to those reported in these previous studies [6, 7]. A secondary outcome of our work is the collection of 162 phenotyping variables that can be reused in other studies. It is worth noting that more than 50% of patients could be characterized by variables constructed from clinical textual data and using NLP extraction methods. Most of the social and behavioral determinants concerning STIs, which in the field of sexual health are considered essential to identify populations at risk, came mainly from narrative data included in the patient EHR [9]. An USA study showed that combining textual and structured data improves the predictive performance of HIV risk prediction models [10].

This study has some limitations. As this was a monocentric study, results cannot be generalized to the whole hospital population. However, in future works we plan to extend our approach to other hospital CDWs to better capture the patient heterogeneity, thus increasing the model performance. Regarding the data sources, only hospital data (generated during patient stays and visits) were used. Primary care data also would probably be useful to generalize our model. Like most studies that reused real world clinical data, data quality was an issue, especially missing information, for instance concerning the normality of clinical findings, or lack of socio-economical and behavior

information that would have been useful to better define the case group. In conclusion, the combined model (LASSO, RF and GLM) is efficient and could be used in the clinical practice to help practitioners to offer PrEP to the right people and at the right time. This raises the question of the ethical aspect of this approach to detect people at risk. In this context, it could be interesting to develop a risk score that would allow practitioners to identify people at high risk when they come to the hospital and propose to these people, if they want it, the intervention of a prevention counselor in sexual health (present in free centers for information, screening and diagnosis of STIs, attached to hospitals)[11]. We plan to improve our model using multicentric data coming from the western Clinical Data networks regrouping 8 hospitals combined with longitudinal data from the French National Health Data System [12]. Our final objective is to implement this model in a decision support system that will be evaluated in terms of performance and usability in a randomized clinical trial.

References

[1] Sante Publique France. Bulletin de sante publique VIH/sida (Octobre 2019). https://www.santepubliquefrance.fr/maladies-et-traumatismes/infections-sexuellement-transmissibles/vih-sida/documents/bulletin-national/bulletin-de-sante-publique-vih-sida.-octobre-2019 (accessed June 6, 2020).

[2] Molina JM, et al. On-Demand Preexposure Prophylaxis in Men at High Risk for HIV-1 Infection. N Engl J Med. 2015;373(23):2237-46.

[3] Billioti de Gage S, Le Tri T, Dray-Spira R (Groupement d'Intérêt Scientifique EPI-PHARE). Suivi de l'utilisation de Truvada® ou génériques pour une prophylaxie pré-exposition (PrEP) au VIH à partir des données du Système National des Données de Santé (SNDS). Nov 2019.

[4] Velter A, Saboni L, Bouyssou A, Semaille C. Comportements sexuels entre hommes à l'ère de la prévention combinée-Résultats d'Enquête presse gays/lesbiennes 2011.BEH. 2013 Nov 26:39-40.

[5] McCormack SM, Noseda V, Molina JM. PrEP in Europe - expectations, opportunities and barriers. J Int AIDS Soc. 2016:19(7(Suppl 6)): 21103.

[6] Marcus JL, et al. Use of electronic health record data and machine learning to identify candidates for HIV preexposure prophylaxis: a modelling study. Lancet HIV. 2019;6(10): e688-95.

[7] Krakower DS, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. Lancet HIV. 2019;6(10): e696-704.

[8] Madec J, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019;264: 1536-7.

[9] SO. Aral. Determinants of STD epidemics: implications for phase appropriate intervention strategies. Sex Transm Infect 78 Suppl 1 (2002), i3-13.

[10] Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. J Acquir Immune Defic Syndr. 2018;77(2):160-6.

[11] Ridgway JP, Et al. Which Patients in the Emergency Department Should Receive Preexposure Prophylaxis? Implementation of a Predictive Analytics Approach. AIDS Patient Care STDS. 2018;32(5);202-7.

[12] Tuppin P, et al. Value of a national administrative database to guide public decisions: From the systeme national d'information interregimes de l'Assurance Maladie (SNIIRAM) to the systeme national des donnees de santé (SNDS) in France. Rev Epidemiol Sante Publique. 2017;65(4):S149-S167.