Published in final edited form as:
AIDS Behav. 2014 December ; 18(12): 2325–2337. doi:10.1007/s10461-014-0838-4.

Evaluating outcome-correlated recruitment and geographic recruitment bias in a respondent-driven sample of people who inject drugs in Tijuana, Mexico

Abby E. Rudolph[1,2], **Tommi L. Gaines**[2], **Remedios Lozada**[3], **Alicia Vera**[2], and **Kimberly C. Brouwer**[2]

[1]The Calverton Center, Pacific Institute for Research and Evaluation, Calverton, Maryland
[2]Division of Global Public Health, Department of Medicine, University of California San Diego, La Jolla, California [3]Pro-COMUSIDA, Tijuana Baja California, Mexico

Abstract

Respondent-driven sampling's (RDS) widespread use and reliance on untested assumptions suggests a need for new exploratory/diagnostic tests. We assessed geographic recruitment bias and outcome-correlated recruitment among 1048 RDS-recruited people who inject drugs (Tijuana, Mexico).

Surveys gathered demographics, drug/sex behaviors, activity locations, and recruiter-recruit pairs. Simulations assessed geographic and network clustering of active syphilis (RPR titers≥1:8). Gender-specific predicted probabilities were estimated using logistic regression with GEE and robust standard errors.

Active syphilis prevalence was 7% (crude: men=5.7% and women=16.6%; RDS-adjusted: men=6.7% and women=7.6%). Syphilis clustered in the Zona Norte, a neighborhood known for drug and sex markets. Network simulations revealed geographic recruitment bias and non-random recruitment by syphilis status. Gender-specific prevalence estimates accounting for clustering were highest among those living/working/injecting/buying drugs in the Zona Norte and directly/indirectly connected to syphilis cases (men:15.9%, women:25.6%) and lowest among those with neither exposure (men:3.0%, women:6.1%). Future RDS analyses should assess/account for network and spatial dependencies.

Keywords

Respondent-driven sampling; bias; people who inject drugs; network analysis; spatial analysis

INTRODUCTION

Respondent-driven sampling (RDS) was introduced by Heckathorn and colleagues as an approach to sampling design and inference for "hidden" populations, or those without a

Corresponding Author: Abby E. Rudolph, 11720 Beltsville Drive Suite 900, Calverton MD, 20705, Tel: 301-755-2797, Fax: 301-755-2799, arudolph@pire.org.

sampling frame[1]. RDS is a modified form of chain-referral sampling whereby peer-recruitment is initiated by a small group of “seeds” selected by the research staff. Seeds are given a limited number of coupons to recruit peers; those eligible are enrolled and asked to recruit their peers. This process of recruits becoming recruiters continues until the desired sample size is reached. Many researchers adopted this recruitment strategy because of its ability to recruit large samples of “hidden” populations (e.g., people who inject drugs (PWID), men who have sex with men, female sex workers) relatively quickly and because it was considered to be an improvement over other available recruitment strategies (e.g., snowball sample, convenience sampling, etc.). By January 2013, RDS had been used by researchers in over 80 countries [2]. As an analytic approach, RDS uses post-stratified weights to offset non-random sampling and generate asymptotically unbiased population estimates. The RDS estimator which is part of RDS Analysis Tool (RDSAT) v7.1 generates individualized weights based on respondents’ degree (e.g., the number of people in the target population whom he/she reports knowing) and the partition variable (e.g., dependent variable)[3]. The validity of the RDS estimator relies on several assumptions[4]. One which has been frequently evaluated is that RDS participants randomly recruit peers from their network. Two potential sources of non-random recruitment are 1) geographic recruitment bias and 2) outcome-correlated recruitment.

Geographic recruitment bias

Geographic recruitment bias is the preferential recruitment of peers from a particular geographic area. The preferential recruitment of geographically proximal peers could lead to the over-recruitment of individuals in certain geographic areas and the under-recruitment of those in others. When this happens, individuals sampled will be more likely to share the same social environment due to their shared geography. Consequently, they may be more similar to one another than those selected independent of their geographic location. If not accounted for in the analysis, prevalence estimates could have artificially narrow confidence intervals. Further, the resulting estimates may be biased if there is also considerable geographic variation in the outcome of interest. For example, spatial analyses of sexually transmitted infections (STIs) have demonstrated that STIs are not equally distributed across geographic areas; studies have reported geographic clustering of gonorrhea [5–7], syphilis [8], and HIV cases [9, 10]. With respect to RDS, the presence of both 1) geographic variation in the outcome of interest and 2) preferential recruitment of peers from the same geographic area could lead to biased population estimates. For example, if individuals in geographic areas characterized by a high disease prevalence tend to recruit others from the same geographic area, the population prevalence will be over-estimated even when sample convergence on key variables has been achieved. A few studies have noted the potential for bias resulting from the preferential recruitment of geographically proximal peers [11–13], but current estimation procedures do not account for this potential source of bias.

Outcome-correlated recruitment behavior

The accuracy of RDS prevalence estimates is affected by the structure of the underlying social network, the distribution of traits within the network, and recruitment dynamics [14]. Differential recruitment based on the outcome of interest may occur when 1) the outcome clusters in a network or 2) network members cluster in space and the outcome is spatially clustered. There are several reasons why an outcome may cluster within a network. First, STIs (including HIV) are transmitted through high-risk sex and/or drug use behaviors. Because members of the same network may have similar risk behaviors, networks comprised of higher risk individuals will likely have a higher disease burden and those comprised of lower risk individuals will typically have a lower prevalence of disease. Further, when network members engage in high risk behaviors together, diseases can be directly transmitted. The rate of disease transmission within a network of individuals depends on the rate at which uninfected individuals engage in high risk behaviors with those infected, the efficiency of disease transmission between individuals, and the infection duration.

If the outcome of interest is spread through high-risk sex behaviors, individuals who recruit sex partners will be more likely to recruit peers with the same outcome status. Consequently, if recruitment penetrates a network with a high disease burden and many sexual partnerships, the population prevalence may be overestimated. On the contrary, if peer recruitment is more common among networks with a low disease burden and/or few sexual partnerships, the population prevalence may be underestimated. Further, the tendency to recruit similar network members (e.g., those with similar risk behaviors and consequently a similar outcome status), rather than a random selection of peers, could result in artificially narrow confidence estimates if not accounted for in the analysis.

Findings from simulations with RDS data suggest that RDS-adjusted estimates may be biased when recruitment is based on characteristics correlated with study outcomes[15]. Outcome-correlated recruitment has also been demonstrated empirically; RDS studies have reported high recruitment homophily [16–18], differential recruitment behavior by HIV serostatus [19–21], and clustering by HIV-status within RDS recruitment chains [22]. While Lu’s doctoral thesis reports that outcome-correlated recruitment behavior is one of the most harmful violations [2], RDS analyses typically do not assess or account for this potential bias.

A need for better diagnostic techniques

Gile and colleagues noted, “the widespread use of RDS for important public health problems combined with its reliance on untested assumptions, creates a pressing need for exploratory and diagnostic techniques for RDS data”[23]. This paper addresses this need by examining network and spatial factors that the RDSAT v. 7.1 RDS estimator[3] does not currently account for. We present a diagnostic tool that uses spatial and network simulations to assess geographic recruitment bias and outcome-correlated recruitment. We additionally present one way to account for these biases in the analysis and compare our resulting prevalence estimates with those obtained using standard RDSAT weighting procedures (e.g., the RDS II estimator).

METHODS

Between April 2006 and June 2007, 1,056 PWID were recruited via RDS to participate in a prospective study, *El Cuete*, which aimed to characterize the epidemiology of HIV, syphilis and tuberculosis among PWID in Tijuana, Mexico. Additional information on recruitment procedures, tolerance, sample convergence, and sample characteristics has previously been reported [10, 24]. In brief, a diverse group of seeds (e.g., heterogeneous by age, gender, drug preference, and neighborhood) were recruited to initiate peer recruitment (N=32). Tijuana residents who were at least 18 years of age who reported injecting an illicit drug ≥ once in the past month and who reported no plans of permanently moving out of the city in the next 18 months were eligible to participate. Of 32 seeds, 24 recruited eligible peers and seven seeds (each extending ≥6 waves) recruited 89% of the sample. One RDS recruitment chain extended 17 waves and accounted for 44% of all study participants (N=457). The network size variable used to generate weights was the sum of the number of family members, friends, and acquaintances who live in Tijuana and inject drugs. Participants were compensated for both participating in the study and for referring eligible PWID.

Blood samples were obtained through venipuncture. The Determine rapid HIV antibody test (Abbott Laboratories) was administered to detect the presence of HIV antibodies. Reactive samples were confirmed using an HIV-1 enzyme immunoassay and immunofluorescence assay. Syphilis serology used the rapid plasma reagin (RPR) test (Macro-Vue; Becton Dickinson). RPR-positive samples were confirmed with the *Treponema pallidum* particle agglutination assay (TPPA; Fujirebio). Specimen testing was conducted at the San Diego County Health Department. In the absence of clinical data to confirm diagnoses, titer has been proposed as an alternative priority marker [25]. RPR titers ≥1:8 (suggestive of active syphilis) were classified as active syphilis cases, which is consistent with variable definitions previously reported using this data [10, 24]. Individuals with RPR titers ≥1:8 were referred to the Tijuana municipal health clinic for free care.

Interviews were conducted by trained outreach workers in high-drug-use neighborhoods using a storefront office located in the Zona Norte and a mobile van which rotated between the following noncontiguous neighborhoods dispersed throughout Tijuana: El Mapa, 3 de Octubre, La Postal, Sanchez Taboada, and El Florido. The mobile van travelled to each recruitment location 1–2 times/week depending on the number of individuals recruited there at the last visit. Locations were visited regularly and each time, the van parked in the same location to ensure that individuals could easily refer peers to the study.

Interviewer-administered surveys gathered socio-demographic characteristics, sexual and drug use behaviors, RDS recruiter-recruit ties, and the locations where individuals lived, worked, bought drugs, and injected drugs. While participants were interviewed at baseline and at 6-month intervals for 18 months, this analysis is restricted to baseline data. Additionally, because this analysis focuses partly on network correlations via RDS-recruitment ties, seeds not recruiting any peers were removed from this analysis (Final N=1,048). Simulations were used to assess network and spatial clustering of active syphilis among RDS participants and to derive relevant measures of each for use in future analyses. Study methods were approved by the institutional review board of the University of California San Diego and the Ethics Board of the Tijuana General Hospital.

Spatial dependence

We mapped those with syphilis titers≥1:8 (referred to as cases, hereafter) and those without syphilis titers or with syphilis titers<1:8 (referred to as controls, hereafter) using residential coordinates. To measure the extent of spatial clustering for those with and without active syphilis, we used *K*-function analysis, which measures the expected number of events within a range of distances, *h*, from observed events. To examine differences in the extent and resolution of spatial clustering for each, we tested the null hypothesis, $H0{:}K_{cases}(h)=K_{controls}(h)$. Monte Carlo simulations were used to generate 95% confidence envelopes for the difference in the K functions, $K_{cases}(h)\text{-}K_{controls}(h)$, for a range of distances, *h*, based on randomly permuting disease status location labels to provide the corresponding distribution under the null hypothesis[26](Figure 1). The R-language statistical computing environment [27] with the SPLANCS contributed software package was used for *K*-function analysis.

Network influence

We visualized the distribution of active syphilis among RDS recruits using Cytoscape[28] (Figure 2). To determine whether clustering by syphilis status could be explained by chance, the observed network was compared with a null distribution (1,000 randomly generated networks with the same network topology and overall prevalence of syphilis, but with syphilis status distributed randomly)[29]. If syphilis clusters more than what would be expected by chance, the probability of an ego having syphilis given that an alter has syphilis would be higher in the observed network than in the null distribution and would not be included within the 95% confidence interval (CI) for the null distribution ($P<0.05$) (Table I). The expected and observed risk of active syphilis for the ego given the syphilis status of his/her alters were calculated for ego-alter pairs separated by 1–6 degrees in R.[27]

Recruitment based on geography

Tijuana’s red light district, or ‘zona roja’, is well known for both sex tourism and drug markets. Prostitution is quasi-legal in the zona roja as long as a work permit is obtained and sex workers undergo mandated periodic medical exams; however, in practice over half operate without permits[30]. To evaluate recruitment preferences based on geography, we created a binary variable to serve as a proxy for spending time in the Zona Norte, the neighborhood containing the zona roja (and within walking distance of its center). Individuals who reported living, working, injecting, or buying drugs in the Zona Norte were compared to those who did not do any of the above in the Zona Norte. Of note, the decision to create this variable was informed by the clustering pattern observed in this neighborhood (Figure 1). To determine whether individuals who participated in activities in this region clustered within RDS recruitment networks, we compared the observed distribution with the null distribution (as described above). If individuals who participated in activities in the Zona Norte are more clustered within RDS recruitment chains than what we would expect by chance, the probability of an ego participating in activities in the Zona Norte given that his/her alter participated in activities in the Zona Norte would be higher in the observed network than in the null distribution and would not be included within the 95%CI for the null distribution ($P<0.05$) (Table I). The observed and null distributions were compared for individuals separated by 1–6 degrees in R.[27]

Prevalence Estimation

We estimated the RDS-adjusted prevalence of active syphilis by gender using RDSAT Version 7.1 and confidence intervals were calculated using 15,000 bootstrap re-samples [3] (Table III). Descriptive statistics and population average logistic regression models with generalized estimating equations (GEE) (clustering on RDS-recruitment chain) and Huber-White robust standard errors were conducted using STATA 10 to estimate the prevalence of active syphilis. Findings from our network simulations (see Table I and Figure 2) demonstrate that both spatial location and active syphilis status cluster within RDS recruitment chains. To account for these observed correlations among individuals in the same RDS recruitment chain, we clustered individuals according to their RDS recruitment chain membership and accounted for possible model misspecification with Huber-White robust standard errors. Because men and women differed significantly with respect to the prevalence of active syphilis, their spatial proximity to the zona roja (mean Euclidean distance from the zona roja centroid to his/her residence was 1.6 km for women and 2.4 km for men (data not shown here)) and their sex-related risk behaviors (Table II), all analyses were stratified by gender.

RESULTS

Overall, the sample was 85% male, the average age was 37 years and the majority earned < $1,000 pesos/month (approximately $77 US dollars). The prevalence of lifetime syphilis was 15% (women=36% and men=12%) and active syphilis was 7% (women=16% and men=6%). Gender differences likely reflect significant differences in high-risk sexual behaviors (Table II). For example, women were younger than men (median age for women=34.3 (IQR:28.1–40.9) vs. median age for men=36.8 (IQR:31.6–42.8); $P<0.001$) and reported more sex partners in the past 6 months (median among women=5 (IQR:1–21) vs. median among men=1 (IQR:0–3); $P<0.001$). In the last 6 months, women were also more likely to use methamphetamine (86% vs. 77%; $P=0.015$), report exchanging sex for money or drugs (39% vs. 6%; $P<0.001$), and report having been forced to have sex (6% vs. <1%; $P<0.001$). Women were also more likely to be HIV seropositive (10% vs. 3%; $P<0.001$). More women than men reported living, working, injecting, or buying drugs in the Zona Norte (75% vs. 51%, respectively; $P<0.001$). Among those participating in activities in the Zona Norte, 65% of women compared with only 46% of men reported doing all four activities in the Zona Norte. For both men and women, the most commonly reported activity in this region was injecting drugs, followed by buying drugs. Further, while both genders reported living and working in this region, more women than men reported living and working in the Zona Norte. Women were also more likely to be directly or indirectly connected to another study participant with active syphilis (46% vs. 27%; $P<0.001$).

Spatial simulations

As seen in Figure 1, active syphilis cases clustered in the Zona Norte, the area surrounding Tijuana’s red light district, or ‘zona roja’. The 95% confidence envelopes suggest that cases were significantly more clustered than controls for individuals separated by ≤4.75 kilometers ($P<0.05$). Those with active syphilis tended to live closer to one another than those without active syphilis.

Network simulations – active syphilis

As seen in Table I, the network influence of syphilis within this respondent-driven sample is significant for ≤ 2 degrees of separation, or those separated by ≤ 2 recruitment linkages. For example, respondents with active syphilis were three times more likely than respondents without active syphilis to recruit or be recruited by PWIDs with active syphilis (*P*<0.0001). Based on these findings, we operationalized the network influence of syphilis as the presence (yes/no) of a direct or indirect link to a participant with active syphilis. Individuals who are directly connected to one another (e.g., one degree of separation) are recruiter-recruit pairs. Indirect linkages (e.g., 2 degrees of separation) are those who share the same recruiter or who are directly connected to one's recruiter or recruit (Figure 2).

Network simulations – participating in activities in the Zona Norte

As seen in Figure 2 and Table I, individuals participating in activities in the Zona Norte were clustered in RDS recruitment chains. Overall, RDS chains appear to be comprised of predominately individuals who participate in activities in the Zona Norte or of individuals who do not participate in activities in this neighborhood. For example, an individual was 2.65 times more likely to participate in activities in the Zona Norte if his/her recruit or recruiter participated in activities in this region (*P*<0.05). Because the strength and significance of this association remained for individuals separated by 1–6 degrees, we used a binary variable to represent participation in activities in the Zona Norte (yes vs. no) for all subsequent analyses. To account for the shared social environment among RDS participants in the same recruitment chain in our prevalence estimates, we created a cluster variable for RDS recruitment chain membership.

Gender-specific correlates of active syphilis

For both men and women, HIV positive serostatus, being directly or indirectly connected to another study participant with active syphilis and participating in activities (e.g., living, working, buying drugs, or injecting drugs) in the Zona Norte were significantly associated with active syphilis (Table II). Among men, using methamphetamine in the past 6 months and younger age were also positively associated with active syphilis (*P*<0.001).

Syphilis prevalence estimates

As seen in Table III, the unadjusted prevalence of active syphilis was 5.7% among men and 16.6% among women; the RDS-adjusted prevalence of active syphilis was 6.7% among men and 7.6% among women. The estimated prevalence of active syphilis was higher for women than for men for every combination of network and spatial exposures. For example, the estimated prevalence of active syphilis among men with neither network nor spatial exposures or with only one of these exposures was 3–5%. The estimated prevalence of active syphilis for women with neither exposure was twice as high as that in men (6.1% vs. 3.0%, respectively). For women with only one of these exposures, the prevalence was ~13%. This corresponds with a 4-fold higher risk of active syphilis among women compared with men who were directly or indirectly linked to another study participant with active syphilis and >2.5-fold higher risk of active syphilis among women compared with men who lived, worked, bought drugs, or injected drugs in the Zona Norte. For both men and women, the prevalence of active syphilis was highest in the group with both network and spatial exposures to active syphilis (15.9% among men and 25.6% among women), but the interaction was significant only for men.

DISCUSSION

Among this sample of PWID, active syphilis cases clustered in the Zona Norte, a neighborhood known for its drug and sex markets. Network simulations revealed geographic recruitment bias; respondents who live, work, inject drugs, or buy drugs in the Zona Norte preferentially recruited other PWID who participate in activities in this neighborhood. Due to this preferential recruitment of peers who participate in activities in the Zona Norte by other participants who participate in activities in this neighborhood, the sample variability is likely less than it should be because individuals in the same RDS recruitment chain share the same social environment. Network simulations also identified non-random recruitment by active syphilis; those with active syphilis were more likely to recruit other PWID infected with active syphilis. Due to the unequal distribution of active syphilis cases across the sampled area and the presence of both outcome-correlated recruitment and geographic recruitment bias (focused in the region where active syphilis clusters), prevalence estimates may not reflect the prevalence of active syphilis among PWID in all of Tijuana, but rather a select subgroup with an elevated prevalence.

As seen in Table II, women were more likely than men to report living, working, injecting, or buying drugs in the Zona Norte (where active syphilis clusters in this sample) and were more likely to be directly/indirectly connected through RDS linkages to other active cases (*P*<0.0001 for both comparisons). This may reflect that 1) female PWID are more likely than their male counterparts to participate in activities in this region, 2) women with active syphilis may have been over-sampled due to the increased percentage of women with ties to the Zona Norte and to syphilis infected network members, or 3) a combination of the two. As the population is "hidden" and female PWID are particularly hard to recruit, it is difficult to know the geographic distribution of the underlying target population (e.g., PWID in Tijuana). Ethnographic and qualitative research can be used to guide inferences with respect to whether the RDS sample reflects the geographic distribution of the target population or whether it reflects a subset of the target population (e.g., PWID in specific neighborhoods of Tijuana). For example, if PWID are located in other regions of Tijuana but were not represented in the final sample, the target population may reflect a sub-group of PWID in Tijuana and corresponding prevalence estimates should not be over-generalized. On the contrary, recruitment of PWID in specific neighborhoods may accurately reflect the distribution of PWID. In the latter case, resulting estimates are likely to be more representative. Based on qualitative research conducted by our research group, PWID do engage in high risk behaviors in several areas occupying the southern region of Tijuana, including the colonias Azteca, Presidentes, Chamizal, Florido, Postal, Sanchez, and Taboada. Additionally, our research group recently identified neighborhoods where PWID encountered adverse police interactions (e.g., syringe confiscation, money extortion, physical/sexual abuse). Statistically significant hotspots for these adverse interactions not only occurred in and around the Zona Norte but also in colonias in southeastern Tijuana (unpublished data). Although we chose very geographically diverse seeds, the final sample tended to cluster around the areas where we regularly parked our mobile van for interviews, likely due to spatial biases in recruitment behaviors. One potential barrier to traveling longer distances to the study office was the increased likelihood of being stopped by the police.

Because active syphilis is not randomly distributed within networks or in space, preferential recruitment of peers with the same outcome status or from the same social environment (where active syphilis clusters) could lead to inaccurate prevalence estimates if not accounted for in the analysis. As seen in Table III, accounting for network and spatial correlates of syphilis and accounting for the shared social environment among those in the same RDS recruitment chain with GEE appears to provide a more complete picture of syphilis among PWID in this region. For example, the estimated prevalence of active syphilis is consistently higher among women than among men in each strata and the prevalence is highest for those with both network and spatial exposures.

From an RDS analytic perspective, our findings underscore the importance of considering both spatial and network dependencies and recruitment patterns when estimating disease prevalence in key populations. We demonstrate that the outcome status of RDS recruiters and recruits (active syphilis) is not independent. Instead, the peer recruitment process is not random with respect to syphilis status. Additionally, active syphilis clusters in Tijuana’s red light district and is more common among those who live, work, buy drugs or inject in the surrounding area. Consequently, RDS estimates which do not account for these network and spatial dependencies (and recruitment biases) could lead to inaccurate prevalence estimates with artificially narrow confidence intervals. Due to the strong association between active syphilis and HIV in this sample (P<0.001 for men and P=0.0007 for women), HIV prevalence estimates could also be affected if these factors are not accounted for in the analysis.

Limitations

There are several limitations of this analysis. First, individuals were asked to recruit PWID. Consequently, RDS recruitment ties measured here are more likely to reflect injecting ties than sexual ties. Only 20 individuals reported that they were recruited by a sex partner, boyfriend, or spouse. Most reported being recruited by a friend (62%), acquaintance (18%), or stranger (12%). While individuals may have had sex with his/her recruiter/recruit, individuals were not asked to recruit sex ties. As a result, the number of sexual ties (which are responsible for syphilis transmission) are likely under-reported in this analysis. Had individuals been asked to exclusively recruit sex partners, we would have likely observed a stronger influence of networks on one’s syphilis status in both men and women. Because 1) there are fewer female than male injectors, 2) women reported more sexual partners and riskier sexual behaviors than men, and 3) there were very few men reporting same sex sexual behaviors, recruiter-recruit pairs of the opposite sex were more likely to be sexual than those between individuals of the same sex. As a result, direct and indirect recruitment linkages are probably more likely to reflect sexual ties for the women sampled than for the men sampled, which may partially explain why the network influence was stronger for women than for men.

Because the network data used in our network simulations are based on RDS recruitment linkages, the network data are incomplete and cannot be interpreted as a complete sociometric network. Because RDS study participants can only be recruited to participate in the study once, we are likely underestimating the total number of connections that exist between individuals sampled. For example, individuals who are connected to one another by two recruitment linkages may actually be directly connected, but because of limitations in our sampling strategy and the fact that we did not ascertain additional network connections in the survey, we were unable to discern these ties. It is also possible that individuals separated by two degrees in our data set may be directly connected to one another through sexual behaviors but because individuals were asked to recruit other PWID, these ties are not reflected in our data set. To address this limitation, we operationalized network influence of syphilis as the presence (yes/no) of a direct or indirect link to a participant with active syphilis.

While the network influence we observed could reflect actual transmission ties, it may also reflect recruitment homophily on high-risk sex behaviors associated with syphilis acquisition. Either way, the network influence observed in this respondent-driven sample could introduce bias to the RDS prevalence estimates if not accounted for, as current RDS adjustment procedures assume no correlation between recruiters and recruits on the outcome of interest. When recruitment is dependent on the study outcome or based on characteristics which are correlated with study outcome, population estimates may be biased[15]. Consequently, the potential for biased estimates in this sample exists for both HIV and active syphilis due to the strong association between active syphilis and HIV status observed here.

Several other researchers have examined non-random recruitment related to demographic characteristics, drug use and/or sexual behaviors, and outcome status by comparing the distribution of these attributes among alters who would have been eligible for recruitment with those of actual peer recruits [18, 31, 32]. A separate study used dyadic analyses to compare RDS recruitment dyads with non-recruitment network dyads; the authors examined non-random recruitment based on drug use similarity, geographic proximity, demographic similarity and relationship-level characteristics (e.g., duration of relationship, frequency of communication, kinship, social/financial support, trust, drug use, and sex) and reported that RDS participants were significantly more likely to recruit kin and those with whom they had more frequent communication[33]. In this analysis, we used network simulations to assess outcome-correlated recruitment and geographic-correlated recruitment. Unfortunately, ego-centric network data were not collected as part of this study, so we were unable to compare our findings with those comparing actual RDS recruits with recruitment-eligible networks who were not recruited. However, in the absence of ego-centric network data, this is a valid alternative for identifying non-random recruitment.

Because the data are cross-sectional, we cannot determine whether ties between individuals existed before transmission occurred or if they were formed later. There may also be limitations associated with using titers≥1:8 to define active syphilis cases. While some individuals may have been misclassified by using this cutoff, false positives are unlikely (e.g., up to 90% of false-positive reactions have titers below this cutoff [34]).

Finally, while we observed extensive cross-neighborhood recruitment by residential location in our sample, we cannot rule out the presence of recruitment bottlenecks as a possible alternative explanation for the observed findings.

CONCLUSION

This paper makes two major contributions. First, we present an analysis which incorporates both social network and spatial analytic approaches, which is important because network and spatial correlates often overlap. As we demonstrate here, the space where people live, work, inject drugs, and buy drugs (e.g., their social environment), also has a network component. Understanding these intersecting relationships is important to advance research on health disparities and to more accurately assess disease burden. Second, it is an important contribution to the RDS literature because it identifies two potential sources of bias (recruitment preferences based on geography and the outcome status) and provides methods for determining whether they are present. RDS is used to determine the disease burden in key populations. Understanding how measures may be biased and the limitations on the representativeness of estimates has important implications for accurately estimating the prevalence of HIV/STIs for disease surveillance and epidemiologic research. We present one approach for 1) assessing the presence of geographic recruitment bias and outcome-correlated recruitment, 2) presenting prevalence estimates which account for these dependencies and 3) interpreting the representativeness of the findings. We encourage other researchers to apply this approach to RDS data from different geographic areas, among different key populations, and with different disease outcomes to see whether similar patterns emerge. We recommend that other researchers conduct similar diagnostics and use ethnographic and qualitative findings to assist with the interpretation of their findings. Where appropriate, we recommend accounting for correlations in the data that might provide misleading estimates of disease prevalence.

Acknowledgments

This research was supported by National Institute on Drug Abuse Grants R01DA019829 and R37DA019829 (PI: Strathdee); A.E.R. was funded by NIH/NIDA Grant K01DA033879; T.L.G. was funded by NIH/NIDA Grant K01DA034523.

Abbreviations

CI	Confidence Interval
GEE	Generalized Estimating Equations
HIV	Human Immunodeficiency Virus
PWID	People who inject drugs
RDS	Respondent Driven Sampling
RDSAT	Respondent Driven Sampling Analysis Tool
RPR	Rapid Plasma Reagin
TPPA	*Treponema pallidum* particle agglutination assay

References

1. Heckathorn D. Respondent-Driven Sampling: A new approach to the study of hidden populations. Social Problems. 1997; 44(2):174–99.
2. Lu, X. Respondent-Driven Sampling: Theory, Limitations & Improvements. 2013. available at: http://hdl.handle.net/10616/41378
3. Volz, E.; Wejnert, C.; Cameron, C.; Spiller, M.; VB; Degani, I., et al. Respondent-driven sampling analysis tool (RDSAT) version 7.1. Ithaca, NY: Cornell University; 2012.
4. Heckathorn D. Respondent-Driven Sampling II. Deriving valid population estimates from chain-referral samples of hidden populations. Social Problems. 2002; 49(1):11–34.
5. Ellen JM, Hessol NA, Kohn RP, Bolan GA. An investigation of geographic clustering of repeat cases of gonorrhea and chlamydial infection in San Francisco, 1989–1993: evidence for core groups. Journal of Infectious Diseases. 1997; 175(6):1519–22. [PubMed: 9180198]
6. Bernstein KT, Curriero FC, Jennings JM, Olthoff G, Erbelding EJ, Zenilman J. Defining core gonorrhea transmission utilizing spatial data. American journal of epidemiology. 2004; 160(1):51–8. [PubMed: 15229117]
7. Rothenberg RB. The geography of gonorrhea: empirical demonstration of core group transmission. American journal of epidemiology. 1983; 117(6):688–94. [PubMed: 6859024]
8. Hamers F, Peterman T, Zaidi A, Ransom R, Wroten J, Witte J. Syphilis and gonorrhea in Miami: similar clustering, different trends. American Journal of Public Health. 1995; 85(8_Pt_1):1104–8. [PubMed: 7625504]
9. Tanser F, Bärnighausen T, Cooke GS, Newell M-L. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. International Journal of Epidemiology. 2009; 38(4):1008–16. [PubMed: 19261659]
10. Brouwer KC, Rusch ML, Weeks JR, Lozada R, Vera A, Magis-RodrÃ-guez C, et al. Spatial epidemiology of HIV among injection drug users in Tijuana, Mexico. Annals of the Association of American Geographers. 102(5):1190–9. [PubMed: 23606753]
11. Burt RD, Hagan H, Sabin K, Thiede H. Evaluating respondent-driven sampling in a major metropolitan area: Comparing injection drug users in the 2005 Seattle area national HIV behavioral surveillance system survey with participants in the RAVEN and Kiwi studies. Annals of Epidemiology. 2010; 20(2):159–67. [PubMed: 20123167]
12. McCreesh N, Johnston LG, Copas A, Sonnenberg P, Seeley J, Hayes RJ, et al. Evaluation of the role of location and distance in recruitment in respondent-driven sampling. International journal of health geographics. 2011; 10(1):56. [PubMed: 22008416]
13. Jenness SM, Neaigus A, Wendel T, Gelpi-Acosta C, Hagan H. Spatial Recruitment Bias in Respondent-Driven Sampling: Implications for HIV Prevalence Estimation in Urban Heterosexuals. AIDS and Behavior. 2013:1–8. [PubMed: 23054037]
14. Goel S, Salganik MJ. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences. 2010; 107(15):6743–7.
15. Lu, X.; Bengtsson, L.; Britton, T.; Camitz, M.; Kim, BJ.; Thorson, A., et al. statAP. 2010. The Sensitivity of Respondent-Driven Sampling Method.
16. Gwadz MV, Leonard NR, Cleland CM, Riedel M, Banfield A, Mildvan D. The Effect of Peer-Driven Intervention on Rates of Screening for AIDS Clinical Trials Among African Americans and Hispanics. Am J Public Health. 2011 Jun; 101(6):1096–102. [PubMed: 21330587]
17. Iguchi M, Ober A, Berry S, Fain R, Heckathorn D, Gorbach P, et al. Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications. Journal of Urban Health. 2009; 86(1):S5–S31.
18. Rudolph AE, Crawford ND, Latkin C, Heimer R, Benjamin EO, Jones KC, et al. Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: implications for research and public health practice. Annals of epidemiology. 2011 Apr; 21(4): 280–9. [PubMed: 21376275]

19. Abramovitz D, Volz EM, Strathdee SA, Patterson TL, Vera A, Frost SDW, et al. Using Respondent-Driven Sampling in a Hidden Population at Risk of HIV Infection: Who Do HIV-Positive Recruiters Recruit? Sex Transm Dis. 2009 Dec; 36(12):750–6. [PubMed: 19704394]
20. Broadhead RS, Heckathorn DD, Weakliem DL, Anthony DL, Madray H, Mills RJ, et al. Harnessing peer networks as an instrument for AIDS prevention: Results from a peer-driven intervention. Public Health. 1999 Jun.113:42–57. [PubMed: 9722809]
21. Ramirez-Valles J, Heckathorn DD, Vazquez R, Diaz RM, Campbell RT. From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino gay men. AIDS and behavior. 2005 Dec; 9(4):387–402. [PubMed: 16235135]
22. Rudolph AE, Crawford ND, Latkin C, Fowler JH, Fuller CM. Individual and neighborhood correlates of membership in drug using networks with a higher prevalence of HIV in New York City (2006–2009). Annals of epidemiology. 2013
23. Gile, KJ.; Johnston, LG.; Salganik, MJ. Diagnostics for Respondent-driven Sampling. 2012. arXiv preprint arXiv:12096254
24. Strathdee SA, Lozada R, Pollini RA, Brouwer KC, Mantsios A, Abramovitz DA, et al. Individual, social, and environmental influences associated with HIV infection among injection drug users in Tijuana, Mexico. Journal of acquired immune deficiency syndromes (1999). 2008; 47(3):369. [PubMed: 18176320]
25. Samoff E, Koumans EH, Gibson JJ, Ross M, Markowitz LE. Pre-treatment syphilis titers: distribution and evaluation of their use to distinguish early from late latent syphilis and to prioritize contact investigations. Sexually transmitted diseases. 2009; 36(12):789–93. [PubMed: 19773682]
26. Ripley B. Modeling spatial patterns (with discussion). Journal of the Royal Statistical Society. 1977; 39:172–212.
27. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003; 13(11):2498–504. [PubMed: 14597658]
29. Szabo, G.; Barabasi, A. Network Effects in Service Usage. 2007.
30. Bucardo J, Semple SJ, Fraga-Vallejo M, Davila W, Patterson TL. A qualitative exploration of female sex work in Tijuana, Mexico. Archives of sexual behavior. 2004; 33(4):343–51. [PubMed: 15162080]
31. Liu H, Li J, Ha T, Li J. Assessment of Random Recruitment Assumption in Respondent-Driven Sampling in Egocentric Network Data. Social Networking. 2012; 1(2):13–21. [PubMed: 23641317]
32. McCreesh N, Frost S, Seeley J, Katongole J, Tarsh MN, Ndunguse R, et al. Evaluation of respondent-driven sampling. Epidemiology (Cambridge, Mass). 2012; 23(1):138.
33. Young AM, Rudolph AE, Quillen D, Havens JR. Spatial, temporal, and relational patterns in respondent driven sampling: evidence from a social network study of rural drug users. Journal of Epidemiology and Community Health. 2014 Published Online First: 1 April, 2014. 10.1136/jech-2014-203935
34. Ratnam S. The laboratory diagnosis of syphilis. The Canadian Journal of Infectious Diseases & Medical Microbiology. 2005; 16(1):45. [PubMed: 18159528]

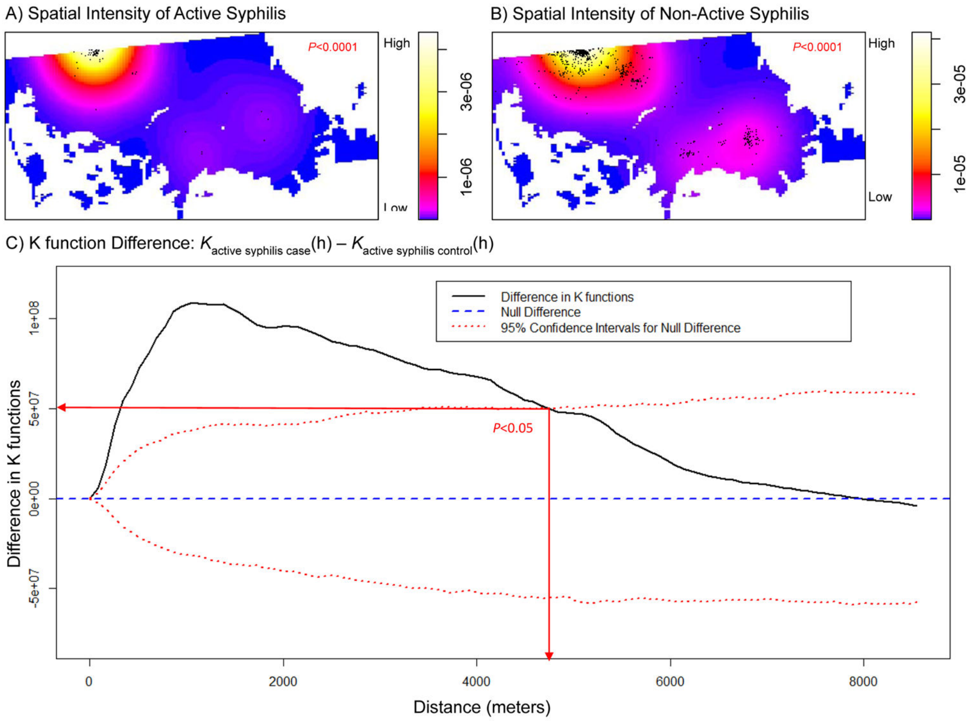

Figure 1.

Spatial intensity of active syphilis cases and controls (Figures 1a and 1b, respectively) and difference between *K* functions for active syphilis cases and controls (Figure 1c) in Tijuana, Mexico (2006–2007). As seen in Figures 1a and 1b, the areas with the most dense point pattern are in white, followed by yellow and then orange. Areas with the lowest density point pattern are blue. Both cases and controls cluster significantly more than what would be expected under the assumption of complete spatial randomness (P<0.0001). Of note, Zona Norte contains the highest concentration of participants in both 1a and 1b. However, there is also a region in the southeast portion of Tijuana which has a higher concentration of participants without active syphilis (Figure 1b). As seen in Figure 1c, the solid black line represents the observed difference in K functions for RDS participants separated by a range of distances, *h*. When the difference is positive, observed clustering among active syphilis cases is greater than that observed among active syphilis controls. 95% confidence envelopes (red dotted line) are based on 1000 Monte Carlo simulations. Confidence envelopes represent the set of confidence intervals over the range of values of spatial distance examined (1 mile = 1609.34 meters). When the difference is not included within the confidence interval, the difference in clustering is significant (P<0.05). Based on the figures above, both active syphilis cases and controls are clustered around the Zona Roja region (P<0.0001). However, the active syphilis cases are significantly more clustered than controls for most of the distances examined here (P<0.05). For example, the spatial location of participants was more spatially dependent on the spatial location of other participants for active syphilis cases than for active syphilis controls at distances less than 4750 meters (P<0.05). However, no significant differences were observed for individuals separated by more than 4750 meters.

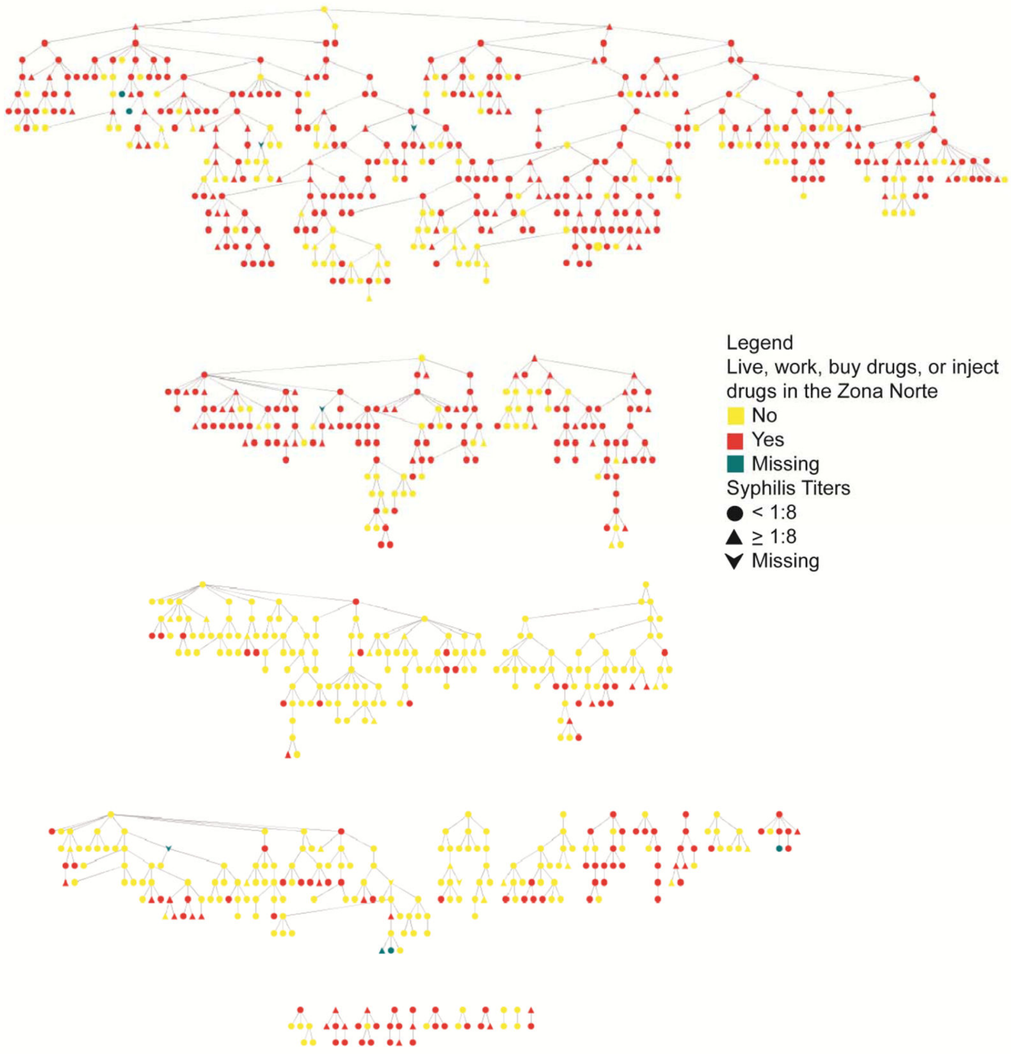

Figure 2.

This figure represents 1,052 RDS-recruited PWID participants in Tijuana, Mexico (2006–2007). Of note, 1,052 nodes were used in this figure to preserve all network connections; however information on those individuals not included in the final sample size (N=1,048) were given missing values for all attributes. Each node represents an individual (triangle= syphilis titers≥1:8, ellipse=syphilis titers<1:8 (including those without syphilis titers), and vee=missing or indeterminate syphilis test result; red=live, work, buy drugs or inject drugs in the Zona Norte, yellow=do not live, work, buy drugs or inject drugs in the Zona Norte, grey=missing data). Lines between nodes represent RDS recruiter-recruit relationships. There are 24 different RDS recruitment chains (median n=8.5; range: 2–459). As seen in the figure above, individuals in the same RDS recruitment chain are predominately those who either “Do” or “Do Not” participate in activities in the Zona Norte.

Table I

Network Influence of Active Syphilis and Participation of Activities in the Zona Norte Among RDS-recruited People Who Inject Drugs (PWID) in Tijuana, Mexico (2006–2007)

N degrees of separation [a]	Number of N-degree paths	Number of individuals separated by N degrees	Observed Risk Ratio of Syphilis	95% CI for the expected Risk Ratio of Syphilis from 1000 random samples	Observed Risk Ratio of Participating in Activities in the Zona Norte	95% CI for the expected Risk Ratio of Participating in Activities in the Zona Norte from 1000 random samples
1	2,056	1,052	3.042	0.234, 1.854	2.647	0.886, 1.109
2	3,790	1,042	2.141	0.413, 1.671	2.353	0.919, 1.087
3	5,530	1,025	0.993	0.519, 1.534	2.284	0.931, 1.062
4	7,568	1,001	1.016	0.585, 1.442	2.206	0.945, 1.057
5	9,818	979	1.362	0.630, 1.402	2.116	0.949, 1.050
6	11,454	967	0.913	0.648, 1.354	1.895	0.952, 1.045

[a] Individuals separated by one degree are recruiter-recruit pairs. Individuals separated by two degrees are indirectly connected to one another through one other person (either his/her recruiter or recruit). For example, individuals separated by two degrees may share the same recruiter, his/her recruiter may have been recruited by this person, or this person may have been recruited by his/her recruit. Of note, 1,052 nodes were used in this analysis to preserve all network connections; however information on those individuals not included in the final sample size (N=1,048) were given missing values for all attributes.

Table II

Correlates of Active Syphilis among PWID Stratified by Gender in Tijuana, Mexico (2006–2007).

	Male (N=892)			Female (N=156)		
	No active syphilis N=839 N(%)[a]	Active syphilis N=51 N(%)[a]	2-sided P- value[b]	No active syphilis N=126 N(%)[a]	Active syphilis N=25 N(%)[a]	2-sided P-value[b]
Age, median (interquartile range)	36.9 (31.7–42.8)	35.8 (28.7–42.6)	<0.001	34.2 (28.3–41.0)	34.1 (25.0–39.2)	0.333
Number of PWID he/she reports knowing, median (interquartile range)	0 (0–12)	2 (0–20)	0.306	1 (0–12)	0 (0–7)	0.943
Interview location			0.706			0.305
El Mapa[c]	281 (33.5)	16 (31.4)		42 (33.3)	6 (24.0)	
PrevaCasa[c]	409 (48.8)	30 (58.8)		72 (57.1)	19 (76.0)	
El Florido	84 (10.0)	3 (5.9)		10 (7.9)	0 (0.0)	
Sanchez Taboado	0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)	
3 de Octubre	47 (5.6)	2 (3.9)		0 (0.0)	0 (0.0)	
La Postal	18 (2.2)	0 (0.0)		2 (1.6)	0 (0.0)	
Live, work, buy drugs, or inject drugs in the Zona Norte	418 (49.8)	38 (74.5)	<0.001	92 (73.0)	22 (88.0)	0.030
Time spent in the Zona Norte by activity						
Live in the Zona Norte	254 (60.8)	26 (68.4)	<0.001	66 (71.7)	17 (77.3)	0.016
Work in the Zona Norte	243 (58.1)	24 (63.2)	<0.001	72 (78.3)	19 (86.4)	0.011
Buy drugs in the Zona Norte	391 (93.5)	35 (92.1)	<0.001	82 (89.1)	22 (100)	0.016
Inject drugs in the Zona Norte	408 (97.6)	37 (97.4)	<0.001	91 (98.9)	22 (100)	0.026
Total number of activities in Zona Norte, median (interquartile range)	0 (0–3)	3 (0–4)	0.005	3 (0–4)	4 (3–4)	0.049
≥ 1 case of active syphilis ≤ 2 degrees of separation	212 (25.7)	26 (51.0)	<0.001	52 (41.6)	16 (64.0)	<0.001
Men who have sex with men (MSM)	226 (26.9)	13 (25.5)	0.826	N/A	N/A	N/A
≥ 2 sex partners in the past 6 months (yes vs. no)	129 (61.4)	7 (58.3)	0.329	53 (76.8)	13 (92.9)	0.067
Total number of sex partners in the past 6 months, median (interquartile range)	1 (0–3)	1 (0–3)	0.069	5 (1–15)	7 (2–26)	0.637
Exchanged sex for money or drugs in the past 6 months (yes vs. no)	53 (6.3)	2 (3.9)	0.252	47 (37.9)	10 (45.5)	0.323
Was forced to have sex in the past 6 months (yes vs. no)	2 (0.2)	0 (0.0)	1.0	6 (5.3)	2 (8.7)	0.454
Used a condom with regular partners in the past 6 months (ever vs. never)	23 (22.6)	4 (66.7)	0.075	14 (34.2)	4 (50.0)	0.504
Used a condom with casual partners in the past 6 months (ever vs. never)	64 (53.8)	4 (66.7)	0.362	36 (83.7)	6 (60.0)	0.366
Had sex while drunk with casual sex partner (ever vs. never)	33 (27.7)	3 (50.0)	0.073	14 (32.6)	1 (10.0)	0.163

	Male (N=892)			Female (N=156)		
	No active syphilis N=839 N(%)[a]	Active syphilis N=51 N(%)[a]	2-sided P- value[b]	No active syphilis N=126 N(%)[a]	Active syphilis N=25 N(%)[a]	2-sided P-value[b]
Had sex while high with casual sex partner (ever vs. never)	110 (92.4)	6 (100.0)	1.0	39 (90.7)	10 (100.0)	1.0
Used methamphetamines in the past 6 months (yes vs. no)	562 (76.6)	37 (88.1)	<0.001	99 (85.3)	22 (91.7)	0.361
HIV positive	23 (2.7)	7 (13.7)	<0.001	10 (7.9)	5 (20.0)	0.007
Marital status						
Single/never married	433 (51.6)	22 (43.1)	0.955	39 (31.0)	8 (32.0)	0.500
Married/common law	229 (27.3)	20 (39.2)		64 (50.8)	12 (48.0)	
Divorced	62 (7.4)	3 (5.9)		5 (4.0)	1 (4.0)	
Separated	99 (11.8)	6 (11.8)		14 (11.1)	1 (4.0)	
Widowed	16 (1.9)	0 (0.0)		4 (3.2)	3 (12.0)	
Years of education, median (interquartile range)	7 (6–9)	7 (5.5–9)	0.061	8 (6–10)	6 (5–9)	0.095
Income (average/month)						
No income	91 (10.9)	6 (11.8)	0.266	17 (13.5)	1 (4.0)	0.665
< $1,000 pesos	633 (75.5)	40 (78.4)		84 (66.7)	19 (76.0)	
$1,000–$1,499 pesos	102 (12.2)	5 (9.8)		21 (16.7)	5 (20.0)	
$1,500–$1,999 pesos	10 (1.2)	0 (0).0		4 (3.2)	0 (0.0)	
$2,000–$2,499 pesos	2 (0.2)	0 (0.0)		0 (0.0)	0 (0.0)	

[a] Values and percentages may not reflect column totals for some variables because of missing data and skip patterns.

[b] *P*-values obtained through logistic regression models (each covariate independently regressed on syphilis) with GEE (clustered on RDS recruitment chain) and Huber-White robust standard errors

[c] These two interview locations were within 10 blocks of one another in the Zona Norte. PrevaCasa was the stationary study office and El Mapa was a mobile van site.

Table III

Estimated Prevalence of Active Syphilis across Network and Spatial Characteristics and Stratified by Gender in Tijuana, Mexico (2006–2007)

		Male (N=876)		Female (N=150)	
		%	95%CI	%	95%CI
Unadjusted prevalence		5.7%	4.2, 7.3	16.6%	10.6, 22.6
RDS weighted prevalence [a]		6.7%	3.8, 9.9	7.6%	3.8, 13.4
Live, work, buy drugs, or inject drugs in the Zona Norte	≥ 1 case of active syphilis ≤ 2 degrees of separation				
No	No	3.0%	2.2, 4.1 [b]	6.1%	3.0, 12.2 [c]
No	Yes	3.2%	1.3, 7.9 [b]	12.8%	5.4, 27.5 [c]
Yes	No	4.9%	2.7, 8.7 [b]	13.2%	9.2, 18.7 [c]
Yes	Yes	15.9%	12.6, 19.8 [b]	25.6%	20.0, 32.2 [c]

[a] Stratified RDS-weighted prevalences were estimated in RDSAT version 7.1 and

confidence intervals were calculated using 15,000 bootstrap re-samples.[3].

[b] Predicted probabilities were calculated using logistic regression with GEE (clustered on RDS recruitment chain) and robust standard errors and did not adjust for other significant correlates of active syphilis

$$\text{logit}\ \left(\frac{E(y_{ij})}{1-E(y_{ij})}\right)=-3.475\ -0.513\ x_{1\,ij}+0.074\ x_{2\,ij}+1.219\ x_{3\,ij},$$

where x1= Live, work, buy drugs, or inject drugs in the Zona Norte (1=Yes; 0=No), x2=

≥ 1 case of active syphilis ≤ 2 degrees of separation (1=Yes; 0=No), and x3=x1*x2 (1= Yes to x1 and x2; 0= No to x1 and/or x2). Individuals are represented by *i* and RDS recruitment chains are represented by *j*.

[c] Predicted probabilities were calculated using logistic regression with GEE (clustered on RDS recruitment chain) and robust standard errors and did not adjust for other significant correlates of active syphilis

$$\text{logit}\ \left(\frac{E(y_{ij})}{1-E(y_{ij})}\right)=-2.731\ -0.850\ x_{1\,ij}+0.814\ x_{2\,ij},$$

where x1= Live, work, buy drugs, or inject drugs in the Zona Norte (1=Yes; 0=No) and x2= ≥ 1 case of active syphilis ≤ 2 degrees of separation (1=Yes; 0=No). Individuals are represented by *i* and RDS recruitment chains are represented by *j*.