BMC Women's Health

RESEARCH Open Access

Development of a predictive model for identifying women vulnerable to HIV in Chicago

Eleanor E. Friedman[1*], Shivanjali Shankaran[2], Samantha A. Devlin[1], Ekta B. Kishen[2], Joseph A. Mason[1], Beverly E. Sha[2] and Jessica P. Ridgway[1]

Abstract

Introduction Researchers in the United States have created several models to predict persons most at risk for HIV. Many of these predictive models use data from all persons newly diagnosed with HIV, the majority of whom are men, and specifically men who have sex with men (MSM). Consequently, risk factors identified by these models are biased toward features that apply only to men or capture sexual behaviours of MSM. We sought to create a predictive model for women using cohort data from two major hospitals in Chicago with large opt-out HIV screening programs.

Methods We matched 48 newly diagnosed women to 192 HIV-negative women based on number of previous encounters at University of Chicago or Rush University hospitals. We examined data for each woman for the two years prior to either their HIV diagnosis or their last encounter. We assessed risk factors including demographic characteristics and clinical diagnoses taken from patient electronic medical records (EMR) using odds ratios and 95% confidence intervals. We created a multivariable logistic regression model and measured predictive power with the area under the curve (AUC). In the multivariable model, age group, race, and ethnicity were included a priori due to increased risk for HIV among specific demographic groups.

Results The following clinical diagnoses were significant at the bivariate level and were included in the model: pregnancy (OR 1.96 (1.00, 3.84)), hepatitis C (OR 5.73 (1.24, 26.51)), substance use (OR 3.12 (1.12, 8.65)) and sexually transmitted infections (STIs) chlamydia, gonorrhoea, or syphilis. We also a priori included demographic factors that are associated with HIV. Our final model had an AUC of 0.74 and included healthcare site, age group, race, ethnicity, pregnancy, hepatitis C, substance use, and STI diagnosis.

Conclusions Our predictive model showed acceptable discrimination between those who were and were not newly diagnosed with HIV. We identified risk factors such as recent pregnancy, recent hepatitis C diagnosis, and substance use in addition to the traditionally used recent STI diagnosis that can be incorporated by health systems to detect women who are vulnerable to HIV and would benefit from preexposure prophylaxis (PrEP).

Keywords HIV, HIV vulnerability, Prediction model, Electronic Medical Record (EMR), PrEP, Risk factors

*Correspondence:
Eleanor E. Friedman
efriedman@medicine.bsd.uchicago.edu
[1]Department of Medicine, University of Chicago, 5841 S. Maryland Ave, MC 5065, Chicago, IL 60637, USA
[2]Rush University Medical Center, Chicago, IL, USA

Introduction

The HIV epidemic remains a major public health challenge in the United States, with 36,801 people newly diagnosed with HIV in 2019. The majority of these new cases occurred among those born male who identify as male (79%) and via male-to-male sexual contact (66%) [1]. Despite representing lower absolute numbers of new HIV cases, women constituted 19% of incident HIV cases, with the majority of these cases attributed to heterosexual contact [2]. Women, particularly African American/Black women, remain at risk for HIV due to complicated interactions between multiple levels of HIV risk, including community and structural level factors, such as widespread poverty and constrained sexual networks [3, 4]. These complex risk factors among women make it difficult to identify those who are most vulnerable to HIV and in need of pre-exposure prophylaxis (PrEP), a group of medications that are highly effective at preventing HIV [5]. Consequently, a large number of the PrEP-eligible indications for women are based not on their individual behavior, but rather on the behavior of their partners, behavior that may not be known by these women [6, 7].

The difficulty in readily identifying HIV risk factors for women also has implications for technology designed to aid in HIV prevention efforts. In the last decade, multiple models have been built to identify persons at risk for HIV, persons who have not been diagnosed with HIV, or persons who are eligible for PrEP. These models have used data from a variety of countries and have included a wide range of cultural and region-specific factors that influence both HIV care and HIV prevention efforts to identify those most in need of intervention [8–13]. One subset of these models has examined using data from electronic medical records (EMR) to identify persons eligible for PrEP. Models from the United States that either used HIV diagnosis as their outcome or were later validated among people who became HIV positive have shown poor predictive performance among women [8, 14–17]. A consequence of using all new HIV diagnoses or using persons eligible for PrEP is that the models are composed mostly of men, and thus the risk factors identified by these models are also biased toward men. This can be seen in the inclusion of variables such as being male, MSM sexual behavior, and medications for erectile dysfunction [8, 16]–[17]. This finding is particularly unfortunate and paradoxical given that women are more likely than men to have a history of preventative healthcare use, [18] which would result in the creation of EMRs that could be used in HIV risk models.

Obstetrics and Gynecology (OBGYN) and emergency department (ED) settings may be of particular importance in identifying women at high risk for HIV. A recent study found that 82% of women newly diagnosed with HIV had prior healthcare encounters that represented missed opportunities for PrEP initiation, 84% of which occurred in the ED [19]. Most women utilize family or internal medicine doctors for primary care, but OBGYNs are more often used as a type of primary provider among people who are uninsured versus people with insurance (12% vs. 7%) and Black women versus white women (12% vs. 6%) [20]. Additionally, between 2007 and 2010, OBGYN appointments represented 44% of all preventative care visits among women, focusing primarily on reproductive health-related services. Accordingly, OBGYN is a setting well suited to deliver PrEP counseling [21, 22].

The lack of predictive ability for women from previous models that incorporate both men and women may be due to sex-specific HIV risks, as well as the relatively smaller number of women who became HIV positive. The creation of an HIV prevention model specifically among vulnerable women may discover risk factors and result in increased ability to identify women. To determine significant HIV risk factors specifically among women, we established a cohort of women to create an EMR-based risk assessment model for women who underwent routine HIV testing.

Methods

Data were drawn from two large opt-out HIV testing programs embedded within hospitals in Chicago, Illinois [23]. Inclusion criteria for this study included: (1) female sex (i.e., a person whose legal sex was female) (2) underwent testing for HIV between 1/1/2014 to 3/31/2020, and (3) had either an outpatient OBGYN or ED visit from 1/1/2014 to 3/31/2020 at Rush University Medical Center (RUMC) or University of Chicago Medicine (UCM). This study was approved by the Institutional Review Board of UCM. The Institutional Review Board of UCM served as the IRB of record for RUMC. A waiver of consent was sought and given to obtain retrospective EMR data from women tested for HIV at these institutions. We collected information regarding sociodemographic characteristics, risk behaviors, infectious diseases and other diseases, and HIV testing from patients' EMRs. We also included data from laboratory results, medical encounters, and social history forms. Both institutions use the same EMR system (Epic), making data extraction and combination straightforward. Only variables that were consistent (i.e., in terms of both presence/absence and how they were measured) across both systems were combined and included in analysis.

To examine differences in sociodemographic, behavioral, and medical history between those who were newly diagnosed with HIV and those who were not, we selected a subset of women who tested negative for HIV using propensity score matching based on site of care and

number of prior encounters in the healthcare system. Four HIV-negative patients were chosen for each woman newly diagnosed with HIV using optimal fixed ratio matching with a caliper difference of 0.25 of the logit of the propensity score. Propensity score matching was done to minimize differences in healthcare utilization, as well as amount of information available in the EMR between those newly diagnosed with HIV and those without HIV. To ensure that we were using only women newly diagnosed with HIV, we confirmed HIV status (i.e., newly diagnosed vs. an existing case) with the Chicago Department of Public Health (CDPH) [24, 25].

After the analytic sample was chosen, we examined the following variables: sociodemographic variables, including race, ethnicity, age, education level, and zip code; behavioral information, including self-reported gender of sexual partners, condom use, and being sexually active; infectious disease information, including diagnosis with hepatitis C and diagnoses of chlamydia, gonorrhea, or syphilis (hereafter referred to as sexually transmitted infections (STIs)), and HIV testing. Other health diagnoses that we examined included mental health disorders (e.g., mood disorders, personality disorders, psychosis, and anxiety), substance use (e.g., sedatives, stimulants, opiates, and cannabis), alcohol use, and pregnancy. All diagnoses were examined using ICD-9/10 CM codes, with the absence of particular diagnosis codes within the EMR considered a lack of those associated diseases (Additional File Table 1). Ultimately for all variables, we limited data to a two-year retrospective period prior to their last medical encounter (for HIV negative individuals) or two years prior to HIV diagnosis date. However, we did explore multiple time intervals, particularly for pregnancy, which has a progression that can be difficult to identify using some ICD 9 pregnancy diagnosis codes.

Differences in characteristics between women newly diagnosed with HIV and women without HIV were described using chi-square tests or Fisher's exact test as necessary. Bivariate and multivariate logistic regression models were created to examine associations, reporting odds ratios (or adjusted odds ratios) and 95% confidence intervals (OR, aOR, 95%CI). For all models, complete case analysis was used. Several variables were entered into the model a priori based on their importance in contributing to disparities between women newly diagnosed with HIV and women without HIV. These variables included age category, race, and ethnicity. We also included variables that were used in the matching process (healthcare site and maximum number of encounters) and any variables that were found to be statistically significant (p-value ≤ 0.05) at the bivariate level. The top performing final logistic regression model was chosen based on both parsimony as well as having a high area under the curve (AUC) value and was compared to simpler models using the receiver operating curve (ROC) contrast test. All methods were carried out in accordance with relevant guidelines and regulations. All analyses were conducted using SAS version 9.4 (SAS Institute INC. Cary, North Carolina). This manuscript meets the "strengthening the reporting of observational studies in epidemiology" (STROBE) guidelines.

Results

Overall, we identified 55,736 women who underwent HIV testing between 1/1/2014 and 3/31/2020. This included 27,965 women at RUMC and 27,771 women at UCM. Out of this population, we identified 48 women newly diagnosed with HIV, and using propensity score matching, matched 192 women without HIV to these cases by healthcare site and number of prior medical encounters. Propensity score matching reduced the variability in the overall number of previous medical encounters ever from a median of 49 (range 1-5982) to a median of 47.5 (range 5-602).

After matching by site and number of encounters, the analytic sample contained 240 women. These women were mostly African American/Black (67.9%), non-Hispanic/Latina (85.4%), and from either the West (34.2%) or the South (30.8%) sides of Chicago (Table 1). Most women did not have information on their educational attainment (65.0%); however, the most commonly reported level was some college education (12.5%). In terms of medical diagnoses, pregnancy (26.3%) and mental health disorders (24.6%) were seen in about a quarter of the study population; STI diagnoses (0.8%) and hepatitis C diagnosis (2.9%) were much less common. Among women diagnosed with hepatitis C, 71.4% also had diagnosis code indicating substance abuse.

When examining alternate time intervals before HIV diagnosis, we discovered that similar numbers of women had a pregnancy associated diagnosis code zero to six months prior to their HIV diagnosis (17/18, 94.4%) as had a pregnancy diagnosis code zero to 12 months or zero to 24 months before their HIV diagnosis (18/18, 100.0%).

In our sample, African American/Black women had nearly five times the odds of being newly diagnosed with HIV compared to white women (OR 4.98 95%CI (1.47, 16.90)). Women who were of Hispanic or Latina heritage had lower odds of being newly diagnosed with HIV versus those who were not Hispanic/Latina (OR 0.33 95%CI (0.10, 1.14)), although this result was not statically significant. Pregnant women had nearly twice the odds of acquiring HIV than those who were not pregnant (OR 1.96 95%CI (1.00, 3.84)). Similarly, those with hepatitis C had five times the odds of being diagnosed with HIV than those without hepatitis C (OR 5.73 95%CI (1.24, 26.51)). Having been diagnosed with a bacterial STI was significant in Fisher's exact testing (Fisher's exact p-value 0.04).

Table 1 Characteristics of propensity score matched sample, comparing those who are HIV negative and those who are HIV positive (N = 240)

Variable	Total population	HIV- (N = 192)	HIV+ (N = 48)	Chi square p-value
Race				
White	46 (19.7%)	43 (22.4%)	3 (6.3%)	0.005
African American	163 (67.9%)	121 (63.0%)	42 (87.5%)	
Other/Unknown	31 (12.9%)	28 (14.6%)	3 (6.3%)	
Ethnicity				
Non-Hispanic/Latino	205 (85.4%)	160 (83.3%)	45 (93.8%)	0.07*
Hispanic/Latino	35 (14.6%)	32 (16.7%)	3 (6.3%)	
Age				
18–26	63 (26.3%)	46 (24.0%)	17 (35.4%)	0.21
27–35	63 (26.3%)	53 (27.6%)	10 (20.8%)	
36–47	55 (22.9%)	42 (21.9%)	13 (27.1%)	
48 and older	59 (24.6%)	51 (26.6%)	8 (16.7%)	
Education				
Null	156 (65.0%)	126 (65.6%)	30 (62.5%)	0.22*
Some high school	8 (3.3%)	5 (2.6%)	3 (6.3%)	
Complete High School	24 (10.0%)	16 (8.3%)	8 (16.7%)	
Some college	30 (12.5%)	26 (13.5%)	4 (8.3%)	
College degree or higher	22 (9.2%)	19 (9.9%)	3 (6.3%)	
Side of city				
Not Chicago/Unknown	62 (25.85)	51 (26.6%)	11 (22.9%)	0.90*
Central/Northside	22 (9.2%)	18 (9.4%)	4 (8.3%)	
Southside	74 (30.8%)	60 (31.3%)	14 (29.2%)	
Westside	82 (34.2%)	63 (32.8%)	19 (39.6%)	
Healthcare site				
Rush	145 (60.4%)	116 (60.4%)	29 (60.4%)	0.99
UCM	95 (39.6%)	76 (39.6%)	19 (39.6%)	
Number of encounters				
(Median, IQR)	47.5 (15–175)	47.5 (15–175)	47.5 (15–175)	0.99
Hepatitis C				
No	233 (97.1%)	189 (98.4%)	44 (91.7%)	0.03*
Yes	7 (2.9%)	3 (1.6%)	4 (8.3%)	
Substance use				
No	223 (92.9%)	182 (94.8%)	41 (85.4%)	0.05
Yes	17 (7.9%)	10 (5.2%)	7 (14.6%)	
Alcohol use				
No	236 (98.3%)	190 (99.0%)	46 (95.8%)	0.18*
Yes	4 (1.7%)	2 (1.0%)	2 (4.2%)	
Mental health disorder				
No	181 (75.4%)	145 (75.5%)	36 (75.0%)	0.94
Yes	59 (24.6%)	47 (24.5%)	12 (25.0%)	
STIs				
No	238 (9.2%)	192 (100.0%)	46 (95.8%)	0.04*
Yes	2 (0.8%)	0 (0.0%)	2 (4.2%)	
Pregnancy				
No	177 (73.8%)	147 (76.6%)	30 (62.5%)	0.05
Yes	63 (26.3%)	45 (23.4%)	18 (37.5%)	
Male partner				
Not noted/no	152 (63.3%)	121 (63.0%)	31 (64.6%)	0.84
Yes	88 (36.7%)	71 (37.0%)	17 (35.4%)	
Condom use				
Not noted/no	224 (93.3%)	181 (94.3%)	43 (89.6%)	0.24
Yes	16 (6.7%)	11 (5.7%)	5 (10.4%)	
Sexually active data present				
No	109 (45.4%)	88 (45.8%)	21 (43.8%)	0.80
Yes	131 (54.6%)	104 (54.2%)	27 (56.3%)	

*Fisher's exact test was used due to small sample size.

Women who had diagnoses of substance use had three times the odds of being newly diagnosed with HIV compared to women who did not have substance use in their medical history (OR 3.12 95%CI (1.12, 8.65)) (Table 2).

The final model included the following variables: STIs, substance use, hepatitis C, pregnancy, race, ethnicity, age group, healthcare site and number of encounters, with an AUC of 0.74 95%CI (0.67, 0.81). This final model performs significantly better than a baseline model consisting only of matching factors and STI diagnoses in the past two years (AUC 0.54 95%CI (0.44, 0.63) (ROC contrast test p-value=0.004)), as well as a model consisting only of matching factors and demographic factors including race, age group, and ethnicity (AUC 0.69 95%CI (0.61, 0.77) (ROC contrast test p-value=0.03)) (Fig. 1).

Discussion

In this study we created a sex-specific model that identified factors associated with HIV incidence among women. Our final model had modest performance, with an AUC of 0.74 95%CI (0.67, 0.81), and compared favorably to both the baseline model and a secondary model. Most notably, the final model identified three relatively novel factors - pregnancy in the last two years, hepatitis C diagnosis in the last two years, and diagnosis of substance use in the last two years - that may assist in identifying women at increased risk for HIV using EMR data. This information may also assist individual providers in identifying women who may need targeted and more frequent HIV screening or women who may be good candidates for PrEP.

Our paper is not the first to use data from women to create a specific model for HIV. Tang et al. created three different machine learning models for populations with different risk factors: injection drug users (IDU), MSM, and female sex workers. Tang et al. also found that hepatitis C test results were important in predicting HIV status among female sex workers [13]. Similarly, Burns et al. created two machine learning models to identify factors associated with HIV among women. These models identified historical drug use but not hepatitis C positive viral testing, as positively associated with incident HIV diagnoses [26].

There may be multiple reasons for the identification of recent pregnancy and recent hepatitis C diagnosis as significant, but relatively novel risk factors for HIV

Table 2 Unadjusted and adjusted odds ratios and 95% confidence intervals for a priori variables, or those that were significant in chi-square or Fisher's exact testing

Variable	Odds ratio OR (95% CI)	Wald chi-square p - value	Adjusted odds ratio aOR (95% CI)	Wald chi-square p - value
Healthcare site				
RUMC	REF	0.99	REF	0.67
UCM	1.00 (0.52, 1.91)		1.20 (0.52, 2.76)	
Number of encounters	1.00 (1.00, 1.00)	0.99	1.00 (1.00, 1.00)	0.59
Race				
White	REF	0.003	REF	0.03
African American	4.98 (1.47, 16.90)	0.59	4.87 (1.21, 19.53)	0.57
Other/Unknown	1.54 (0.29, 8.16)		1.39 (0.23, 8.35)	
Ethnicity				
Non-Hispanic/Latino	REF	0.08	REF	0.89
Hispanic/Latino	0.33 (0.10, 1.14)		0.89 (0.16, 4.97)	
Age				
18–26	REF	0.42	REF	0.83
27–35	0.51 (0.21, 1.23)	0.37	0.62 (0.24, 1.65)	0.37
36–47	0.84 (0.36, 1.93)	0.17	0.88 (0.34, 2.31)	0.12
≥48	0.42 (0.17, 1.08)		0.37 (0.12, 1.17)	
Substance use				
No	REF	0.03	REF	0.34
Yes	3.12 (1.12, 8.65)		1.94 (0.49, 7.64)	
Hepatitis C				
No	REF	0.02	REF	0.15
Yes	5.73 (1.24, 26.51)		4.54 (0.59, 34.88)	
Pregnancy				
No	REF	0.05	REF	0.21
Yes	1.96 (1.00, 3.84)		1.69 (0.74, 3.86)	
STIs				
No	NA	NA	NA	NA
Yes				

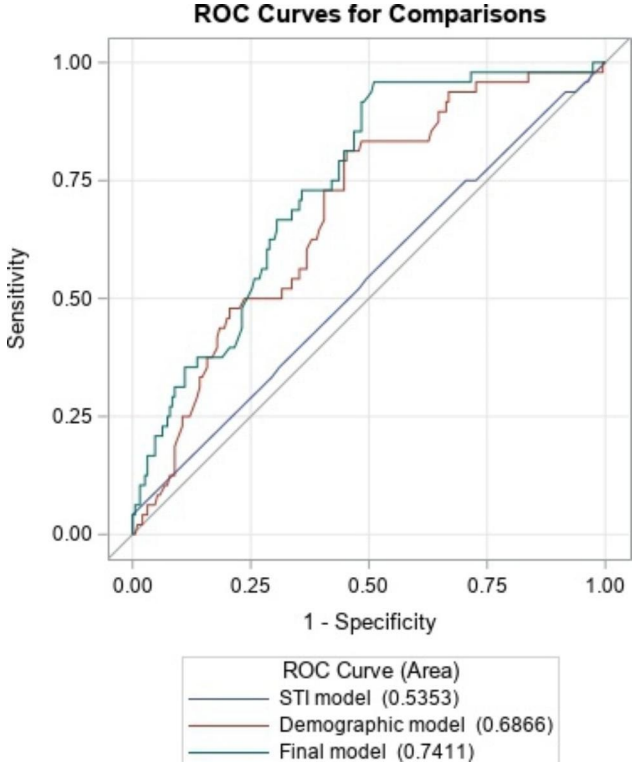

Fig. 1 Area under the curve results from various logistic regression models for the outcome of being newly diagnosed with HIV. (STI model: STIs, healthcare site, number of encounters. Demographic model: Race, ethnicity, age group, healthcare site, number of encounters. Final model: STIs, hepatitis C, pregnancy, race, ethnicity, age group, healthcare site, number of encounters)

acquisition among women. Pregnancy is a high specificity variable that could serve as a proxy for condomless (or condom failure) sex that increases the risk of potential HIV transmission. Late pregnancy and the postpartum period have also been found to be times when HIV transmission per sex act is increased among serodiscordant couples, [27] with the suggestion of a biologic reason for increased risk. It is also possible that pregnancy and early childrearing time periods capture changes in partner behavior that increase HIV risk, such as less frequent condom use and men being more likely to seek other partners during pregnancy, both of which have been reported in data from sub-Saharan Africa [28]. The other identified risk factor, hepatitis C, may be a proxy for IDU, with historical data suggesting that up to 90% of chronic injection drug users (i.e., those who have injected drugs for 10 or more years) are diagnosed with hepatitis C [29, 30]. Hepatitis C diagnosis being linked to IDU is supported by the fact that the majority of hepatitis C cases within our cohort also had EMR diagnosis codes indicating substance use. Additionally, hepatitis C infection may also be a proxy for either anal sex, or rough vaginal unprotected sex, as hepatitis C is transmitted via blood. Previous work conducted in a sample of mostly Black women in Chicago found 16% of participants self-reported anal sex in the last six months, most of which was condomless [31]. Sexual transmission of hepatitis C has been shown to occur among MSM, although documentation of sexual transmission to women is not as robust [32]. Our finding that hepatitis C is a risk factor for new HIV diagnosis may indicate that women in our study had partners who injected drugs [33]. It is also possible that because hepatitis C transmission appears more common among sexual partners with HIV, hepatitis C diagnosis in our model is somewhat collinear with HIV infection itself [34].

The recognition of recent pregnancy, in particular, as a factor associated with HIV acquisition may be helpful for expanding PrEP discussions to more women who are engaging in unprotected sex. This expansion also fits with the new Centers for Disease Control and Prevention (CDC) PrEP guidelines that suggest all sexually active adult and adolescent patients should receive information about PrEP [35]. Public health practitioners should consider pregnant women and women who have recently given birth to be especially vulnerable to HIV infection. HIV testing during first and third trimesters has been recommended by the American College of Obstetricians and Gynecologists (ACOG) for women who are negative after the first test and "known to be at high risk of acquiring HIV infection," including those "who reside in jurisdictions with elevated HIV incidence." [36]. This has relevance for our population, as Chicago is within Cook County, a priority area under the Ending the HIV Epidemic (EHE) plan, identifying it as a high incidence area [37]. Based on this ACOG recommendation, it is likely that the women in our study underwent both first and third trimester HIV screening/rapid screening during labor and delivery. Although we do not know what gestational week women in our population received pregnancy diagnosis codes at RUMC or UCM, the fact that so few HIV cases were gained with expansion of the lookback period suggests that the majority of these women were either pregnant or had recently given birth at the time of their HIV diagnosis. Further work should examine if an additional screening for women in the postnatal period is needed to identify those who may still be vulnerable to HIV. Lastly, although there is evidence to suggest that pregnancy is associated with HIV transmission, it is also possible that due to ACOG screening recommendations pregnancy is serving as an indicator of increased HIV testing rather than HIV transmission itself. Regardless, this finding reemphasizes the importance of HIV screening during pregnancy to identify persons who have been newly infected with HIV or who have been undiagnosed until pregnancy screening.

Our model also identified risk factors that have been previously found to be associated with acquisition of

HIV. Recent history of a bacterial STI has been used as an eligibility criterion for PrEP by the CDC since 2017, [38] due to plausible social and biologic mechanisms by which bacterial STIs could increase HIV risk. Although our lookback period was longer than that used by the CDC (two years vs. six months), we found that women with a history of STIs were more likely to be eventually diagnosed with HIV. In fact, both the women in our study who were diagnosed with STIs were also eventually diagnosed with HIV. We also found that drug use was associated with being newly diagnosed with HIV, a finding that has been previously reported among urban women at risk for HIV who suffer from the syndemic of drug use, violence and sexual risk behaviors [39, 40]. Although all of these syndemic components are stigmatized and unlikely to be revealed by patients during a visit, drug use is perhaps the most medically observable and the easiest to record in structured EMR fields. It is therefore possible that drug use in our model serves as a marker for one or more of these syndemic components.

Our findings also reinforce previously established racial disparities in incident HIV cases in the United States. In our study, the women newly diagnosed with HIV were more likely to be Black. Additionally, our general patient population was more likely to be from the West or South sides of Chicago, which are areas with increased minority populations. This inequity reinforces that even though the CDC recommends that all sexually active adults receive information about PrEP, some populations have an increased need, including minority women within the city of Chicago.

This study has some limitations. Despite combining HIV screening information from two major urban hospital opt-out screening programs, the number of incident cases among women was relatively small. The small sample size limited the complexity of the models we could create, as well as the number of risk factors that could be entered into models. We only included three STIs (chlamydia, gonorrhea, and syphilis), but there are other STIs like Human papillomavirus (HPV) that may increase risk for HIV transmission [41]. Unfortunately, we did not measure HPV infection in our study population. Inclusion of other STIs into our model would likely have increased the number of people with diagnosed STIs, and may have either strengthened or weakened our finding that STI diagnoses are a risk factor for new HIV diagnoses. Our data was based on women who had an ED or OBGYN visit with HIV screening at two academic medical centers. Our results may not be generalizable to different patient populations (e.g., migrant women) or to women who receive HIV screening in other settings. It is possible that our study identified risk factors that are different from those that would be identified had we included women who use their primary care provider for STI screening or who are tested for HIV in other hospital departments (e.g., inpatient). It is difficult to determine the directionality of this bias on risk factors we did identify. Our data also contained variables that were poorly reported or completely missing from the EMR, particularly variables on condom use, male sexual partners, sexual activity, and education level. This incomplete information made it difficult to assess these factors, although they may have added greatly to the predictive power of the model. Unfortunately, it is likely that the low level of documentation seen in our EMR systems is reflective of larger unwillingness and lack of time among healthcare providers to document this information, especially for women who appear to be at low risk or during visits that are unrelated to sexual health. To increase ability to use these variables in EMR modeling applications, interventions that normalize discussions of PrEP and sexual health should be promoted. Lastly, although we believe that the majority of our sample was composed of cisgender women (women both born female and who currently identify as female), the lack of EMR variables that specify both birth sex and current gender identity prevented us from ensuring our sample contained only ciswomen.

Additional studies to identify persons at high risk for HIV should consider the use of stratified models, as they may be necessary to fully determine the ways in which different people are vulnerable to HIV. Future work to identify risk factors for women should be conducted among a consortium of healthcare centers or a large healthcare network that would permit the creation of a larger cohort of newly positive women. A larger sample size would allow for the use of a machine learning approach, and also support a more thorough examination of all possible EMR risk factors.

Conclusion

Overall, this study created a model with acceptable discrimination to determine women with new HIV diagnoses (AUC of 0.74 95%CI (0.67, 0.81)). We identified risk factors including pregnancy, hepatitis C diagnosis, substance use diagnosis and STI diagnosis in a two-year period prior to HIV diagnosis that can be used to identify women who are vulnerable to HIV and would benefit from PrEP.

Abbreviations

MSM	Men who have sex with men
AUC	Area under the curve
STI	Sexually transmitted infections
EMR	Electronic Medical Record
PrEP	Pre-exposure prophylaxis
OBGYN	Obstetrics and Gynecology
ED	Emergency Department

UCM	University of Chicago Medicine
RUMC	Rush University Medical Center
CDPH	Chicago Department of Public Health
OR	Odds ratios
aOR	Adjusted odds ratios
95%CI	95% Confidence interval
EHE	Ending the HIV Epidemic
IDU	Injection drug use
CDC	Centers for Disease Control and Prevention
ACOG	American College of Obstetricians and Gynecologists

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12905-023-02460-7.

Additional File Table 1: ICD9/10 codes used to determine disease diagnoses

Acknowledgements
Not applicable.

Authors' contributions
E.F., J.R., S.D., S.S., and B.S., designed the research study. E.F, J.M. and E.K. performed the data analysis. E.F and S.D. wrote the paper.

Funding
This work was supported in part by the University of Chicago Institute of Translational Medicine (ITM) 5UL1TR002389-04. ITM had no role in the design, of the study nor in the collection, analysis, and interpretation of data or in the writing of the manuscript.

Data availability
The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
All experimental protocols were approved by the ethics committee/Institutional Review Board of the University of Chicago. The need for informed consent was waived by the ethics committee/Institutional Review Board of the University of Chicago, because of the retrospective nature of the study (IRB20-0274-AM005).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 October 2022 / Accepted: 3 June 2023
Published online: 16 June 2023

References

1. CDC. *HIV Surveillance Report, 2019* 2021.
2. CDC. HIV and Women. 2022.
3. Brawner BM. A multilevel understanding of HIV/AIDS disease burden among african american women. J Obstet Gynecol Neonatal Nurs Sep-Oct. 2014;43(5):633–43.
4. Frew PM, Parker K, Vo L, et al. Socioecological factors influencing women's HIV risk in the United States: qualitative findings from the women's HIV SeroIncidence study (HPTN 064). BMC Public Health Aug. 2016;17(1):803.
5. CDC, About. PrEP https://www.cdc.gov/hiv/basics/prep/about-prep.html. Accessed July 15, 2022.
6. Raifman J, Sherman SG. US Guidelines that Empower Women to prevent HIV with Preexposure Prophylaxis. Sex Transm Dis Jun. 2018;45(6):e38–9.
7. Bradley ELP, Hoover KW. Improving HIV Preexposure Prophylaxis implementation for women: Summary of Key Findings from a discussion series with Women's HIV Prevention experts. Womens Health Issues Jan-Feb. 2019;29(1):3–7.
8. Xu X, Ge Z, Chow EPF et al. A machine-learning-based risk-prediction Tool for HIV and sexually transmitted Infections Acquisition over the Next 12 months. J Clin Med Mar 25 2022;11(7).
9. 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 Jul. 2021;31(1):159.
10. Hailu TG. Comparing Data Mining Techniques in HIV Testing Prediction. Intell Inform Manage. 2015;7(3):153–80.
11. Haukoos JS, White DAE, Rowan SE, et al. Development of a 2-step algorithm to identify emergency department patients for HIV pre-exposure prophylaxis. Am J Emerg Med Jan. 2022;51:6–12.
12. Ahlström MG, Ronit A, Omland LH, Vedel S, Obel N. Algorithmic prediction of HIV status using nation-wide electronic registry data. eClinicalMedicine 2019;17.
13. Tang D, Zhang M, Xu J et al. Application of Data Mining Technology on Surveillance Report Data of HIV/AIDS High-risk group in Urumqi from 2009 to 2015. Complexity 2018/12/10 2018;2018:9193248.
14. Ridgway JP, Almirol EA, Bender A, et al. Which patients in the Emergency Department should receive Preexposure Prophylaxis? Implementation of a Predictive Analytics Approach. AIDS Patient Care STDS May. 2018;32(5):202–7.
15. Ridgway JP, Friedman EE, Bender A, et al. Evaluation of an electronic algorithm for identifying Cisgender Female Pre-Exposure Prophylaxis candidates. AIDS Patient Care STDS Jan. 2021;35(1):5–8.
16. 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 Oct. 2019;6(10):e688–95.
17. Krakower DS, Gruber S, Hsu K, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. Lancet HIV Oct. 2019;6(10):e696–e704.
18. Hing E, Albert M. State Variation in Preventive Care visits, by patient characteristics. NCHS Data Brief Jan. 2012;2016(234):1–8.
19. Smith DK, Chang MH, Duffus WA, Okoye S, Weissman S. Missed Opportunities to prescribe Preexposure Prophylaxis in South Carolina, 2013–2016. Clin Infect Dis Jan. 2019;1(1):37–42.
20. Long M, Frederiksen B, Ranji U, Salganicoff A. *Women's Health Care Utilization and Costs: Findings from the 2020 KFF Women's Health Survey* KFF 2021.
21. Stormo AR, Saraiya M, Hing E, Henderson JT, Sawaya GF. Women's clinical Preventive Services in the United States: who is doing what? JAMA Intern Med. 2014;174(9):1512–4.
22. Ralph JA, Westberg SM, Boraas CM, Terrell CA, Fischer JR. PrEP-aring the General Gynecologist to offer HIV pre-exposure Prophylaxis. Clin Obstet Gynecol 9900:https://doi.org/10.1097/GRF.0000000000000713.
23. Almirol EA, McNulty MC, Schmitt J, et al. Gender differences in HIV Testing, diagnosis, and linkage to Care in Healthcare Settings: identifying african American Women with HIV in Chicago. AIDS Patient Care STDS Oct. 2018;32(10):399–407.
24. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc Dec. 2013;20(e2):e311–318.
25. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. Bmj Apr. 2018;30:361:k1479.
26. Burns CM, Pung L, Witt D, et al. Development of a human immunodeficiency Virus Risk Prediction Model using Electronic Health Record Data from an Academic Health System in the Southern United States. Clin Infect Dis Jan. 2023;13(2):299–306.
27. Thomson KA, Hughes J, Baeten JM, et al. Increased risk of HIV Acquisition among Women throughout pregnancy and during the Postpartum period: a prospective per-coital-act analysis among women with HIV-Infected Partners. J Infect Dis Jun. 2018;5(1):16–25.
28. Graybill LA, Kasaro M, Freeborn K, et al. Incident HIV among pregnant and breast-feeding women in sub-saharan Africa: a systematic review and meta-analysis. Aids Apr. 2020;1(5):761–76.
29. Thomas DL, Vlahov D, Solomon L, et al. Correlates of hepatitis C virus infections among injection drug users. Med (Baltimore) Jul. 1995;74(4):212–20.

30. Lorvick J, Kral AH, Seal K, Gee L, Edlin BR. Prevalence and duration of hepatitis C among injection drug users in San Francisco, Calif. Am J Public Health Jan. 2001;91(1):46–7.
31. Hirschhorn LR, Brown RN, Friedman EE, et al. Black Cisgender Women's PrEP Knowledge, Attitudes, Preferences, and experience in Chicago. J Acquir Immune Defic Syndr Aug. 2020;15(5):497–507.
32. Tohme RA, Holmberg SD. Is sexual contact a major mode of hepatitis C virus transmission? Hepatol Oct. 2010;52(4):1497–505.
33. Frederick T, Burian P, Terrault N, et al. Factors associated with prevalent hepatitis C infection among HIV-infected women with no reported history of injection drug use: the Women's Interagency HIV Study (WIHS). AIDS Patient Care STDS Nov. 2009;23(11):915–23.
34. Alipour A, Rezaianzadeh A, Hasanzadeh J, Rajaeefard A, Davarpanah MA. Sexual transmission of Hepatitis C Virus between HIV infected subjects and their main Heterosexual Partners. Hepat Mon. 2013;13(11):e13593.
35. CDC. *Preexposure Prophylaxis for the Prevention of HIV Infection in the United States – 2021 Update Clinical Practice Guideline* 2021.
36. ACOG. ACOG Committee Opinion No. 752: prenatal and Perinatal Human Immunodeficiency Virus Testing. Obstet Gynecol Sep. 2018;132(3):e138–42.
37. CDC, Jurisdictions. https://www.cdc.gov/endhiv/about.html. Accessed July 15, 2022.
38. CDC. *Preexposure Prophylaxis for the Prevention of HIV Infection in the United Staes – 2017 Update Clinical Practice Guideline* 2017.
39. Lounsbury AWB, Palma DW. Importance of substance use and violence in psychosocial syndemics among women with and at-risk for HIV. AIDS Care Oct. 2016;28(10):1316–20.
40. Koblin BA, Grant S, Frye V, et al. HIV sexual risk and syndemics among women in three urban Areas in the United States: analysis from HVTN 906. J Urban Health Jun. 2015;92(3):572–83.
41. Zayats R, Murooka TT, McKinnon LR. HPV and the risk of HIV Acquisition in Women. Front Cell Infect Microbiol. 2022;12:814948.

Publisher's note