A Machine Learning–Based Risk Stratification Score for Anal Cancer Screening in People With HIV

LB #642

Jordi Aceiton[1†], **Raquel Martín-Iguacel**[1,2,3†], Georgia Escaramis[1], Robert Zangerle[4], Giota Touloumi[5], Linda Wittkop[6], Ferdinand Wit[7], Inma Jarrin[8], Sophie Grabar[9], Christoph Stephan[10], Antonella d'Arminio Monforte[11], Michael J Silverberg[12], Jessica L. Castilho[13], Martina Cusinato[14], Josep M Llibre[1]

[1] Centre of Epidemiological Studies of HIV/AIDS and STI of Catalonia (CEEISCAT), Health Department, Generalitat de Catalunya, Badalona, Spain.; [2] Department of Infectious Diseases, Odense University Hospital, Odense, Denmark; [3] University of Southern Denmark; [4] Medizinische Universität Innsbruck, Innsbruch, Austria; [5] Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens; [6] Univ. Bordeaux, INSERM, Institut Bergonié, BPH, U1219, CIC-EC 1401, Bordeaux, France. INRIA SISTM team, Talence, France. CHU de Bordeaux, Service d'information médicale, INSERM, Institut Bergonié, CIC-EC 1401, Bordeaux, France; [7] Stichting HIV Monitoring Foundation, Amsterdam, The Netherlands. Department of Infectious Diseases, Amsterdam Institute for Immunology and Infectious Disease, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands; [8] Centro Nacional de Epidemiología, CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain; [9] Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital St Antoine, Paris, France; [10] Department of Internal Medicine 2, Infectious Diseases Unit, University Hospital Frankfurt, Goethe University Frankfurt, Frankfurt, Germany; [11] ICONA Foundation, Milan, Italy; [12] Division of Research, Kaiser Permanente Northern California, Oakland, California, USA; [13] Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [14] Population Health Sciences, University of Bristol, Bristol, UK.

† These authors contributed equally to this work.

BACKGROUND

People with HIV (PWH) have a higher risk of anal cancer than the general population. Current international guidelines recommend screening PWH aged ≥35 years who are men who have sex with men (MSM) or transgender women, and ≥45 years otherwise; however, access to screening is often limited. Improved risk stratification is needed to better target prevention and screening efforts. We aimed to develop and validate a prognostic model to identify PWH at highest risk of anal cancer.

METHODS

We used data from the Antiretroviral Therapy Cohort Collaboration (ART-CC) to develop a binary model to predict incident anal cancer in PWH in Europe and North America. Candidate predictors routinely available at clinic visits were selected a priori: age, sex at birth, HIV transmission group, years in HIV care, nadir CD4 count, prior AIDS-defining events (excluding cancer), history of other cancers and smoking status. Data were harmonized across cohorts. We evaluated several machine-learning algorithms, including penalized logistic regression, support vector machines (SVM) , random forests, and gradient boosting (XGBoost). Models were trained using stratified 10-fold cross-validation with nested hyperparameter tuning, optimizing the F1-score (combining sensitivity and positive predictive value). Class imbalance was handled using combined under-sampling and SMOTENC oversampling. Model selection prioritized high sensitivity while minimizing the proportion classified as screening eligible.

Table 1: Baseline characteristics

	No Anal cancer N = 157725	Anal cancer N = 1120
Sex:		
Male	117492 (74.5%)	978 (87.3%)
Female	39565 (25.1%)	138 (12.3%)
Unknown	668 (0.42%)	4 (0.36%)
Risk group:		
MSM	61611 (39.1%)	751 (67.1%)
Heterosexual men	30286 (19.2%)	136 (12.1%)
Heterosexual women	39565 (25.1%)	138 (12.3%)
Other	26263 (16.7%)	95 (8.48%)
Age, years (IQR)	49.0 [40.0;58.0]	51.0 [44.0;57.0]
Years in HIV (IQR)	12.0 [6.00;20.0]	16.0 [9.00;22.0]
Prior CD4 Nadir, (IQR)	207 [86.0;342]	100 [25.0;219]
Prior AIDS event (no cancer):		
Yes	44626 (28.3%)	524 (46.8%)
No	113099 (71.7%)	595 (53.1%)
Unknown	0 (0.00%)	1 (0.09%)
Prior cancer event:		
AIDS cancer	9775 (6.20%)	155 (13.8%)
Non AIDS cancer	7702 (4.88%)	49 (4.38%)
No cancer	140248 (88.9%)	916 (81.8%)
Smoking status:		
Smoker	30397 (19.3%)	275 (24.6%)
Former smoker	65604 (41.6%)	402 (35.9%)
Never smoked	16411 (10.4%)	72 (6.43%)
Unknown	45313 (28.7%)	371 (33.1%)

Risk-based Machine-Learning screening improves anal cancer screening sensitivity while reducing unnecessary screening compared with current guidelines

RESULTS

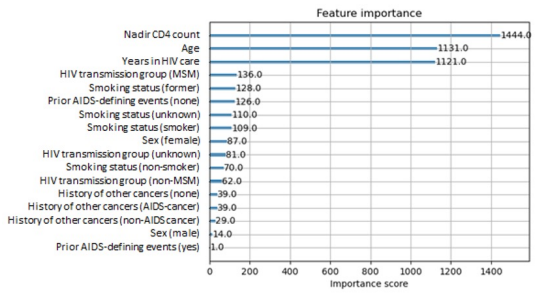
The analysis included 157,725 PWH, of whom 1,120 developed incident anal cancer (Table 1).

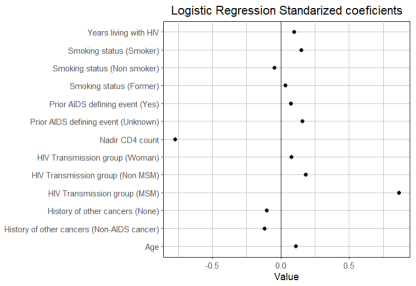
Across all approaches, prediction models outperformed clinical guideline–based screening. Clinical guidelines achieved a sensitivity of 88% while classifying 72% of individuals as screening-eligible. Among prediction models, XGBoost showed the best overall performance. At a predicted-risk probability threshold of 0.5, XGBoost achieved a sensitivity of 94.2% while classifying 68.8% of individuals as screening-eligible, reducing the screening burden ~3% compared with clinical guidelines while improving sensitivity (Table 2).

Optimized thresholds for logistic regression and SVM also increased sensitivity to 93.8% and 94.2%, respectively, but at the cost of screening a higher proportion of individuals (72.4% and 73.5%).

Feature importance analyses identified nadir CD4 count, age, and years in HIV care as the strongest predictors, followed by MSM status and former smoking (Fig. 1), highlighting the role of cumulative immunosuppression and long-term HIV exposure in anal cancer risk.

Figure 1: Feature importance of predictors in the XGBoost model for anal cancer risk in people with HIV
Feature importance reflects each variable's contribution to the model's classification performance.





Author Contact Information: Raquel Martin Iguacel: raquel@bisaurin.org

CONCLUSIONS

- A machine-learning-based risk stratification score using routinely collected, easily accessible clinical data accurately identifies PWH at highest risk of anal cancer.
- This approach can reduce the number of individuals requiring screening while maintaining high sensitivity, thereby improving feasibility and resource allocation in settings with limited screening capacity.
- Although the absolute differences may appear modest, the combined gains in sensitivity and reduction in screening burden translates into meaningful clinical and public health benefits at the population level.

Table 2. Screening performance of clinical guidelines and prediction models across probability thresholds.

Approach	Model	Probability threshold	Sensitivity (%)	Screening eligibility (%)
Clinical guidelines		--	88%	72%
Prediction model	Logistic regression	Conventional (0.5)	70.5%	31.9%
		Optimised (0.25)	93.8%	72.4%
	SVM	Conventional (0.5)	76.3%	37.1%
		Optimised (0.27)	94.2%	73.5%
	XGBoost	**Conventional (0.5)**	**94.2%**	**68.8%**
		Optimised (0.5)	**94.2%**	**68.8%**

Acknowledgements: We thank the patients who participated in this cohort and our colleagues in the clinical departments for their continued contribution and dedication.

PLAIN LANGUAGE SUMMARY

Using routine clinical data, a machine learning–based tool identified people with HIV at higher risk of anal cancer more accurately than current screening guidelines. This approach could help target screening to those most likely to benefit, while reducing unnecessary screening in lower-risk individuals.

ART COHORT COLLABORATION

PISCIS POPULATIONAL HIV COHORT FROM CATALONIA AND BALEARIC ISLANDS

Region Syddanmark OUH Odense Universitetshospital Svendborg Sygehus SDU