OPEN

Systematic identification of correlates of HIV infection: an X-wide association study

Chirag J. Patel[a], Jay Bhattacharya[b,c], John P.A. Ioannidis[d] and Eran Bendavid[b,e]

Objective: Better identification of at-risk groups could benefit HIV-1 care programmes. We systematically identified HIV-1 risk factors in two nationally representative cohorts of women in the Demographic and Health Surveys.

Methods: We identified and replicated the association of 1415 social, economic, environmental, and behavioral factors with HIV-1 status. We used the 2007 and 2013–2014 surveys conducted among 5715 and 15 433 Zambian women, respectively (688 shared factors). We used false discovery rate criteria to identify factors that are strongly associated with HIV-1 in univariate and multivariate models of the entire population, as well as in subgroups stratified by wealth, residence, age, and past HIV-1 testing.

Results: In the univariate analysis, we identified 102 and 182 variables that are associated with HIV-1 in the two surveys, respectively (79 factors were associated in both). Factors that were associated with HIV-1 status in full-sample analyses and in subgroups include being formerly married (adjusted OR 2007, 2.8, $P < 10^{-16}$; 2013–2014 2.8, $P < 10^{-29}$), widowhood (aOR 3.7, $P < 10^{-12}$; and 4.2, $P < 10^{-30}$), genital ulcers within 12 months (aOR 2.4, $P < 10^{-5}$; and 2.2, $P < 10^{-6}$), and having a woman head of the household (aOR 1.7, $P < 10^{-7}$; and 2.1, $P < 10^{-26}$), while owning a bicycle (aOR 0.6, $P < 10^{-6}$; and 0.6, $P < 10^{-8}$) and currently breastfeeding (aOR 0.5, $P < 10^{-9}$; and 0.4, $P < 10^{-26}$) were associated with decreased risk. Area under the curve for HIV-1 positivity was 0.76–0.82.

Conclusion: Our X-wide association study identifies under-recognized factors related to HIV-1 infection, including widowhood, breastfeeding, and gender of head of the household. These features could be used to improve HIV-1 identification programs.

AIDS 2018, **32**:933–943

Keywords: demographic and health surveys, environment-wide association study, epidemiology, HIV-1, sub-Saharan Africa, X-wide association study

Introduction

Approaching the public health goals for global HIV-1 such as '90–90–90' (90% of those with HIV-1 aware of their status, 90% in regular treatment, and 90% of those on treatment virally suppressed) requires effective identification of at-risk and infected individuals [1,2]. Despite large efforts to expand testing, treatment, and retention services, only 45% of HIV-infected individuals in Sub-Saharan Africa knew their HIV-1 status in 2013, and estimated antiretroviral therapy (ART) coverage in 2016 exceeded 60% of all those infected in only five

[a]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, [b]Center for Health Policy and the Center for Primary Care and Outcomes Research, Stanford University, Stanford, California, [c]National Bureau of Economic Research, Cambridge, Massachusetts, [d]Stanford Prevention Research Center, and [e]Division of Primary Care and Population Health, Department of Medicine, Stanford University, Stanford, California, USA.

Correspondence to Eran Bendavid, MD MS, Division of General Medical Disciplines, Stanford University, Stanford, CA 94305, USA.

Tel: +1 650 723 2363; fax: +1 650 723 8596; E-mail: ebd@stanford.edu

Received: 17 July 2017; revised: 10 January 2018; accepted: 22 January 2018.

DOI:10.1097/QAD.0000000000001767

ISSN 0269-9370

countries [3–6]. A potential approach to improving HIV-1 testing and diagnosis is to better target individuals and populations for testing and care. Existing HIV-1 control programs, such as the US President's Emergency Plan for AIDS Relief, increasingly use data-driven approaches to align resources towards high-burden populations [7,8].

Current understanding of HIV-1 risk factors in Sub-Saharan Africa commonly come from nationally representative surveys such as the Demographic and Health Surveys (DHS) that include HIV-1 testing and report prevalence stratified by prespecified groups such as age, education, place of residence, and number of sexual partners [9–11]. Such HIV-1 testing and epidemiologic stratification was carried out in two nationally representative DHS surveys conducted among samples of nearly 6000 (in 2007) and 15 000 (in 2013–2014) Zambian women and men [12,13]. However, selective identification of risk factors by testing one or only a few factors at a time may lead to incomplete understanding of or even misleading notions about possible risk factors [14,15]. Traditional risk factors (age, sex, education, place of residence, and number of sexual partners) explain less than 10% of the variation in HIV-1 infection, and represent only a small fraction of the information available in the surveys [12,13]. Although risk factors such as age and sex are intuitive and important, unintuitive or under-recognized correlates that may identify novel high-risk groups and generate new hypotheses for further study and intervention design.

We present an approach for systematically assessing the relationship between HIV-1 status and many putative risk factors. We exploit the breadth of DHS surveys to conduct an X-wide association study (XWAS) of HIV-1 risk, where X stands for all social, behavioral, environmental, and economic factors that are reliably available in DHS. This approach systematically associates each available variable with HIV-1 status, as is done in current-day genomics investigations [e.g. genome-wide association studies (GWASs)]. We have previously utilized the approach to systematically study the association of environmental exposures, dietary indicators, clinical biomarkers, and micronutrient blood tests associated with outcomes such as type II diabetes, blood pressure, mortality, and income [16–19]. An advantage of this approach is that variables are examined using a systematic approach, thus avoiding selective reporting bias, while controlling for the rate of false positives [20,21].

Methods

Overview

We used the 2007 and 2013–2014 DHS surveys from Zambia, where HIV-1 prevalence among women 15–49 years old was estimated at 16.1 and 15.1%, respectively [13]. (We analyzed both men and women, separately, and focus on women with additional results for men in the Appendix.) We linked HIV-1 status with all the indicators in the individual women's surveys. We split the data in each survey into training ('discovery') and replication data, analyzed the association of each variable with HIV-1 status in univariate and adjusted analyses, and examined the stability of the findings over time and in population subgroups. Figure 1 illustrates the analysis steps.

HIV-1 status

The HIV-1 testing procedure in the surveys involved identifying eligible household members, obtaining consent, collecting dried blood spots, and testing in a centralized lab. Two ELISA tests were used for screening and confirmation of HIV-positive tests, with western blot confirmation for discordant results. In both surveys, every test was definitively identified as positive or negative (i.e. there were no indeterminate tests). We then linked the HIV-1 test results with the individual survey data.

Selection of social, behavioral, environmental, and economic indicators

We used the following process to identify and create the variables for the XWAS (Table 1 includes the variable selection process metadata). Starting with the raw data after removal of placeholder variables (e.g. birthdates of children 6–20 for mothers with 5 children), we recast all variables with 30 or fewer levels as binary variables. Variables with 30 or more levels were treated as continuous. This decision rule aimed to preserve meaningfully continuous variables and discretize nonordinal variables. Then, we kept only those variables with at least 90% complete data to avoid what some consider unacceptable levels of missingness [22]. This led to dropping over 40% of the variables in each survey. We removed variables with no variation (e.g. an indicator variable for completion of the survey), and kept the first occurrence in any pair of collinear variables with correlation coefficient at least 0.99. The entire set of variables is in Supplementary Table 1, http://links.lww.com/QAD/B226

Association study procedures

We divided each survey (5715 women in 2007 and 15 433 women in 2013–2014), randomly into two equally sized (± 1) datasets for discovery and replication. We conducted three XWAS analyses in the discovery dataset (Fig. 1d): a univariate analysis; an analysis adjusted for known HIV-1 correlates (ex ante analysis); and an analysis that, in addition to the ex ante factors, adjusted for the 10 variables that explained the greatest portion of the variation in the univariate analysis (ex post analysis).

In the first step we estimated univariate logistic regression models of the following form:

$$\text{log}it(HIV_p) = \alpha + \beta^i X_p^i \quad (1)$$

where HIV_p represents the HIV-1 status of person 'p', and X_p^i denotes the ith variable for person 'p.' This procedure is repeated for each of the variables in the 2007

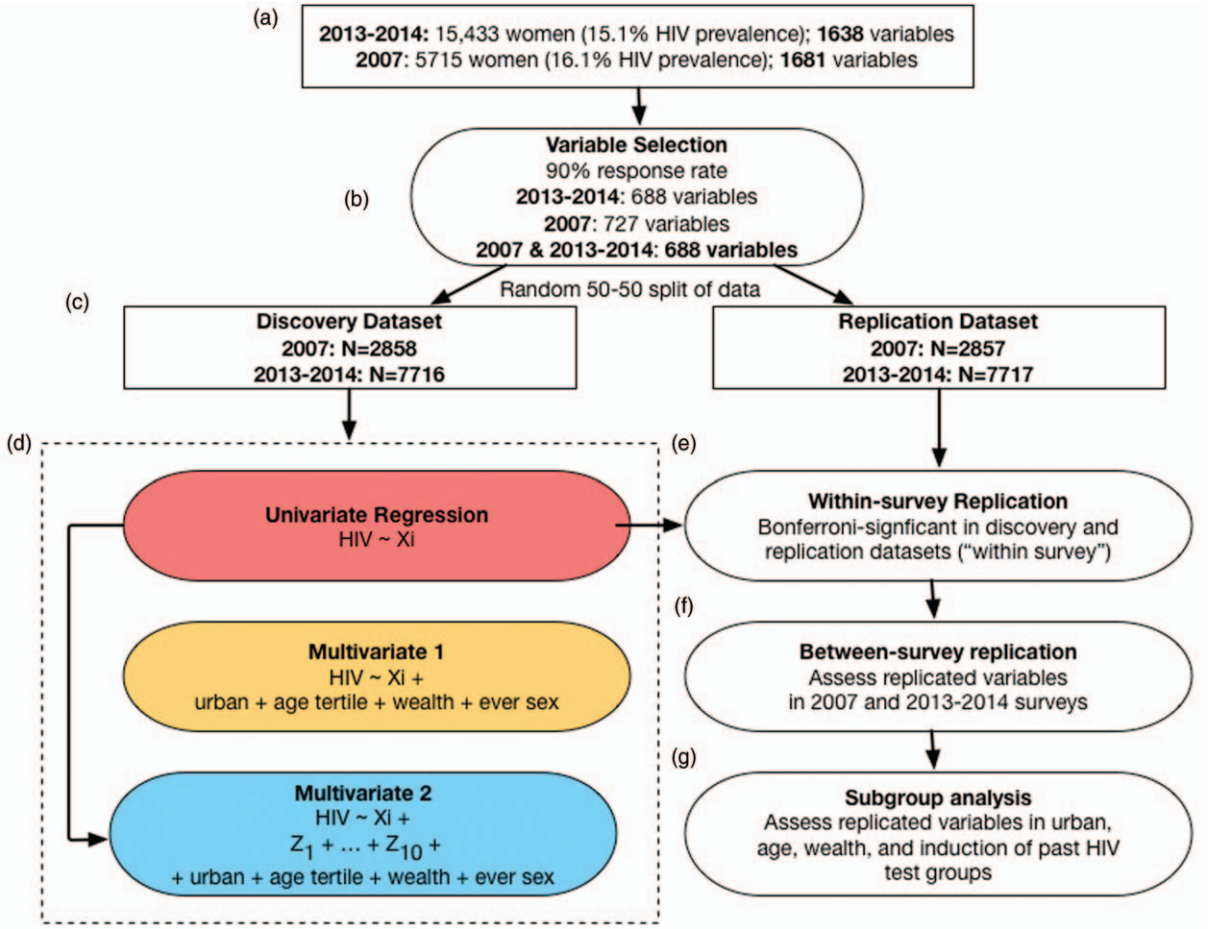

Fig. 1. Schematic overview of XWAS process. (a) The Demographic and Health Study consists of 15 433 women (15% HIV-1 prevalence) in 2013–2014 and 5715 women (16.1 prevalence) in 2007. (b) We selected variables that had at least 90% response rate and were nonredundant resulting in 688 total variables in both the 2007 and 2013–2014 surveys. (c) Split the data into two random subsets for discovery and replication ($N = 7716$ and 7717, respectively, for 2013–2014 and 2858 and 2857 for 2007). (d) We ran three model configurations: a univariate (red); a multivariate with adjustment variables selected a priori (yellow), including age, urban resident, wealth index, and ever had sex prior to survey; and a multivariate model consisting of variables identified in the univariate analysis (blue). X_i denotes the *i*th variable out of 688 in 2013–2014 and 727 variables in 2007, respectively (688 overlapping). (e) We attempted to replicate results within surveys from models in the independent replication dataset. (f) We identified variables replicated between the 2007 and 2013–2014 surveys. (g) We executed subgroup analysis for each variable identified in the univariate regression.

and 2013–2014 surveys. The exponentiation of β^i corresponds to the odds ratio for HIV-1 per unit change for each variable X^i. To control for multiple hypothesis testing, we calculated the Benjamini–Hochberg (BH) false discovery rate (FDR), the estimated proportion of discoveries made that were false [23]. The Benjamini–Hochberg method assumes independence between statistical tests, and therefore, counts correlated variables as

Table 1. Variable preparation process.

	2007 survey		2013–2014 survey	
Variable list, preparation stage	Change (*n*)	Remaining (*n*)	Change	Remaining
Initial dataset		976		957
Discretize categorical variables	705	1681	681	1638
Remove variables with more than 10% missingness	−676	1005	−667	971
Remove indicator and no-variation variables	−84	921	−82	889
Remove one from any pair of variables that are 99% or more correlated	−194	727[a] (Final sample)	−201	688[a] (Final sample)

[a]In the 2007 survey, 160 of the 727 variables in the final sample (22%) were unchanged from the original dataset. In the 2003–2014 survey, 160 (23%) were unchanged from the original dataset.

independent for determining the discovery threshold (median absolute correlation between all pairs of replicated variables 0.06 [interquartile range (IQR) 0.02–0.12] in 2007 and 0.04 (IQR 0.02–0.09) in 2013). Throughout, we used the HIV-1 sampling weights and Huber–White robust standard errors [24].

In the ex ante analysis, we adjusted for five predetermined (ex ante) controls: urban or rural residence, DHS wealth index (a five-level scale from poorest to wealthiest, with poorest as reference), age, whether or not the respondent indicated, she had previously been tested for HIV, and whether she indicated that she never had sex at the time of the interview [25]. Specifically, the model was implemented as follows:

$$\begin{aligned}\text{log}it(HIV_p) = \alpha + \beta^i X^i_p + \gamma\, URBAN_p \\ + \varpi WEALTH_p + \xi\, AGE_p \\ + \delta TEST_p + \tau\, SEX_p \end{aligned} \quad (2)$$

where covariates for urban residence, wealth, age, past testing, and sexual debut are indexed by person (p), and X^i_p again denotes the ith exposure variable for person 'p.'

In the ex post analysis, we adjusted for the 10 variables that had the highest explanatory power (using Nagelkerke R^2) among those replicated in the univariate analysis. The purpose of this analysis was to improve the identification of strong correlates that identify HIV-1 status even after controlling for the most explanatory variables [26]. The 10 variables were selected separately in the 2007 and 2013–14 surveys.

We had two levels of replication, within-survey replication and between-survey replication (Fig. 1e and f). We deemed a within-survey 'replicated finding' for β^i as one that had FDR less than 5% in the discovery dataset and the sign for β^i was in the same direction in the replication dataset with a nominal P value less than 0.05 (Fig. 1e). The second level of replication is between-survey replication where we sought within-survey replication in both the 2007 and 2013–2014 surveys (Fig. 1f).

Next, we assessed the predictive capability of HIV positivity of the variables found in all three modeling scenarios. For example, if variables X^a. X^b. X^c were tentatively replicated in the univariate modeling scenario in the 2007 survey, we fit a model:

$$\text{log}it(HIV_p) = \alpha + \beta^a X^a_p + \beta^b X^b_p + \beta^c X^c_p \quad (3)$$

and assessed the Nagelkerke R^2 and the area under the curve for the model.

We then assessed pairwise correlations among all of the replicated variables to assess the clustering of HIV-1 risk factors and variables. That is, we wanted to identify the clusters of variables that potentially measure a latent HIV-1 risk factor (e.g. if a host of household possession variables are all related to HIV-1 and are correlated among themselves, that may indicate that wealth, a likely latent variable they measure, is a risk factor for HIV). We visualized pairwise correlations in a heatmap [27,28].

Finally, we tested the association of all replicated univariate findings in nine subgroups (Fig. 1g): (1–3) three age bins (15 to <23; 23 to <33; and 33–49); (4 and 5) two wealth groups [wealth quintiles 1–3 (poorer) and wealth quintiles 4 and 5 (wealthier)]; (6 and 7) two residence groups (urban and rural); and (8 and 9) two groups based on whether or not the respondent indicated that they had ever received an HIV-1 test.

To promote reproducibility of this work, the analytic code is available in a Github repository, and the figures can be accessed at www.chiragjpgroup.org/hiv_zambia; all analyses were performed using Stata 14 (Statacorp, College Station, Texas, USA) and R v3.2.2 (http://cran.r-project.org/).

Results

Our surveys included information on 5715 Zambian women with HIV-1 test results in 2007 and 15 433 in 2013–2014. In the univariate analysis, 102 (out of 727, 14%) variables were replicated and associated with HIV-1 in 2007, and 182 (out of 688, 26%) in 2013–2014. Figure 2 shows a plot of P values versus odds ratios of the association with HIV-1 of the variables tested in 2007 and 2013–2014. A total of 79 variables were associated with HIV-1 status in the univariate analysis, 30 variables in the ex ante analysis, and 8 variables in the ex post analysis in both 2007 and 2013–2014. Table 2 shows all the variables that were replicated *in both* surveys in at least one analysis. All replicated variables (in any analysis) appear in Supplementary Tables SA1.2–SA1.4, http://links.lww.com/QAD/B226. Supplementary Table SA1.5, http://links.lww.com/QAD/B226 shows the associations between the control variables and HIV in the ex post models.

Several variables stand out for their association with HIV-1 and for raising potentially useful targets for future investigation. Three variables were associated with HIV-1 in all three analyses and both surveys: having exactly one birth in the past 5 years (increased risk), currently breastfeeding (decreased risk), and desiring to delay having children for more than 2 years (decreased risk). Eleven additional variables were associated with HIV-1 positivity in all but one of the ex post analyses (including several variables that were used as ex post control): being formerly and not currently married, including divorce and widowhood (three variables, all increase risk), variables related to being the head of the household (three variables, all increase risk), the number of children

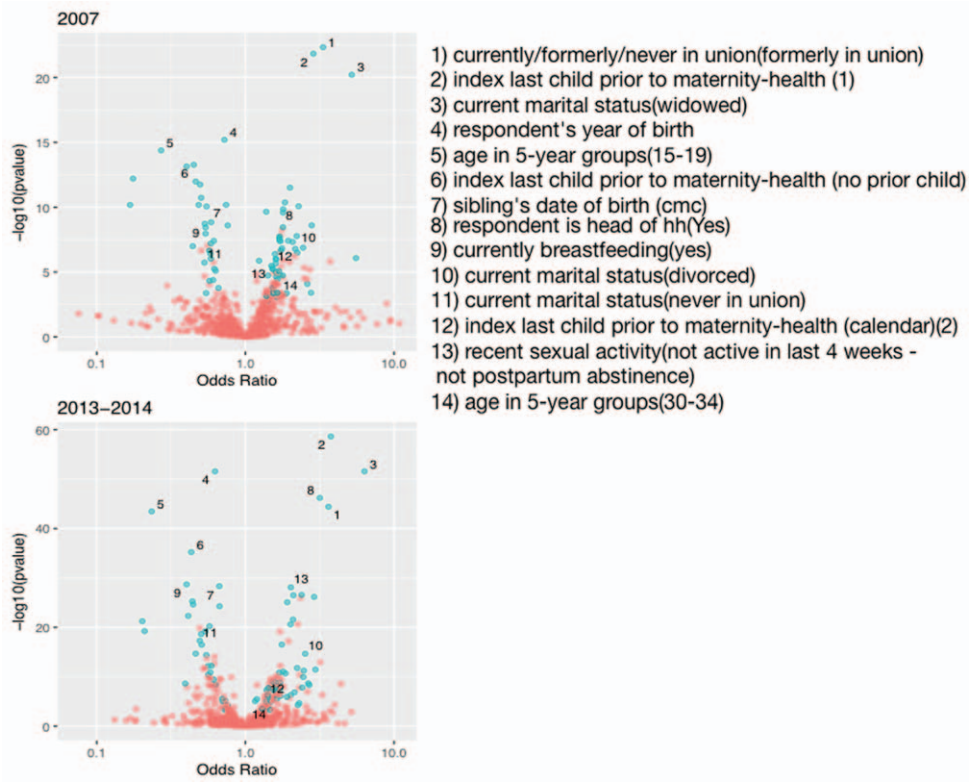

Fig. 2. Volcano plot from univariate analysis depicting odds ratio versus −log10(*P* value) of association. Teal points denote replicated findings in each dataset (2007 and 2013–2014).

Table 2. Univariate and adjusted associations with HIV-1 status.

		Univariate		Ex ante		Ex post	
Variable name	Value	OR (2007; 2013–2014)	−log10 (*P*) (2007; 2013–2014)	OR (2007; 2013–2014)	−log10 (*P*) (2007; 2013–2014)	OR (2007; 2013–2014)	−log10 (*P*) (2007; 2013–2014)
Desire for more children	Wants after 2+ years	0.5; 0.4	10.7[a]; 24.6[a]	0.5; 0.5	7.9[a]; 13.1[a]	0.5; 0.5	6.9[a]; 11.7[a]
Currently breastfeeding	Yes	0.5; 0.4	8.7[a]; 28.7[a]	0.5; 0.4	9.5[a]; 26.4[a]	0.5; 0.3	6.9[a]; 22.3[a]
Number of children 5 and under in household (de jure)	1	1.7; 2.0	7.4[a]; 20.6[a]	1.7; 2.0	6.8[a]; 15.4[a]	1.9; 2.3	5.6[a]; 17.3[a]
Currently/formerly/never in union	Formerly in union/ living with a man	3.3; 3.6	22.3[a]; 44.4[a]	2.8; 2.8	15.7[a]; 29.0[a]	–; 2.7	–; 18.5[a]
Index last child prior to maternity-health (calendar)	1	2.9; 3.8	21.8[a]; 58.6[a]	2.2; 2.6	9.1[a]; 22.6[a]	1.8; 2.6	3.6; 15.0[a]
Current marital status	Widowed	5.2; 6.3	20.2[a]; 51.6[a]	3.7; 4.2	12.6[a]; 31.0[a]	1.7; 4.5	1.3; 22.6[a]
Respondent's line number	1	2.3; 3.2	10.1[a]; 46.2[a]	1.9; 2.4	5.7[a]; 27.3[a]	0.9; 2.3	0.1; 19.5[a]
Number of living children	1	1.8; 1.3	9.6[a]; 2.9[a]	1.7; 1.6	6.3[a]; 4.9[a]	1.4; 1.6	2.4; 4.4[a]
Number of children 5 and under in household (de jure)	3	0.5; 0.6	8.4[a]; 10.9[a]	0.6; 0.6	7.0[a]; 8.7[a]	0.6; 0.6	3.3; 6.5[a]
Current contraceptive method	Condom	2.2; 3.0	7.8[a]; 11.5[a]	1.8; 2.5	4.6[a]; 8.5[a]	1.5; 2.4	1.4; 5.9[a]
Household has: bicycle	Yes	0.6; 0.6	7.4[a]; 9.5[a]	0.6; 0.6	6.2[a]; 8.1[a]	–; 0.7	–; 4.6[a]
Sex of household head	Female	1.7; 2.1	7.4[a]; 26.5[a]	1.7; 2.1	7.1[a]; 26.1[a]	0.9; 2.1	0.2; 20.0[a]
Current marital status	Divorced	2.2; 2.5	6.8[a]; 14.7[a]	1.9; 1.9	4.6[a]; 8.0[a]	0.8; 1.9	0.4; 5.6[a]
Number of household members (listed)	2	2.2; 2.5	6.5[a]; 10.0[a]	2.2; 2.3	6.6[a]; 7.3[a]	2.3; 2.6	4.3; 7.4[a]
Age in 5-year groups	15–19	0.3; 0.2	14.4[a]; 43.5[a]	0.5; 0.5	4.3; 6.9[a]	0.5; 0.5	2.6; 5.2[a]
Daughters who have died	1	2.0; 1.7	11.5[a]; 7.4[a]	1.8; 1.2	7.4[a]; 1.3	2.3; 1.2	8.3[a]; 0.5
Unmet need	Never had sex	0.2; 0.2	10.2[a]; 19.2[a]	0.4; 0.8	0.7; 0.1	–; –	42.5[a]; 89.5[a]
Had genital sore/ulcer in last 12 months	Yes	2.8; 2.4	8.6[a]; 7.8[a]	2.4; 2.2	5.9[a]; 6.1[a]	1.8; 2.0	1.8; 2.9
Sons who have died	0	0.6; 0.6	7.2[a]; 12.2[a]	0.7; 0.8	3.7[a]; 1.8	0.6; 0.8	5.0[a]; 0.9
Recent sexual activity	Not active in last 4 weeks - not postpartum abstinence	1.5; 2.0	5.3[a]; 28.1[a]	1.2; 1.8	0.9; 17.4[a]	0.9; 1.8	0.4; 13.3[a]
Desire for more children	Wants within 2 years	1.6; 2.4	4.5[a]; 26.6[a]	1.6; 2.2	3.9; 19.8[a]	1.7; 2.3	3.3; 15.0[a]
Number of sex partners, excluding spouse, in last 12 months	0	0.6; 0.6	4.4[a]; 12.2[a]	0.6; 0.5	2.8; 12.9[a]	0.9; 0.5	0.2; 9.7[a]
Number of living children	0	0.5; 0.5	11.7[a]; 18.6[a]	0.9; 1.3	0.2; 2.0	1.3; 1.5	1.0; 3.3[a]
Daughters who have died	0	0.5; 0.6	10.1[a]; 10.4[a]	0.6; 0.8	6.1; 1.9	0.5; 0.8	8.5[a]; 1.4
Main roof material	Thatch/palm leaf	0.5; 0.5	8.0[a]; 14.4[a]	0.8; 0.7	0.7; 3.8[a]	0.9; 0.7	0.2; 2.8
Tuberculosis can be cured	Yes	2.4; 2.2	6.9[a]; 11.8[a]	1.8; 1.6	3.3; 4.9[a]	–; –	–; –
Sons who have died	1	1.7; 1.8	6.7[a]; 11.0[a]	1.5; 1.3	3.8[a]; 2.0	1.6; 1.3	3.9; 1.2
Respondent's occupation (grouped)	Sales	1.7; 2.1	5.1[a]; 21.6[a]	1.2; 1.3	0.7; 3.4[a]	1.0; 1.3	0.0; 2.9

Table 2 (*continued*)

Variable name	Value	Univariate		Ex ante		Ex post	
		OR (2007; 2013–2014)	−log10 (*P*) (2007; 2013–2014)	OR (2007; 2013–2014)	−log10 (*P*) (2007; 2013–2014)	OR (2007; 2013–2014)	−log10 (*P*) (2007; 2013–2014)
Had genital discharge in last 12 months	Yes	2.6; 2.6	4.1[a]; 8.6[a]	2.1; 2.2	2.5; 6.1[a]	2.0; 2.0	1.6; 3.3
Respondent's year of birth		0.7; 0.6	15.2[a]; 51.6[a]	–; –	–; –	–; –	–; –
Index last child prior to maternity-health (calendar)	No prior child	0.4; 0.4	13.3[a]; 35.2[a]	0.5; 0.8	5.6; 1.2	0.6; 1.0	2.9; 0.0
Total children ever born	0	0.4; 0.4	13.1[a]; 22.3[a]	0.8; 1.0	1.1; 0.0	0.9; 1.2	0.1; 0.5
Age at first sex	Not had sex	0.2; 0.2	12.2[a]; 21.3[a]	–; –	–; –	–; –	–; –
Type of place of residence	Rural	0.5; 0.4	12.0[a]; 25.3[a]	–; –	–; –	–; –	–; –
Fertility preference	No more	1.8; 1.9	10.4[a]; 25.1[a]	1.3; 1.1	1.7; 1.1	1.2; 1.1	1.0; 0.4
Sibling's date of birth (CMC)		0.7; 0.7	10.2[a]; 28.3[a]	1.0; 1.0	0.4; 0.1	–; –	–; –
Living children + current pregnancy	0	0.5; 0.5	10.2[a]; 16.5[a]	1.0; 1.4	0.1; 2.3	1.3; 1.6	0.8; 3.1
Ever been tested for HIV	Yes	1.8; 2.9	9.8[a]; 26.2[a]	–; –	–; –	–; –	–; –
Wealth index factor score (5 decimals)		1.4; 1.2	9.7[a]; 5.5[a]	1.5; 1.0	1.5; 0.0	1.3; 1.1	0.6; 0.4
Fertility preference	Have another	0.6; 0.6	8.8[a]; 20.2[a]	0.8; 1.0	1.8; 0.1	0.7; 1.1	2.2; 0.6
Sibling's date of birth (CMC)		0.8; 0.7	8.6[a]; 24.3[a]	1.0; 1.0	0.0; 0.2	–; 0.6	–; 1.6
Knows method	Yes	1.8; 2.5	8.5[a]; 11.2[a]	1.2; 1.1	1.1; 0.5	1.2; 1.2	0.7; 0.4
Source for condoms: shop	Yes	1.7; 1.4	7.7[a]; 7.7[a]	1.4; 1.2	3.2; 2.4	1.3; 1.2	1.4; 1.1
Knows method	Yes	1.7; 1.5	7.6[a]; 9.0[a]	1.1; 0.8	0.8; 1.6	1.0; 0.8	0.1; 1.7
Source for condoms: pharmacy	Yes	1.9; 1.5	7.4[a]; 3.3[a]	1.4; 1.1	2.5; 0.5	1.2; 1.1	0.7; 0.5
Source of drinking water	Piped to yard/plot	2.1; 1.7	7.3[a]; 6.1[a]	1.6; 1.4	3.0; 2.5	1.4; 1.5	1.3; 2.4
Wealth index	Poorest	0.4; 0.5	7.0[a]; 14.7[a]	–; –	–; –	–; –	–; –
Knows method	Yes	1.8; 1.8	6.8[a]; 16.5[a]	1.1; 1.1	0.5; 0.9	1.1; 1.0	0.3; 0.2
Sons elsewhere	0	0.6; 0.5	6.7[a]; 19.2[a]	0.8; 0.9	1.4; 1.2	0.9; 1.0	0.6; 0.2
Flag for v531	No flag	0.6; 0.7	6.4[a]; 5.5[a]	0.8; 0.9	2.1; 1.0	0.7; 0.9	1.4; 0.5
Primary caregiver of children under age 18	Yes	1.6; 1.6	6.4[a]; 7.2[a]	1.1; 1.0	0.3; 0.2	1.1; 0.9	0.6; 0.5
Respondent's occupation	Personal and protective services workers	5.5; 2.7	6.1[a]; 8.3[a]	3.6; 1.9	3.6; 3.5	4.0; 2.1	2.0; 3.6
Main floor material	Earth, sand	0.6; 0.6	6.0[a]; 8.4[a]	0.9; 0.8	0.3; 1.6	1.0; 0.7	0.1; 1.9
Knows method	Yes	1.6; 2.0	6.0[a]; 6.3[a]	1.2; 1.2	1.1; 0.9	1.0; 1.2	0.1; 0.4
Getting medical help for self: distance to health facility	Not a big problem	1.6; 1.4	6.0[a]; 6.3[a]	1.2; 1.1	1.1; 0.6	1.1; 1.1	0.2; 0.6
Time to get to water source		1.2; 1.2	5.9[a]; 5.1[a]	1.1; 1.1	2.0; 1.5	1.1; 1.1	0.6; 1.5
Current marital status	Never in union	0.5; 0.5	5.7[a]; 17.3[a]	0.9; 1.1	0.2; 0.6	0.9; 1.3	0.2; 1.2
Index last child prior to maternity-health (calendar)	2	1.6; 1.4	5.7[a]; 5.8[a]	1.3; 0.9	1.6; 0.4	1.1; 0.8	0.6; 2.0
Knows method	Yes	1.5; 1.4	5.5[a]; 7.3[a]	1.0; 1.0	0.2; 0.0	1.0; 0.8	0.2; 1.5
Educational attainment	Incomplete primary	0.6; 0.7	5.3[a]; 5.2[a]	0.8; 0.8	1.8; 3.6	0.9; 0.7	0.3; 2.2
Source of drinking water	Public tap/standpipe	1.5; 1.5	5.2[a]; 5.1[a]	1.0; 1.1	0.1; 0.5	1.0; 1.0	0.1; 0.1
Tuberculosis spread by: don't know	Yes	0.6; 0.7	5.1[a]; 5.0[a]	0.8; 0.9	1.4; 0.9	0.9; 1.0	0.3; 0.1
Knows method	Yes	1.6; 1.7	5.0[a]; 10.8[a]	1.1; 1.1	0.7; 1.0	1.2; 1.1	0.7; 0.7
Age in 5-year groups	30–34	1.7; 1.4	4.8[a]; 4.3[a]	1.4; 1.1	2.1; 0.5	1.4; 1.1	1.5; 0.7
Sons elsewhere	1	1.8; 1.9	4.7[a]; 10.7[a]	1.4; 1.3	1.6; 1.7	1.1; 1.2	0.4; 0.7
Would want HIV infection in family to remain secret	Yes	1.4; 1.3	4.7[a]; 3.5[a]	1.3; 1.2	2.9; 2.5	1.3; 1.1	2.4; 0.7
Cohabitation duration (grouped)	15–19	1.6; 1.7	4.6[a]; 8.8[a]	1.3; 1.2	1.6; 1.5	1.2; 1.1	0.7; 0.7
Do not know any source for condoms	Yes: no source known	0.6; 0.4	4.3[a]; 8.6[a]	0.8; 0.7	0.6; 1.5	1.1; –	0.2; –
Heard family planning on TV last few months	Yes	1.5; 1.6	3.8[a]; 8.9[a]	1.0; 1.2	0.0; 2.1	0.9; 1.3	0.2; 2.3
Literacy	Cannot read at all	0.7; 0.8	3.8[a]; 4.1[a]	0.8; 0.8	0.8; 2.2	0.9; 0.9	0.2; 0.7
Getting medical help for self: not wanting to go alone	Not a big problem	1.6; 1.6	3.4[a]; 5.6[a]	1.2; 1.2	0.9; 1.4	1.1; 1.1	0.4; 0.4
Willing to care for relative with AIDS	Yes	2.8; 2.3	3.4[a]; 4.7[a]	2.0; 1.6	1.8; 1.5	1.8; 2.2	0.8; 2.0
Drugs to avoid HIV transmission to baby during pregnancy	Yes	1.6; 1.9	3.4[a]; 5.9[a]	1.2; 1.4	0.7; 1.8	–; –	–; –
Source of drinking water	Unprotected well	0.5; 0.7	3.4[a]; 3.4[a]	0.8; 0.9	0.5; 0.9	1.1; 1.0	0.1; 0.0
Living children with current pregnancy	3	1.5; 1.6	3.4[a]; 7.7[a]	1.3; 1.3	1.7; 3.0	1.2; 1.1	0.9; 0.4
Knows method	Yes	1.9; 2.3	3.4[a]; 4.3[a]	0.9; 0.8	0.1; 0.5	0.8; 0.6	0.5; 1.0
Respondent's occupation	Sales and services elementary occupations	1.6; 2.1	3.4[a]; 6.8[a]	1.1; 1.6	0.2; 2.6	0.9; 1.4	0.2; 1.4
Tuberculosis spread by: air when coughing or sneezing	Yes	1.4; 1.4	3.1[a]; 5.0[a]	1.1; 1.2	0.4; 1.3	1.0; 1.0	0.1; 0.0
Main wall material	Cement	1.6; 1.4	3.0[a]; 3.9[a]	1.4; 1.3	1.3; 1.7	1.2; 1.2	0.7; 1.3
Living children with current pregnancy	1	1.7; 1.3	7.8[a]; 2.0	1.6; 1.5	5.3[a]; 4.7[a]	1.4; 1.5	2.4; 3.9[a]
Living children with current pregnancy (grouped)	6+	0.5; 0.6	4.7; 9.6[a]	0.3; 0.3	13.3[a]; 34.2[a]	0.3; 0.2	6.2; 23.4[a]
Births in last 3 years	0	1.4; 1.7	3.6; 19.1[a]	1.6; 1.9	5.6[a]; 22.6[a]	1.5; 2.0	3.2; 14.7[a]
Births in last 3 years	1	0.8; 0.6	2.6; 14.0[a]	0.7; 0.6	4.5[a]; 17.2[a]	0.7; 0.5	2.6; 13.0[a]
Births in last 5 years	No births	1.2; 1.4	1.3; 8.6[a]	1.5; 1.9	3.8[a]; 19.5[a]	1.5; 2.0	2.6; 14.7[a]
Relationship to household head[b]	Wife	0.8; 0.8	1.0; 3.3[a]	0.6; 0.5	5.7[a]; 23.1[a]	1.0; 0.5	0.0; 18.4[a]
Ideal number of children	7	0.3; 0.5	4.0[a]; 2.8	0.3; 0.5	3.1[a]; 4.3[a]	0.3; 0.4	2.1; 3.3
Rohrer's index		0.9; 0.9	0.6; 0.9	0.7; 0.7	6.0[a]; 9.9[a]	0.8; 0.7	4.6[a]; 9.4[a]
Daughters at home	0	1.1; 1.0	0.4; 0.1	1.6; 1.9	4.3[a]; 17.8[a]	1.8; 1.9	4.8[a]; 15.0[a]
BMI		1.0; 1.0	0.1; 0.1	0.8; 0.8	6.2[a]; 10.1[a]	0.8; 0.7	4.7[a]; 9.5[a]
Current marital status	Married	0.8; 0.8	1.7; 2.6	0.5; 0.5	7.3[a]; 24.6[a]	0.9; 0.5	0.1; 19.6[a]
Currently/formerly/never in union	Currently in union/ living with a man	0.8; 0.8	1.1; 2.3	0.5; 0.5	6.0[a]; 24.4[a]	1.1; 0.5	0.2; 20.0[a]
Sons at home	0	1.1; 1.0	0.3; 0.2	1.6; 1.8	5.3[a]; 15.8[a]	1.7; 2.0	5.1; 15.4[a]
Daughters at home	3	0.7; 0.8	1.5; 1.3	0.5; 0.6	3.2[a]; 5.8[a]	0.5; 0.5	2.5; 4.3
Age in 5-year groups	45–49	0.9; 1.2	0.4; 1.1	0.4; 0.4	6.0[a]; 9.0[a]	0.4; 0.5	2.7; 4.6
Number of children 5 and under in household (de jure)	7	1.1; 0.6	0.0; 0.2	2.1; 0.6	0.2; 0.2	0.0; 0.0	22.8[a]; 55.5[a]

[a]Replicated in analysis. Specifically, these variables were above the false discovery rate 5% threshold in the discovery dataset, and had a nominal *P* value less than 0.05 in the replication dataset.
[b]Variables indicated with double asterisk were used as controls in the ex ante and/or post analysis, and were tested in a model with only the ex post variables as independent predictors.

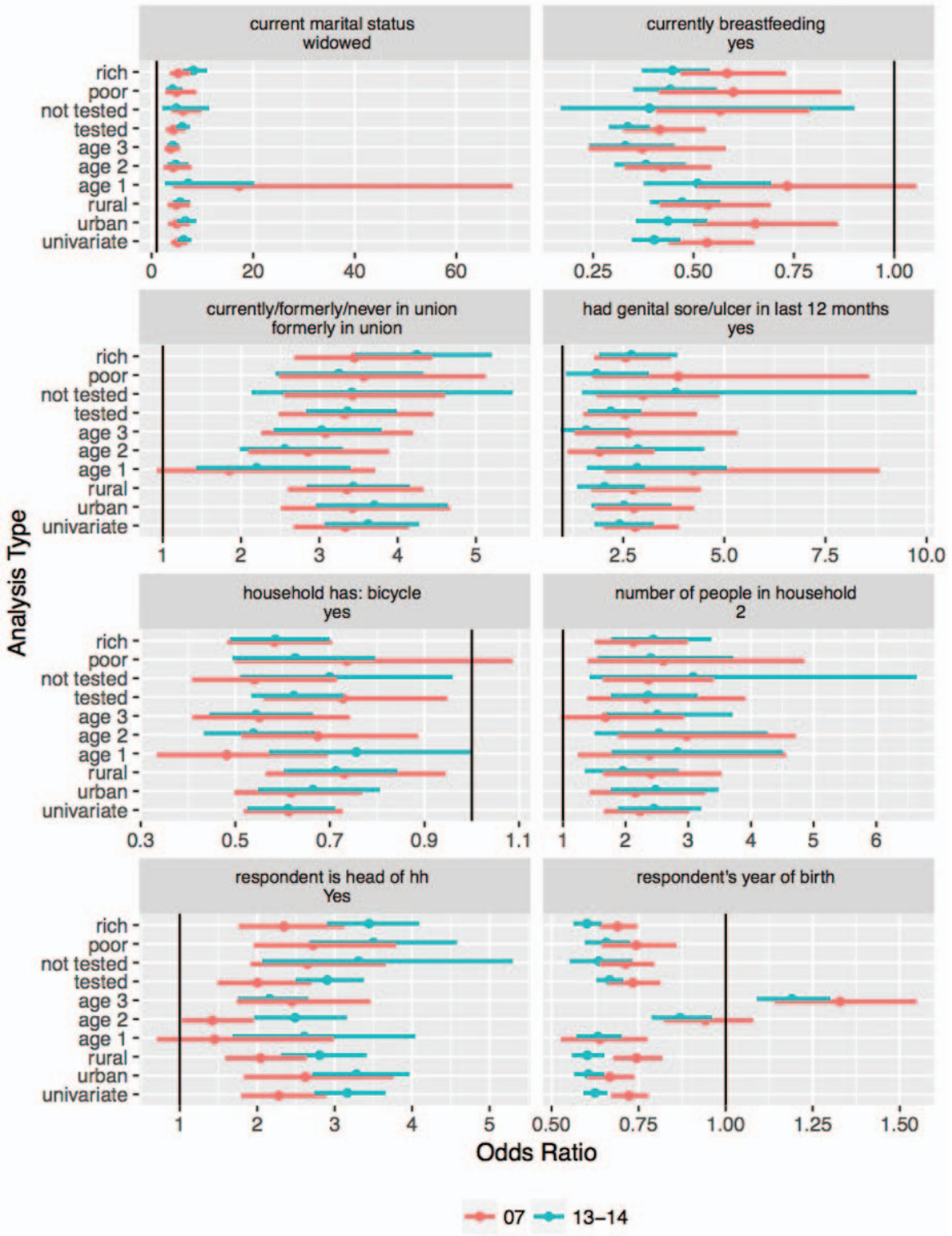

Fig. 3. Strength of association for univariate overall and population subgroups. Stratas include wealth index less than or equal to 3 and greater than 3 ('poor' and 'rich,' respectively), individuals that have had or have not had an HIV-1 test ('tested' and 'not tested'), individuals living a rural ('rural') or urban ('urban') areas, and of ages less than 23 ('age 1'), between 23 and 33 ('age 2'), and older than 33 ('age 3').

(three variables, fewer children confer increased risk), currently using a condom for contraception (increased risk), and an indicator for ownership of a bicycle in the household (decreased risk).

Figure 3 shows the sub-group associations for the variables that were replicated in univariate analyses in the overall sample in both surveys and in at least 17 out of the 18 subgroups examined (nine subgroups, as noted

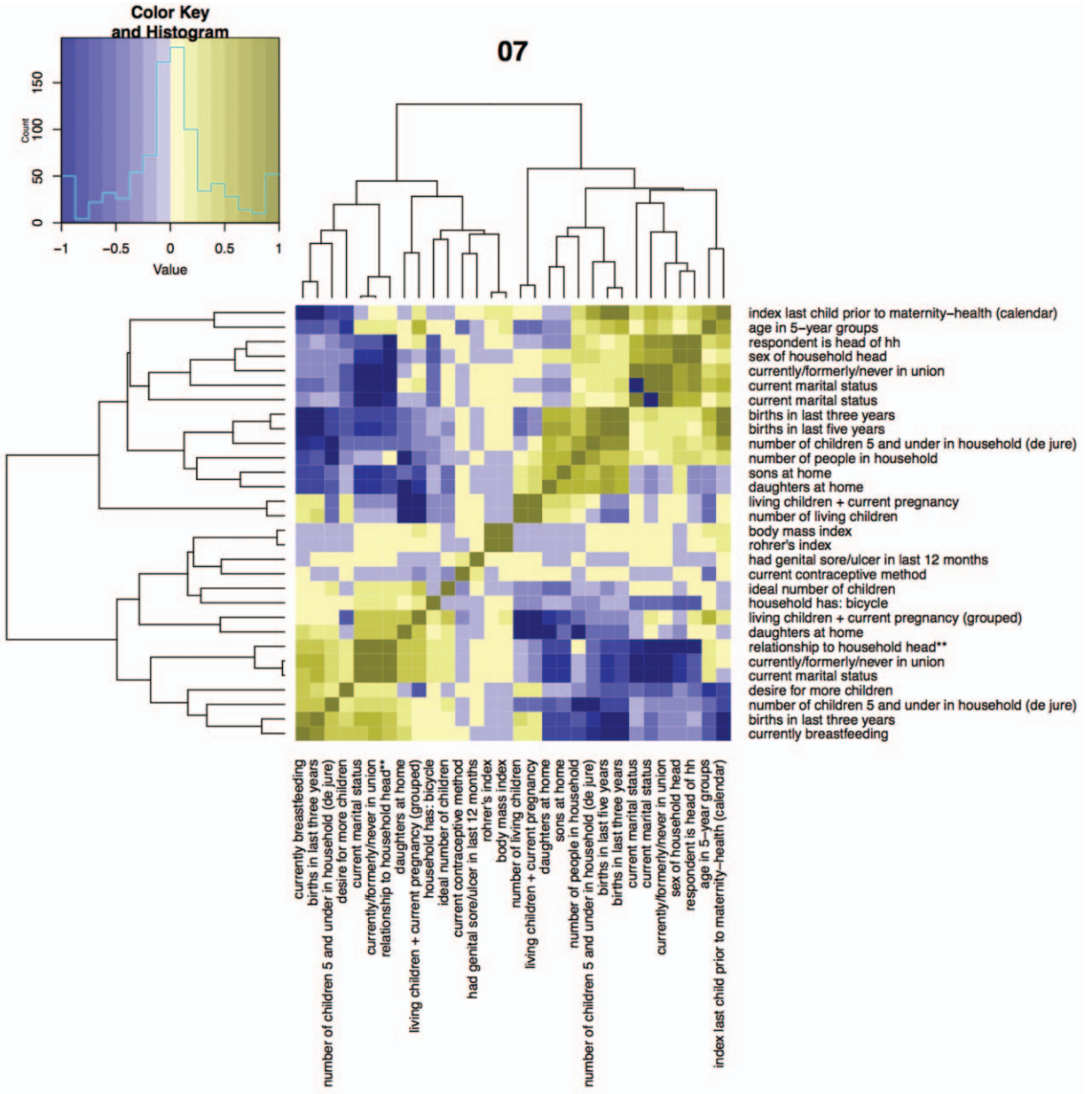

Fig. 4. Correlation matrix of variables replicated in the ex ante analysis. Top panel shows results for the 2007 survey and the bottom panel for the 2013–2014 survey. Variable clusters representing similar constructs appear in both surveys, such as variables that characterize marital status.

above, in each of the two surveys). A total of eight variables met these criteria (including several that were associated with HIV-1 in all full-sample analyses, denoted with the symbol asterisk (*): the indicators for widowhood and being formerly in a union*, being the head of the household*, having exactly two people in the household (relative to all other household sizes), age, reporting a genital ulcer in the past 12 months, owning a bicycle*, and currently breastfeeding*.

Two general categories of variables were associated with HIV-1 in the ex ante and ex post analyses but not in the univariate analyses (i.e. variables whose association with HIV-1 was 'uncovered' after adjustment, shown at the bottom variables in Table 2). These include variables related to the number of children who were different from those replicated in all or nearly all analyses (again, fewer children confer increased risk), and the anthropometric measurements BMI and Rohrer's index (higher index associated with lower HIV-1 risk).

Figure 4 shows the extent to which the variables that were replicated in the ex ante analysis are correlated and clustered among themselves. We observed that the correlation pattern between the 2007 and 2013–2014 surveys were strikingly similar. We hypothesized that this

may be partly a reflection of the variable construction process (e.g. being married would be expected to be strongly negatively correlated with being divorced), and partly of the likely stable social and environmental patterns in Zambia over this time period (Supplementary Figures SA1 and SA2, http://links.lww.com/QAD/B226).

In the three analyses and the two surveys, the variance explained in HIV-1 status ranged from 0.21 (in the ex post analysis of the 2007 survey) to 0.32 (in the univariate analysis of the 2013–2014 survey). The area under the curve in the six analyses ranged from 0.76 to 0.82 (specificity 0.75 at sensitivity 0.75; Supplementary Figure SA3, http://links.lww.com/QAD/B226).

Discussion

We describe the findings from the first XWAS of HIV-1 risk in nationally representative sero-surveys. Out of all the variables tested (688 in 2007 and 727 in 2013–2014, of which 688 were overlapping), we identify several candidate variables that are associated with HIV-1 whenever examined in multiple analyses and may present opportunities for identifying previously under-recognized risks. These include positive associations with widowhood/divorce/being formerly in union, being the head of the household, having a small household size, and reporting a genital ulcer in the past 12 months; and negative associations with breastfeeding and bicycle ownership. The reasons for these consistent associations may have different implications. A causal relationship may have implications for targeting and design of prevention interventions. A noncausal relationship (that is observed because of confounding or reverse causation) may still have benefits for testing programs that are interested in increasing testing among high-prevalence groups. The nature of the associations, therefore, deserves further discussion.

The reason for the positive association of widowhood with HIV-1 may be because of widows' engagement in high-risk behaviors for basic income and sustenance; it may also be partly caused by HIV-1 positivity among the widows' now-deceased husbands. Our study cannot tease apart the dominant causal pathway, and both may contribute to the association. The similar effect among divorced women is more consistent with risky behaviors following the loss of a spouse. Recent evidence also supports a causal role: a nationally representative survey of HIV-1 incidence in Rwanda from 2013 to 2014 found elevated rate of new infections in widows [29]. If widowhood and divorce lead to increased HIV-1 risk, then targeting of prevention interventions such as preexposure prophylaxis may mitigate the associated risk. If HIV-1 prevalence is higher among these women because of preexisting risk, then this finding may still help in guiding HIV-1 identification for early treatment and care that may reduce their risk of infecting others.

The relationship between current breastfeeding and HIV-1 risk is also notable. It is not a known factor that decreases risk of HIV-1 acquisition [30,31]. This association may indicate the decreased propensity to breastfeed among HIV-positive women. Although public health guidelines for breastfeeding among HIV-infected women has shifted over the past decade, breastfeeding has been recommended by the World Health Organization since 2010 [32,33]. As we find decreased risk of HIV-1 among women who breastfeed in both 2007 and 2013–2014 (in all three analyses and 17 subgroups), this finding may indicate the challenges of changing breastfeeding behaviors and the importance of finding effective approaches to behavior change in this domain.

The variables that we highlight were replicated in multiple analyses, but this study also identifies factors whose less consistent association with HIV-1 may nevertheless warrant additional consideration. Several variables related to method of contraception were positively associated with HIV-1 status in the univariate analyses, including hormonal contraceptives, condoms, and female condoms. Wealth was associated with HIV-1 in the univariate analysis (higher risk among wealthier women), but not in the adjusted models. No variable identifying educational attainment was associated with HIV-1 in the adjusted models. These assessments improve on the extant assessment of epidemiologic risk that are commonly presented along with the DHS data (and commonly used by the Joint United Nations Program on HIV and AIDS and others) [4,13]. The DHS stratifies HIV-1 risk by age, residence, marital status, education, and wealth. XWAS improves on such stratifications by reducing potential bias from failure to consider other relevant covariates, and by using an FDR that accounts for multiple comparisons.

The extent to which our findings are generalizable to other contexts is unknown. Extending HIV-1 XWAS to additional surveys across sub-Saharan Africa and over time, however, is readily feasible and will enable greater understanding of the generalizability and stability over time of our findings. We note that the putative variables we identified in common in the 2007 and 2013–2014 surveys had similar association sizes in both surveys. These similar association sizes and correlations point to the stability of social, behavior, and environment over time in Zambia.

The limitations of this study deserve explicit mention. First, we only tested variables with at least 90% complete data. Although we retained approximately 700 variables for analysis, some important variables could have been excluded because of missingness. Second, the error rate among self-reported variables may also bias results. Errors are more likely for some variables than for others. Any

nondifferential bias (e.g. individuals that report inaccurately in both HIV-1-positive and HIV-1-negative individuals) will lead to loss of power and correlations that are closer to null; however, we emphasize that sample sizes in our investigation are large. Third, self-reported variables may exhibit differential bias if participants answer differentially based on HIV-1 status. Differential bias may distort effects in unpredictable ways. Fourth, we could not assess association with incident or recent infection to mitigate chances of reverse causality. Although some DHS surveys also measure $CD4^+$ cell counts (that may proxy for duration of infection), the Zambia surveys did not, and we did not control for duration of infection (except through some indirect controlling by adjusting for age). It is plausible, for example, that a decrease in BMI is a consequence of HIV-1 rather than a cause. Challenges to causal identification are a generic issue in large-scale cross-sectional association studies, but such analyses nevertheless remain an important method to identify potential risk factors [34].

In conclusion, we report the findings from the first XWAS of HIV-1 risk from nationally representative surveys of social, economic, environmental, and behavioral factors in Zambia. We identify strong and consistent associations with widowhood, breastfeeding, and several other self-reported indicators that may be amenable to further investigations and interventions and that may be used to guide screening policies.

Acknowledgements

Funding: This work was supported in part by grants R01-AI127250 from the National Institute of Allergy and Infectious Diseases, R01-DA15612 from the National Institute on Drug Abuse, R00 ES023054 and R21 ES025052 from the National Institutes of Environmental Health Sciences, and U54 HG007963 from NIH Common Fund. The sponsors had no role in the design, interpretation or conclusions of this study.

Author contributions: E.B. and C.J.P. conceived the work and carried out the analyses. J.B. and J.P.A.I. critically assessed the methods and findings, and contributed to the study conceptualization and preparation of the manuscript.

Conflicts of interest

There are no conflicts of interest.

References

1. 90–90–90 - An ambitious treatment target to help end the AIDS epidemic. Available at: http://www.unaids.org/en/resources/documents/2014/90-90-90. [Accessed 10 December 2015]
2. Deeks SG, Lewin SR, Havlir DV. **The end of AIDS: HIV infection as a chronic disease.** *Lancet* 2013; **382**:1525–1533.
3. Staveteig S, Bradley S, Nybro E, Wang S. Demographic patterns of HIV testing uptake in sub-Saharan Africa. DHS Comparative Reports No. 30. ICF International. 2013.
4. UNAIDS AIDSinfo: Epidemiological status. Available at: http://aidsinfo.unaids.org/. [Accessed 31 April 2016]
5. UNAIDS 2014 Gap Report. Available at: http://www.unaids.org/en/resources/documents/2014/20140716_UNAIDS_gap_report. [Accessed 31 August 2015]
6. Demographic and Health Surveys. ICF International. Available at: http://www.measuredhs.com [Accessed 3 August 2015]
7. Opening Statement From Ambassador Deborah L. Birx, M.D., at the UNAIDS 37th Programme Coordinating Board Meeting. Available at: http://www.pepfar.gov/press/releases/2015/248739.htm. [Accessed 10 August 2016]
8. PEPFAR's Dr Deborah Birx urges sharper focus to halt HIV globally. Available at: https://www.fic.nih.gov/News/GlobalHealthMatters/january-february-2016/Pages/deborah-birx-pepfar-global-hiv-control.aspx. [Accessed 20 August 2016]
9. WHO HIV/AIDS strategic information: surveillance. Available at: http://www.who.int/hiv/strategic/surveillance/en/. [Accessed 31 August 2015]
10. De Cock KM, Rutherford GW, Akhwale W. **Kenya AIDS Indicator Survey 2012.** *J Acquir Immune Defic Syndr* 2014; **66 (Suppl 1)**:S1–S2.
11. Anderson S-J, Cherutich P, Kilonzo N, Cremin I, Fecht D, Kimanga D, *et al*. **Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study.** *Lancet* 2014; **384**:249–256.
12. Zambia DHS, 2007 - final report. Available at: http://dhsprogram.com/publications/publication-FR211-DHS-Final-Reports.cfm. [Accessed 13 June 2016]
13. Zambia DHS, 2013–14 - Final Report. Available at: http://dhsprogram.com/publications/publication-FR304-DHS-Final-Reports.cfm. [Accessed 31 August 2015]
14. Patel CJ, Ioannidis JP. **Studying the elusive environment in large scale.** *J Am Med Assoc* 2014; **311**:2173–2174.
15. Ioannidis J. **Why most published research findings are false.** *PLoS Med* 2005; **2**:e124.
16. Tzoulaki I, Patel CJ, Okamura T, Chan Q, Brown IJ, Miura K, *et al*. **A nutrient-wide association study on blood pressure.** *Circulation* 2012; **126**:2456–2464.
17. Patel CJ, Rehkopf DH, Leppert JT, Bortz WM, Cullen MR, Chertow GM, Ioannidis JP. **Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey.** *Int J Epidemiol* 2013; **42**:1795–1810.
18. Patel CJ, Cullen MR, Ioannidis JP, Butte AJ. **Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels.** *Int J Epidemiol* 2012; **41**:828–843.
19. Patel CJ, Bhattacharya J, Butte AJ. **An environment-wide association study (EWAS) on type 2 diabetes mellitus.** *PloS One* 2010; **5**:e10746.
20. Ioannidis JP, Tarone R, McLaughlin JK. **The false-positive to false-negative ratio in epidemiologic studies.** *Epidemiology* 2011; **22**:450–456.
21. Patel CJ, Ioannidis JP. **Placing epidemiological results in the context of multiplicity and typical correlations of exposures.** *J Epidemiol Community health* 2014; **68**:1096–1100.
22. Dong Y, Peng C-YJ. **Principled missing data methods for researchers.** *Springerplus* 2013; **2**:222.
23. Benjamini Y, Hochberg Y. **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Series B Stat Methodol* 1995; **57**:289–300.
24. White H. **A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.** *Econometrica* 1980:817–838.
25. Rutstein SO, Johnson K. *The DHS wealth index*. DHS Comparative Reports no. 6. *Calverton: ORC Macro*; 2004.
26. Yang J, Ferreira T, Morris AP, Medland SE. **Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits.** *Nat Genet* 2012; **44**:369–375.
27. Patel CJ, Cullen MR, Ioannidis JPA, Rehkopf DH. **Systematic assessment of the correlation of household income with infectious, biochemical, physiological factors in the United States.** *Am J Epidemiol* 2014; **181**:171–179.

28. Johnson RA, Wichern DW. *Applied multivariate statistical analysis*:. NJ: Prentice Hall Englewood Cliffs; 1992.
29. Remera E, Kanters S, Mulidabigwi A, *et al.* 2013-14 Rwanda HIV incidence household survey: understanding HIV epidemic in Rwanda. *CROI*, 2016, Boston, USA.
30. Serwadda D, Wawer MJ, Musgrave SD, Sewankambo NK, Kaplan JE, Gray RH. **HIV risk factors in three geographic strata of rural Rakai District, Uganda.** *AIDS* 1992; **6**:983–990.
31. Cain D, Simbayi L, Kalichman S, Cherry C, Jooste S, Mfecane S. Risk factors for HIV-AIDS among youth in Cape Town, South Africa. 2015.
32. World Health Organization. HIV and infant feeding: update. 2006. Available at: http://apps.who.int/iris/bitstream/10665/43747/1/9789241595964_eng.pdf. [Accessed 17 November 2016]
33. World Health Organization. Guidelines on HIV and infant feeding: principles and recommendations for infant feeding in the context of HIV and a summary of evidence. 2016. Available at: https://www.unicef.org/aids/files/hiv_WHO_guideline_on_-HIV_and_IF.pdf. [Accessed 17 November 2016]
34. Ioannidis J. **Exposure wide epidemiology: revisiting Bradford Hill.** *Statistics in medicine* 2015; **35**:1749–1762.