frontiers | Frontiers in Reproductive Health

TYPE Perspective
PUBLISHED 22 December 2022
DOI 10.3389/frph.2022.1062387

OPEN ACCESS

EDITED BY
Samanta Tresha Lalla-Edward,
Ezintsha, a division of the Wits Health Consortium, South Africa

REVIEWED BY
Birgitta Dresp-Langley,
Centre National de la Recherche Scientifique (CNRS), France

*CORRESPONDENCE
Joshua Fieggen
✉ jfieggen@gmail.com;
fggjos001@myuct.ac.za

SPECIALTY SECTION
This article was submitted to HIV and STIs, a section of the journal Frontiers in Reproductive Health

RECEIVED 05 October 2022
ACCEPTED 05 December 2022
PUBLISHED 22 December 2022

CITATION
Fieggen J, Smith E, Arora L and Segal B (2022) The role of machine learning in HIV risk prediction.
Front. Reprod. Health 4:1062387.
doi: 10.3389/frph.2022.1062387

The role of machine learning in HIV risk prediction

Joshua Fieggen[1,2]*, Eli Smith[2], Lovkesh Arora[2] and Bradley Segal[2,3]

[1]School of Public Health and Family Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [2]Phithos Technologies, Johannesburg, South Africa, [3]Department of Biomedical Engineering, University of the Witwatersrand, Johannesburg, South Africa

Despite advances in reducing HIV-related mortality, persistently high HIV incidence rates are undermining global efforts to end the epidemic by 2030. The UNAIDS Fast-track targets as well as other preventative strategies, such as pre-exposure prophylaxis, have been identified as priority areas to reduce the ongoing transmission threatening to undermine recent progress. Accurate and granular risk prediction is critical for these campaigns but is often lacking in regions where the burden is highest. Owing to their ability to capture complex interactions between data, machine learning and artificial intelligence algorithms have proven effective at predicting the risk of HIV infection in both high resource and low resource settings. However, interpretability of these algorithms presents a challenge to the understanding and adoption of these algorithms. In this perspectives article, we provide an introduction to machine learning and discuss some of the important considerations when choosing the variables used in model development and when evaluating the performance of different machine learning algorithms, as well as the role emerging tools such as Shapely Additive Explanations may play in helping understand and decompose these models in the context of HIV. Finally, we discuss some of the potential public health and clinical use cases for such decomposed risk assessment models in directing testing and preventative interventions including pre-exposure prophylaxis, as well as highlight the potential integration synergies with algorithms that predict the risk of sexually transmitted infections and tuberculosis.

KEYWORDS
HIV, machine learning, risk prediction, artificial intelligence, prevention, PrEP

Introduction

Since the start of the HIV epidemic, the virus has infect an estimated 76 million people worldwide, roughly 33 million of whom have died (1). While there has been a roughly 60% reduction in estimated AIDS-related annual deaths this progress has not been reflected in HIV incidence with only a 17% decrease in HIV incidence across a similar period leading to a significant rise in the number of people living with HIV (1–3). In recognition of the limited successes in reducing HIV infection incidence globally, the UNAIDS "Fast-track" targets of 95–95–95 have become accepted as foundational for accelerating HIV incidence reductions to achieve the goal of ending the HIV epidemic by 2030 (4). The updated targets seek to have 95% of people living with HIV know their status (diagnosis), 95% of those diagnosed on antiretroviral

therapy (ART), and 95% of those on ART virally suppressed by 2030 (4). Fundamental to achieving this goal in Sub-Saharan Africa is comprehensive diagnosis including populations unaware that they are at high risk of having HIV (5). However, a major challenge to this is that the relative importance of different at-risk and missed groups varies significantly both between and within different countries.

Beyond the Fast-track targets other preventative strategies remain vital to the global efforts to end the HIV epidemic, with Pre-Exposure Prophylaxis (PrEP), behaviour change communication, and early ART as prevention considered to be the three most effective strategies for preventing HIV transmission (6). When taken correctly, both PrEP and ART as prevention has been shown to be up to 100% effective in preventing HIV transmission (7–10). Critical to directing both HIV testing campaigns (required to meet the first goal of 95–95–95) as well as PrEP prescription and other targeted preventative strategies is a capacity for granular HIV risk estimation. This unmet need coupled with the initial successes of more traditional modelling techniques in delineating HIV risk (11), has led to a growing interest in the role machine learning (ML) and artificial intelligence (AI) could play in helping quantify individual risk of HIV infection. To this end, various ML models and AI algorithms have been developed using diverse datasets from both data-rich high income settings and more data-sparse low-to-middle income countries (LMICs) (12–19). In this perspective article we seek to describe the benefits and limitations to using ML for HIV risk prediction as well as discuss some of the potential future use cases of ML-guided HIV risk prediction algorithms in both meeting the UNAIDS targets as well as guiding the roll-out of other preventative strategies such as PrEP.

Machine learning for HIV risk prediction

Machine learning (ML) can be described as a collection of scientific techniques that focus on how computers learn relationships between data (20, 21). The automated pattern recognition of ML has found growing utility in medical statistics owing to the increasing size and complexity of medical data (22). ML can be classified into supervised or unsupervised learning by whether the algorithm is trained on labelled data or the algorithm self-defines the data structure from unlabelled data (20, 23, 24). Supervised learning can then be further subclassified into classification and regression algorithms based on whether the outcome being predicted is a categorical or continuous variable respectively (24). Common examples of classification problems include email spam filters (25), movie or online shopping recommendations (26), differentiating malignant and benign skin lesions (27), modelling the risk of ICU admission (28), and chest radiograph pneumonia detection (29).

The output of a classification algorithm is typically interpreted as a probability which is then binarized by means of a threshold that can be altered to increase either the sensitivity or specificity based on the model's clinical requirements (30). This makes classification models particularly useful in risk prediction. Given the persistent global burden of infectious diseases, there has been growing interest in the use of ML in risk prediction in this field (31, 32). Within HIV specifically there has been substantial attempt to try and identify individuals at high risk of infection. Initially people were classified using single risk factors, such as sero-discordant spouses (12). Subsequent approaches have largely focused on risk scores calculated *via* traditional clinical prediction tools based on regression modelling (33), with different models attempting to quantify risk of HIV seroconversion among different risk groups including men-who-have-sex-with-men (MSM), women, and sero-discordant couples (11, 34–38). Most recently, various authors have used ML approaches to attempt to quantify the complex relationships between risk factors that contribute to HIV risk (12–19). Balzer et al. directly evaluated these three approaches by comparing traditional risk factors, a risk score estimated by logistical regression, and an ML model estimated using the Super Learner algorithm and showed that ML significantly improved both the efficiency and sensitivity in identifying HIV seroconversions (12).

ML: model development and evaluation

Feature selection and model building

The predictor variables used in a ML model are called features. While the ability to handle higher numbers of features and learn the complex associations between them is one of the inherent advantages of ML, in general fewer features reduces the risks of model overfitting and leads to improved generalisability of the algorithm (39). Model overfitting is where a model's predictions are too finely tuned to the statistical noise or spurious statistical correlations in the dataset used to build the model and leads to significant limitations with generalising the model's output to new data (40). For this reason, parsimonious inclusion of features is important. Some supervised learning algorithms inherently only select the most predictive features, while for other algorithms this process needs to be made explicit (41). In addition to careful feature selection and design, another fundamental component of ML model building that helps prevent overfitting is the random splitting of the initial dataset into training, validation, and test datasets (24). Different models are then developed (trained) and compared using the validation dataset with the final model applied to the features of the holdout test dataset to evaluate

the model's performance. This explicit separation of data attempts to select for models that have extracted useful features that at a minimum generalise across unseen subsets of the dataset. Variants of this such as a bootstrapping or cross validation exist with the overall gold standard being the use of an external dataset for testing (42).

Within HIV, exact risk factors for transmission vary significantly by population but are typically behavioural and socio-demographic in nature (43). This has led to such factors featuring prominently in ML algorithms attempting to estimate HIV risk (19). The propensity of these features to have complex and poorly understood interactions has given ML-based approaches a distinct advantage in adjusting and quantifying aggregate infection risk but simultaneously introduces particular risk of overfitting. In addition, a significant challenge with modelling HIV risk is that incidence varies significantly by population impacting the prior or baseline probability of infection. A possible way to manage this challenges across difference populations is to include geography as a feature either as a ZIP/postal code (13) or as a longitude, latitude, and altitude (44). Finally, ML is increasingly used in combination with other AI techniques such natural language processing to assist with extracting important features from the narrative text of electronic health records to further enhance an automated process of HIV risk prediction (17).

Performance metrics

In contrast to traditional statistics where the primary purpose is inference, in classification ML the primary focus is on accurate prediction (24). Given that prediction is the primary concern, the confusion matrix (**Figure 1**) is a useful method to assess the performance of a given model. To generate a confusion matrix, the model is applied to the features of the holdout test dataset and the predicted outcomes (at a set probability threshold) are compared to the actual outcomes seen in the dataset. From there the performance metrics of sensitivity, specificity, accuracy, precision, and negative predictive value are calculated. However, one of the challenges with the confusion matrix is the fact the predictions are made at a particular threshold probability and thus it is difficult to assess what would happen at different probability cut-offs. In this respect receiver operator curves (ROCs), and particularly the area under the ROC (AUROC), provides a useful way of visualising and describing the trade-off between sensitivity and specificity in a model at all probability thresholds and is considered among the gold standard measures of ML model performance when applied to clinical risk prediction (45). The AUROC is especially relevant to healthcare applications as the results are not dependent on the relative prevalence of the outcome.

Understanding the model

ML's advantage in predictive performance often comes at the expense of the more typical research goal of interpretability. A common heuristic in estimating this trade-off is in the number of parameters a model utilises to make predictions (46). A simple logistic regression model has a single parameter per predictive feature whereas large-scale modern deep neural networks may have several billion (47). This presents a clear challenge to utilising ML models in practice as it becomes difficult to trust predictions that are based off unknown combinations of features, especially with concerns that models may automatically learn specific biases inherent to the dataset

	Predicted Positive	**Predicted Negative**	
Positive	True Positive (TP)	False Negative (FN)	*Sensitivity* $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP)	True Negative (TN)	*Specificity* $\frac{TN}{(TN+FP)}$
Accuracy $\frac{(TP+TN)}{(TP+FN+TN+FP)}$	*Precision* $\frac{TP}{(TP+FP)}$	*Negative Predictive Value* $\frac{TN}{(TN+FN)}$	

FIGURE 1
Performance metrics calculated from a confusion matrix.

(48–50). Interpretable ML is the domain interested in combining these two paradigms by providing techniques that enable explanations to be extracted from models several orders of magnitude more complex than is typically feasible (51).

A method that has gained substantial popularity in recent years for this task is the Shapley Additive exPlanations (SHAP) framework (52). These tools utilise an approach rooted in game theory to provide so called 'SHAP values'. These values indicate the influence each predictive feature exerted on the final model prediction and can be used to gain substantial insight into the most discriminative features a model may utilise. In addition, this assists in model trust by enabling an individual to sanity check the model's feature attribution in making a final determination. This decomposition allows for model utilisation beyond simple prediction, and can be employed to, for example, provide separate estimates of modifiable and non-modifiable risk factors despite the use of a more complex model.

The authors have developed a ML model using socio-demographic and behavioural data collected prospectively with a digital survey as described in the published protocol (53). The cohort is described in detail in a manuscript currently under review and **Figure 2** is a sub-analysis of this data presented as a visual explanation of the potential utility of SHAP-based metrics in decomposing HIV risk at both an aggregate (**Figure 2A**) and individual (**Figure 2B**) level. Given that the social and behavioural risk factors for HIV vary by context (43), the relative predictive value of these features seen in **Figure 2** is specific to this cohort and likely varies across cultures and regions. This underscores the importance of local validation and fine-tuning of any risk prediction algorithm that utilises socio-behavioural features prior to deployment emphasises the importance of model decomposition in understanding the contributors to risk in a given population.

The main limitation of SHAP-based metrics is that while they provide explanations of how a model reached a particular prediction, they do not quantify how accurate that prediction is (52, 54). Multiple methods exist that attempt to determine the relationship between the inclusion of a variable and overall model performance. Permutation Importance (PIMP) is one such tool that attempts to provide a structured approach to determine variable importance (54). This method randomly shuffles one column of the dataset at a time for several thousand iterations, one set with the outcome preserved and another with the outcome also shuffled. This provides two distributions the overlap of which provides a measure of significance and scale to which a variable improves a model. The main challenges with this methodology are that it can be computationally intensive to run enough replications and that certain variables may be highly correlated and may need to be shuffled together to gain an accurate estimation of importance (54).

Discussion

Modifying the public health response: community and individual orientated care

A major limitation to the use of ML models in HIV risk prediction thus far has been the limited interpretability of

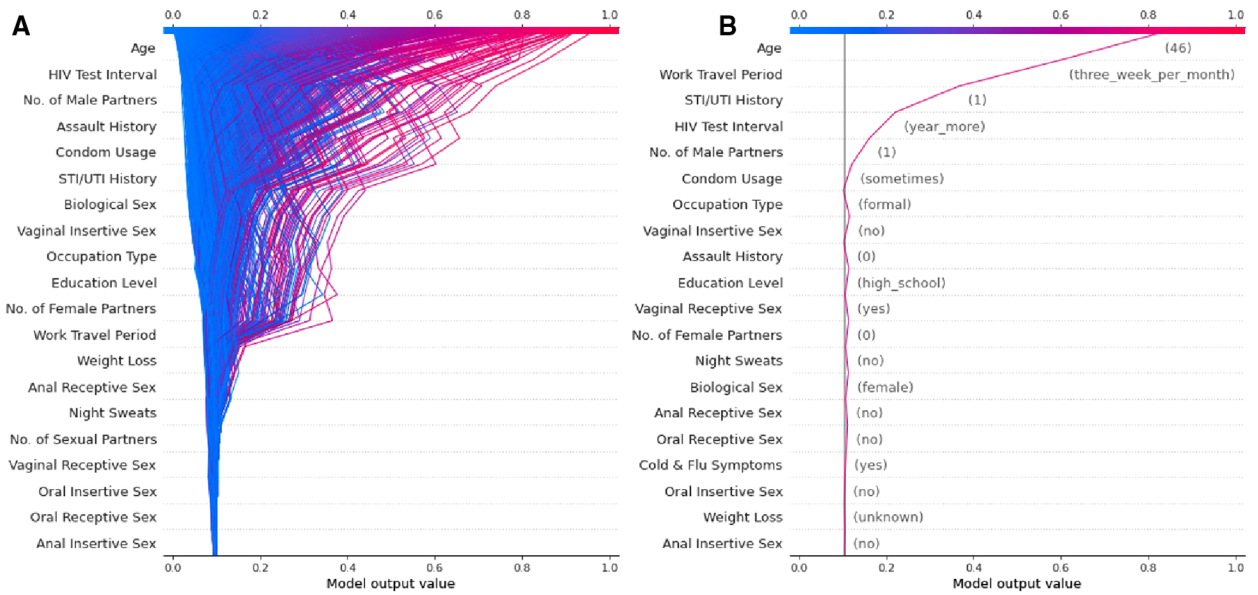

FIGURE 2
Examples of a SHAP-based feature explanation for a ML model showing (A) overall model risk output for all individuals in the test dataset and (B) an example individual at high risk for contracting HIV.

these models beyond their predictive capacity. However, we argue that tools such as PIMP and particularly SHAP allow these models to have clinical implications beyond simple prediction. Specifically, these tools allow for the decomposition of the features that make up "risk" at both an aggregate level (**Figure 2A**) and an individual level (**Figure 2B**). We believe this can translate into clinical practice by facilitating more efficient and targeted use of the interventions currently available.

If a model is appropriately contextualised and locally validated, a feature decomposition such as that presented in **Figure 2A** should provide an overview of the most important contributors to risk in a given community. In this example, age, duration since last HIV test, and the number of male sexual partners appears to convey the largest risk component. These features can then be considered in terms of modifiable risk (e.g., low levels of condom usage) or non-modifiable risk (e.g., high rates of work travel) and the public health response tailored towards either behaviour change communication or PrEP as guided by a given population's overall risk distribution. Similarly, by identifying specific risks at the individual level (**Figure 2B**), one is able to offer directed counselling and personalised interventions for risk factors that are most impacting the individual's chance of contracting HIV. For example, the major contributors to risk for the example individual in **Figure 2B** are non-modifiable and thus they may well be a good candidate for PrEP and counselling for this intervention can be directed by the risk profile generated. By identifying these factors, public health interventions could be targeted at specific issues rather than attempting to solve a heterogenous problem with blanket solutions that are not necessarily applicable to specific individuals or communities.

Utility in PrEP initiation

PrEP is widely regarded as one of the most effective strategy in the prevention of HIV transmission (55–57) and has been shown to be a cost-effective method to address the HIV epidemic (58). The recent advent of an injectable PrEP preparation containing cabotegravir heralds much excitement as the drug persists for long periods in those exposed allowing long intervals between dosing (56) which is required only every second month. The ease of administration this enables promises to alleviate some of the adherence issues faced in PrEP strategies (56). The agent has recently been approved by the Federal Drug Administration and is currently under review by various local agencies including in some LMICs, providing an opportunity to renew efforts to promote large-scale global uptake of PrEP.

Much of the current discussion around PrEP strategies centres around the issue of to whom PrEP should be offered (14, 16, 55). Identification of individuals at risk forms the basis of this discussion. Thus far strategies have directed PrEP administration at particular population groups such as MSM or particular geographical regions known to have a high prevalence of HIV (14), however there is a need to better identify candidates for PrEP in order to optimize its benefit (16). Methods in this area have aimed to identify individuals that would glean the greatest benefit from PrEP administration by identifying individuals at the greatest risk of HIV acquisition or seroconversion (14). Recognition of various individual level data as conferring risk for seroconversion has been the topic of much literature. These factors include non-modifiable factors such as age, sex, sexual-orientation and behaviour, as well as modifiable factors such as number of sexual partners or condom use (14, 16). Combinations of various factors of this type have been used to identify the most at-risk individuals and therefore those that would benefit most from PrEP. The complex matrix of data points that arises from analysis of this data is not always captured by simple calculations of risk. As such, there is significant benefit to ML as a method to augment the use of such data (14, 16). These strategies allow the capturing of the intricate interaction between factors and better identifies individuals at risk of contracting HIV and seroconverting. By using these methodologies, the efficient use of PrEP is increased as its administration is targeted at individuals with a greater likelihood of contracting HIV.

Future uses of ML in HIV associated conditions

Given the important interactions between the risk factors for sexually transmitted infections (STIs) and HIV and ML's strength in this area, it is logical to build an integrated tool that predicts the risk of both conditions. Xu et al. (2022) have recently built such a tool with a web-based interface that delivered reasonable predictive performance for HIV (AUROC = 0.72), syphilis (AUROC = 0.75), gonorrhoea (AUROC = 0.73), and chlamydia (AUROC = 0.67) (18). Given the biologically-based increased risk of HIV infection conferred by STIs (59), as well as the persistent use of syndromic management in treating STIs in many LMICs (60), algorithms that incorporate both conditions likely have significant potential synergies and utility in LMIC settings. In addition, tuberculosis (TB) is perhaps the most importance HIV-associated disease as it is estimated be responsible for around a third of deaths among people living with HIV (61). ML has been shown to be effective in assisting with both the screening and diagnosis of TB as well as predicting the risk of TB drug resistance (62). However, to the best of our knowledge ML-based TB and HIV risk assessment models have not yet been integrated into a single tool. While the

combination of TB and HIV prediction algorithms offers less potential predictive synergy there is valuable overlap in possible clinical utility.

Conclusion

As authors, we believe ML as applied to HIV risk prediction has the potential to make a significant contribution towards ending the HIV epidemic. Specifically, we see it as a critical tool in directing testing, behaviour change communication, and PrEP towards individuals and communities at high risk of infection in a resource efficient manner. Yet, while these models have been shown to be scientifically valid there remain significant barriers to them having a tangible impact. The most important of these challenges include establishing the tools for the collection of socio-demographic and behavioural data, the appropriate contextualisation and local validation of models, and the successful integration of such systems into routine HIV prevention services, particularly in resource constrained LMIC settings.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by University of the Witwatersrand Medical Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JF, ES, and BS contributed to writing the manuscript. All authors contributed to reviewing and editing the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

Phithos Technologies is currently building an HIV risk assessment tool and all contributing authors are involved in the development of that tool.

The handling editor SL-E declared a past co-authorship with the author LA.

Publisher's note

References

1. UNAIDS. Global HIV & AIDS statistics — 2020 fact sheet. 2020 [cited 2020 Aug 13]. Available at: https://www.unaids.org/en/resources/fact-sheet

2. Frank TD, Carter A, Jahagirdar D, Biehl MH, Douwes-Schultz D, Larson SL, et al. Global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2017, and forecasts to 2030, for 195 countries and territories: a systematic analysis for the global burden of diseases, injuries, and risk factors study 2017. *lancet HIV*. (2019) 6(12):e831–59. doi: 10.1016/S2352-3018(19)30196-1

3. Pandey A, Galvani AP. The global burden of HIV and prospects for control. *Lancet HIV*. (2019) 6(12):e809–11. doi: 10.1016/S2352-3018(19)30230-9

4. UNAIDS. 95-95-95 Fast-track targets.

5. Lebelonyane R, Bachanas P, Block L, Ussery F, Alwano MG, Marukutira T, et al. To achieve 95-95-95 targets we must reach men and youth: high level of knowledge of HIV status, ART coverage, and viral suppression in the Botswana combination prevention project through universal test and treat approach. *PLoS One*. (2021) 16(8):e0255227. doi: 10.1371/journal.pone.0255227

6. McGillen JB, Anderson S-J, Dybul MR, Hallett TB. Optimum resource allocation to reduce HIV incidence across sub-saharan Africa: a mathematical modelling study. *lancet HIV*. (2016) 3(9):e441–8. doi: 10.1016/S2352-3018(16)30051-0

7. Rodger AJ, Cambiano V, Bruun T, Vernazza P, Collins S, Degen O, et al. Risk of HIV transmission through condomless sex in serodifferent gay couples with the HIV-positive partner taking suppressive antiretroviral therapy (PARTNER): final results of a multicentre, prospective, observational study. *Lancet*. (2019) 393 (10189):2428–38. doi: 10.1016/S0140-6736(19)30418-0

8. Rodger AJ, Cambiano V, Bruun T, Vernazza P, Collins S, van Lunzen J, et al. Sexual activity without condoms and risk of HIV transmission in serodifferent couples when the HIV-positive partner is using suppressive antiretroviral therapy. *JAMA*. (2016) 316(2):171–81. doi: 10.1001/jama.2016.5148

9. McCormack S, Dunn DT, Desai M, Dolling DI, Gafos M, Gilson R, et al. Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. *Lancet*. (2016) 387(10013):53–60. doi: 10.1016/S0140-6736 (15)00056-2

10. Grant RM, Anderson PL, McMahan V, Liu A, Amico KR, Mehrotra M, et al. An observational study of preexposure prophylaxis uptake, sexual practices, and HIV incidence among men and transgender women who have sex with men. *Lancet Infect Dis*. (2014) 14(9):820. doi: 10.1016/S1473-3099 (14)70847-3

11. Wand H, Reddy T, Naidoo S, Moonsamy S, Siva S, Morar NS, et al. A simple risk prediction algorithm for HIV transmission: results from HIV prevention trials in KwaZulu natal, South Africa (2002–2012). *AIDS Behav*. (2018) 22(1):325–36. doi: 10.1007/s10461-017-1785-7

12. Balzer LB, Havlir D V, Kamya MR, Chamie G, Charlebois ED, Clark TD, et al. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. *Clin Infect Dis*. (2020) 71(9):2326–33. doi: 10.1093/cid/ciz1096

13. Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *lancet HIV*. (2019) 6(10): e688–95. doi: 10.1016/S2352-3018(19)30137-7

14. Zheng W, Balzer L, van der Laan M, Petersen M, Collaboration S. Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies. *Stat Med*. (2018) 37(2):261–79. doi: 10.1002/sim.7296

15. Orel E, Esra R, Estill J, Marchand-Maillet S, Merzouki A, Keiser O. Prediction of HIV status based on socio-behavioural characteristics in East and Southern Africa. *PloS one*. (2022) 17(3):e0264429. doi: 10.1371/journal.pone.0264429

16. Krakower DS, Gruber S, Hsu K, Menchaca JT, Maro JC, Kruskal BA, 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. doi: 10.1016/S2352-3018(19)30139-0

17. 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. doi: 10.1097/QAI.0000000000001580

18. Xu X, Ge Z, Chow EPF, Yu Z, Lee D, Wu J, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. *J Clin Med*. (2022) 11(7):1818. doi: 10.3390/jcm11071818

19. Mutai CK, McSharry PE, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-saharan Africa. *BMC Med Res Methodol*. (2021) 21(1):1–11. doi: 10.1186/s12874-021-01346-2

20. Deo RC. Machine learning in medicine. *Circ*. (2015) 132(20):1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593

21. Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The elements of statistical learning: Data mining, inference, and prediction*. Vol. 2. New York: Springer (2009).

22. Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med*. (2020) 104:101822. doi: 10.1016/j.artmed.2020.101822

23. Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. *Cell*. (2020) 181(1):92–101. doi: 10.1016/j.cell.2020.03.022

24. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. (2019) 19(1):1–18. doi: 10.1186/s12874-018-0650-3

25. Mansoor R, Jayasinghe ND, Muslam MMA. *In: 2021 international conference on information networking (ICOIN). IEEE. A comprehensive review on email spam classification using machine learning algorithms* (2021). p. 327–32

26. Zheng Y, Mobasher B, Burke R. *In: 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT). IEEE. Context recommendation using multi-label classification* (2014). p. 288–95

27. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542(7639):115–8. doi: 10.1038/nature21056

28. Escobar GJ, Turk BJ, Ragins A, Ha J, Hoberman B, LeVine SM, et al. Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals. *J Hosp Med*. (2016) 11:S18–24. doi: 10.1002/jhm.2652

29. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv Prepr ArXiv*. (2017) 1711:05225.

30. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol*. (2018) 138(7):1529–38. doi: 10.1016/j.jid.2018.01.028

31. Agrebi S, Larbi A. Use of artificial intelligence in infectious diseases. In: *Artificial intelligence in precision health*. Cambridge: Elsevier (2020). p. 415–38.

32. Chiu H-YR, Hwang C-K, Chen S-Y, Shih F-Y, Han H-C, King C-C, et al. Machine learning for emerging infectious disease field responses. *Sci Rep*. (2022) 12(1):1–13. doi: 10.1038/s41598-021-99269-x

33. Laupacis A, Sekar N. Clinical prediction rules: a review and suggested modifications of methodological standards. *Jama*. (1997) 277(6):488–94. doi: 10.1001/jama.1997.03540300056034

34. Kahle EM, Hughes JP, Lingappa JR, Grace J-S, Celum C, Nakku-Joloba E, et al. An empiric risk scoring tool for identifying high-risk heterosexual HIV-1 serodiscordant couples for targeted HIV-1 prevention. *J Acquir Immune Defic Syndr*. (2013) 62(3):339. doi: 10.1097/QAI.0b013e31827e622d

35. Pintye J, Drake AL, Kinuthia J, Unger JA, Matemo D, Heffron RA, et al. A risk assessment tool for identifying pregnant and postpartum women who may benefit from preexposure prophylaxis. *Clin Infect Dis*. (2017) 64(6):751–8. doi: 10.1093/cid/ciw850

36. Balkus JE, Brown E, Palanee T, Nair G, Gafoor Z, Zhang J, et al. An empiric HIV risk scoring tool to predict HIV-1 acquisition in African women. *J Acquir Immune Defic Syndr*. (2016) 72(3):333. doi: 10.1097/QAI.0000000000000974

37. Wahome E, Thiong'o AN, Mwashigadi G, Chirro O, Mohamed K, Gichuru E, et al. An empiric risk score to guide PrEP targeting among MSM in coastal Kenya. *AIDS Behav*. (2018) 22(1):35–44. doi: 10.1007/s10461-018-2141-2

38. Wahome E, Fegan G, Okuku HS, Mugo P, Price MA, Mwashigadi G, et al. Evaluation of an empiric risk screening score to identify acute and early HIV-1 infection among MSM in coastal Kenya. *AIDS*. (2013) 27(13):2163. doi: 10.1097/QAD.0b013e3283629095

39. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Pearson Education India; 2016.

40. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. (2016) 375(13):1216. doi: 10.1056/NEJMp1606181

41. Bisaso KR, Anguzu GT, Karungi SA, Kiragga A, Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. *Comput Biol Med*. (2017) 91:366–71. doi: 10.1016/j.compbiomed.2017.11.001

42. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine learning and modeling: data, validation, communication challenges. *Med Phys*. (2018) 45(10):e834–40. doi: 10.1002/mp.12811

43. Ordóñez CE, Marconi VC. Understanding HIV risk behavior from a sociocultural perspective. *J AIDS Clin Res*. (2012) 3(7):1–3. doi: 10.4172/2155-6113.1000e108

44. Cuadros DF, Li J, Branscum AJ, Akullian A, Jia P, Mziray EN, et al. Mapping the spatial variability of HIV infection in sub-saharan Africa: effective information for localized HIV prevention and control. *Sci Rep*. (2017) 7(1):1–11. doi: 10.1038/s41598-017-09464-y

45. Narkhede S. Understanding auc-roc curve. *Towar Data Sci*. (2018) 26 (1):220–7.

46. Johansson U, Sönströd C, Norinder U, Boström H. Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Med Chem*. (2011) 3(6):647–63. doi: 10.4155/fmc.11.23

47. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. *arXiv Prepr ArXiv*. (2020) 2001::08361.

48. Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L-C, Correa R, et al. Reading Race: aI recognises Patient's Racial identity in medical images. *arXiv Prepr ArXiv*. (2021) 2107:10356. doi: 10.48550/arXiv.2107.10356

49. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-)*. (2019) 366 (6464):447–53. doi: 10.1126/science.aax2342

50. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. (2021) 54(6):1–35. doi: 10.1145/3457607

51. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci*. (2019) 116(44):22071–80. doi: 10.1073/pnas.1900654116

52. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. (2017) 30:1–10.

53. Majam M, Phatsoane M, Hanna K, Faul C, Arora L, Makthal S, et al. Utility of a machine-guided tool for assessing risk behavior associated with contracting HIV in three sites in South Africa: protocol for an in-field evaluation. *JMIR Res Protoc*. (2021) 10(12):e30304. doi: 10.2196/30304

54. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. (2010) 26(10):1340–7. doi: 10.1093/bioinformatics/btq134

55. Okwundu CI, Uthman OA, Okoromah CAN. Antiretroviral pre-exposure prophylaxis (PrEP) for preventing HIV in high-risk individuals. *Cochrane Database Syst Rev*. (2012) 7:1–41. doi: 10.1002/14651858.CD007189.pub3

56. Clement ME, Kofron R, Landovitz RJ. Long-acting injectable cabotegravir for the prevention of HIV infection. *Curr Opin HIV AIDS*. (2020) 15(1):19. doi: 10.1097/COH.0000000000000597

57. Jiang J, Yang X, Ye L, Zhou B, Ning C, Huang J, et al. Pre-exposure prophylaxis for the prevention of HIV infection in high risk populations: a meta-analysis of randomized controlled trials. *PLoS One*. (2014) 9(2):e87674. doi: 10.1371/journal.pone.0087674

58. Pretorius C, Stover J, Bollinger L, Bacaër N, Williams B. Evaluating the cost-effectiveness of pre-exposure prophylaxis (PrEP) and its impact on HIV-1 transmission in South Africa. *PLoS One*. (2010) 5(11):e13646. doi: 10.1371/journal.pone.0013646

59. Cohen MS, Council OD, Chen JS. Sexually transmitted infections and HIV in the era of antiretroviral treatment and prevention: the biologic basis for epidemiologic synergy. *J Int AIDS Soc*. (2019) 22:e25355. doi: 10.1002/jia2.25355

60. Garrett NJ, Osman F, Maharaj B, Naicker N, Gibbs A, Norman E, et al. Beyond syndromic management: opportunities for diagnosis-based treatment of sexually transmitted infections in low-and middle-income countries. *PLoS One*. (2018) 13(4):e0196209. doi: 10.1371/journal.pone.0196209

61. Global WHO. tuberculosis report 2017. Geneva, Switz World Heal Organ. 2017.

62. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure F-X, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect*. (2020) 26 (5):584–95. doi: 10.1016/j.cmi.2019.09.009