Is the place of residence predictive of HIV acquisition in rural South Africa? Results from an ongoing population-based cohort in KwaZulu-Natal

Adrian Dobra¹*, Till Bärnighausen²,³,⁴,⁵, Alain Vandormael²,⁶, Frank Tanser²,³,⁷,⁸

¹ Department of Statistics, Department of Biobehavioral Nursing and Health Informatics, Center for Statistics and the Social Sciences and Center for Studies in Demography and Ecology, University of Washington, Seattle, WA, USA
² Africa Health Research Institute, Durban, South Africa
³ Institute of Epidemiology and Health Care, University College London, London, UK
⁴ Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, USA
⁵ Institute for Public Health, University of Heidelberg, Heidelberg, Germany
⁶ Research Innovation and Sequencing Platform (KRISP), Durban, South Africa
⁷ Centre for the AIDS Programme of Research in South Africa (CAPRISA), University of KwaZulu-Natal, Durban, South Africa
⁸ School of Nursing and Public Health, University of KwaZulu-Natal, South Africa
* adobra@uw.edu

Abstract

Very little is known at the present time about the role of the place of residence of mobile individuals in the transmission of HIV outside high transmission areas such as mining settlements, transport corridors, or poor urban and periurban communities. The main objective of this study is to bridge a widening knowledge gap that is caused by new mobility dynamics of men and women that live in rural South Africa.

This study makes use of data from one of the most comprehensive demographic surveillance site in Africa that is characterized by high adult HIV prevalence, high levels of poverty and unemployment and frequent residential changes. Its main objective was to determine which places of residence are predictive of HIV acquisition. Between 2004 and 2016, residence changes were recorded for 21,015 individuals over 105,614 person-years. These individuals were HIV negative at baseline. This is one of the largest HIV incidence cohorts in the world in terms of the number of individuals under surveillance, and the number of person-years of surveillance. Over the study duration, there were a total of 3,264 HIV seroconversions. We organized our data in two 48-dimensional contingency tables, one for men and one for women that cross-classify the study participants with respect to the locations of their residencies, their age, and whether they seroconverted. We used state of the art Bayesian methods for structural learning of graphical loglinear models to identify mobility graphs which encode the strongest multivariate predictive relationships supported by the data. Our analysis of the mobility graphs shows that whether men move farther away from their original places of residence is predictive of their likelihood of HIV seroconversion (OR = 2.003, 95% CI = [1.718,2.332]), but similar residential changes do not seem to predictive of HIV seroconversion in women given their age. The location of the original place of residence is not a strong predictor for HIV acquisition in both men and women given knowledge of age and whether residential moves over longer distances have occurred.
The results of this study which is one of the largest individual-level longitudinal study of mobility patterns and HIV to date, provide evidence that geodemographic segmentation based on the history of residential locations, gender and age can constitute a reliable, objective, cost-effective way to ensure optimal allocation of HIV prevention intervention strategies.

Introduction

Historically, human mobility has been one of the key drivers in the spread of HIV at a global scale [1-13]. Many studies have provided significant evidence linking increased population mobility with multiple sexual partners, reduced condom use, increased risky behavior (e.g., encounters with commercial sex workers, engaging in transactional sex) [1-10], increased sexual behavior [1-10], and increased likelihood of HIV acquisition [11,12]. The sociodemographic, geographic and economic characteristics of the places in which mobile individuals establish their residence for shorter or longer periods of time seem to influence their likelihood of HIV seroconversion. Mining settlements, transport corridors, or poor urban and periurban communities exacerbate the effect of the risk factors of HIV acquisition, and are referred to as high transmission areas [12,25,26]. It has been empirically demonstrated that an individual’s risk of acquisition of HIV is strongly driven by community-level HIV prevalence [27], community-level migration intensity [28], mean number of sexual partners in the surrounding local community [29], as well as ART coverage and population viral load in the local community, respectively [27,30]. These community-level risk factors confer substantial additional risk of new HIV infection after controlling for a suite of well-established individual-level risk factors.

In this study we focus on South Africa, and examine the role of place of residence as a driver of HIV transmission risk in a comprehensive population-based demographic surveillance site in the KwaZulu-Natal Province – the Africa Centre, now Africa Health Research Institute (AHRI). This is one of the largest HIV incidence cohorts in the world in terms of the number of individuals under surveillance, and the number of person-years of surveillance [31]. Specifically, we analyze mobility patterns of 21,015 individuals who were HIV negative at baseline, and were registered in the AHRI demographic surveillance system. Their mobility patterns are defined by residential histories over the study period. The AHRI site is characterized by high adult HIV prevalence (24% in adults aged 15 years 30 and older in 2011), and high levels of poverty and unemployment (in 2010, 67% of adults over the age of 18 in the rural study area were unemployed) [27].

In South Africa, the risk of HIV infection has been shown to be increased by human mobility [3,32,33]. South Africa is one of the countries with the highest burden of HIV, and has a long history of internal labor migration of men that periodically leave their areas of permanent residence to seek temporary employment in mines and factories due to the scarcity of local employment [34]. During the apartheid era which imposed travel restrictions for blacks, women were typically left behind to take care of families, while men submitted remittances back to their households. Because of economic conditions, this way of life continues to exist in poor rural regions of South Africa including this rural study community. However, as opposed to the apartheid era, in the last decade both men and women from the AHRI surveillance site frequently establish residences for various periods of time to work or for many other reasons in locations within the KwaZulu-Natal Province (e.g., Richards Bay or Durban), or in other more distant locations in South Africa (e.g., Johannesburg, Pretoria or Cape Town) [24]. Given these new mobility dynamics, very little is known at the present time about the role of the place of residence of mobile individuals in the transmission of HIV. It is important to fill...
this persistent knowledge gap, because only once it is known how the place of residence affects the transmission of HIV outside the well characterized high transmission areas can policy and programs be designed that are effectively targeted at the segments of the population that are at the highest risk of HIV acquisition. The main objective of this study is to contribute to this understanding. The geographical location of the AHRI demographic surveillance area is ideal for our aim.

Methods

Setting

The study was conducted in the Africa Health Research Institute (AHRI) Population Intervention Platform Study Area (PIPSA), formerly the Africa Centre Demographic Information System (ACDIS), in uMkhanyakude District, KwaZulu-Natal Province. PIPS A was commissioned in 2000 by the Wellcome Trust as a platform for longitudinal population based studies of epidemiology and intervention research. This rural study area covers 438 km$^2$, and comprises approximately 11,000 households with 100,000 individuals. This community is characterized by high HIV prevalence (29% of the adult population aged 15–49 are infected with HIV) [31], frequent migration (38% of men and 32% of women were non-resident in 2008) [35], low marital rates (only 23% of men and 31% of women have ever been married) [36], late marriage especially for men, polygamous marriages (about 14% of all marriages for men and 12% of all marriages for women) [36] and multiple sexual partnerships, as well as by poor knowledge and disclosure of HIV status [24]. Incidence peaked at 6.6 per 100 person-years in women aged 24 years, and at 4.1 per 100 person-years in men aged 29 years over the same period [23].

For over 15 years, PIPS A has continuously collected longitudinal surveillance data on a range of health care and social intervention exposures, as well as health, socio-economic and behavioral outcomes [31]. Population-based HIV surveillance and sexual behavior surveys take place annually. Since 2003, annual HIV testing became part of household surveillance. Eligible participants are interviewed in private by trained fieldworkers, who also extract blood from consenting individuals by finger-prick for HIV testing. The longitudinal dynamics of participation in the HIV survey is described in detail elsewhere [37].

Ethics approval for all surveillance data collection activities was obtained from the Biomedical Research Ethics Committee of the University of KwaZulu-Natal (Durban, South Africa).

Cohort description

From the entire population under surveillance in PIPS A between January 1, 2004 and December 31, 2016, we selected those individuals who consented to test at least twice for HIV after the age of 15, and whose first test was negative. A number of 8,857 men and 12,158 women satisfy these inclusion criteria. The date of HIV seroconversion was assumed to occur according to a uniformly random distribution between the date of the last negative and first positive HIV test [38].

PIPS A collects data about all the individuals that are members of a family unit or a household in the rural study area irrespective of the current residence status. It collects longitudinal residential information about the exact periods of time each study participant spent living in each location. Fieldworkers record changes in residence as the origin place of residence, the destination place of residence and the date of the move. Residencies can be located inside or outside the rural study area. The residential
locations inside the rural study area have been comprehensively geolocated to an accuracy of <2m [39]. The repeat-testers can change their place of residence multiple times: they can move between two residencies located inside the rural study area, between two residencies located outside the rural study area, or between a residence inside the rural study area and another residence outside the rural study area. The relevance of looking whether repeat-testers have resided outside the rural study area comes from the findings of [24]: their results indicate that, for the same rural study area, the risk of HIV acquisition is significantly increased for both men and women when they spend more time outside the rural study area, or when they change their residencies over longer distances.

For the purpose of this study, the geolocations of the homesteads have been mapped into 45 non-overlapping communities that cover the rural study area – see Figures S1 Fig and S2 Fig in Supporting information. The division of the rural study area into communities is motivated by the results of [40]. Their study identified a significant geographical variation in HIV incidence in the same rural study area. Specifically, they identified three large irregularly-shaped clusters of new HIV infections. Although these clusters cover only 6.8% of the rural study area, about 25% of the sero-conversions that occurred over this study’s period are associated with residencies in them. This suggests the existence of clear corridors of HIV transmission inside the rural study area. Together, the results of [24] and [40] indicate that men and women who reside outside the rural study area, or occupy residencies that are located in the corridors of HIV transmission inside the rural study area are at an increased risk of acquiring HIV.

We note that the exposure period for a repeat-tester starts at the time of their first HIV test, and ends at their HIV seroconversion date for seroconverters, or at the time of their last HIV negative test for those that did not seroconvert. The residential locations occupied before seroconversion could have contributed to changes in sexual behavior that led to HIV acquisition, while residential locations occupied after seroconversion could be associated with repeat-testers seeking family support, health care or moving away to avoid social stigma [6,22]. For this reason, the residential locations occupied by seroconverters after they acquired HIV were discarded.

**Statistical analyses**

We determined in which of the 45 communities each of the 8,857 men and 12,158 women lived during the study period. This information was recorded as binary variables C1, C2, . . . , C45 with levels “yes” or “no”. We also determined whether a repeat-tester moved outside the rural study area. This information was recorded as a binary variable Outside with levels “yes” or “no”. Furthermore, we determined whether a repeat-tester has seroconverted, and whether a repeat-tester was younger than 30 years at start of their observation period. This information was recorded as two additional binary variables Seroconverted and Young with levels “yes” or “no”. For example, a repeat-tester that lived in communities C1 and C2, moved outside the rural study area, was older than 30 years at baseline, and has seroconverted, would have C1 = C2 = Outside = Seroconverted = yes and C3 = . . . = C45 = Young = no.

These 48 binary variables define two dichotomous contingency tables with $2^{48}$ cells, one table for men and another table for women. These tables which we call mobility tables are hyper-sparse: most of their counts are zero. The mobility table for men has only 598 positive counts – see the data file S1 Data in Supporting information. Among these counts, there are 292 (48.83%) counts of 1, 48 (8.03%) counts of 2, 30 (5.02%) counts of 3, 28 (4.68%) counts of 4, and 13 (2.17%) counts of 5. The top five largest counts are 192, 186, 180, 177 and 168, respectively. They correspond with men that were less than 30 years old at the start of their observation period, did not seroconvert by the end of their observation period, never moved outside the rural study area, and
lived in exactly one of these communities: C7, C37, C40, C39 and C22. The mobility
table for women has only 939 positive counts – see the data file S2 Data in Supporting
information. Among these counts, there are 534 (56.87%) counts of 1, 98 (10.44%)
counts of 2, 30 (3.19%) counts of 3, 15 (1.60%) counts of 4 and 15 (1.60%) counts of 5.
The top five largest counts are 185, 176, 175, 172 and 171. They correspond with
women that were less than 30 years old at the start of their observation period, did not
seroconvert by the end of their observation period, never moved outside the rural study
area, and lived in exactly one of these communities: C22, C10, C25, C39, and C7,
respectively. Tables 1 and 2 give the cross-classification of the men and women in the
study with respect to the binary variables Seroconverted, Young and Outside. These
tables have 2^3 = 8 cells, and are the three dimensional marginal tables of the 48
dimensional contingency tables in the data files S1 Data and S2 Data in Supporting
information.

Table 1. Cross-classification of the 8,857 men repeat-testers.

<table>
<thead>
<tr>
<th>Seroconverted</th>
<th>Young</th>
<th>Outside</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>889</td>
<td>209</td>
<td>5368</td>
</tr>
<tr>
<td>Yes</td>
<td>73</td>
<td>26</td>
<td>439</td>
</tr>
</tbody>
</table>

The men repeat-testers are cross-classified by their HIV seroconversion status
(Seroconverted: Yes/No), whether they moved outside the study area (Outside: Yes/No)
and whether they were less than 30 years old at the start of the study (Young: Yes/No).

Table 2. Cross-classification of the 12,158 women repeat-testers.

<table>
<thead>
<tr>
<th>Seroconverted</th>
<th>Young</th>
<th>Outside</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>2232</td>
<td>242</td>
<td>5379</td>
</tr>
<tr>
<td>Yes</td>
<td>210</td>
<td>35</td>
<td>1559</td>
</tr>
</tbody>
</table>

The women repeat-testers are cross-classified by their HIV seroconversion status
(Seroconverted: Yes/No), whether they moved outside the study area (Outside: Yes/No)
and whether they were less than 30 years old at the start of the study (Young: Yes/No).

Statistical modeling framework

Graphical models represent our main modeling framework. These models characterize
the multivariate dependency structure (e.g., independence or conditional independence)
among random variables using graphs [41]. Graphical models have received considerable
attention in the literature, and have a vast domain of applicability that encompasses all
the scientific fields in which the analysis of multivariate datasets is key, e.g. biology,
neuroscience, social sciences, and economics. A key step in data analysis with graphical
models is estimating the underlying graph which is called the structural learning
problem [42,43]. This is a very difficult computational problem when many random
variables are involved. Bayesian methods provide a flexible framework for incorporating
uncertainty of the graph structure: inference and estimation are based on averages of
the posterior distributions of quantities of interest, weighted by the corresponding
posterior probabilities of graphs.
In this paper we make use of a Bayesian framework for solving the structural learning problem that is suitable for the analysis of hyper-sparse contingency tables with \( p = 48 \) variables. This framework was developed in [44], and determines graphical loglinear models that are a special type of hierarchical loglinear models [41, 45]. A graphical model for a random vector \( \mathbf{X} = (X_1, X_2, \ldots, X_p) \) is specified by an undirected graph \( G = (V, E) \) where \( V = \{1, \ldots, p\} \) are vertices or nodes, and \( E \subset V \times V \) are edges or links [41]. A vertex \( i \in V \) of \( G \) corresponds with variable \( X_i \). The absence of an edge between vertices \( i \) and \( j \) in \( G \) means that \( X_i \) and \( X_j \) are conditional independent given the remaining variables \( X_{V \setminus \{i,j\}} \). The graph \( G \) also has a predictive interpretation. Denote by \( \text{nbd}_G(i) = \{j \in V : (i, j) \in E\} \) the neighbors of vertex \( i \) in \( G \). Then \( X_i \) is conditionally independent of \( X_{V \setminus \{\text{nbd}_G(i) \cup \{i\}\}} \) given \( X_{\text{nbd}_G(i)} \) which implies that, given \( G \), a mean squared optimal prediction of \( X_i \) can be made from the neighboring variables \( X_{\text{nbd}_G(i)} \). The structural learning problem estimates the structure of \( G \) (i.e., which edges are present or absent in \( E \)) from the available data \( \mathbf{x} = (x^{(1)}, \ldots, x^{(n)}) \) by sampling from the posterior distribution of \( G \) conditional on the data \( \mathbf{x} \), i.e.

\[
P(G \mid \mathbf{x}) = \frac{P(G)P(\mathbf{x} \mid G)}{\sum_{G' \in \mathcal{G}_p} P(G')P(\mathbf{x} \mid G')},
\]

where \( P(G) \) is a prior distribution on the graph space \( \mathcal{G}_p \) with \( p \) variables, and \( P(\mathbf{x} \mid G) \) is the marginal likelihood of the data conditional on \( G \) [42]. We use a prior on the space of graphs \( \mathcal{G}_p \) that encourages sparsity by penalizing for the inclusion of additional edges in the graph \( G = (V, E) \) [42].

\[
P(G) \propto \left( \beta \frac{1}{1-\beta} \right)^{|E|} = \left( \prod_{i=1}^{p} \left( \beta \frac{1}{1-\beta} \right)^{|\text{nbd}_G(i)|} \right)^{1/2},
\]

where \( \beta \in (0, 1) \) is set to a small value, e.g. \( \beta = 1/1128 \approx 0.00089 \). Under this prior, the expected number of edges for a graph is 1. This means that sparser graphs with few edges receive larger prior probabilities compared with denser graphs in which most edges are present.

Determining the graphs with the highest posterior probabilities Eq (1) is a complex problem since the number of possible undirected graphs \( 2^\binom{p}{2} \) becomes large very fast as \( p \) increases. For example, our two mobility tables involve \( p = 48 \) variables, and the number of possible undirected graphs in \( \mathcal{G}_{48} \) is approximately \( 10^{125} \). This motivated the development of computationally efficient search algorithms for exploring large spaces of graphs that have the ability to move quickly towards high posterior probability regions by taking advantage of local computation. Among them, the birth-death Markov chain Monte Carlo (BDMCMC) algorithm [44] determines graphical loglinear models. BDMCMC is a trans-dimensional MCMC algorithm that is based on a continuous time birth-death Markov process [46]. Its underlying sampling scheme traverses \( \mathcal{G}_p \) by adding and removing edges corresponding to the birth and death events. This algorithm is implemented in the package BDgraph [47, 48] for R [49].

By employing the BDgraph package, we ran the BD-MCMC algorithm for 250,000 iterations to sample graphs from the posterior distribution (1) on \( \mathcal{G}_{48} \) for the mobility tables for men and women. Figures S3 Fig and Fig S4 Fig in Supporting information give the estimated posterior inclusion probabilities of the \( \binom{p}{2} = 1128 \) edges across iterations. We see that, after about 50,000 iterations, the subsequent posterior edge inclusion estimates stabilize. For this reason, the first 50,000 sampled graphs were discarded as burn-in, and the remaining 200,000 sampled graphs were used to estimate posterior edge inclusion probabilities.
Results

Descriptive summaries

We recorded residence changes for 21,015 repeat-testers over 105,614 person-years. Specifically, there were 8,857 men repeat-testers under surveillance for a total of 40,659 person-years with a median observation period of 3.72 years (IQR = 4.00). There were 12,158 women repeat-testers under surveillance over 64,955 person-years with a median observation period of 4.41 years (IQR = 5.47). During the study period, a number of 3,264 HIV seroconversions have been recorded: 806 men and 2,458 women. Table 3 gives seroconversion rates stratified by gender, age (younger or older than 30 years at baseline), and residence outside the rural study area. The largest seroconversion rate 22.47% (95% CI: 21.49-23.45) is for young women who resided in the rural study area for their entire exposure period. The seroconversion rate for young women who resided outside the study area is slightly lower: 19.20% (95% CI: 17.88-20.52). The largest seroconversion rate for men is 13.24% (95% CI: 11.76-14.72), and corresponds to the young group that moved outside the rural study area. The seroconversion rate for young men who did not move outside the rural study area is significantly lower: 7.56% (95% CI: 6.88-8.24). Table 3 also shows that the seroconversion rates for both men and women in the older age group are higher for the repeat-testers that moved outside the study area as compared to the repeat-testers that did not move outside the study area.

Table 3. Seroconversion rates and 95% CIs for repeat-testers.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Young Outside</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Men</td>
<td>7.59</td>
<td>11.06</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td>(5.91,9.26)</td>
<td>(7.05,15.07)</td>
<td>(6.88,8.24)</td>
</tr>
<tr>
<td>Women</td>
<td>8.60</td>
<td>12.64</td>
<td>22.47</td>
</tr>
<tr>
<td></td>
<td>(7.49,9.71)</td>
<td>(8.72,16.55)</td>
<td>(21.49,23.45)</td>
</tr>
</tbody>
</table>

The 8,857 men and 12,158 women repeat-testers are cross-classified by whether they moved outside the rural study area (Outside: Yes/No) and whether they were less than 30 years old at the start of the study (Young: Yes/No).

We determined the number of men and women that moved their residence between any two communities, or between a community and a location outside the rural study area. The resulting mobility flow diagrams are shown in Fig 1 and Fig 2. We see that, while men and women move between the 45 communities, substantially larger flows are associated with changes of residencies to and from locations outside the rural study area.

Table 4 gives a summary of the frequency of residential movements inside the rural study area, and also between a location outside the rural study area and another location inside or outside the rural study area by age group and gender. Women in the 20-24 age group move outside the rural study area more often than men in the same age group (26.56% vs. 23.31%). Residential movements outside the rural study area become less frequent for women in the 25-29 age group, but are comparable in frequency with residential movements of men in the 25-29 age group. However, men in the 30-34 age group move to and from locations outside the rural study area more frequently than women in the 30-34 age group. Residential movements outside the rural study area of women become significantly less frequent in the age groups 35-39, 40-44 and older than 45 as compared to residential movements of men in the same age group. Residential movements inside the rural study area of both men and women are substantially less frequent than residential movements to and from a location outside the rural study area.
Fig 1. Mobility flows for men repeat-testers. These are the flows between the 45 communities labeled C1, C2, ..., C45 and locations outside the rural study area associated with men repeat-testers who changed their residencies.
Fig 2. Mobility flows for women repeat-testers. These are the flows between the 45 communities labeled C1, C2, ..., C45 and locations outside the rural study area associated with women repeat-testers who changed their residencies.
Table 4. Residential mobility of the repeat-testers by gender and age group.

<table>
<thead>
<tr>
<th>Age Stratum (years)</th>
<th>Men At least once (95% CI)</th>
<th>Men Two or more times (95% CI)</th>
<th>Women At least once (95% CI)</th>
<th>Women Two or more times (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15–19</td>
<td>9.92 (9.13,10.71)</td>
<td>2.53 (2.11,2.95)</td>
<td>13.99 (13.13,14.84)</td>
<td>4.05 (3.56,4.54)</td>
</tr>
<tr>
<td>20–24</td>
<td>23.31 (22.00,24.62)</td>
<td>9.31 (8.41,10.21)</td>
<td>26.56 (25.32,27.80)</td>
<td>12.30 (11.37,13.22)</td>
</tr>
<tr>
<td>25–29</td>
<td>23.45 (21.46,25.45)</td>
<td>9.65 (8.26,11.04)</td>
<td>22.12 (20.47,23.77)</td>
<td>8.81 (7.69,9.94)</td>
</tr>
<tr>
<td>30–34</td>
<td>18.49 (16.02,20.95)</td>
<td>7.14 (5.51,8.78)</td>
<td>11.60 (10.02,13.18)</td>
<td>3.87 (2.91,4.82)</td>
</tr>
<tr>
<td>35–39</td>
<td>16.11 (13.37,18.86)</td>
<td>7.55 (5.57,9.52)</td>
<td>8.82 (7.40,10.25)</td>
<td>3.42 (2.39,4.34)</td>
</tr>
<tr>
<td>40–44</td>
<td>10.48 (8.08,12.87)</td>
<td>3.97 (2.44,5.49)</td>
<td>5.73 (4.66,6.81)</td>
<td>2.39 (1.69,3.10)</td>
</tr>
<tr>
<td>≥45</td>
<td>12.75 (10.69,14.80)</td>
<td>6.92 (5.35,8.48)</td>
<td>5.94 (5.02,6.87)</td>
<td>3.09 (2.41,3.77)</td>
</tr>
</tbody>
</table>

Outside residence changes

Inside residence changes

Percentages by age group of men and women repeat-testers who changed residences between a location outside the rural study area and another location inside or outside the rural study area (outside residence changes, upper panel), or between two locations inside the rural study area (inside residence changes, lower panel).

Graphical loglinear models for mobility tables

Fig 3 shows a heatmap of the estimated posterior inclusion probabilities of edges connecting the 48 binary variables cross-classified in the mobility tables for men and women. These estimates are based on the 200,000 graphs sampled with the BDMCMC algorithm. For the men’s mobility table, 115 (10.20%) posterior edge inclusion probabilities are 0, and 993 (88.03%) are 1. A number of 18 and 2 edges have estimated posterior inclusion probabilities in (0, 0.5) and [0, 0.5], respectively. For the women’s
mobility table, 100 (8.87%) posterior edge inclusion probabilities are 0, and 1,013 (89.80%) are 1. A number of 6 and 9 edges have estimated posterior inclusion probabilities in (0, 0.5), [0.5, 1), respectively. We use the median graph which includes the edges with estimated posterior inclusion probabilities greater than 0.5 as our estimate of the conditional independence graph. The median graph for men’s mobility table has 995 edges, while the median graph for women’s mobility table has 1,022 edges. We refer to these two graphs as men’s and women’s mobility graphs.

Fig 3. Heatmap of the estimated posterior probabilities of edge inclusion. The matrix entries below the main diagonal show the estimated posterior edge inclusion probabilities for men’s mobility, while the matrix entries above the main diagonal show the estimated posterior edge inclusion probabilities for women’s mobility. Darker shades of blue indicate smaller (closer to 0) matrix entries, while darker shades of red indicate larger (closer to 1) matrix entries.

The overall structure of the two mobility graphs is remarkably similar. In the men’s mobility graph, the vertex associated with the variable Outside is connected with the
vertices associated with 33 out of the 45 communities – see the map S1 Fig in Supporting information. The subgraph that involves vertices associated with the 45 communities is dense: it has 961 edges – 97.07% of the 990 possible edges. In the women’s mobility graph, the vertex Outside is connected with vertices associated with 39 out of 45 communities – see the map S2 Fig from Supporting information. The subgraph associated with the 45 communities is also dense: it has 981 edges – 99.09% of the 990 possible edges. In both graphs, there is no edge between the vertices associated with variables Seroconverted and Young, and the community vertices. This implies that, conditional on the variable Outside, the variables Seroconverted and Young are independent of the community variables C1, . . ., C45 for both men and women.

The most relevant differences between the two mobility graphs are related to the edges that link the variables Outside, Seroconverted and Young. For men (Fig 4), vertex Outside is connected with vertex Seroconverted, but the edges between vertices Outside and Young, and between vertices Seroconverted and Young are missing. For women (Fig 5), the situation is reversed: the edges between vertices Outside and Young, and between vertices Seroconverted and Young are present, but the edge between vertices Outside and Seroconverted is missing. This has the following implications: (a) for men, variable Young is independent of variables Outside and Seroconverted; (b) for men, only variable Outside is predictive of variable Seroconverted; (c) for women, variable Young is predictive of variable Seroconverted; and (d) for women, given variable Young, variable Seroconverted is independent of variable Outside.

Fig 4. Subgraph of men’s mobility graph. This subgraph shows the edges between vertices Outside, Seroconverted and Young.

Since the structure of interactions among variables Outside, Seroconverted and Young is essential for our understanding of the mobility data, we performed a second statistical analysis of the three-way tables cross-classifying these variables – see Tables 1 and 2. However, this time we followed a classical approach to hierarchical loglinear model determination [50,51] that also solves the structural learning problem, but is conceptually different from the Bayesian approach implemented in the BDMCMC algorithm. We note that this classical approach is suitable for analyzing these two tables because they involve only three variables and they do not contain any counts of 0.
Fig 5. Subgraph of women’s mobility graph. This subgraph shows the edges between vertices Outside, Seroconverted and Young.

However, this approach is not feasible for analyzing the 48-dimensional mobility tables for men and women due to sparsity and the number of variables involved. Specifically, we fitted the eight hierarchical loglinear models that contain main effects for variables Outside, Seroconverted and Young, and also one, two or all three of the pairwise interactions between these variables. The results are presented in Tables 5 and 6.

For men, the loglinear model that contains interactions between variables Seroconverted and Outside, and between variables Outside and Young, and the loglinear model that contains all three pairwise interactions do not fit the data well: the p-values for the likelihood ratio test against the saturated loglinear model are 0.348 and 0.215, respectively. The other six hierarchical models fit the data well at the significance level $\alpha = 0.05$. To select the most relevant model among the remaining six models, we calculated their AIC and BIC. The smallest values for both AIC and BIC are realized for the model that contains the interaction between Seroconverted and Outside, and no interaction involving variable Young. This is precisely the graphical loglinear model we determined using the BDMCMC algorithm – see Fig 4. For women, the loglinear model that contains all three pairwise interactions does not fit the data well (p-value = 0.264). The other seven hierarchical models fit the data well at the significance level $\alpha = 0.05$. Among these seven models, the model that has the minimum value for both AIC and BIC contains interactions between variables Outside and Young, and between variables Seroconverted and Young. As for men, we found the same graphical loglinear model as we did before using the BDMCMC algorithm for the women’s mobility table – see Fig 5.

Discussion

We have used a population-based cohort to identify the predictive power of the place of residence for HIV acquisition in a hyper-endemic, rural sub-Saharan African context. With 21,015 repeat-testers under continuous surveillance over 105,614 person-years, and 3,264 HIV seroconversions, this represents one of the largest individual-level
Table 5. The fit of three-way loglinear models for men’s mobility data from Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>DF</th>
<th>p-value</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Seroconverted Outside Young</td>
<td>89.35</td>
<td>4</td>
<td>&lt;0.001</td>
<td>22095.90</td>
<td>22117.17</td>
</tr>
<tr>
<td>2 Seroconverted – Outside Young</td>
<td>14.48</td>
<td>3</td>
<td>0.002</td>
<td>22023.03</td>
<td>22051.39</td>
</tr>
<tr>
<td>3 Seroconverted Outside – Young</td>
<td>76.98</td>
<td>3</td>
<td>&lt;0.001</td>
<td>22085.54</td>
<td>22113.89</td>
</tr>
<tr>
<td>4 Seroconverted – Young Outside</td>
<td>88.17</td>
<td>3</td>
<td>&lt;0.001</td>
<td>22096.72</td>
<td>22125.08</td>
</tr>
<tr>
<td>5 Seroconverted – Young – Seroconverted</td>
<td>13.30</td>
<td>2</td>
<td>0.001</td>
<td>22023.85</td>
<td>22059.30</td>
</tr>
<tr>
<td>6 Seroconverted – Outside Young – Seroconverted</td>
<td>2.11</td>
<td>2</td>
<td>0.348</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>7 Outside – Young – Seroconverted</td>
<td>1.25</td>
<td>2</td>
<td>&lt;0.001</td>
<td>22086.36</td>
<td>22121.80</td>
</tr>
<tr>
<td>8 Seroconverted – Outside – Young – Seroconverted</td>
<td>1.53</td>
<td>1</td>
<td>0.215</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

“Dev” stands for deviance, and “DF” stands for degrees of freedom. The loglinear models are specified by their maximal interaction terms.

Table 6. The fit of three-way loglinear models for women’s mobility data from Table 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>DF</th>
<th>p-value</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Seroconverted Outside Young</td>
<td>649.72</td>
<td>4</td>
<td>&lt;0.001</td>
<td>38238.54</td>
<td>38260.75</td>
</tr>
<tr>
<td>2 Seroconverted – Outside Young</td>
<td>649.43</td>
<td>3</td>
<td>&lt;0.001</td>
<td>38195.24</td>
<td>38224.87</td>
</tr>
<tr>
<td>3 Seroconverted Outside – Young</td>
<td>331.25</td>
<td>3</td>
<td>&lt;0.001</td>
<td>37877.07</td>
<td>37906.69</td>
</tr>
<tr>
<td>4 Seroconverted – Young Outside</td>
<td>381.61</td>
<td>3</td>
<td>&lt;0.001</td>
<td>37927.43</td>
<td>37957.05</td>
</tr>
<tr>
<td>5 Seroconverted – Outside Young – Seroconverted</td>
<td>336.32</td>
<td>2</td>
<td>&lt;0.001</td>
<td>37884.14</td>
<td>37921.17</td>
</tr>
<tr>
<td>6 Seroconverted – Outside – Young – Seroconverted</td>
<td>285.96</td>
<td>2</td>
<td>&lt;0.001</td>
<td>37833.78</td>
<td>37870.81</td>
</tr>
<tr>
<td>7 Outside – Young – Seroconverted</td>
<td>18.15</td>
<td>2</td>
<td>&lt;0.001</td>
<td>37565.96</td>
<td>37602.99</td>
</tr>
<tr>
<td>8 Seroconverted – Outside – Young – Seroconverted</td>
<td>1.25</td>
<td>1</td>
<td>0.264</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

“Dev” stands for deviance, and “DF” stands for degrees of freedom. The loglinear models are specified by their maximal interaction terms.

longitudinal study of mobility patterns and HIV to date. The residential locations occupied by every study participant were classified as outside or inside the rural study area. The residential locations inside the rural study area was further classified as
belonging to one of 45 non-overlapping communities that fully cover the rural study area. We also included age (younger or older than 30 years at the start of the exposure period) as another predictor for HIV acquisition. Our interest in place of residence, gender and age stems from their nearly universal availability for everyone. They can be identified without major investments, and can be objectively measured. Other psychometric measures, behaviours, and attitudes that have been shown to be predictive of HIV acquisition are more expensive to collect, and, for the vast majority, are self-reported, hence suffer from various biases \[12,28\]. We found that, for both men and women, substantially larger flows are associated with residential movements that involve locations outside the rural study area compared to flows associated with residential movements between locations inside the study area. Thus, households in the rural study area are typical net-senders of mobile individuals to destinations in the KwaZulu-Natal province, or to other, more distant places throughout South Africa \[24\]. This circular migration stream effectively links a poor, rural community with more affluent urban centers where many employment opportunities are usually available, and also with other rural areas that offer more specialized types of employment (e.g. mining). This finding is consistent with the theoretical literature on the influence of spatial heterogeneity in resources on migration patterns \[25\].

Our main objective was to identify whether places of residence are predictive of HIV seroconversion. Multivariate predictive relationships are revealed in the mobility graphs for men and women we identified using state of the art Bayesian structural learning methods for graphical loglinear models. The fact that the structure of the mobility graphs we determined for men and women repeat-testers is very similar represents a good indication of the validity of the mobility graphs we selected despite the extremely large ($\approx 10^{325}$) possible undirected graphs with 48 variables. In both mobility graphs, in order to reach any of the communities vertices C1, C2, \ldots, C45 from the vertex Seroconverted by following paths of adjacent edges, we must first pass through the vertex Outside. Once we know whether a man or a woman moved outside the rural study area, also knowing which communities inside the rural study area they lived in becomes less relevant for the purpose of predicting whether they seroconverted. For this reason the location of the communities in which a repeat-tester resides seems to play a lesser role in predicting HIV seroconversion as compared with moving outside the rural study area. This finding is surprising because this rural study area has three large irregularly-shaped clusters of new HIV infections near a national road and in a rural node bordering a recent coal mine development \[40\]. These spatial areas are characterized by HIV incidence rates higher then in the other surrounding regions. We expected at least some of the communities spanned by these three clusters to be linked by an edge with vertex Seroconverted. However, no such edges are present in the two mobility graphs. Consequently, while the places of residence inside the rural study area certainly play a role in predicting HIV acquisition risk given the significant clustering of HIV infections in this rural community, their predictive power vanishes when taking into account whether a study participant moved outside the rural study area. While this is true for both men and women, the predictive importance of having a residence outside the rural study area differs for men as compared to women. These differences are evidenced in the subgraphs of the two mobility graphs associated with variables Outside, Young and Seroconverted – see Fig 4 and Fig 5.

The presence of an edge between Outside and Seroconverted in the subgraph for men means that whether a man moved outside the rural study area is predictive of whether he seroconverts (OR = 2.003, 95% CI = [1.718,2.332]). The absence of an edge between Young and Seroconverted in the same subgraph means that age has less predictive power for the HIV seroconversion of a man given that we know whether this man had a residence outside the rural study area. We point out that this does not imply that age
is predictive of HIV acquisition in men. For women, the relative predictive importance of moving outside the rural study area and age is reversed: the edge between Outside and Seroconverted is missing, while the edge between Young and Seroconverted is present. Whether a woman is younger than 30 years is predictive of whether she seroconverts (OR = 3.091, 95% CI = [2.693,3.561]). However, given that we know the age of a woman, also knowing whether she moved outside the rural study area has less predictive power for HIV seroconversion. As such, the location of place of residence, and residential moves have less predictive power for HIV seroconversion of women given their age. As an aside, we mention that the presence of an edge that links vertices Outside and Young in the women’s mobility graph makes sense: women younger than 30 years are more likely to move outside the rural study area (OR = 3.176, 95% CI = [2.787,3.633]). This edge is missing in men’s mobility graph because the relationship between variables Young and Outside is weaker (OR = 1.306, 95% CI = [1.124,1.523]).

These results shed light on the differential patterns of sexual behavior of mobile men and women. When establishing a residence outside the rural study area, men are likely to weaken their ties with the partners they leave behind which disconnects them from the local sexual network for shorter or longer periods of time. This uprooting process makes men more susceptible to exhibiting sexual behavior such as multiple sexual partnering, limited condom use, client to sex workers/engaging in casual sex, that significantly increases their vulnerability to acquiring HIV. Women, on the other hand, do not seem to follow similar behavioral transformations when establishing a residence outside the rural study area. Even when mobile women alter their sexual behavior due to uprooting and potentially engage in commercial/transactional sex or other risky sexual behavior, the effect of these changes seems to be less important than the age group they belong to. Our results indicate that, even if the frequency, duration and distance traveled associated with residential moves is similar for men and women who live in this rural study area [24], there are key differences between the behavioral processes that lead to HIV seroconversion of mobile men and women. Therefore, the socio-ecological drivers of mobility lead to differentiated sexual behavioral consequences for men and women. In order to formulate gender-specific combination HIV prevention strategies for high-risk mobile individuals, particularly in the light of attaining the UNAIDS 90-90-90 treatment targets [52], it is of paramount importance to understand these differences with respect to the complex network of structural, biological and socio-demographic factors that characterize places of residence outside the rural study area, and significantly alter the social context of mobile individuals [28].

The results of our study help with our understanding of how social marketing can be used to influence health behavior [53]. We are contributing knowledge to guide segmenting, targeting and positioning of HIV prevention interventions. While segmenting based on sociodemographics, attitudes, psychographics and behavior is possible, we argue that geodemographic segmentation based on the history of residential locations, gender and age can constitute a more reliable, objective, cost-effective way to to ensure effectiveness and optimal allocation of HIV prevention intervention strategies.

Supporting information

S1 Fig. Map of the communities that are linked with an edge with the vertex Outside in the estimated conditional independence graph for men’s mobility. A number of 33 communities (gray) are linked with Outside, while 12 communities (white) are not linked with an edge with Outside.

S2 Fig. Map of the communities that are linked with an edge with the vertex Outside in the estimated conditional independence graph for
women’s mobility. A number of 39 communities (gray) are linked with Outside, while 6 communities (white) are not linked with an edge with Outside.

S3 Fig. Convergence plot of the BD MCCM C algorithm for men’s mobility table. The plot shows the estimated posterior inclusion probabilities of edges in graphs associated with men’s mobility across iterations.

S4 Fig. Convergence plot of the BD MCCM C algorithm for women’s mobility table. The plot shows the estimated posterior inclusion probabilities of edges in graphs associated with women’s mobility across iterations.

S1 Data. Mobility table for men. This is a 48-dimensional dichotomous contingency table with 598 positive counts.

S2 Data. Mobility table for women. This is a 48-dimensional dichotomous contingency table with 939 positive counts.

Acknowledgments

The work of AD was supported by the National Science Foundation Grant DMS/MPS-1737746 to University of Washington. The work of TB was supported by the Alexander von Humboldt Foundation through the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research. The work of FT was supported by National Institute of Health (R01HD084233 and R01AI124389) and South African Medical Research Council Flagship (MRC-RFA-UFSP-01–2013/UKZN HIVEPI) grants as well as a UK Academy of Medical Sciences Newton Advanced Fellowship (NA150161). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References


52. UNAIDS. 90–90–90 - An ambitious treatment target to help end the AIDS epidemic; 2014. UNAIDS.