

RESEARCH

Open Access



Epidemiological and viral characteristics of undiagnosed HIV infections in Botswana

Lynnette Bhebhe¹, Sikhulile Moyo^{1,2}, Simani Gaseitsiwe^{1,2}, Molly Pretorius-Holme², Etienne K. Yankinda¹, Kutlo Manyake¹, Coulson Kgathi¹, Mompoti Mmalane¹, Refeletswe Lebelonyane³, Tendani Gaolathe¹, Pamela Bachanas⁴, Faith Ussery⁴, Mpho Letebele⁵, Joseph Makhema^{1,2}, Kathleen E. Wirth^{4,5}, Shahin Lockman^{1,2,6}, Max Essex², Vlad Novitsky^{2,7} and Manon Ragonnet-Cronin^{8,9*}

Abstract

Background: HIV-1 is endemic in Botswana. The country's primary challenge is identifying people living with HIV who are unaware of their status. We evaluated factors associated with undiagnosed HIV infection using HIV-1 phylogenetic, behavioural, and demographic data.

Methods: As part of the Botswana Combination Prevention Project, 20% of households in 30 villages were tested for HIV and followed from 2013 to 2018. A total of 12,610 participants were enrolled, 3596 tested HIV-positive at enrolment, and 147 participants acquired HIV during the trial. Extensive socio-demographic and behavioural data were collected from participants and next-generation sequences were generated for HIV-positive cases. We compared three groups of participants: (1) those previously known to be HIV-positive at enrolment ($n = 2995$); (2) those newly diagnosed at enrolment ($n = 601$) and (3) those who tested HIV-negative at enrolment but tested HIV-positive during follow-up ($n = 147$). We searched for differences in demographic and behavioural factors between known and newly diagnosed group using logistic regression. We also compared the topology of each group in HIV-1 phylogenies and used a genetic diversity-based algorithm to classify infections as recent (< 1 year) or chronic (≥ 1 year).

Results: Being male (aOR = 2.23) and younger than 35 years old (aOR = 8.08) was associated with undiagnosed HIV infection ($p < 0.001$), as was inconsistent condom use (aOR = 1.76). Women were more likely to have undiagnosed infections if they were married, educated, and tested frequently. For men, being divorced increased their risk. The genetic diversity-based algorithm classified most incident infections as recent (75.0%), but almost none of known infections (2.0%). The estimated proportion of recent infections among new diagnoses was 37.0% ($p < 0.001$).

Conclusion: Our results indicate that those with undiagnosed infections are likely to be young men and women who do not use condoms consistently. Among women, several factors were predictive: being married, educated, and testing frequently increased risk. Men at risk were more difficult to delineate. A sizeable proportion of undiagnosed infections were recent based on a genetic diversity-based classifier. In the era of "test and treat all", pre-exposure prophylaxis may be prioritized towards individuals who self-identify or who can be identified using these predictors in order to halt onward transmission in time.

Keywords: HIV, Undiagnosed infection, Phylogenetics, Recent HIV infection

*Correspondence: manonragonnet@uchicago.edu

⁸ MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK

Full list of author information is available at the end of the article

Background

Botswana was among the first countries to reach the Joint United Nations Programme on HIV/AIDS (UNAIDS) 90-90-90 targets, defined as 90% of people



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

living with human immunodeficiency virus (HIV) aware of their status, 90% of those diagnosed on antiretroviral treatment and 90% of those on treatment, virally suppressed [1]. Botswana reached 93-93-98 in 2021 [2]. However, the estimated prevalence of human immunodeficiency virus type 1 (HIV-1) ranges between 18.5 and 21% among adults in the general population and HIV-1 incidence ranges from 0.59 to 1.35% [3–5]

One of the main challenges in curbing transmission is identifying individuals with HIV who are unaware of their HIV status in order to link them to care. Early HIV symptoms are not sufficiently clear or specific to warrant testing, thus there is often a delay between HIV infection and diagnosis. This delayed diagnosis poses an obstacle to the elimination of HIV, as early HIV infection is associated with a disproportionate amount of onward transmission [6–9]. In Sub-Saharan Africa, several factors have been associated with poor HIV testing behaviours and undiagnosed HIV infection [2, 10–16]. These include age, sex, marital status, religion, education, employment status, sexual experience, condom usage [10–12, 17–19]. However, many prior studies have focused on deliberately delineated epidemiological groups such as adolescents, older adults (>50 years old), or public sector clinic attendees.

To reduce the incidence of HIV-1 in Botswana, it would help to deploy targeted testing strategies for individuals at high risk of being infected or of transmitting the virus. Phylogenetic analysis can provide insights into recent and ongoing transmission events. Thus, it may be possible to utilise phylogenetic analysis to identify high-risk individuals, then describe their demographic or behavioural characteristics. In addition, viral genetic data are informative regarding stage of HIV infection [20–22], so that transmissions in a phylogeny can be interpreted in the context of time since HIV infection.

The Botswana Combination Prevention Project (BCPP) evaluated whether a wide-ranging strategy would reduce HIV incidence over time [3, 4]. 20% of the households across 30 communities were systematically tested for HIV between 2013 to 2018. The study found 601 infections among participants who reported no prior positive test at the start of the study, and 147 individuals seroconverted during follow-up [3]. All study participants completed detailed questionnaires, and HIV-1 full genome sequences were generated from all participants with HIV. Using data from the BCPP study, we sought to evaluate factors associated with undiagnosed HIV-1 infection in Botswana, and to estimate time since infection among new diagnoses.

Methods

Study population

The BCPP study enrolled participants from 2013 to 2018 [3]. Approximately 20% of households within 30 communities were randomly selected and eligible participants were enrolled. Each community had an average population size of 5855 for a total trial population of 175,664 [23]. Individuals aged 16 or older completed questionnaires and had blood drawn [3]. We collected questionnaire data in two broad categories: (1) Socio-demographics and community environment: participant and community variables, residency and mobility, education, employment, (2) HIV risk behaviour: HIV testing history and sexual behaviour (see [24] for protocols, questionnaires and data). Negative participants were followed up with yearly HIV tests.

HIV-1 full genome sequencing

HIV samples from all positively diagnosed participants were sequenced, regardless of antiretroviral therapy (ART) status and viral load. The majority of people with known HIV were already on ART with viral suppression defined as viral load ≤ 400 copies/mL; in these instances, integrated virus was sequenced from viral RNA and proviral DNA templates. Next-generation sequencing (NGS) was performed by the BioPolymers Facility at Harvard Medical School [25] and through collaboration with the PANGEA HIV consortium [26–28] using Illumina platforms MiSeq and HiSeq, as previously described [29–31]. In brief, the nucleic acids were reverse-transcribed and PCR amplified. Amplicons were pooled in equimolar amounts for Illumina library preparation. Sequence assembly was performed de novo using SPAdes version 2.4.0. The HIV-1 reference strain, HXB2 (NC_001802), was used for sequence alignment and a consensus sequence was generated using Abacas version 1.3.1 and MUMmer version 3.2. Next, sequence reads were mapped against the consensus sequence using SMALT version 0.5.0 [31].

HIV consensus sequences were subtyped using COMET [32], and only HIV-1 subtype C sequences were included in our analysis (accounting for >99% of BCPP sequences). Based on the NGS reads, we were provided with nucleotide frequency files for each patient, detailing the relative frequency of each nucleotide at each site in the alignment. HIV sequences and basic demographic and clinical data are available upon request to the PANGEA consortium [28].

Statistical analysis

We stratified participants with HIV-1 into three groups as follows: 2995 with previously diagnosed HIV (henceforth referred to as “known cases”), 601 persons with

newly diagnosed HIV at enrolment (“new cases”), and 147 persons who seroconverted to HIV-positive during follow-up (“incident cases”). As study participants who tested negative at enrolment were then tested yearly, incident cases were known to have been infected for ≤ 1 year when they were diagnosed. Baseline characteristics and descriptive statistics for all participants have been previously described [3].

We compared responses to the questionnaire across our three groups. We included variables known to be associated with HIV infection and variables known to be associated with undiagnosed infection. We employed logistic regression to compare the characteristics of new cases versus known cases to identify predictors associated with undiagnosed infection. For each variable of interest, we performed univariate analyses comparing the two groups (new cases versus known cases). A single multivariable analysis was fit including demographic and behavioural predictors that were significant ($p < 0.05$) in the univariate analyses. We did not adjust for missing data in the multivariable analysis, and the proportion of complete cases in the data was 53.9%. Demographic predictors evaluated included sex, age, marital status, number of children, religious affiliation, education, and employment status. The behavioural predictors assessed were previous number of HIV tests, sexual activity (yes/no), number of partners, partner concurrency, condom use, condom use frequency in the past year, number of nights spent away from home and partner’s HIV status. Logistic regression was adjusted by community ($n = 30$) with a random effect using a model with robust standard errors in R. All variables were analysed as categorical variables except for age, which was analysed as a continuous predictor in the primary analysis. A sensitivity analysis was run with age as a categorical variable with five age ranges: 16–24, 25–34, 35–44, 45–54 and 55–64 years old. Finally, we reran the analysis disaggregated by sex, in order to disentangle differential predictors for men and women.

We sought to use viral genetic data to determine whether the time from infection to diagnosis varied significantly between the three groups (new, incident, and known cases). First, we compared the assigned stage of infection based on within-host genetic diversity ($<$ or ≥ 1 year) for each group, using a χ^2 test. We then compared the recency probability distributions between the three groups using a Kruskal–Wallis test. These comparisons were conducted for participants for whom deep sequencing nucleotide frequency data were available ($n = 1867$). Finally, we compared the length of terminal branch lengths across the groups, and across each HIV-1 gene (*gag*, *pol*, *env*), in a pairwise manner, using Kolmogorov–Smirnov (KS) tests. These comparisons were

conducted for participants for whom consensus genetic sequences were available for at least one gene ($n = 2872$). This dataset included 2339 known cases, 399 new cases and 134 incident cases. To account for multiple non-independent comparisons (across different genes), we used Bonferroni’s correction to assess statistical significance where appropriate. All statistical analyses were conducted in R (version 3.6.0) [33].

Phylogenetic analysis

For participants for whom viral genetic data were available ($n = 2872$), we constructed phylogenies separately for each HIV gene region: *gag*, *pol* and *env*. Sequences were available for 2339 known cases, 399 new cases and 134 incident cases. We compared the viral characteristics between those two groups (known vs. new cases), and a third: those diagnosed with incident infections during the BCPP trial ($n = 147$). Because this latter group were negative at the start of the trial, then tested yearly, we knew their infections were < 1 year when they were diagnosed. We wanted to use our two reference groups (known older cases, and incident cases) to evaluate how long those diagnosed at the start of trial were likely to have been infected before they were diagnosed.

Maximum likelihood phylogenies were reconstructed using RaxML [34] under a GTR model with four gamma rates. Phylogenies included sequences from an additional 3277 patients from Botswana clinics, to serve as local controls. Final phylogeny sizes were: *gag* ($n = 5631$), *pol* ($n = 6084$) and *env* ($n = 5840$). We time-resolved the phylogenies using the treedater package, available in R [35], using sample times as tip dates. For each tip, we extracted terminal branch lengths (measured in time) using the “Analyses of Phylogenetics and Evolution” (APE) [36] and phytools [37] R packages. We compared the distributions of terminal branch lengths across our three groups (known cases, new cases, incident cases) using Kolmogorov–Smirnov tests.

Inference of stage of HIV infection

For each study participant for whom deep sequencing nucleotide frequency data were available ($n = 1867$), we calculated the probability (0–1) of their infection being recent (< 1 year) based on within-host genetic diversity, demographic (age, sex) and clinical (treatment status, viral load) predictors using an xgboost gradient boosting [38] machine learning algorithm. The machine learning classifier is trained on a dataset of known recent (< 1 year) and chronic (≥ 1 year) infections to classify stage of HIV infection for individuals in each of our three groups. This analysis was conducted on all individuals for whom NGS coverage was sufficient to derive site-specific nucleotide frequency distributions ($n = 1867$). We selected the

threshold that optimized for accuracy (the highest number of overall correct classifications). The algorithm has been previously developed for and trained on the BCPP dataset on participants with known duration of infection [39]. For each participant in the present study, we used the algorithm to predict the probability of recency. We carefully excluded individuals comprised in the present study from the dataset used to train the classifier.

Results

Participant demographics

In total, 12,610 people were enrolled in BCPP. Of these, 8050 (63.8%) were women and 4560 (36.2%) were men. A total of 3596 (29%) participants tested positive for HIV-1 at enrolment and 147 participants seroconverted during the study.

Predictors associated with undiagnosed HIV infection at baseline

Demographic predictors of undiagnosed HIV infection at baseline We compared new cases to known cases to identify factors associated with undiagnosed infection. The analysis was conducted on the group as a whole and

repeated on women and men separately, demonstrating stark differences by sex. Men were more likely than women to have an undiagnosed HIV infection (Additional file 1: Table S1). As participants of both sexes aged, they were less likely to have an undiagnosed HIV infection (aOR=0.94 per year, $p < 0.001$; Additional file 1: Table S1). But for women, only those aged 16–24 had an increased chance of an undiagnosed infection (aOR=4.63, $p < 0.001$), while for men age groups 16–24, 25–34, 35–44 were all more likely to harbour undiagnosed infections than the reference group (55–64; Table 1). For women, being single (aOR=0.73, $p < 0.001$) and divorced or widowed (aOR=0.47, $p < 0.001$) significantly reduced the odds of having an undiagnosed infection as compared to being married. This effect was inversed for men, where being single (aOR=1.72, $p < 0.001$) or divorced (aOR=3.95, $p < 0.05$) increased the odds of having an undiagnosed infection as compared to being married. Next, not having children increased the odds of a woman harbouring an undiagnosed infection compared to having children (OR=0.53, $p < 0.001$), and the more children a woman had the less likely she was to have an undiagnosed infection. However, due to collinearity with sex (the ques-

Table 1 Logistic regression for demographic characteristics associated with undiagnosed infection by sex

Variable	Category	Women (n = 383 newly diagnosed, n = 2252 known cases)		Men (n = 218 newly diagnosed, n = 743 known cases)	
		OR (95% CI) ^a	aOR (95% CI) ^b	OR (95% CI)	aOR (95% CI)
Age	16–24 years	5.42*** (4.29–6.89)	4.63*** (2.63–8.18)	4.58*** (2.73–7.98)	12.7** (1.32–121.7)
	25–34 years	0.99 (0.83–1.17)	0.81 (0.51–1.28)	4.57*** (3.25–6.43)	5.68*** (2.54–12.7)
	35–44 years	0.29*** (0.25–0.35)	0.36*** (0.22–0.57)	2.64*** (1.94–3.59)	3.16** (1.45–6.87)
	45–54 years	0.54*** (0.45–0.65)	0.95 (0.66–1.39)	0.98 (0.7–1.35)	1.0 (0.48–2.14)
	55–64 years	Ref		Ref	
Marital status	Married	Ref		Ref	
	Single/Never married	0.73*** (0.63–0.84)	0.45*** (0.33–0.6)	2.58*** (2.0–3.27)	1.72 (1.0–2.05)
	Divorced/Widowed	0.47*** (0.38–0.58)	0.25*** (0.16–0.39)	1.31 (0.79–2.19)	3.95* (1.3–11.98)
Number of children ^c	None	Ref			
	1–5 children	0.53** (0.33–0.87)	–	–	–
	≥ 5 children	0.36*** (0.21–0.64)	–	–	–
Religious affiliation	Affiliated with religion	Ref		Ref	
	No religious affiliation	1.1 (0.98–1.25)	1.33* (1.06–1.67)	0.85 (0.7–1.04)	0.8 (0.57–1.11)
Education level	Tertiary	Ref		Ref	
	Senior secondary	0.93 (0.73–1.18)	0.64* (0.44–0.92)	1.5 (0.93–2.45)	1.55 (0.75–3.19)
	Junior secondary	0.29*** (0.24–0.35)	0.37*** (0.26–0.53)	1.91*** (1.31–2.75)	1.42 (0.78–2.60)
	Primary	0.21*** (0.17–0.26)	0.38*** (0.23–0.61)	0.71 (0.49–1.04)	1.19 (0.61–2.29)
	Non-formal	0.27*** (0.22–0.34)	0.44** (0.26–0.75)	0.53*** (0.35–0.79)	0.94 (0.4–2.2)

n number of participants, OR odds ratio, aOR adjusted odds ratio for clustering by community, CI confidence interval, Ref reference group

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^a Univariate logistic regression

^b Multivariable logistic regression

^c Excluded from the multivariable due to interaction with other variables

tion was asked only to women), the effect of child number could not be evaluated in men. Women who were not affiliated with any religion were more likely to have undiagnosed HIV infections (aOR=1.33, $p < 0.05$), but there was no such effect in men. For women, the odds of an undiagnosed infection increased slightly with attaining higher education beyond Senior Secondary level ($p < 0.05$), but this was not the case for men.

Behavioural predictors of undiagnosed HIV infection at baseline Women who tested for HIV more frequently were more likely to have undiagnosed infections (Table 2, aOR=18.88 for those with ten or more tests compared to those with 1 or 2, $p < 0.001$) but no such effect was observed in men. In the multivariable model, number of sexual partners and concurrent partners were not significantly predictive ($p > 0.05$). Women who reported knowing their partner’s HIV status to be negative were more likely to have undiagnosed HIV infection than those who reported their partners to be HIV positive ($p < 0.001$). Condom use in the past year was a significant predictor

of undiagnosed infection among women in the univariate logistic regression (OR=2.37, $p < 0.001$), but it was excluded from the multivariable analysis due its interaction with condom frequency. Participants who never used condoms, and those who used them only sometimes, were more likely to have undiagnosed infection than those who always used them ($p < 0.001$ for women and $p < 0.05$ for men). Increased time away from home increased the risk of undiagnosed infection in the univariate analysis for women (OR=1.24 and 1.54, $p < 0.001$) but not in the multivariable analysis.

Differences in assigned timing of HIV infection

Results were aligned with those expected: incident cases were nearly all classified as recent infections (75%), whereas 37% of newly diagnosed cases and only 2% of known cases were classified as recent (Table 3, $p < 0.001$ based on the Pearson χ^2 test). The classifier produces a probability of recency for each individual, rather than assigning a categorical stage, and we observed the same pattern when we compared the probability distributions

Table 2 Logistic regression for behavioural characteristics associated with undiagnosed infection by sex

Variable	Category	Women (n = 383 newly diagnosed, n = 2252 known cases)		Men (n = 218 newly diagnosed, n = 743 known cases)	
		OR (95% CI) ^a	aOR (95% CI) ^b	OR (95% CI)	aOR (95% CI)
Number of HIV tests	1–2	Ref		Ref	
	3–4	3.74*** (3.28–4.27)	4.81*** (3.72–6.23)	0.76 (0.58–1.0)	0.74 (0.49–1.1)
	4–9	6.41*** (5.43–7.61)	10.9*** (7.82–15.2)	0.73 (0.51–1.07)	0.49* (0.28–0.85)
	10+	9.8*** (6.77–14.74)	18.8*** (8.41–42.1)	1.1 (0.48–3.03)	1.0 (0.36–3.0)
Number partners in 1 year	1 partner	Ref		Ref	
	2 partners	1.15 (1.0–1.32)	0.88 (0.15–5.1)	1.32*** (1.0–1.68)	0.67 (0.02–20.0)
	3 partners	2.54*** (1.43–4.93)	4.6 (0.25–84.6)	1.17 (0.5–2.68)	0.19 (0.004–9.8)
	4 partners	1.34 (0.66–3.0)	0.52 (0.09–3.0)	1.87 (0.73–4.61)	1.0 (0.03–35.1)
	None	0.71*** (0.62–0.82)	–	1.36 (1.0–1.86)	–
Concurrency in past year	No concurrent partners	Ref		Ref	
	Concurrent partners	1.17* (1.0–1.34)	1.2 (0.21–6.9)	1.42*** (1.12–1.79)	1.68 (0.06–46.8)
Partner’s HIV status	HIV negative	Ref		Ref	
	HIV positive	0.06*** (0.05–0.07)	0.09*** (0.07–0.12)	0.8 (0.6–1.07)	0.8 (0.6–1.06)
Condom use in past year	No	Ref		Ref	
	Yes	2.37*** (2.11–2.66)	–	0.8 (0.65–1.0)	–
Condom frequency	Always	Ref		Ref	
	Sometimes	1.98*** (1.75–2.25)	1.64*** (1.32–2.03)	0.83 (0.66–1.05)	1.56* (1.07–2.25)
	Never	3.07*** (2.53–3.75)	2.9*** (2.15–3.89)	0.77 (0.53–1.14)	2.17** (1.25–3.76)
Nights away from home	Zero	Ref		Ref	
	< 1 week–< 1 month	1.24*** (1.09–1.42)	0.88 (0.67–1.15)	0.81 (0.64–1.03)	1.0 (0.65–1.68)
	> 1 month	1.54*** (1.38–1.73)	0.96 (0.77–1.20)	0.85 (0.69–1.06)	0.95 (0.56–1.59)

n number of participants, OR odds ratio, aOR adjusted odds ratio for clustering by community, CI confidence interval, Ref reference group

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

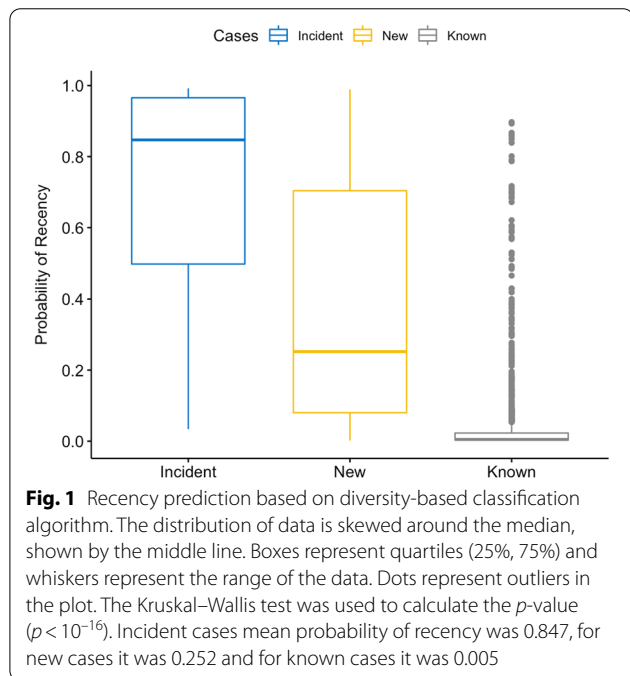
^a Univariate logistic regression

^b Multivariable logistic regression

Table 3 Timing of HIV infection among all participants

Variables	Incident cases N = 136 [#]	New cases N = 284 [#]	Known cases N = 1447 [#]	p-value
Recent, n (%)	102 (75.0%)	105 (37.0%)	29 (2.0%)	
Chronic, n (%)	34 (25.0%)	179 (63.0%)	1418 (98.0%)	< 0.001

[#] Analysis for participants with available data, n—number of participants. Non-independence between groups was evaluated using Pearson χ^2 test for categorical data



across the three groups (Fig. 1). Incident cases had the highest probability of recency ($p=0.847$) while known cases had the lowest probability ($p=0.005$). New cases had an intermediary probability of being recently infected ($p=0.252$).

Differences in terminal branch lengths

The statistical analysis for terminal branch lengths across three groups and three genes amounted to nine non-independent comparisons, therefore we set our threshold for significance to $p < 0.0055$. We observed no consistent pattern of difference in terminal branch lengths between groups.

Discussion

Among a large representative population of adults with HIV in rural/peri-urban Botswana, we compared the demographic and behavioural characteristics of those with known, diagnosed HIV-infection compared with those newly diagnosed with HIV at the start of a clinical

trial. Participants with undiagnosed HIV infections were more likely to be male, young, and not to consistently use condoms. Among women, being married, educated and testing frequently increased the risk of an undiagnosed HIV infection. For men, there was a wider range of ages among undiagnosed HIV cases and being divorced increased risk.

Viral genetic sequences are informative regarding stage of HIV infection because individuals are usually infected with a single virus and genetic diversity then increases with time [20–22]. Results from the genetic diversity-based classifier were concordant with epidemiological data: most known cases were classified as chronic (97.9%) while most incident cases were classified as recent (75.0%). This means that 25% of incident infections were misclassified as chronic, a high false negative rate due to a high proportion of participants being on ART [39]. Based on the classifier, over a third (37.0%) of newly diagnosed cases were recently infected. Given the high false negative rate of the classifier this proportion may be an underestimate—therefore a sizeable proportion of newly diagnosed cases were likely to be recent infections. Another genetic signature of time between infection and diagnosis is the length of the terminal branch leading to a sequence [22, 40, 41]. Short terminal branch lengths indicate that sampling occurred shortly after the last transmission event. However, we found no consistent differences in the distribution of terminal branch lengths among our three groups. This lack of signal may be due to the low sample proportion of our population as a whole (12,610 of a total trial area population of 175,664, 7.2%): the relationship between terminal branch length and time to diagnosis will be disrupted if too many transmissions are missed in the phylogeny [42]. The relationship between time to diagnosis and terminal branch length has been demonstrated in simulations [41] but its utility in the real world will have to be further ascertained in datasets with higher sampling proportions. One study that successfully used root to tip branch length in determining time since infection demonstrated its use for estimating HIV incidence at a population level [43].

In our study, more women (63.8%) were enrolled than men (36.2%). A similar study in Zambia and South Africa (HPTN-071) showed similar enrolment patterns for women (70%) and men (30%) [44]. Surveys across sub-Saharan Africa have consistently demonstrated that HIV testing uptake is higher among women than among men [2, 13, 45]. One reason women are more likely to know their HIV status is that most countries, including Botswana, screen for HIV during pregnancy as part of prevention of mother to child transmission [3, 46, 47]. This strategy may explain why in our study, women with more children were less likely to have undiagnosed HIV

infections. Furthermore, fewer undiagnosed infections were observed as participants increased in age, in agreement with others [2, 13, 19, 45], and testing increased with age. HIV testing rates among men below 25 years are low [17]. In Botswana, people above 50 living with HIV are more likely to be aware of their HIV status and to be on antiretroviral therapy [36]. Older women in particular are more likely to be aware of their positive HIV status than older men [48], and concordantly, in our study the effect of age was much stronger in women than in men. Men up to age 44 had an increased risk for an undiagnosed infection, while this was only true for women up to age 24.

Our data shows that being married increased the risk of an undiagnosed HIV infection for women. The association between marriage and undiagnosed HIV infection has been noted previously [10, 11, 49]. In contrast, a survey conducted in South Africa showed that married people living with their spouses were *less* likely to be HIV positive [50], but even in that study, HIV infections were highest among those who were married but who spent extended periods away from home. Knowing one's partner to be HIV-positive decreased the odds of an undiagnosed HIV infection. This finding could indicate that sero-discordant couples take precautions to prevent transmission and to get diagnosed rapidly but it seems reasonable to assume that this is most likely due to demonstrate the success of treatment as prevention. In our study, we found that those women with a higher education were more likely to have undiagnosed HIV infections, despite education improving HIV testing behaviours [10, 46, 49]. Previous cross-sectional analyses have demonstrated a positive association between educational attainment and HIV positivity across sub-Saharan African countries [51]; educational attainment may increase the likelihood of sexual opportunities or risk-taking.

The behavioural factors associated with undiagnosed HIV were repeated HIV testing (for women) and inconsistent condom use (for both sexes). At first glance, this first finding seems contradictory: frequent testing increases the chances of diagnosing HIV within the early stages of infection and should decrease the probability that a person has an undiagnosed HIV infection. Most participants (72.8% of known, previously diagnosed cases and 44.6% newly diagnosed cases) were tested for the first-time at enrolment or had tested only once previously. It is possible that women who test frequently do so because they are aware of being at risk for HIV. For example, they may use HIV testing as a greenlight for unprotected sex, or they may get tested following a risk event. However, higher numbers of partners and concurrency were not identified among these women. They may

have risk behaviours that they did not disclose within this study. Further research is required to make specific recommendations regarding optimal HIV testing timings and frequency for this group, who already test regularly. We note that once a person tests HIV positive, they will not continue to get tests. Therefore, there is some censoring in our data on number of tests, however this censoring should not bias our interpretation. Health seeking behaviours such as frequent HIV testing and safe sexual practices by using a condom reduce HIV transmission [12]. Consistent condom use offered protection against HIV infection and participants who never used condoms were 4 times more likely to have undiagnosed HIV than those who always did.

Our analysis was subject to several limitations. First, the BCPP study enrolled only 20% of households in target areas. While this proportion is high for a single study, it is low when compared to HIV testing and sequencing coverage in countries such as the UK. A similar study, PopART, conducted in Zambia and South Africa sampled 4.8% of their population (48,301/100,000) [44] while BCPP reached 7.2% (12,610/175,664). It may be due to this low sampling proportion that we were not able to see differences in terminal branch lengths, because too many transmissions were missed in the phylogeny. Lastly, although the BCPP questionnaires were thorough, many answers were not complete. Our analyses were affected by missing data in some variables of interest ranging from 1–40%. Our analysis yielded significant results, but if some information was deliberately obscured, we may be missing important associations. Further studies are warranted to further investigate these variables with complete datasets.

Conclusion

Despite the tremendous success of the “Test and Treat all” strategy in Botswana, 16.7% (601/3596) of BCPP participants with HIV at enrolment did not know their positive status. Taken together, our results indicate that adults with undiagnosed infections are likely to be young (especially women), and to not consistently use condoms with their partners. Women were more likely to have undiagnosed infections if they were married, educated, and tested frequently. Men are more likely to have undiagnosed HIV infections and being divorced increased their risk but otherwise they did not stand out in obvious ways from men with diagnosed infections. Notably, a sizeable proportion of undiagnosed infections were likely to be recent based on a genetic-diversity-based classifier, suggesting they are aware of their risk. Our results stress the importance of targeting interventions towards men in a range of places where they might be useful; for example, by offering HIV self-testing or testing in workplaces,

sporting events, barber shops or places where men congregate regularly. Clearer recommendations may be needed as to how frequently, and under what circumstances, this group should get tested for HIV. With high rates and coverage of HIV testing and antiretroviral therapy initiation, incorporating the identified predictors to prioritize HIV testing and pre-exposure prophylaxis (PreP) will help reduce national HIV incidence in Botswana.

Abbreviations

HIV: Human immunodeficiency virus; BCPP: Botswana Combination Prevention Project; NGS: Next generation sequencing; KS: Kolmogorov–Smirnov; n: Number of participants; OR: Odds ratio; aOR: Adjusted odds ratio; CI: Confidence interval; Ref: Reference group.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12879-022-07698-4>.

Additional file 1: Table S1 Demographic and behavioural factors associated with newly diagnosed HIV-1 infections ($n = 601$) compared to known HIV-cases ($n = 2995$) in Botswana. **Figure S1.** Differences in terminal branch lengths for three gene regions: A. Gag, B. Polymerase and C. Envelope genes of the HIV-1C virus. Incident cases (blue), new cases (yellow) and known cases (grey) are shown in each plot. The Student's t test was used to generate the p -values: ns -not significant, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. After Bonferroni's correction, no statistical difference was observed in terminal branch lengths across the HIV genes.

Acknowledgements

We acknowledge the Botswana Combination Prevention Project study team, the research staff at Botswana Harvard AIDS Institute Partnership and, most importantly, the study participants. We thank all members of the PANGEA consortium, and in particular Tanya Golubchik for bioinformatic support.

Disclaimer

The views expressed in this publication are those of the authors and not necessarily those of NIH, Bill and Melinda Gates Foundation, MRC UK, AAS, NEPAD Agency, Wellcome Trust, PEPFAR, or the U.K. government. The funders had no role in the study design, data collection and decision to publish, or in the preparation of the manuscript.

Author contributions

LB, MRC, SM, conceived and designed the study. TG, EK, MM and JM conducted field work and quality assurance. SM and SG performed the lab work and generated viral data. PB, FU, SM, SG, MPH, EKY, MM, TG, JM, KEW, SL, ME, VN collected, cleaned, and compiled data. LB and MRC performed analyses and interpreted results. SM, SG, SL and VN provided support for analysis. LB wrote the first draft of the manuscript. All authors read and approved the final manuscript.

Funding

We acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. PANGEA consortium is funded by the Bill & Melinda Gates Foundation (OPP1084362 and OPP1175094). Additional support was provided by the Research England Global Challenges Research Fund to MRC. The President's Emergency Plan for AIDS Relief (PEPFAR) funded the implementation and evaluation of the BCPP through the Centers for Disease Control and Prevention under the terms of

cooperative agreements U01 GH000447, U2G GH001911, U2G GH000073, U2G GH002027, and U2G GH000419.

SG was partially supported by H3ABioNet. H3ABioNet is supported by the National Institutes of Health Common Fund [U41HG006941]. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) programme of the African Academy of Science (AAS). SM & SG were supported by the Fogarty International Center and National Institute of Mental Health, of the National Institutes of Health under Award Number D43TW010543. SM & SG were partly supported through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant #107752/Z/15/Z] and the U.K. government. The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, or the U.K. government. SL received funding from US NIH K24AI131928. The funders had no role in the study design, data collection and decision to publish, or in the preparation of the manuscript.

Availability of data and materials

Sequence data and basic demographics for this study are available upon request to the PANGEA HIV consortium (www.pangea-hiv.org). BCPP protocols and collected data are made available at <https://data.cdc.gov/Global-Health/Botswana-Combination-Prevention-Project-BCPP-Publi/qcw5-4m9q>. For full access, please use the data request form.

Declarations

Ethics approval and consent to participate

The BCPP study was approved by the Human Research Development Committee (HPDME 13/18/1) (the Institutional Review Board based at the Botswana Ministry of Health and Wellness), and the US Centers for Disease Control and Prevention (Protocol 6475). The study is registered at ClinicalTrials.gov (NCT01965470). All participants provided written informed consent and participants aged 16 to 18 years provided written assent (with parental or guardian written permission). The data collection procedures were performed in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

SM is an editorial Board Member for BMC Infectious Diseases. All other authors report no competing interests.

Author details

¹Botswana-Harvard AIDS Institute Partnership, Gaborone, Botswana. ²Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ³Ministry of Health and Wellness, Gaborone, Botswana. ⁴Centers for Disease Control and Prevention, Atlanta, GA, USA. ⁵Centers for Disease Control and Prevention, Gaborone, Botswana. ⁶Department of Medicine, Division of Infectious Diseases Brigham and Women's Hospital, Boston, MA, USA. ⁷Brown University, Providence, RI, USA. ⁸MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK. ⁹Present Address: Department of Ecology and Evolution, University of Chicago, Chicago, USA.

Received: 14 March 2022 Accepted: 17 August 2022

Published online: 28 August 2022

References

- Essex M, Makhema J, Lockman S. Reaching 90–90–90 in Botswana. *Curr Opin HIV AIDS*. 2019;14(6):442–8.
- Lebelonyane R, 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–e0255227.
3. Gaolathe T, et al. Botswana's progress toward achieving the 2020 UNAIDS 90-90-90 antiretroviral therapy and virological suppression goals: a population-based survey. *Lancet HIV*. 2016;3(5):e221–30.
 4. Makhema J, et al. Universal testing, expanded treatment, and incidence of HIV infection in Botswana. *N Engl J Med*. 2019;381(3):230–42.
 5. Statistics B. Botswana AIDS impact survey (BAIS) IV. Gaborone: Statistics Botswana; 2013.
 6. Pao D, et al. Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS*. 2005;19(1):85–90.
 7. Brenner BG, et al. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis*. 2007;195(7):951–9.
 8. Fisher M, et al. Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *AIDS*. 2010;24(11):1739–47.
 9. Ragonnet-Cronin M, et al. Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics. *J Acquir Immune Defic Syndr*. 2010;55(1):102–8.
 10. Mtowa A, et al. Socio-demographic inequalities in HIV testing behaviour and HIV prevalence among older adults in rural Tanzania, 2013. *AIDS Care*. 2017;29(9):1162–8.
 11. Budambula V, et al. Socio-demographic and sexual practices associated with HIV infection in Kenyan injection and non-injection drug users. *BMC Public Health*. 2018;18(1):193–193.
 12. Auvert B, et al. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial.
 13. Alwano MG, et al. Increasing knowledge of HIV status in a country with high HIV testing coverage: results from the Botswana Combination Prevention Project. *PLoS ONE*. 2019;14(11):e0225076.
 14. Mkandawire P. Assessing factors associated with HIV testing among adolescents in Malawi. *Glob Public Health*. 2017;12(7):927–40.
 15. Shanaube K, et al. What works—reaching universal HIV testing: lessons from HPTN 071 (PopART) trial in Zambia. *AIDS (Lond, Engl)*. 2017;31(11):1555–64.
 16. Pettifor A, et al. Factors associated with HIV testing among public sector clinic attendees in Johannesburg, South Africa. *AIDS Behav*. 2010;14(4):913–21.
 17. Marcus U, et al. Recency and frequency of HIV testing among men who have sex with men in Germany and socio-demographic factors associated with testing behaviour. *BMC Public Health*. 2015;15:727–727.
 18. Billings JD, et al. Factors associated with previously undiagnosed human immunodeficiency virus infection in a population of men who have sex with men and male-to-female transgender women in Lima, Peru. *Medicine*. 2016;95(42):e5147–e5147.
 19. Mustanski B, et al. Factors associated with HIV testing in teenage men who have sex with men. *Pediatrics*. 2020;145(3):e20192322.
 20. Carlisle LA, et al. Viral diversity based on next-generation sequencing of HIV-1 provides precise estimates of infection recency and time since infection. *J Infect Dis*. 2019;220(2):254–65.
 21. Ragonnet-Cronin M, et al. Genetic diversity as a marker for timing infection in HIV-infected patients: evaluation of a 6-month window and comparison with BED. *J Infect Dis*. 2012;206(5):756–64.
 22. Shankarappa R, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol*. 1999;73(12):10489–502.
 23. Magosi LE, et al. Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial—BCPP/Ya Tsie trial. *medRxiv*. 2021. p. 2021.06.19.21259186.
 24. CDC. Botswana Combination Prevention Project (BCPP)—Public Release Data <https://data.cdc.gov/Global-Health/Botswana-Combination-Prevention-Project-BCPP-Publi/qcw5-4m9q>. Accessed on 09 Mar 2022.
 25. Biopolymers Facility. <https://genome.med.harvard.edu>. Accessed on 9 Mar 2022.
 26. Pillay D, et al. PANGEA-HIV: phylogenetics for generalised epidemics in Africa. *Lancet Infect Dis*. 2015;15(3):259–61.
 27. Abeler-Dorner L, et al. PANGEA-HIV 2: phylogenetics and networks for generalised epidemics in Africa. *Curr Opin HIV AIDS*. 2019;14(3):173–80.
 28. PANGEA HIV. <https://www.pangea-hiv.org>. Accessed on 09 Mar 2022.
 29. Ratmann O, et al. HIV-1 full-genome phylogenetics of generalised epidemics in sub-Saharan Africa: impact of missing nucleotide characters in next-generation sequences. *AIDS Res Hum Retroviruses*. 2017;33(11):1083–98.
 30. Novitsky V, et al. Long-range HIV genotyping using viral RNA and proviral DNA for analysis of HIV drug resistance and HIV clustering. *J Clin Microbiol*. 2015;53(8):2581–92.
 31. Gall A, et al. Complete genome sequence of the WHO international standard for HIV-1 RNA determined by deep sequencing. *Genome Announc*. 2014;2(1):e01254-13.
 32. Struck D, et al. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res*. 2014;42(18):e144.
 33. Team R.C. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019.
 34. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22(21):2688–90.
 35. Volz EM, Frost SF. Scalable relaxed clock phylogenetic dating. *Virus Evol*. 2017;3(2).
 36. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2018;35:526–8.
 37. Revell L. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol*. 2012;3:217–23.
 38. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016*. New York, NY, USA.
 39. Ragonnet-Cronin M, et al. HIV genetic diversity informs stage of HIV-1 infection among patients receiving antiretroviral therapy in Botswana. *J Infect Dis*. 2021;225:1330–8.
 40. Poon AF. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evol*. 2016;2(2):vew031.
 41. Moshiri N, Smith DM, Mirarab S. HIV care prioritization using phylogenetic branch length. *JAIDS J Acquir Immune Defic Syndr*. 2021;86(5):626–37.
 42. Lewis F, et al. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med*. 2008;5(3):e50.
 43. Golubchik T, et al. HIV-phyloTSL: subtype-independent estimation of time since HIV-1 infection for cross-sectional measures of population incidence using deep sequence data. *medRxiv*. 2022. p. 2022.05.15.22275117.
 44. Hayes RJ, et al. Effect of universal testing and treatment on HIV incidence—HPTN 071 (PopART). *N Engl J Med*. 2019;381(3):207–18.
 45. Staveteig L, et al. Demographic patterns of HIV testing uptake in sub-Saharan Africa. DHS comparative reports. Calverton; 2013.
 46. Singh K, Luseno W, Haney E. Gender equality and education: Increasing the uptake of HIV testing among married women in Kenya, Zambia and Zimbabwe. *AIDS Care*. 2013;25(11):1452–61.
 47. Centers for Disease Control and Prevention. Introduction of routine HIV testing in prenatal care—Botswana. *MMWR Morb Mortal Wkly Rep*. 2004;53:1083–6.
 48. Matlho K, et al. HIV prevalence and related behaviours of older people in Botswana—secondary analysis of the Botswana AIDS Impact Survey (BAIS) IV. *Afr J AIDS Res*. 2019;18(1):18–26.
 49. Brima N, et al. Factors associated with HIV prevalence and HIV testing in Sierra Leone: findings from the 2008 Demographic Health Survey. *PLoS ONE*. 2015;10(10):e0137055.
 50. Shisana O, et al. Does marital status matter in an HIV hyperendemic country? Findings from the 2012 South African National HIV prevalence, incidence and behaviour survey. *AIDS Care*. 2016;28(2):234–41.
 51. Fortson JG. The gradient in sub-Saharan Africa: socioeconomic status and HIV/AIDS. *Demography*. 2008;45(2):303–22.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.