

BACKGROUND

The first of the UN 95-95-95 targets for 2025 is to achieve that 95% of people living with HIV know their serological status.

An estimated 7.5% of people with HIV (PWH) in Spain are unaware of their HIV status, potentially contributing to 54-65% of ongoing transmissions.

In Spain (2023), 48.7% of new HIV diagnoses were classified as late.

Late diagnosis is associated with advanced disease progression, increased health costs, and increased morbidity and mortality.

OBJECTIVE

Our aim was to develop an easily implementable tool based on electronic health records (EHR) through HIV indicator conditions, sex and age that could help identify people to test and bring to light undiagnosed HIV and reduce late diagnosis.

METHODS

A retrospective analysis was conducted using Telotrón®, an EMA-registered database containing anonymized data from 2.2 million individuals from the Spanish national healthcare system, covering the period from 2012 to October 2023¹. To ensure representativeness of the database for HIV modeling, PWH alive (2023) in Telotrón® were compared to the Spanish national hospital survey on HIV published by the Ministry of Health².

Two cohorts were randomly selected (Figure 1). The first cohort was divided in 70% for model training and 30% for internal validation. The second cohort was allocated for external validation. Index date was defined as last record date in database or first HIV diagnosis date. Age at index date and birth sex were recorded.

59 HIV indicator conditions were selected based on ECDC and Spanish Ministry of Health guidelines, supplemented by expert opinions. The variables considered for model inclusion were

specific age group (20-59 years old), sex and the presence of the HIV indicator conditions in the five years prior to index date as predictor variables. Due to low HIV prevalence, an elastic net-regularized logistic regression model with SMOTE was applied (Figure 2).

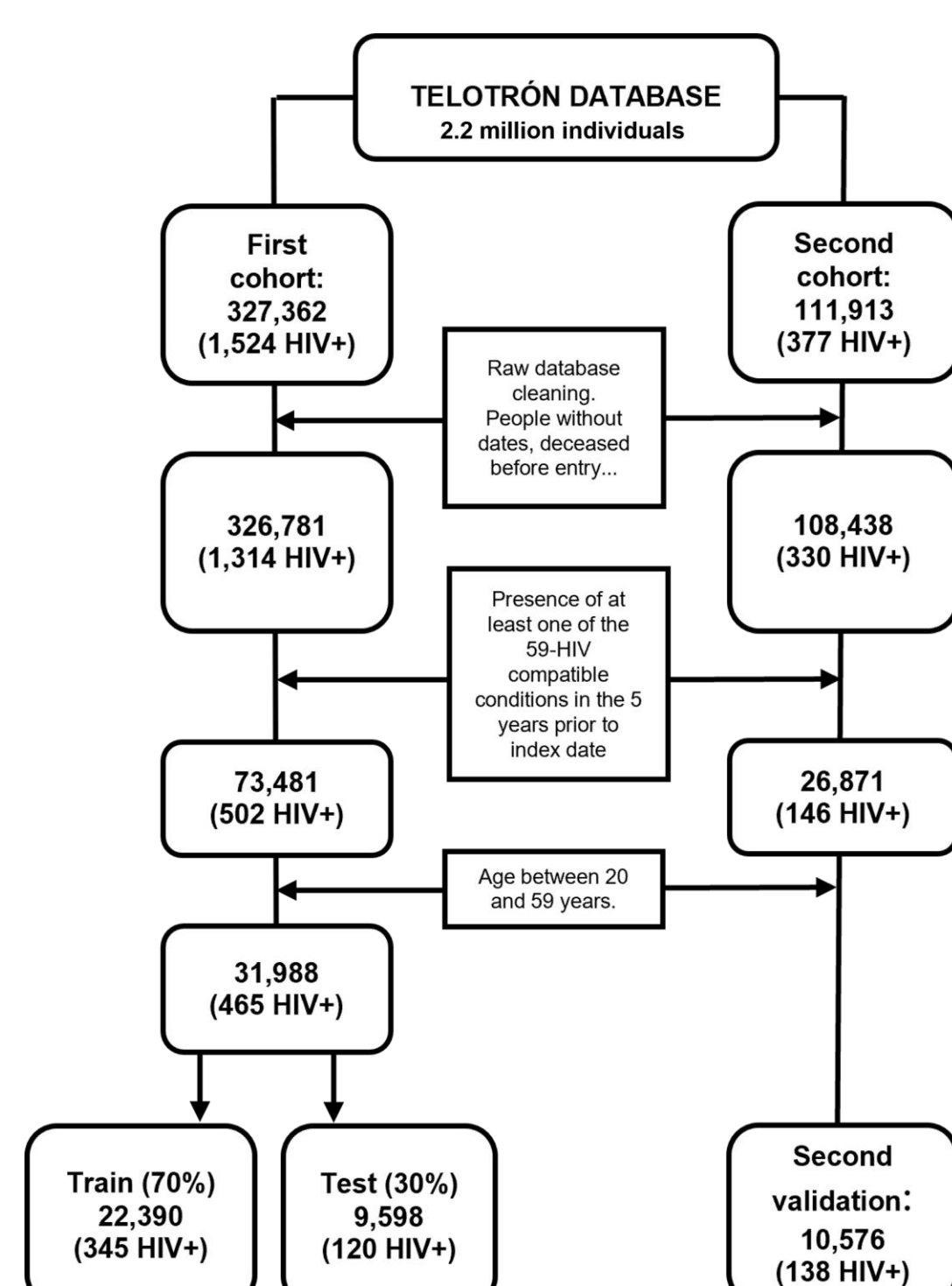


Figure 1. Population Flowchart

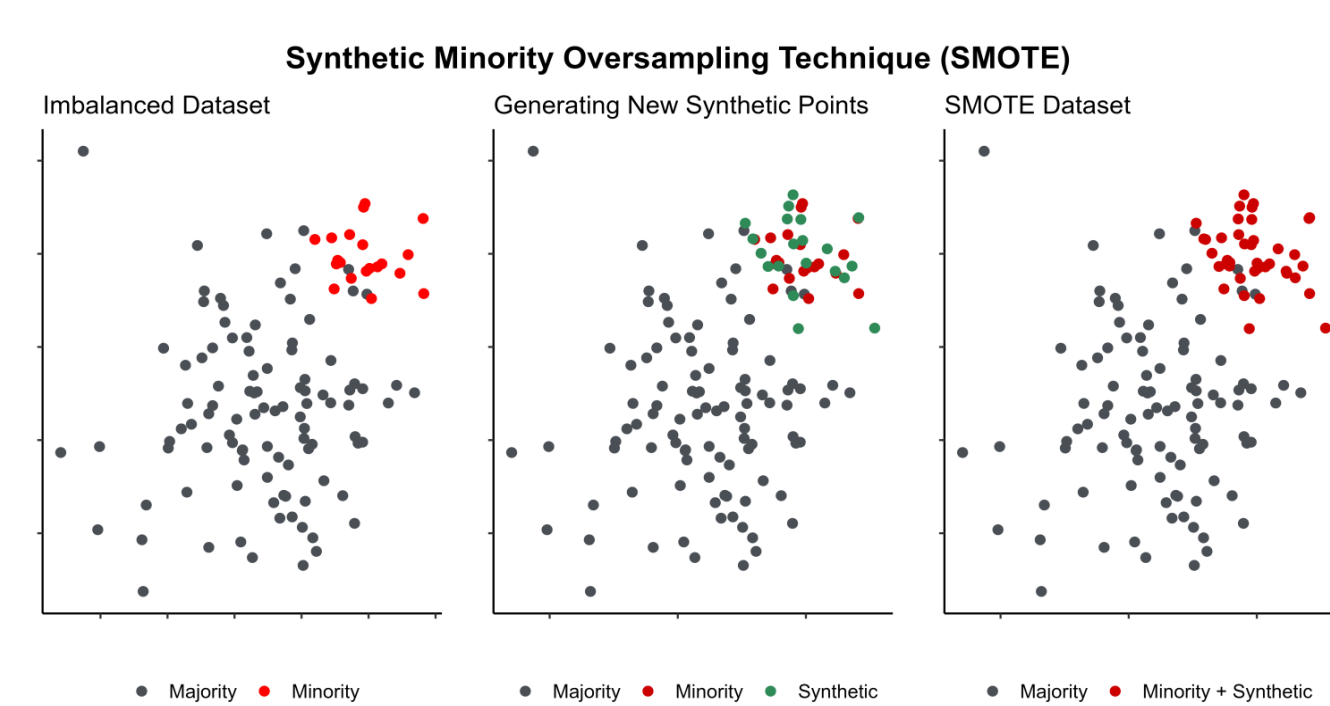


Figure 2. SMOTE Technique

The model identified over 70% of HIV cases, screening less than 4% of a health department size population and achieved a prevalence rate above 2.6%.

RESULTS

The predictive model identified 10 clinical conditions with the highest predictive risk value, 6 different age groups and sex (Table 1).

The model achieved strong performance metrics at training (sensitivity **82.96%**, specificity **64.54%**).

Validation demonstrated consistent performance, with AUC values ranging from 0.77 to 0.78 across the test and external validation cohorts, confirming the model's robustness. Sensitivity was **77.50%** in the internal validation and **71.74%** in the external validation, while specificity ranged from **64.05%** to **65.51%**.

Table 1. Model Results for HIV-Predictors

Predictor	Adjusted Odds Ratio (aOR) 95% IC	Predictor	Adjusted Odds Ratio (aOR) 95% IC
Hepatitis C	20.02 (17.63–22.83)	Sex	
Hepatitis A	16.25 (8.91–32.26)	Male	4.59 (4.37 – 4.82)
Gonococcal infection	8.02 (7.01–9.21)	Age groups	
Syphilis	6.03 (5.21–7.02)	20-24	Ref
Infectious proctitis	4.82 (3.25–7.32)	25-29	1.56 (1.33 – 1.82)
Urethritis	2.26 (2.07–2.48)	30-34	2.57 (2.23 – 2.96)
*Risk of exposure to HIV/STI	1.61 (1.31–1.98)	35-39	7.54 (6.60 – 8.62)
Drug intoxication	1.58 (1.45–1.73)	40-49	9.13 (8.06 – 10.38)
Candidiasis	1.27 (1.15–1.40)	50-59	5.26 (4.63 – 6.00)
Condyloma acuminatum	1.24 (1.11–1.39)		

*ICD codes related to contact or risk of STI/HIV

To aid testing efficiency, prediction results were stratified into four equal cumulative groups (Table 2) with a risk index of having HIV: ≥ 50 (cluster 1), ≥ 62.5 (cluster 2), ≥ 75 (cluster 3) and ≥ 87.5 (cluster 4). At risk index clustering 4, testing 216 individuals yielded an HIV+ prevalence of 11.11%, requiring only 15 tests per positive case. Across both validation groups, HIV prevalence exceeded 2.6%, and the number needed to test (NNT) to identify one positive case ranged between 12 and 48, depending on the probability risk index clustering. This approach highlights the model's utility in optimizing resource allocation.

Table 2. Stratified HIV-Detection Analysis of Model Performance at Different Risk Clusters Thresholds

Validation cohorts	Risk index clustering	People to be tested	People within cohort to be tested (%)	HIV Prevalence among tested (%)	95% CI	NNT
Model Validation Primary cohort (98,209)	Cluster 1	3,500	3.6	2.66	2.12 – 3.19	48
	Cluster 2	1,877	1.9	3.57	2.73 – 4.41	37
	Cluster 3	431	0.4	8.12	5.54 – 10.70	18
	Cluster 4	216	0.2	11.11	6.92 – 15.30	15
External validation Secondary cohort (111,913)	Cluster 1	3,699	3.3	2.68	2.16 – 3.20	47
	Cluster 2	1,990	1.8	4.02	3.16 – 4.88	32
	Cluster 3	467	0.4	9.42	6.77 – 12.07	15
	Cluster 4	235	0.2	13.19	8.86 – 17.52	12

Abbreviations: CI: Confidence Interval; NNT: number needed to test to identify a positive result

CONCLUSIONS

The model demonstrated the ability to identify more than 70% of HIV cases by screening less than 4% of a health department-sized population, achieving a prevalence rate above 2.6%.

By focusing on people with high probability of having HIV, this approach allows for better prioritization of testing, ensuring efficient allocation of resources without neglecting any group. Its validation on multiple datasets confirmed its robustness and scalability. Designed for integration into EHR systems, the model enables automation through associated conditions-alerts and/or proactive testing, reducing missed opportunities for diagnosis and subsequent HIV transmission. This transformative tool aligns with global initiatives, such as the UNAIDS 95-95-95 target and the 2030 goal, offering a promising and scalable solution to reduce late diagnosis and contribute to ending AIDS as a public health threat.

REFERENCES & ACRONYMS

- Alamillo ML, Díaz Y, Enriquez JL, León M. External Validity and Representativeness of TELOTRON® Database: A Reliable Source for Real-World Evidence Research in Spain. *Value in Health* 2024; 27.
- Centro Nacional de Epidemiología - Instituto de Salud Carlos III / División de control de VIH IH virales y T. Encuesta Hospitalaria de pacientes con infección por el VIH: Resultados 2023. Madrid, 2023.

Acronyms: ECDC: European Centre for Disease Prevention and Control; SMOTE: elastic net regularization alongside the synthetic minority oversampling technique

*Predictive model owned by Gilead Sciences Spain. Model named "PREDICE". (Developed by Telómera)

Author contact: Arkaitz Imaz Vacas, aimaz@bellvitgehospital.cat

