Algorithmic prediction of HIV status using nation-wide electronic registry data

Ahlström, Magnus G; Ronit, Andreas; Omland, Lars Haukali; Vedel, Søren; Obel, Niels

Published in:
EClinicalMedicine

DOI:
10.1016/j.eclinm.2019.10.016

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
https://doi.org/10.1016/j.eclinm.2019.10.016
Algorithmic prediction of HIV status using nation-wide electronic registry data

Magnus G. Ahlström, Andreas Ronit, Lars Haukali Omland, Søren Vedel, Niels Obel

Department of Infectious Diseases 8632, Rigshospitalet, Copenhagen University Hospital, Blegdamsvej 9B, Copenhagen DK-2100, Denmark

A R T I C L E   I N F O

Article history:
Received 25 April 2019
Revised 17 October 2019
Accepted 22 October 2019
Available online 5 November 2019

Keywords:
HIV prevention
HIV diagnosis
Machine learning
Registry data
Big data

A B S T R A C T

Background: Late HIV diagnosis is detrimental both to the individual and to society. Strategies to improve early diagnosis of HIV must be a key health care priority. We examined whether nation-wide electronic registry data could be used to predict HIV status using machine learning algorithms. Methods: We extracted individual level data from Danish registries and used algorithms to predict HIV status. We used various algorithms to train prediction models and validated these models. We calibrated the models to mimic different clinical scenarios and created confusion matrices based on the calibrated models. Findings: A total 4,384,178 individuals, including 4,350 with incident HIV, were included in the analyses. The full model that included all variables that included demographic variables and information on past medical history had the highest area under the receiver operating characteristics curves of 88.4% (95%CI: 87.5% – 89.4%) in the validation dataset. Performance measures did not differ substantially with regards to which machine learning algorithm was used. When we calibrated the models to a specificity of 99.9% (pre-exposure prophylaxis (PrEP) scenario), we found a positive predictive value (PPV) of 8.3% in the full model. When we calibrated the models to a sensitivity of 90% (screening scenario), 384 individuals would have to be tested to find one undiagnosed person with HIV. Interpretation: Machine learning algorithms can learn from electronic registry data and help to predict HIV status with a fairly high level of accuracy. Integration of prediction models into clinical software systems may complement existing strategies such as indicator condition-guided HIV testing and prove useful for identifying individuals suitable for PrEP. Funding: The study was supported by funds from the Preben and Anne Simonsens Foundation, the Novo Nordisk Foundation, Rigshospitalet, Copenhagen University, the Danish AIDS Foundation, the Augustinus Foundation and the Danish Health Foundation.

© 2019 Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Introduction

Over half of individuals diagnosed with HIV in the European Region are diagnosed in a late stage of infection (i.e. with a CD4 count below 350 cells/μL at time of diagnosis) [1]. Early diagnosis is important for several reasons; it improves life expectancy among people living with HIV (PLWH) [2], reduces risk of onward HIV transmission [3], and lowers associated health care costs [4]. Moreover, there is now a large body of evidence from randomized controlled trials that early initiation of antiretroviral therapy (ART) lowers risk of acquired immunodeficiency syndrome (AIDS) and non-AIDS events [5,6].

One way to diagnose PLWH at an early stage may be through universal screening in which all individuals, that encounter the health care system, are offered testing. In the Western European Region, where the prevalence of HIV is low, such an approach may be associated with a poor cost-benefit ratio [7]. Thus, alternative strategies have been proposed, such as indicator-condition guided testing. Such guidance, which in part builds on studies on HIV prevalence within specific medical conditions [8], recommends HIV testing in any condition, which may indicate the presence of HIV [9]. Although this approach offers several advantages, it does not consider all the prior health data of a given individual. Moreover, it does not consider age, sex or any temporal patterns of prior medical diseases, and it does not include those medical conditions that would be associated with a lower risk of HIV transmission. Including this kind of information into decision algorithms of HIV pre-
diction may not only improve discriminative abilities but may po-
tentially also help identify individuals at increased risk of acquiring
HIV, including individuals who might be pre-exposure prophylaxis
(PrEP) candidates.

Machine learning algorithms, a set of mathematical or algorithm-
tic tools which extract generalizable patterns from large data sets
in order to make predictions about the outcome in new, unseen
cases, are rapidly growing research areas which have also found
their way into HIV research. It has so far primarily been used
to predict progression of HIV and ART resistance and for ART opti-
mization [10–12]. In the present study, we applied machine learn-
ing algorithms that learn from nation-wide historic electronic reg-
istry data to make prediction of HIV status and evaluated the po-
tential of such prediction models as a screening tool and to iden-
tify suitable candidates for PrEP. Focusing on assessing the general
applicability of these types of methods and thus establishing a ref-
ereence baseline performance, we focused our investigation on sim-
pler general-purpose modeling frameworks over detailed models of
higher performance but reduced ability to generalize.

Methods

Setting

As of December 31, 2016, Denmark had an adult population of
approximately 4.6 million people, 6200 (~0.1%) people diagnosed
with HIV, and an estimated 600 (~0.01%) individuals with undi-
agnosed HIV infection [13]. PLWH are treated in eight specialized
HIV-care centers and are typically seen as outpatients at intended
intervals of 12–24 weeks. Combination antiretroviral treatment is
provided free of charge at the HIV-centers.

Data sources

The unique 10-digit personal identification number assigned to
all Danish residents at birth or immigration was used to avoid
multiple registrations and to track individuals in the following reg-
istraries.

The Danish National Hospital Registry (DNHR) was established in
1977 and stores information on all inpatient and outpatient admis-
sion to hospitals in Denmark. Diagnoses are classified according to
the ICD 8th revision (ICD8) until December 31, 1993, and 10th re-
vision (ICD10) thereafter. The registry has been expanded gradu-
ally so that it now also contains outpatient data, data from emergency
wards, and data from psychiatric admissions dating back to 1995,
which led us to use data on admission dating back to Jan 1, 1995.

The Danish Civil Registration System was established in 1968 and
stores information on all Danish residents. From this registry we
deported data on date of birth, sex, residency, marital status and
migration. The Employment Classification Module provides information on
the occupation and employment status of the population charge-
able with tax. From this registry we extracted data on main source of income

Educational Classification Module includes data on successfully
completed educational attainments collected directly from all Dan-
ish educational institutions for all Danish residents. Ninety-seven
percent of the Danish-born population and 85–90% of the immi-
grant population has non-missing data. Educational attainment is
classified according to a Danish modified version of the Interna-
tional Standard Classification of Education 2011 [14].

Study population

We included all Danish individuals that: i) had at least one hos-
pital visit between January 1, 1995, until December 31, 2016, ii) had
a Danish personal identification number, and iii) were aged ≥ 16
years at the hospital visit. We excluded individuals with i) an HIV
diagnoses registered before 16 years of age and ii) an HIV diagnosis
before Jan 1, 1995.

Statistics

We defined an index visit which was the last registered hospital
visit or the first registered hospital visit with an HIV-diagnosis. For
the index visit we computed 167 variables associated with med-
ical history. These variables were a slightly modified version of
categories used previously [15]. The categories and the associated
ICD8/ICD10 codes are provided as online supplementary material.
Some ICD codes were omitted from the chosen disease categories
used for the models. All codes have been highlighted in online
supplementary material Table S1. For each index visit we identi-
fied all hospital visits up until three months before the index visit.
Based on these visits we calculated a score for each disease cate-
ogy that was 10 minus the number of years between each hospi-
tal visit and the index visit, e.g. if an individual was hospitalized
five years prior to the index visit with a diagnosis of pneumococ-
cal pneumonia (PP) the variable PP was assigned a value of 10 −
5 = 5, if the time was equal to or exceeded 10 years the variable
was assigned the value 0. We also included a variable that indi-
cated how many hospital visits the individual had in the last year
prior to the index visit. Additional variables included were age
(>16, 16–25, 25–35, 35–45, 45–55 and 55+), sex (male and fe-
male), place of birth (Denmark, Scandinavia, other and unknown),
highest educational attainment (primary school, high school, voca-
tional internships and main course, short higher education, middle
length higher education, bachelor and long length higher educa-
tion, PhD programs, unknown), marital status (married, divorced,
widowhood, registered partnership, cancelled registered partnership,
longest living of two partners and unknown), main source of income
(self-employed, employed spouse, wage earner with own
business, wage earner without business, wage earner with support,
pre senior citizen with own business, senior citizen, other and
unknown) and place of residency (capital region, large city region,
hinterland region, provincial region, rural region and unknown).
For more details on the variables please refer to Table 1 and Ta-
ble S1.

The response variable in all models was an HIV diagnosis dur-
ing the index visit. HIV was defined as one a of the following di-
agnoses: ICD8: 07983 or ICD10: B20-B24.9. The validity of an HIV
diagnosis in LPR has previously been shown to be very high [16].

We used logistic regression with an elastic net penalty function
(GLMnet) as regularizer to fit models with increasing variable com-
plexity. A regularizer is a tool which automatically penalizes com-
plex models over simpler ones during model fitting, and the partic-
ular type used here (elastic net) allows for the automatic exclusion
of variables with little impact on model performance (these vari-
bles are marked with a “−” in Table S1); thus, the role of the reg-
ularizer is to find the model that best fits the data with the mini-
mal number of variables. The elastic net is a regularized regression
method that combines lasso and ridge regression penalties, the hy-
perparameter α controls which mix between the two penalties is
used, i.e. if α = 0 pure ridge penalty is used and if α = 1 pure lasso
penalty is used [17]. We used 10-fold cross validation to determine
the value of λ in the penalty function and used 10-fold cross vali-
idation to determine the mix between the penalty functions (the
α-value). When a model was fit, we used the model to calculate a
risk score of HIV for each individual included. The risk scores were
used to create receiver operating characteristics (ROC) curve for all
the models. Furthermore, we calculated area under receiver operat-
ing characteristics curve (AUROC) in order to compare the models.
The ROC curves were used to calibrate the models to certain val-
ues of sensitivity and specificity by calculating their corresponding risk scores. We additionally calculated sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

To evaluate how much information was gained by including an increasing number of variables, we fitted three different models with the elastic net penalty function. Each model was fitted independently, i.e. regularization was done separately for each model and parameters in the model calculated independently. The variables introduced in the three different models were: i) age, sex and sexually transmitted infections (STIs, see Table 1 for a complete list), ii) age, sex, place of birth, educational attainment, marital status, place of residency and main source of income and iii) all included in ii and medical history, this was the final model used. The total number of index visits was randomly divided into a training set (70%) and a validation set (30%). The validation set was used to make out-of-sample validation on the prediction algorithm.

To evaluate whether different machine learning algorithms performed better with regards to discriminative ability, we also fitted simple logistic regression, random forest with random undersampling (1 HIV-positive to 10 HIV-negative individuals) to balance the dataset, logistic regression with lasso regularizer ($\text{GLM}_{\text{lasso}}$) and logistic regression with ridge regularizer ($\text{GLM}_{\text{ridge}}$), and logistic regression with elastic net regularizer with synthetic minority oversampling technique (SMOTE) to balance dataset, on all variables and compared these using AUROC. We tested differences in performance of the models on the models that were calibrated to a sensitivity of 90% with McNemars test.

We created calibrated confusion matrices of the final model to mimic three clinical scenarios. The first scenario is when a high certainty of an HIV diagnosis is required (model calibrated to a specificity of 99.9%), e.g. when evaluating potential candidates of PrEP, the specific score of 99.9% was chosen to be high but other than that is arbitrary. The second scenario is the optimal trade-off between sensitivity and specificity (model optimized according to the highest value of the Youden's index (sensitivity + specificity - 1) [18]. The third scenario is when a large coverage of the diagnosis is required (model calibrated to a sensitivity of 90%), e.g. a screening setting, the 90% sensitivity was chosen to reflect the first 90% in the WHO '90–90–90'. The calibration was done by calculating the risk score that yields the desired performance characteristic (e.g. sensitivity of 90%) in the training data set, and then applying this risk score to the validation data set and then calcu-

### Table 1
Clinical characteristics of the training and validation cohort.

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Validation cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of individuals</td>
<td>2,972,264</td>
<td>1,411,914</td>
</tr>
<tr>
<td>HIV+, n (%)</td>
<td>3,063 (0.1%)</td>
<td>1,287 (0.1%)</td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>1,470,420 (49%)</td>
<td>652,912 (46%)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15–25, n (%)</td>
<td>383,521 (13%)</td>
<td>149,525 (11%)</td>
</tr>
<tr>
<td>25–35, n (%)</td>
<td>33,662 (11%)</td>
<td>232,098 (16%)</td>
</tr>
<tr>
<td>35–45, n (%)</td>
<td>439,223 (15%)</td>
<td>168,687 (12%)</td>
</tr>
<tr>
<td>45–55, n (%)</td>
<td>408,877 (14%)</td>
<td>203,669 (14%)</td>
</tr>
<tr>
<td>55+, n (%)</td>
<td>1,406,981 (47%)</td>
<td>657,935 (47%)</td>
</tr>
<tr>
<td>Country of origin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denmark, n (%)</td>
<td>2,623,960 (88%)</td>
<td>1,288,389 (91%)</td>
</tr>
<tr>
<td>Scandinavia, n (%)</td>
<td>34,075 (1%)</td>
<td>13,315 (1%)</td>
</tr>
<tr>
<td>Other Countries, n (%)</td>
<td>280,223 (9%)</td>
<td>100,519 (7%)</td>
</tr>
<tr>
<td>Unknown, n (%)</td>
<td>34,006 (1%)</td>
<td>9,691 (1%)</td>
</tr>
<tr>
<td>Highest Educational Attainment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary school, n (%)</td>
<td>929,426 (31%)</td>
<td>403,871 (29%)</td>
</tr>
<tr>
<td>High school, n (%)</td>
<td>160,838 (5%)</td>
<td>92,277 (7%)</td>
</tr>
<tr>
<td>Vocational internships and main course, n (%)</td>
<td>829,897 (28%)</td>
<td>417,074 (30%)</td>
</tr>
<tr>
<td>Short higher education, n (%)</td>
<td>89,474 (3%)</td>
<td>44,397 (3%)</td>
</tr>
<tr>
<td>Middle length higher education, n (%)</td>
<td>322,035 (11%)</td>
<td>165,348 (12%)</td>
</tr>
<tr>
<td>Bachelor and long higher education, n (%)</td>
<td>192,771 (6%)</td>
<td>95,981 (7%)</td>
</tr>
<tr>
<td>PhD Programs, n (%)</td>
<td>10,187 (&gt;1%)</td>
<td>4,256 (&gt;1%)</td>
</tr>
<tr>
<td>Unknown, n (%)</td>
<td>437,636 (15%)</td>
<td>188,710 (13%)</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married, n (%)</td>
<td>1,331,231 (45%)</td>
<td>622,066 (44%)</td>
</tr>
<tr>
<td>Divorced, n (%)</td>
<td>313,083 (11%)</td>
<td>154,192 (11%)</td>
</tr>
<tr>
<td>Widowhood, n (%)</td>
<td>364,476 (12%)</td>
<td>163,582 (12%)</td>
</tr>
<tr>
<td>Registered partnership, n (%)</td>
<td>4,280 (&gt;1%)</td>
<td>1,863 (&gt;1%)</td>
</tr>
<tr>
<td>Canceled registered partnership, n (%)</td>
<td>1,322 (&lt;1%)</td>
<td>625 (&lt;1%)</td>
</tr>
<tr>
<td>Longest living of two partners, n (%)</td>
<td>330 (&lt;1%)</td>
<td>125 (&lt;1%)</td>
</tr>
<tr>
<td>Unknown, n (%)</td>
<td>85,289 (3%)</td>
<td>33,911 (2%)</td>
</tr>
<tr>
<td>Main source of income</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-employed, n (%)</td>
<td>93,678 (3%)</td>
<td>41,650 (3%)</td>
</tr>
<tr>
<td>Employed spouse, n (%)</td>
<td>3,273 (&lt;1%)</td>
<td>1,854 (&lt;1%)</td>
</tr>
<tr>
<td>Wage earner with own business, n (%)</td>
<td>39,665 (1%)</td>
<td>18,179 (1%)</td>
</tr>
<tr>
<td>Wage earner without business, n (%)</td>
<td>1,174,574 (40%)</td>
<td>610,584 (43%)</td>
</tr>
<tr>
<td>Wage earner with support, n (%)</td>
<td>6,074 (&lt;1%)</td>
<td>2,989 (&lt;1%)</td>
</tr>
<tr>
<td>Senior citizen with own business, n (%)</td>
<td>21,963 (1%)</td>
<td>9,666 (1%)</td>
</tr>
<tr>
<td>Senior citizen, n (%)</td>
<td>1,059,706 (36%)</td>
<td>499,156 (35%)</td>
</tr>
<tr>
<td>Others, n (%)</td>
<td>524,385 (18%)</td>
<td>211,455 (15%)</td>
</tr>
<tr>
<td>Unknown, n (%)</td>
<td>48,948 (2%)</td>
<td>16,581 (1%)</td>
</tr>
<tr>
<td>Place of residence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capital region, n (%)</td>
<td>761,306 (26%)</td>
<td>363,318 (26%)</td>
</tr>
<tr>
<td>Large city region, n (%)</td>
<td>352,530 (12%)</td>
<td>174,090 (12%)</td>
</tr>
<tr>
<td>Hinterland region, n (%)</td>
<td>467,034 (16%)</td>
<td>219,237 (16%)</td>
</tr>
<tr>
<td>Provincial region, n (%)</td>
<td>657,852 (22%)</td>
<td>311,164 (22%)</td>
</tr>
<tr>
<td>Rural region, n (%)</td>
<td>648,253 (22%)</td>
<td>311,194 (22%)</td>
</tr>
<tr>
<td>Unknown, n (%)</td>
<td>85,289 (3%)</td>
<td>32,911 (2%)</td>
</tr>
</tbody>
</table>
late the actual sensitivity, specificity, NPV and PPV. Of note is that the calibration of the model is made on the training set, and the confusion matrix is calculated on the validation dataset. Therefore, the actual performance values may differ slightly from what the model was calibrated to.

Statistical analyses were performed using R version 3.5.0 (R Development Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria) [19].

Results

We included 4,384,178 individuals from the Danish population with at least one hospital visit. These data were divided into a training set with 2,972,264 hospital visits and a validation set with 1,411,914 hospital visits. We identified 4,350 PLWH that were included in the analyses, in either the training data or the validation data. For exact characteristics of the two datasets please refer to Table 1.

We found that an increasing number of variables included in the model resulted in increasing predictive performance. For the model including age, sex and STIs (model I), we found an AUC of 0.780 (95% CI: 0.769–0.792) in the validation set. For the full model (model III), which included additional information about medical history, we found an AUC of 0.884 (95% CI: 0.874–0.893) in the validation set (Fig. 1a and b).

When we calibrated the models to a sensitivity of 90%, the actual sensitivity of model III was 86.2%. Specificity was 69.9% for model III and only 43.5% for model I. The model, which included information on birth, educational attainment, marital status, place of residency and main source of income, but not medical history (model II), had a sensitivity of 86.1% and a specificity of 65.4% (Table 2a). The PPV of model III was 0.26%, for model II it was 0.23% and for the model I it was 0.15%. Thus, for the full model, 384 individuals would have to be tested to find one undiagnosed PLWH. For model I and model II, 681 and 441 individuals would have to be tested, respectively (Table 2a).

When we calibrated the models to a specificity of 99.9%, model III had the highest sensitivity (8.1%). The specificity did not differ substantially between the three models (Table 2b).

When we calibrated the models with regards to the highest Youden’s Index, model III outperformed the other models in all metrics. Thus, sensitivity was 76.2%, 74% and 71%, specificity was 85.0%, 82.5% and 69.9% and PPV was 0.46%, 0.37% and 0.22% for model III, II and I, respectively. Thus, for the full model, 226 individuals would have to be tested in order to find one undiagnosed PLWH. For model I and model II, 446 and 269 individuals would have to be tested, respectively (Table 2c).

When we compared different machine learning algorithms, we found that the random forest algorithm performed slightly better with an AUC of 0.892 (95% CI: 0.882–0.903), however, the CIs where overlapping with the other machine learning algorithms. Regardless, the improved performance of the random forest could indicate that some covariate interactions are present. The logistic regression with elastic net penalty and SMOTE to balance the data set performed worst with an AUC of 0.846 (95% CI: 0.836–0.857). When we calibrated the model from the random forest algorithm to different sensitivities and specificities the actual sensitivities and specificities in the validation data was substantially different from what they were calibrated to, i.e. the actual sensitivity in the validation data set was 37.5%, and the specificity was 95.4% (Table 3). This was not the case for the other machine learning algorithms (GLMlasso, GLMridge and simple logistic regression). When we calibrated the final models from the different machine learning algorithms to a sensitivity of 90% we found that the ridge regression performed best, and differed highly from all other algorithms (McNemars test: P-value < 0.001) except the simple logistic regression from which we found no statistically significant difference (McNemars test: P-value 0.494). When we compared the models from the machine learning algorithms (GLMlasso and GLMridge) with a simple logistic regression, we found that the simple logistic regression model had performance characteristics that did not differ substantially from any of the machine learning algorithms (Fig. 2).
### Table 2a: Algorithms calibrated to a sensitivity of 0.90 (high coverage - screening) Age, sex and STIs

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
<th>Total</th>
<th>Sensitivity (0.894 - 0.926)</th>
<th>Specificity (0.434 - 0.436)</th>
<th>PPV (0.0014 - 0.0016)</th>
<th>NPV (0.9998 - 0.9998)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test+</td>
<td>1,172</td>
<td>797,138</td>
<td>798,310</td>
<td>0.911</td>
<td>0.433</td>
<td>0.0015</td>
<td>0.9998</td>
</tr>
<tr>
<td>Test-</td>
<td>115</td>
<td>613,489</td>
<td>613,604</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,287</td>
<td>1,410,627</td>
<td>1,411,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Age, sex, origin of birth, educational attainment, marital status, place of residence and main source of income

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
<th>Total</th>
<th>Sensitivity (0.841 - 0.879)</th>
<th>Specificity (0.654 - 0.655)</th>
<th>PPV (0.0023 - 0.0024)</th>
<th>NPV (0.9998 - 0.9998)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test+</td>
<td>1,108</td>
<td>487,880</td>
<td>488,988</td>
<td>0.861</td>
<td>0.699</td>
<td>0.0026</td>
<td>0.9998</td>
</tr>
<tr>
<td>Test-</td>
<td>170</td>
<td>922,747</td>
<td>922,926</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,278</td>
<td>1,410,627</td>
<td>1,411,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2b: Algorithms calibrated to a specificity of 0.999 (high risk population) Age, sex and STIs

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
<th>Total</th>
<th>Sensitivity (0.842 - 0.881)</th>
<th>Specificity (0.698 - 0.700)</th>
<th>PPV (0.0025 - 0.0028)</th>
<th>NPV (0.9998 - 0.9998)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test+</td>
<td>1,110</td>
<td>424,654</td>
<td>425,764</td>
<td>0.862</td>
<td>0.699</td>
<td>0.0026</td>
<td>0.9992</td>
</tr>
<tr>
<td>Test-</td>
<td>177</td>
<td>985,973</td>
<td>986,150</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,287</td>
<td>1,410,627</td>
<td>1,411,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Age, sex, origin of birth, educational attainment, marital status, place of residence, main source of income and medical history

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
<th>Total</th>
<th>Sensitivity (0.999 - 0.999)</th>
<th>Specificity (0.9991 - 0.9992)</th>
<th>PPV (0.0047 - 0.0853)</th>
<th>NPV (0.9999 - 0.9999)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test+</td>
<td>35</td>
<td>1,667</td>
<td>1,702</td>
<td>0.027</td>
<td>0.999</td>
<td>0.0026</td>
<td>0.9992</td>
</tr>
<tr>
<td>Test-</td>
<td>1,252</td>
<td>1,408,960</td>
<td>1,410,212</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,287</td>
<td>1,410,627</td>
<td>1,411,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2c: Algorithms calibrated to maximal Youden's Index (Sensitivity + Specificity – 1) Age, sex and STIs

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
<th>Total</th>
<th>Sensitivity (0.715 - 0.763)</th>
<th>Specificity (0.698 - 0.700)</th>
<th>PPV (0.0021 - 0.0024)</th>
<th>NPV (0.9997 - 0.9998)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test+</td>
<td>952</td>
<td>424,326</td>
<td>425,278</td>
<td>0.740</td>
<td>0.699</td>
<td>0.0022</td>
<td>0.9997</td>
</tr>
<tr>
<td>Test-</td>
<td>335</td>
<td>986,301</td>
<td>986,636</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,287</td>
<td>1,410,627</td>
<td>1,411,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Age, sex, origin of birth, educational attainment, marital status, place of residence, main source of income and medical history

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
<th>Total</th>
<th>Sensitivity (0.689 - 0.739)</th>
<th>Specificity (0.824 - 0.826)</th>
<th>PPV (0.0035 - 0.004)</th>
<th>NPV (0.9996 - 0.9997)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test+</td>
<td>919</td>
<td>246,857</td>
<td>247,776</td>
<td>0.714</td>
<td>0.825</td>
<td>0.0037</td>
<td>0.9997</td>
</tr>
<tr>
<td>Test-</td>
<td>368</td>
<td>1,163,770</td>
<td>1,164,138</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,287</td>
<td>1,410,627</td>
<td>1,411,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The tables depict confusion matrices and actual sensitivities, specificities, positive predictive values (PPVs) and negative predictive values (NPVs) of the best performing GLM_{ridge} model. The models are calibrated according to the risk score that yields the desired value in the training data, i.e. when sensitivities are calculated on the validation set the actual sensitivities and specificities may differ slightly.
Table 3
Confusion matrices and performance characteristics of different algorithms.

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
<th>Total</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a: simple logistic regression algorithm sensitivity of 0.90 (high coverage - screening)</td>
<td>1,107</td>
<td>424,879</td>
<td>425,986</td>
<td>0.860 (0.840 - 0.879)</td>
<td>0.699 (0.698 - 0.700)</td>
<td>0.026 (0.0024 - 0.0027)</td>
<td>0.9998 (0.9998 - 0.9998)</td>
</tr>
<tr>
<td>3b: Random forest algorithm sensitivity of 0.90 (high coverage - screening)</td>
<td>483</td>
<td>68,763</td>
<td>69,246</td>
<td>0.375 (0.349 - 0.402)</td>
<td>0.954 (0.953 - 0.954)</td>
<td>0.0070 (0.0063 - 0.0067)</td>
<td>0.9994 (0.9994 - 0.9995)</td>
</tr>
<tr>
<td>3c: Lasso regression algorithm sensitivity of 0.90 (high coverage - screening)</td>
<td>1,120</td>
<td>439,787</td>
<td>440,907</td>
<td>0.870 (0.851 - 0.888)</td>
<td>0.688 (0.687 - 0.689)</td>
<td>0.0025 (0.0023 - 0.0027)</td>
<td>0.9998 (0.9998 - 0.9999)</td>
</tr>
<tr>
<td>3d: Ridge regression algorithm sensitivity of 0.90 (high coverage - screening)</td>
<td>1,110</td>
<td>424,854</td>
<td>425,764</td>
<td>0.860 (0.840 - 0.879)</td>
<td>0.699 (0.698 - 0.700)</td>
<td>0.0026 (0.0025 - 0.0028)</td>
<td>0.9998 (0.9998 - 0.9998)</td>
</tr>
<tr>
<td>3e: Elastic net penalty regression ($\alpha = 0.992$) algorithm with Synthetic minority oversampling (SMOTE) and sensitivity of 0.90 (high coverage - screening)</td>
<td>1,127</td>
<td>439,787</td>
<td>440,907</td>
<td>0.870 (0.851 - 0.888)</td>
<td>0.688 (0.687 - 0.689)</td>
<td>0.0025 (0.0023 - 0.0027)</td>
<td>0.9998 (0.9998 - 0.9999)</td>
</tr>
</tbody>
</table>

The tables depict confusion matrices and actual sensitivities, specificities, positive predictive values (PPVs) and negative predictive values (NPVs) of four different algorithms. The models are calibrated according to the risk score that yields the desired value in the training data, i.e. when sensitivities are calculated on the validation set the actual sensitivities and specificities may differ slightly.

Fig. 2. Training and validation sample ROC curves using different machine learning algorithms. The figure shows ROC curves for the training (2a) and validation data (2b) of simple logistic regression, best performing random forest algorithm, the best performing GLM$\text{logit}$ algorithm, the best performing GLM$\text{lasso}$ algorithm and the best performing logistic regression with elastic net penalty and synthetic minority oversampling technique (SMOTE) prior to analyses.
We assessed whether there was a change in performance of the model over the years (1995–2016) which did not seem to be the case. However, we did see minor fluctuations in sensitivity and specificity as depicted in supplementary Figure S1.

Discussion

The first of the United Nations’ (UN) 90–90–90 targets to end the HIV epidemic is for 90% of PLWH to know their HIV status. Early diagnosis is a prerequisite for achieving these targets. We examined whether machine learning algorithms could learn from existing electronic registry data and help to predict HIV status. Our study shows that such an approach is feasible and that the algorithms had a fairly high accuracy for the prediction of HIV status.

Routine HIV testing of individuals presenting with certain conditions, which may indicate the presence of HIV, has been implemented in European countries guidance on HIV testing [20]. These indicator conditions are defined as conditions in which the prevalence of HIV is >0.1%, i.e. the level at which testing is assumed to be cost-effective [7]. Although this strategy may have facilitated early diagnosis of PLWH across Europe [21], testing rates of individuals presenting with indicator conditions remain low in many settings [22,23], and the number of diagnosed PLWH in Europe is ranging from 28–98% with only three countries meeting the first “90%” of the UN’s targets [24]. Moreover, a significant number of PLWH do not experience an HIV indicator condition in the years prior to diagnosis of HIV [25,26]. Research aimed at developing complementary risk stratification tools are therefore needed.

To our knowledge, this is the first study to apply machine learning methods on nation-wide electronic registry data for the prediction of HIV status. We have shown that this method may help identify undiagnosed PLWH and potentially individuals at increased risk of acquiring HIV. Moreover, the properties of the models seemed to be favorable, and although PPVs was generally low (e.g. 0.26% in the full model where algorithms were calibrated to a high sensitivity and 83% in the full model where algorithms were calibrated to a high specificity), this PPV was comparable to currently available manual HIV risk prediction tools which also included very detailed information about sexual preferences and illicit drug use [27]. The difference between model 2 and model 3 were smaller than expected, i.e. A fairly high discriminatory value can be achieved by including demographics only, indicating that these variables are very important to include in models that guide clinicians with regards to HIV-testing. However, we did see a substantial difference in that a total number of 441 and 384, respectively, would have to be tested in order to find one PLWH when models were calibrated to a sensitivity of 90%.

There are several barriers to HIV testing including professional barriers (e.g. lack of knowledge of underlying risk factors), and personal barriers (e.g. not disclosing underlying risks due to fear of stigma, discrimination or prosecution) [28]. The present algorithms do not rely on specific information about sexual orientation, whether the individual has a partner with HIV, from a country with high HIV prevalence, or is an intravenous drug user, in which cases a testing may be recommended per se [1]. On the other hand, incorporating such information into the algorithms may potentially improve the discriminative properties of the algorithm.

The use of machine learning algorithms is a growing trend in HIV research. Thus, it has been used to predict progression of HIV and ART resistance and for ART optimization [10–12]. Recent studies have also shown that machine learning algorithms may be used for surveillance, including prediction of new HIV diagnoses and AIDS incidence based on internet search data [29,30], and there is early evidence that social media data may be used to study parameters such as HIV prevention, testing, and treatment efforts [31]. Our study adds to this growing field. However, whether such algorithms have any potential when applied in a real world setting needs to be further studied. We envision that algorithms could be integrated into clinical systems at hospitals, clinics or in primary care, e.g. by triggering a prompt to offer an HIV test to a given individual. One study evaluated a prototype application which prompted clinicians to add an HIV test when other laboratory tests selected suggested that the patient was at higher risk of HIV infection [32]. This application was found to be both feasible and acceptable among the clinicians. Future work should work toward improving these algorithms, and do external validation of currently available algorithms. This may include exploring other machine learning techniques, redefining or including new predictors and exploring interactions in the present models. Previous studies have suggested that behavioral information such as sexual preference and illicit drug use is useful for HIV prediction [30] and may improve performance of these models further. Future studies should try and make such extensions. However, in our nation-wide setting some of these extensions will prove difficult, e.g. we do not have access to behavioral data or clinical measures of disease for all people living in Denmark.

Certain ethical issues may arise when using individual characteristics for clinical prediction including the ability of an individual to consent for the use of their information [33]. Implementation of algorithms into clinical practice should take such issues into account and any benefits should outweigh the risks of harm to patients. As previously discussed, the present algorithms do not include information about substance abuse, sexual orientation or gender minority status which may be more acceptable to patients than if such information was to be included.

There are several limitations to this study. First, the algorithms assume that PLWH and uninfected individuals are classified correctly. This is not the case as an estimated number of 600,580,000 (~0.01%) individuals in Denmark are undiagnosed PLWH [13], under the assumption that undiagnosed individuals with HIV resemble the PLWH in the current study, the sensitivity and specificity of the models should not change, however PPV would decrease and NPV would increase. For obvious reason, we do not know the medical history of those PLWH that are currently undiagnosed. Second, our algorithm, in its current form, does not have universal applicability. The type and extent of electronic registry data will inevitably vary by country, the regions with the highest number of undiagnosed PLWH may not have access to this type of data. Moreover, there may be some issues in relation to the use of ICD codes due to potential systematic biases in their use, and the distribution of risk factors may vary from country to country. Also we only included data from the secondary health care sector, the majority of health care is provided in the primary sector, so exploring possibilities with data from primary health care would be valuable as and some variables look promising e.g. frequency of visits [34]. Furthermore, although Danish registry data provides a unique data capture of all individuals living in Denmark, the quality and completeness of the data may be impeded by a few factors. We only included PLWH diagnosed after 1995 which represents the date the Danish HIV Cohort Study was initiated. Moreover, there may be a slight loss to follow up (<1%) and we may not have complete information about past medical history from immigrants. Moreover, we have not yet studied implementation and we are unaware of the potential epidemiological impact or the cost-effectiveness of introducing such an approach. Thus, the aim of this work is not to convey that these algorithms is a standalone strategy, or should replace existing HIV identification strategies, such as indicator-guided HIV testing, but rather to present a novel and complementary method for early HIV identification. Future studies may try to compare the combined effect of algorithmic prediction and indicator condition guided testing. Even in the case that algorithmic prediction would only provide a little addi-
tional benefit, future studies in the field may reveal novel insights in the risk characterization of PLWH. The strength of the study includes a nation-wide population-based design including PLWH and uninfected individuals with identical electronic registry data from well-validated and comprehensive registries.

In conclusion, machine learning algorithms can learn from nation-wide electronic registry data and help to identify undiagnosed PLWH with a fairly high level of accuracy. Moreover, these algorithms may help identify individuals who may be suitable for PrEP. Future studies are needed to further evaluate the usefulness and effects of these algorithms and should be aimed toward improving the algorithms. This may include exploring other machine learning techniques to improve model performance (e.g. models intrinsically capable of hand When we calibrated the final models ling non-linearities of which neural nets (plain or more advanced) or gradient boosting could be starting points; models catering to the time-series nature of the problem (e.g. survival analysis); or models based on deeper analysis of the covariance structure), re-defining or including new predictors and exploring interactions in the present models.

Declaration of Competing Interest

MGA: No conflicts of interest. AR: No conflicts of interest. LHO: No conflicts of interest. SV: This co-author has taken up full-time employment with the pharmaceutical company Novo Nordisk A/S during the final stages of manuscript preparation. NO: No conflicts of interest.

Acknowledgments

This work was supported by the Preben and Anne Simonsens Foundation, the Novo Nordisk Foundation, Rigshospitalet, Copenhagen University, the Danish AIDS Foundation, the Augustinus Foundation, and the Danish Health Foundation.

Supplementary materials


References


8 M.G. Ahlstrøm, A. Ronit and L.H. Omland et al. /EClinicalMedicine 17 (2019) 100203