A prediction risk score for HIV among adolescent girls and young women in South Africa: identifying those in need of HIV pre-exposure prophylaxis

Reuben Christopher Moyo[1], Darshini Govindasamy[2], Samuel Om Manda[3] and Peter Suwirakwenda Nyasulu[1]

[1]Faculty of Medicine and Family Health, Division of Epidemiology and Biostatistics, Stellenbosch University, Cape Town, South Africa; [2]Health Systems Research Unit, South African Medical Research Council, Cape Town, South Africa; [3]Department of Statistics, University of Pretoria, Pretoria, South Africa

Background: In sub-Saharan Africa (SSA), adolescent girls and young women (AGYW) have the highest risk of acquiring HIV. This has led to several studies aimed at identifying risk factors for HIV in AGYM. However, a combination of the purported risk variables in a multivariate risk model could be more useful in determining HIV risk in AGYW than one at a time. The purpose of this study was to develop and validate an HIV risk prediction model for AGYW.
Methods: We analyzed HIV-related HERStory survey data on 4,399 AGYW from South Africa. We identified 16 purported risk variables from the data set. The HIV acquisition risk scores were computed by combining coefficients of a multivariate logistic regression model of HIV positivity. The performance of the final model at discriminating between HIV positive and HIV negative was assessed using the area under the receiver-operating characteristic curve (AUROC). The optimal cut-point of the prediction model was determined using the Youden index. We also used other measures of discriminative abilities such as predictive values, sensitivity, and specificity.
Results: The estimated HIV prevalence was 12.4% (11.7% – 14.0) %. The score of the derived risk prediction model had a mean and standard deviation of 2.36 and 0.64 respectively and ranged from 0.37 to 4.59. The prediction model's sensitivity was 16. 7% and a specificity of 98.5%. The model's positive predictive value was 68.2% and a negative predictive value of 85.8%. The prediction model's optimal cut-point was 2.43 with sensitivity of 71% and specificity of 60%. Our model performed well at predicting HIV positivity with training AUC of 0.78 and a testing AUC of 0.76.
Conclusion: A combination of the identified risk factors provided good discrimination and calibration at predicting HIV positivity in AGYW. This model could provide a simple and low-cost strategy for screening AGYW in primary healthcare clinics and community-based settings. In this way, health service providers could easily identify and link AGYW to HIV PrEP services.

KEYWORDS: HIV, risk score, adolescent girls and young women & pre-exposure prophylaxis

Background

Adolescents and young people represent a growing share of people living with HIV worldwide. In 2022, it was estimated that over 1.7 million adolescents were living with HIV worldwide with approximately 88% of all adolescents living with HIV residing in Sub-Saharan Africa (SSA).[1] The World Health Organization (WHO) estimated that 30% of all new HIV infections globally are predicted to occur in young people aged 10 to 24. Adolescents aged 10 to 19 years account for approximately 5% of all people living with HIV worldwide, but 10% of all new HIV infections are occurring amongst adolescents.[2] In South Africa, the 2016 South African Demographic and Health Survey (SADHS) showed that AGYW are disproportionately affected by HIV compared to their

Correspondence to: Reuben Christopher Moyo, Faculty of Medicine and Family Health, Division of Epidemiology and Biostatistics, Stellenbosch University, P.O. Box 241, Cape Town 8000, South Africa. Email: reuben.moyo2014@yahoo.com

DOI 10.1080/25787489.2023.2221377

male counterparts. HIV prevalence was approximately 4 times higher in AGYW (12%) compared to their male counterparts (3%).[3]

HIV infection in AGYW is predicted by many factors broadly categorized into structural and sexual behavior factors.[4] Structural drivers of HIV are factors that relate to socio-economic status, education, and organizational factors such as health service delivery points which play a significant role in offering biomedical HIV prevention services including PrEP as well as schools that offer primary HIV prevention messages.[5] Sexual behavior factors that predict HIV are those that relate to multiple and concurrent partnerships, early sexual debut, transactional sex, and low condom use.[6–10] Sexual behavior factors such as history of anal sex, having a partner suspected of having or known to be living with HIV and having concurrent partners have been shown to predict HIV acquisition in AGYW.[11] Some studies have shown a protective effect of higher levels of parental education as well as the AGYW's levels of education on HIV acquisition.[8] Gender inequalities, violence against women (VAW), stigma and discrimination, limited access to sexual and reproductive health information and services are some of the structural factors that hinder AGYW's ability to protect themselves from HIV.[4] AGYW living with HIV experience low school attendance and are associated with high school dropouts due to HIV-related morbidity if they do not adhere to HIV treatment.[12] While HIV infection does not affect school enrolment and retention, a South African study on HIV and educational attainments found that adolescent HIV infection significantly reduced their school progress index.[13] Among women aged 15 – 44, HIV is one of the leading causes of death globally with higher death rates observed in AGYW in SSA.[14]

The South Africa's National HIV, Tuberculosis (TB) and sexually transmitted infections (STIs) strategic plan prioritizes the provision of a comprehensive package of high impact, context tailored and carefully targeted combination prevention interventions.[15] Combination prevention focuses on the combined delivery of structural, biomedical, and behavioral interventions to maximize the impact of interventions on HIV incidence. Combination HIV prevention is a key strategy in achieving the United Nations AIDS 95-95-95 targets set in 2020.[16] The target states that by 2030, 95% of people living with HIV will know their status, further 95% of people diagnosed with HIV will receive sustained antiretroviral therapy and lastly 95% of people receiving antiretroviral therapy will have their viral load suppressed.[17] Most HIV prevention strategies mainly focus on correct and consistent male condom use which leaves women with less power and control in their intimate relationships.[18] Evidence from studies on use of HIV PrEP for HIV prevention has shown that PrEP is an effective additional preventive measure for AGYW.[18,19] In December 2015, South Africa became the first SSA country to start implementing PrEP as a biomedical HIV prevention strategy. As of 2022, it was estimated that 792,000 people were using HIV PrEP in both ongoing and planned projects across South Africa against the target of 250,364.[20] The huge disparity between planned and achieved targets shows that the targets were hugely underestimated. There are three forms of HIV PrEP and these are oral drugs (TDF-FTC), vaginal ring (Dipivefrine) and long acting injectables.[21,22] In South Africa, oral PrEP is the one that is widely used at present.

Risk perceptions are key in determining high-risk AGYW to be initiated on PrEP, however evidence from studies on risk perception and PrEP use have shown that risk perceptions may be inaccurate and driven by incomplete understanding of epidemiologic risk profile often influenced by factors not related to sexual behavior.[23] Evidence from risk perception longitudinal studies among AGYW in South Africa proved that there were no significant differences in HIV positivity between those categorized as 'low' versus 'high-risk' participants which proved that their risk perceptions were inaccurate.[24] Offering PrEP to high-risk populations based on risk scoring may maximize impact and minimize cost by offering PrEP to high-risk populations based on risk stratification. Successful coverage and implementation of PrEP may be affected by health service providers inability to identify candidates to be initiated on PrEP in part due to the limited use of risk scoring tools which have been recommended to maximize PrEP impact.[25] Several barriers that limit PrEP uptake and utilization have been identified. Individual factors such as fear of HIV acquisition, fear of side effects, and burden of taking PrEP daily. Interpersonal factors that limit PrEP uptake are parental influence and absence of a sexual partner. There are also community factors (peer influence, social stigma), institutional factors (long waiting times at clinics, attitudes of health workers as well as structural factors (cost of PrEP and mode of administration, accessibility concerns) that affect utilization of HIV PrEP services.[26] Given the increase in new infections coupled with low PrEP coverage among AGYW, our study aimed at developing and validating a risk prediction model for HIV acquisition in order to identify high-risk AGYW who should be linked to HIV PrEP services.

Methods

Study design, setting and population

The data used in this study came from the HERStory survey conducted by South African Medical Research Council (SAMRC) between 2017 and 2018. We analyzed HIV-related data on 4,399 AGYW from six South African provinces namely the City of Cape Town, Ehlanzeni, OR Tambo, Tshwane, uThungulu, and Zululand. The HERStory survey used community household survey that linked information from the community household survey to participants clinic records. To ensure that participants were representative of the population from where they were drawn, participants in the survey were selected using a stratified probability proportional to size (PPS) sampling design. Sampling frames for the survey were compiled for each district based on the 2011 census small area layers and were limited to the areas targeted for the planned HIV prevention programs for AYGW. Interviewers were trained prior to data collection on how to avoid different types of information bias during data collection. Details of the methodology of the HERStory survey, inclusion and exclusion criteria have been explained in detail in their final report.[4]

Outcome and exposures of interest

Outcome of interest

HIV status: The HIV status of the AGYW was ascertained by HIV testing of dried blood spots. HIV status was coded 1 for HIV positive status and 0 for HIV negative status.

Predictors of the outcome

- Age: This was the exact age of the participant.
- Age at first sex: This variable described the age at which the AGYW first had penetrative sex.
- Condom use: This variable described whether the participant used a condom the last time she had sex.
- History of Sexually transmitted infections (STIs): This variable indicated whether the participant presented with any STI symptoms.
- The number of sexual partners: This variable indicated the participant's number of penetrative sexual partners she ever had.
- Partner HIV status: This variable described the HIV status of the participant's sexual partner.
- Marital status: Whether the AGYW was legally or traditionally married or not.
- Transactional sex: This variable describes whether the participant engaged in sex for money or other items.
- Orphanhood: Whether the AGYW was an orphan or not.
- Use of drugs and substances: This variable described whether the participant used drugs and other substances.
- Partner age: This variable referred to the age of the current or last partner participant's sexual partner.
- Socio-economic status: This variable was derived from the wealth index score of the AGYW which measured participants socio-economic status based on household asset.
- Highest Education: This variable described the highest education level attained.
- District: The is the exact geographical district where the AGYW was captured during data collection.
- Ever pregnant: This variable described whether the AGYW has ever been pregnant or not.
- Rape: A variable that described whether the AGYW was ever raped or not.

Statistical analysis

Statistical analyses were conducted in Stata version 16.1. We conducted descriptive analysis using the frequency procedure to show descriptive statistics in the form of numbers and proportions. To examine association between HIV status and its purported predictor variables, Pearson's chi-square tests (X^2) were conducted. Predictor variables were considered significant at $p < 0.05$. To ensure that estimates produced in this study were representative of the AGYW population from their respective geographical areas, we applied sampling weights to facilitate analysis of survey data which has the ability to correct unequal representation of the sampled population. Missing data was not a concern in this study because the data had only one variable with a missing observation.

Development and validation of a risk prediction model

We used 70% of the data for training the model and 30% for testing the performance of the model at predicting the outcome when applied to an external population of AGYW. To quantify the amount of HIV risk associated with each explanatory variable after controlling for the independent effects of other covariates such as age, education levels and age at sexual debut, a multivariable binary logistic regression model of HIV status on its predictors was used to obtain coefficients for use in deriving the HIV risk scores. Variables in a model were selected using least absolute shrinkage selection operator (LASSO). LASSO is a machine learning feature selection method to maximize prediction accuracy of the model. Age, number of sexual partners, pregnancy, rape, and transactional sex were forced in the model because they were treated as priori confounders and have been shown to mediate HIV infection in AGYW.[23,27] We used likelihood ratio test to select the best preforming model among successive models. The performance of the final model was assessed using discrimination and calibration measures.

We used area under receiver operating characteristic curve (AUC) to assess the performance of the model at discriminating HIV positive versus negative status on both training and testing datasets. Calibration was assessed using Hosmer Lemeshow, Brier score and Pseudo R^2.

Scoring of a risk prediction tool

The HIV risk prediction scores were developed by summing coefficients of the risk variables from the multivariable logistic regression model of HIV positivity (26). The optimal cut-point of the risk score at which AGYW were likely to have an HIV-positive status was determined using the Youden index.[28]

Results

Distribution of study participants

A total of 4,399 AGYW participated in the survey. Overall, most participants were drawn from Zululand (17.9%), Ehlanzeni (17.8%), O. R. Tambo (17.0%), and Tshwane (15.7%). Approximately 57% of the study participants were aged 15 to 19 while 43% were aged 20 to 24 years. 69.2% of the participants reported having ever had sex. Of the participants who ever had sex, 8.8% reported having started sex before or at the age of 15. The proportion of participants who reported that they were ever raped was 6% while the proportion of participants who had been pregnant before accounted for 38%. Only 12.1% of the participants reported that they ever engaged in transactional sex at some point. Table 1 shows the distribution of selected participants' characteristics.

The overall HIV prevalence among the study participants was 12.4% (11.7% − 14.0%). HIV prevalence was high among AGYW aged 20 to 24 (19.7%) compared to participants aged 15–19 (6.75%). AGYW who reported having been involved in transactional sex had a higher prevalence (20.0%) compared to those who never practiced transactional sex (11.6%). HIV prevalence was high in AGYW with low socio-economic status (13.2%) compared to those in the high socio-economic status category (9.2%). The prevalence of HIV was extremely high (57.6%) among AGWY in a relationship with a known HIV-positive partner. The HIV prevalence among AGYW who did not know the partner's HIV status and those who preferred not to reveal their partner's status was approximately 24%. Participants who reported any STI symptom had a slightly higher prevalence of HIV (18.5%) compared to those who were not treated for STIs (10.6%). There were no significant differences in HIV prevalence between AGYW who ever used drugs and substances and those who did not. Table 2 shows a comparison of HIV status by selected predictor variables and their corresponding p-values.

Table 1. Characteristics of the sampled participants (*N* = 4,399).

Characteristic	Frequency	Weighted percentage
District		
Cape Town	377	8.6
Ehlanzeni	803	18.3
O R Tambo	690	15.7
Tshwane	767	17.5
uThungulu	748	17.0
Zululand	1,014	22.1
Age group		
15 – 19	2,515	57.2
20 – 24	1,884	42.8
Ever had sex		
Yes	3,009	68.4
Age at sexual debut		
≤ 15	259	5.9
16 – 24	2,750	62.5
Not applicable	1,390	31.6
Highest education		
None or primary	210	4.8
Secondary	3,892	88.5
Some post-secondary	297	6.8
History of transactional sex		
Yes	424	9.6
Socio-economic status		
High SES	792	18.0
Low SES	3,607	82.0
Relationship status		
Married	39	0.9
Dating	2,693	61.2
Single	1,624	36.9
Prefer not to say	43	1.0
HIV status		
HIV positive	568	12.9
HIV negative	3,831	87.1
Any STI symptom		
Yes	990	22.5
Partner HIV status		
HIV positive	141	3.2
HIV negative	2,013	45.8
Prefer not to say	86	2.0
Do not know	899	20.4
Do not have partner	1,260	28.6
Often use substances		
Yes	415	9.4
Ever pregnant		
Yes	1,680	38.2
Ever raped		
Yes	265	6.0

Independent HIV risk scores for AGYW

Table 3 shows coefficients and their corresponding risk scores from the final multivariable prediction model of HIV risk factors. 13 candidate predictors were used in the final model. Compared to AGYW aged 15 to 19, participants aged 20 to 24 had higher risk of HIV ($\beta = 0.70$, $p < 0.001$). The risk of HIV infection was slightly high for AGYW in the lower socio-economic status ($\beta = 0.17$, $p = 0.010$) compared to those in high socio-economic status. Being in a relationship with a partner known to be living with HIV was associated with more than 11 times the odds of

HIV infection ($\beta = 2.6$, $p < 0.001$). AGYW who did not know the status of the partner and those who preferred not to reveal the status of their partners were also associated with higher risk ($\beta = 1.0$, $p < 0.001$ and $\beta = 1.4$, $p < 0.001$) respectively. The risk score associated any STI symptom was slightly high (0.3) compared to those who did not report any STI symptom. AGYW who lost a parent had slightly higher risk score (0.3) compared to those who did not. Higher levels of education were associated with lower risk of HIV infection compared those with primary education ($\beta = -0.7$, $p = 0{,}020$ and $\beta = -1.11$, $p < 0.001$) for secondary level and post-secondary education levels respectively. There were no significant differentials in the risk of HIV with respect to age at sexual debut. The optimal coefficient cut-off point estimated using the Youden index was 2.43 with a sensitivity cut point of 71%, a specificity cut point of 60% and an AUC of 0.66. Table 4 shows selected cut-points at various levels.

Performance of the risk prediction model

Table 5 shows a summary of the prediction model's classification. The model's sensitivity was 16.7% and a specificity of 98.5%. The positive predictive value was 68.2% while the negative predictive value was 85.8%. The model correctly classified 85.1% of the outcome. The training AUC was 0.78 while the testing AUC was 0.76. Figure 1 shows the training and testing AUC of the prediction model. Table 6 shows a summary of the discrimination and calibration statistics of the prediction model. The prediction model showed good calibration with a non-significant Hosmer Lemeshow test ($p = 0.564$), a brier score of 0.095 and a pseudo R^2 value of 0.150.

Table 2. Comparison of HIV status by predictor variables ($N = 4{,}399$).

Characteristic	HIV positive (Weighted row%)	HIV negative (Weighted row%)	P-Value
Age group (years)			<0.001
15 – 19	185 (6.75)	2,330 (93.25)	
20 – 24	383 (19.7)	1,501 (80.3)	
Age of sexual debut			<0.001
≤ 15	43 (15.9)	216 (84.1)	
16 – 24	456 (15.25)	2294 (84.75)	
Not Applicable	69 (4.5)	1.321(95.5)	
Transactional sex			<0.001
Yes	90 (20.0)	334 (80.0)	
No	478 (11.56)	3,497(88.44)	
Socio-economic status			0.036
High SES	78 (9.20)	714 (90.80)	
Low SES	490 (13.20)	4,117 (86.8)	
Highest education			0.146
None or primary	35 (15.35)	175 (84.65)	
Secondary	501 (12.36)	3,391 (87.64)	
Some post-secondary	32 (10.31)	265 (89.69)	
Partner older than 5yrs			<0.001
Yes	211 (20.28)	839 (78.72)	
No	304 (13.92)	1,785 (86.08)	
Do not know/NA	53 (3.69)	1,207 (96.31)	
Partner HIV status			<0.001
HIV positive	84 (57.63)	57 (42.36)	
HIV negative	244 (7.45)	3,029 (92.55)	
Prefer not to say	21 (23.68)	65 (76.32)	
Do not know	219 (24.24)	680 (75.76)	
Ever treated for STIs			<0.001
Yes	191 (18.45)	799 (81.55)	
No	377 (10.56)	3,032 (89.44)	
Has more than six partners			0.7045
Yes	45 (11.67)	299 (88.33)	
No	523 (12.43)	3,532 (87.57)	
One parent dead			<0.001
Yes	345 (16.76)	1,688 (83.24)	
No	219 (9.0)	2,065 (91.0)	
Often use substances			0.5590
Yes	57 (12.14)	358 (87.86)	
No	511 (12.39)	3,473 (87.61)	
Ever pregnant			<0.001
Yes	338 (20.12)	1,342 (79.88)	
No	223 (8.35)	2,447 (91.65)	
Ever raped			0.003
Yes	50 (18.87)	215 (81.13)	
No	518 (12.53)	3,616 (87.47)	

Table 3. Coefficients and risk scoring for the independent predictors of HIV using the training dataset.

				95% CI	
Predictor	**Coefficient**	**Risk score**	**P-Value**	**Lower**	**Upper**
Age group					
15 – 19	0	0			
20 – 24	0.7013	0.70	<0.001	0.3988	1.0039
Age at first sex					
≤ 15	0				
>15	−0.0371	−0.04	0.883	−0.5305	0.4562
Relationship status					
Single	0.3084	0.31	0.051	−0.0015	0.6183
Dating	0	0			
Married	−0.2600	−0.26	0.716	−1.6592	1.1392
Prefer not to say	0.7405	0.74	0.250	−0.5212	2.0023
Education					
Primary	0	0			
Secondary	−0.7093	−0.71	0.020	−0.3047	−0.1140
Post-secondary	−1.2398	−1.24	<0.001	−2.0175	−0.4620
Socioeconomic status					
Relatively low SES	0.1701	0.17	0.369	−0.2007	0.5409
High SES	0	0			
Partner 5 years older					
Yes	0.1956	0.20	0.145	−0.0675	0.4587
No	0	0			
No partner	−0.3038	−0.30	0.383	−0.9864	0.3837
Any STI symptom					
Yes	0.3240	0.32	0.025	0.0416	0.6063
No	0	0			
Partner HIV positive					
No	0	0			
Yes	2.5827	2.58	<0.001	2.1077	2.0583
Prefer not to say	1.3863	1.39	<0.001	0.6121	2.1605
Do not know	0.9275	0.93	<0.001	0.6568	1.1983
No partner	−1.558	1.56	0.056	−2.2035	0.0859
Used condom at last sex					
Yes	0.1575	0.16	0.277	−0.1267	0.4419
No	0	0			
Prefer not to say	−0.0197	−0.02	0.920	−0.4036	0.3641
Ever engaged in transactional sex					
No	0	0			
Yes	0.1818	0.18	0.328	−0.2064	0.5701
Parent dead					
No	0	0			
Yes	0.3190	0.32	0.016	0.0603	0.5778
Ever pregnant					
No	0	0			
Yes	0.1209	0.12	0.418	−0.1720	0.4139
Ever raped					
No	0	0			
Yes	0.1285	0.13	0.621	−0.3813	0.6384

Table 4. Performance characteristics of the selected risk scores based on the training dataset.

Cut point	Sensitivity	Specificity	Correctly classified	LR+	LR-
>=0	100.00%	0.00%	12.91%	1.0000	
>=0.5	100.00%	0.03%	12.93%	1.0003	0.0000
>=1.0	100.00%	0.57%	13.41%	1.0058	0.0000
>=1.5	98.59%	7.60%	18.35%	1.0670	0.1854
>=2.0	88.03%	33.12%	40.21%	1.3163	0.3614
>=2.5	61.09%	69.15%	68.11%	1.9800	0.5627
>=3.0	38.56%	84.23%	78.34%	2.4455	0.7294
>=3.5	12.32%	96.95%	86.02%	4.0353	0.9044
>=4.0	0.00%	100.00%	87.098		1.0000

The optimal cut-off point of the risk score was 2.43 with a sensitivity of 71% and specificity of 60%.

Key.

Sensitivity: The proportion of actual positives which are correctly identified as such.

Specificity: The proportion of negatives which are correctly identified as such.

LR+: The ratio of the probability of a positive test among the truly positive subjects to the probability of a positive test among the truly negative subjects.

LR-: The ratio of the probability of a negative test among the truly positive subjects.

Discussion

This study aimed at developing and validating a risk prediction model for predicting AGYW with elevated HIV risk based on selected predictors of HIV status. Our risk prediction model showed both good discrimination and calibration at predicting HIV in AGYW. Based on the Youden Index cut-off point score, an optimal risk score cut-off point of 2.43 may be indicative of a positive HIV status. This means that AGYW with a risk score of at least 2.43 should be offered screening and linked to HIV PrEP services. This study has found that AGYW exposed to HIV-positive partners have more than twice the risk of HIV compared to HIV-negative partners. This finding is not unusual because of the increased exposure to the HIV in the absence of protection and HIV PrEP use. Deliberate efforts are required to initiate such AGYW on PrEP to reduce their likelihood of seroconversion. Similarly, the risk score was slightly high in AGYW who reported any STI symptom. This finding supports and strengthens the policy recommendation of offering HIV testing to all clients visiting STI clinics to ascertain their HIV status. The optimal cut point identified in this study does not replace routine screening in clinical care settings. The HIV risk prediction models and their cut-off points are meant to help health care workers stratify AGYW based on risk scoring and provide services according to their risk stratification. These findings inform users of the risk prediction model that HIV programming for AGYW should particularly target AGYW with elevated risk to ensure HIV prevention interventions are impactful and cost-effective.

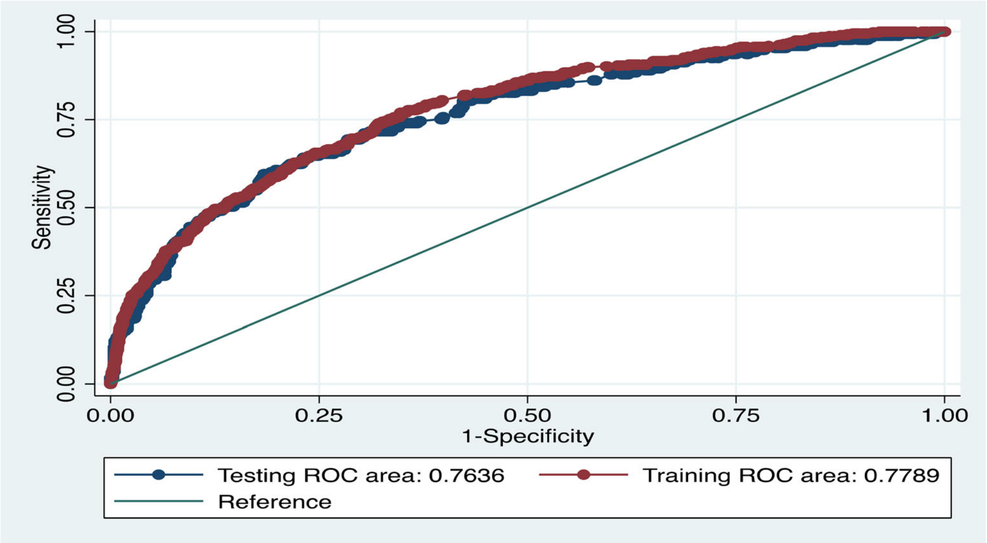
Based on the number of variables selected to assess risk, the risk classification and their optimal cut-off points may change to fit the user's situation. While HIV risk prediction models based on HIV prevalence exist, many available models in literature do not have key predictors of HIV in AGYW such as partner HIV status, any symptoms of STIs, sexual violence and substance use. This study has investigated the contribution of all these predictors and included them in the model. Our risk prediction model further adds a new

Table 5. Prediction model's classification table based on the training dataset.

Sensitivity	16.67%
Specificity	98.48%
Positive predictive value	68.24%
Negative predictive value	85.77%
False positive rate	1.52%
False negative rate	83.33%
False positive rate for classified positive	31.76
False negative rate for classified negative	14.23
Correctly classified	85.07

Table 6. Performance of the risk prediction model on the training and testing datasets.

	Training set		Testing set	
	Value	Remark	Value	Remark
Discrimination				
AUC	0.778	Good	0.764	Good
Calibration				
Hosmer Lemeshow	0.5638	Good	0.3627	Good
Brier Score	0.0954	Good	0.0954	Good
Pseudo R^2	0.1502	Good	0.1727	Good



X-axis is the model's sensitivity while y-axis is 1- Specificity. The blue line represents the testing ROC while red line represents the training ROC.

Figure 1. Training and testing AUC of the prediction model.
X-axis is the model's sensitivity while *y*-axis is 1–Specificity. The blue line represents the testing ROC while red line represents the training ROC.

understanding of the risk prediction approach in HIV epidemiology by presenting the most up-to-date risk assessment tool for AGYW. Our prediction may be applied in similar settings with high prevalence of HIV in sub-Saharan Africa. Some prediction tools are based on proximate predictors of HIV only while others include mediators including socio-demographic and proximate predictors. The risk tool developed in Malawi using data from VOICE trial included 6 predictors in the final model that lacked key predictors of HIV such as partner HIV status, age at first sex and whether the AGYW was involved in transactional sex or not.[29] Similarly, an HIV risk assessments tool for AGYW in South Africa also lacked variables such as partner HIV status and transactional sex.[30] Lack of key variables that strongly predict HIV in a prediction model may reduce the model's performance at discriminating AGYW with an elevated risk of HIV. One feature that has been shown to predict HIV but has not been included in our model is history of TB treatment. History of TB was found to significantly increase the probability of HIV in a study on use of machine learning techniques to identify HIV predictors in sub-Saharan Africa.[31] A recent study on HIV risk score among adult populations in sub-Saharan Africa also showed that younger age, non-cohabiting and recent STIs were consistently identified as predicting future HIV infection with moderate prediction accuracy.[32] This study therefore presents a robust tool that has been developed and validated to accurately capture AGYW with elevated risk. Based on behavior change and other circumstances, risk scores for AGYW may change hence the need to periodically review and follow up risk over time. Clinical prediction models for HIV have the potential to increase the number of AGYW to be initiated on HIV PrEP to reduce their risk of acquiring HIV infection. This, however, depends on the accuracy of the prediction model at identifying high-risk AGYW. If AGYW have been falsely identified as high risk when their actual risk is low, they will be initiated on PrEP when they are not supposed to be initiated on PrEP. Likewise, AGYW falsely identified as low risk when their actual risk is high will not be initiated on PrEP which may expose them to HIV infection. Monitoring risk status over time is important to prevent and correct this misclassification. Studies on use of risk prediction models to identify high risk populations have shown that these models lose their prediction power over time in part due to changes in prevalence of the outcome which may affect the prediction model's performance at predicting the outcome. It is important to continuously monitor the performance of the model and update it when it no longer predicts the outcome. The period for updating and recalibrating the risk prediction model depends on changes on the prevalence of the outcome and the need to add or remove predictors. The model maybe updated by either using new datasets or adding candidate predictors.[33,34] Given the high prevalence of HIV among AGYW, we suggest that programming for PrEP should not only target high risk AGYW but a larger proportion of AGYW with both high epidemiologic risk and those with high perceived risk to increase PrEP coverage. Neglecting those with low epidemiologic risk but with high-risk perception will reduce PrEP coverage and lead to an increase in new infections among AGYW.

Policy implications and applications

Our findings have policy implications and applications in HIV programming. Firstly, risk scores may be used by service providers to supplement health education and counseling to AGYW in high HIV prevalent settings in SSA to increase coverage of both screening and PrEP initiation. Secondly, HIV risk prediction models may also be useful in monitoring changes in risk over time to check if the AGYW's risk score is changing from low to high or vice versa depending on circumstances.[35] Depending on geographical areas where many AGYW are scoring above the cut-off point on a risk score, this tool may be used to allocate resources to such areas so that more resources allocated to areas where high risk scores are likely.

Limitations

This study has limitations. Firstly, this study utilized cross-sectional survey data from six South African provinces only. This can affect generalizability in other settings and countries where HIV prevalence is low. However, results from this study could potentially be relevant and be applied in countries with high HIV prevalence mostly in SSA. The validation of the model was done using the same data set, this may affect the accuracy of the model if there are systematic differences between the sampled population and AGYW from other settings not represented in the study. This study used HIV prevalence data to develop a risk prediction model for AGYW, use of HIV prevalence data at the expense of incident data may affect the model's capacity to predict new infections. There are many methods of developing risk prediction models such as generalized linear models and ensemble methods, this study used generalized linear models at the expense of ensemble models that perform both feature selection and prediction modeling to increase prediction accuracy. The risk tool does better with more in-depth and personal exposure questions, this many affect its use in busy settings where more time is required and when participants are not willing to disclose such information.

Recommendations

We recommend the use of this risk prediction model to supplement clinical decision making to increase coverage of PrEP use. We also recommend a feasibility study of using the risk tool in clinical settings to assess its user-friendliness and its accuracy at identifying high-risk AGYW.

Conclusion

Our risk prediction model has shown good discrimination and calibration at predicting AGYW with elevated risk of acquiring HIV. Our risk prediction model provides a data-driven way of identifying predictors as well as predicting AYGW at high risk of infection to be targeted with HIV PrEP services both at clinic level as well as community level.

Acknowledgements

We would like to acknowledge South African Medical Research Council (SAMRC) for authorizing HERStory data to be used in the development of this risk prediction model.

Authors contributions

Study design: Moyo R, Govindasamy D, Nyasulu PS and Manda S; Data collection and cleaning: Govindasamy D; Data analysis and interpretation: Moyo R, Manda S, Nyasulu PS and Govindasamy D; Writing of the draft manuscript: Moyo R; Read and approve the final manuscript: All authors.

Availability of data and materials

The datasets analyzed during the current study are not publicly available due

Data protection policy of South African Medical Research Council (SAMRC) which only makes the data available on request. Data can be requested from SAMRC through Dr Darshini Govindasamy on darshini.govindasamy@mrc.ac.za

Competing interests

The authors declare no competing interests.

Consent for publication

Not Applicable.

Disclosure statement

No potential conflict of interest was reported by the authors.

Ethics approval

This study was approved by Stellenbosch University health research ethics committee (S21/09/171). There was actual involvement of humans during data collection in the initial HERStory survey. All methods and procedures during data collection were conducted in accordance with relevant guidelines and regulations of research involving humans. Informed consent was obtained from all subjects and or participants representatives.

Abbreviations

AGYW	adolescent girls and young women
ART	antiretroviral therapy
AUC	area under receiver operator characteristic curve
CDC	centers for disease control and prevention
HIV	human immunodeficiency virus
PEP	post exposure
PPS	Probability proportional to Size
PrEP	pre-exposure prophylaxis
SADHS	South African demographic and health survey
SAL	small area layers
SSA	sub-Saharan Africa
STI	sexually transmitted infections
UNAIDS	joint united nations program on acquired human immunodeficiency syndrome
UNICEF	united nations children fund
VAW	violence against women
WHO	world health organization

Funding

This study was not funded.

ORCID

Reuben Christopher Moyo http://orcid.org/0000-0003-4308-0103

References

1 Children's UN. Aldolescent HIV prevention. http://data unicef/topic/hivaids/adolescents-youth-Accessed on 19/10/2020. Published online 2020.

2 W. Maternal, Newborn Child and Adolescent Health.

3 National Department of Health (NDoH), Statistics South Africa (Stats SA), South African Medical Research Council (SAMRC), and ICF. Pretoria, South Africa, and Rockville, Maryland, USA: NDoH, Stats SA, SAMRC and ICF. *South Africa Demographic and Health Survey 2016*, 2019.

4 SAMRC. Evaluation of a South African Combination HIV prevention programme for adolescent girls and young women. 2020 August; 128.

5 Saul J, Bachman G, Allen S, Toiv NF, Cooney C, Beamon T. The DREAMS core package of interventions: A comprehensive approach to preventing HIV among adolescent girls and young women. *PLoS One*. 2018;13(12):e0208167.

6 Lakew Y, Benedict S, Haile D. Social determinants of HIV infection, hotspot areas and subpopulation groups in Ethiopia: evidence from the National Demographic and Health Survey in 2011. *BMJ Open*. 2015;5(11):e008669.

7 Chersich MF, Rees HV. Vulnerability of women in southern Africa to infection with HIV: Biological determinants and priority health sector interventions. *Aids.* 2008;22(Suppl. 4):S27–S40.

8 Skovdal M, Belton S. The Social Determinants of Health as they relate to children and youth growing up with HIV infection in sub-Saharan Africa. *Child Youth Serv Rev.* 2014;45(C):1–8.

9 Underwood C, Skinner J, Osman N, Schwandt H. Structural determinants of adolescent girls' vulnerability to HIV: Views from community members in Botswana, Malawi, and Mozambique. *Soc Sci Med*. 2011;73(2):343–350.

10 Buot MLG, Docena JP, Ratemo BK, et al. Beyond race and place: Distal sociological determinants of HIV disparities. *PLoS One*. 2014;9(4):e91711.

11 Asaolu IO, Gunn JK, Center KE, Koss MP, Iwelunmor JI, Ehiri JE. Predictors of HIV testing among youth in sub-Saharan Africa: A cross-sectional study. *PLoS One*. 2016;11(10): e0164052.

12 Ijumba N. Impact of HIV/AIDS on education and poverty About the author.

13 Fotso AS, Banjo O, Akinyemi JO. Article HIV and adolescents' educational attainment in South Africa: Disentangling the effect of infection in children and household members. *SAJCH* 2018; 12(2):1512.

14 Impact THE, HIV OF, Girls AON, Globally YW. IWHC the impact of HIV and AIDS on girls.

15 SANAC. *South Africa's National Strategic Plan on HIV/Aids, TB and STI's 2017-2022. South African National AIDS Council*. 2017, v.1. Pretoria, South Africa.

16 Frescuraid L, Godfrey-Faussett P, Feizzadeh A, El-Sadr W, Syarif O, Ghys PD. Achieving the 95 95 95 targets for all: A pathway to ending AIDS. 2022;17(8):e0272405.

17 Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA. An ambitious treatment target to help end the AIDS epidemic. *Unaids.* 2016;1–40.

18 Hill LM, Maseko B, Chagomerana M, et al. HIV risk, risk perception, and PrEP interest among adolescent girls and young women in Lilongwe, Malawi: operationalizing the PrEP cascade. *J Intern AIDS Soc*. 2020;23(S3):40–47.

19 Celum CL, Delany-Moretlwe S, McConnell M, et al. Rethinking HIV prevention to prepare for oral PrEP implementation for young African women. *J Int AIDS Soc.* 2015;18(Suppl 3):20227.

20 PrePWatch SA. A snapshot of PrEP scale-up. Registration and Resources for South Africa. 2021.

21 South Africa – PrEPWatch. https://www.prepwatch.org/countries/south-africa/. Accessed May 4, 2023.

22 South Africa to begin piloting injectable PrEP in early 2023 | aidsmap. https://www.aidsmap.com/news/nov-2022/south-africa-begin-piloting-injectable-prep-early-2023. Accessed May 4, 2023.

23 Price JT, Rosenberg NE, Vansia D, et al. Predictors of HIV, HIV risk perception, and HIV worry among adolescent girls and young women in Lilongwe, Malawi. *J Acquir Immune Defic Syndr*. 2018;77(1):53–63.

24 Maughan-Brown B, Venkataramani AS. Accuracy and determinants of perceived HIV risk among young women in South Africa. *BMC Public Health.* 2018;18(1):1–9.

25 Wilton J, Kain T, Fowler S, et al. Use of an HIV-risk screening tool to identify optimal candidates for PrEP scale-up among men who have sex with men in Toronto, Canada: Disconnect between objective and subjective HIV risk. *J Int AIDS Soc.* 2016;19(1): 20777.

26 Muhumuza R, Ssemata AS, Kakande A, et al. Exploring perceived barriers and facilitators of PrEP uptake among young people in Uganda, Zimbabwe, and South Africa. *Arch Sex Behav*. 2021;50(4):1729–1742.

27 Ranganathan M, Heise L, Pettifor A, et al. Transactional sex among young women in rural South Africa: Prevalence, mediators and association with HIV infection. *J Int AIDS Soc.* 2016; 19(1):20749.

28 Pettit AC, Bian A, Schember CO, et al. Development and validation of a multivariable prediction model for missed HIV health care provider visits in a large US Clinical Cohort. *Open Forum Infect Dis.* 2021;8(7):1–9.

29 Rosenberg NE, Kudowa E, Price JT, et al. Identifying adolescent girls and young women at high risk for HIV acquisition: A risk assessment tool fromthe girl power-malawi study. *Sex Transm Dis*. 2020;47(11):760–766.

30 Pasteur L, Koch R. Introduction. *Introduction*. 1934;74:535–546.

31 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.

32 Jia KM, Eilerts H, Edun O, et al. Risk scores for predicting HIV incidence among adult heterosexual populations in sub-Saharan Africa: a systematic review and meta-analysis. *J Int AIDS Soc*. 2022;25(1):e25861.

33 Su TL, Jaki TF, Hickey G, et al. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res*. 2018;27(1):185–197.

34 Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res*. 2021;5(1):1–7.

35 Luo Q, Huang X, Li L, et al. External validation of a prediction tool to estimate the risk of human immunodeficiency virus infection amongst men who have sex with men. *Medicine (United States).* 2019;98(29):e16375.