AIDS Behav (2014) 18:2366–2373
DOI 10.1007/s10461-013-0640-8

ORIGINAL PAPER

Spatial Recruitment Bias in Respondent-Driven Sampling: Implications for HIV Prevalence Estimation in Urban Heterosexuals

Samuel M. Jenness · Alan Neaigus · Travis Wendel · Camila Gelpi-Acosta · Holly Hagan

Published online: 13 October 2013

Abstract Respondent-driven sampling (RDS) is a study design used to investigate populations for which a probabilistic sampling frame cannot be efficiently generated. Biases in parameter estimates may result from systematic non-random recruitment within social networks by geography. We investigate the spatial distribution of RDS recruits relative to an inferred social network among heterosexual adults in New York City in 2010. Mean distances between recruitment dyads are compared to those of network dyads to quantify bias. Spatial regression models are then used to assess the impact of spatial structure on risk and prevalence outcomes. In our primary distance metric, network dyads were an average of 1.34 (95 % CI 0.82–1.86) miles farther dispersed than recruitment dyads, suggesting spatial bias. However, there was no evidence that demographic associations with HIV risk or prevalence were spatially confounded. Therefore, while the spatial structure of recruitment may be biased in heterogeneous urban settings, the impact of this bias on estimates of outcome measures appears minimal.

Keywords Respondent-driven sampling · Survey sampling · HIV/AIDS · Heterosexual

S. M. Jenness (✉)
Department of Epidemiology, University of Washington, 1959 Pacific Street NE, Box 357236, Seattle, WA 98195, USA
e-mail: sjenness@uw.edu

A. Neaigus
New York City Department of Health, HIV Epidemiology Program, New York, NY, USA

T. Wendel
Department of Anthropology, John Jay College of Criminal Justice, New York, NY, USA

C. Gelpi-Acosta
Department of Sociology, The New School, New York, NY, USA

H. Hagan
College of Nursing, New York University, New York, NY, USA

Introduction

Respondent-driven sampling (RDS) is a study design used to investigate "hidden" populations: those for which a probabilistic sampling frame cannot be generated efficiently or at all [1]. An example is injection drug users: there is no enumeration of population members given the illegal nature of their defining activity [2]. Probabilistic sampling may be possible for a potentially enumerated subgroup, such as those visiting medical facilities, but inference to the full population will be biased if the institutionalized systematically differ from the non-institutionalized [3]. RDS addresses this by sampling at convenience an initial set of population members ('seeds'), then having each seed sample a theoretically random set of members from his network of peers in the target population, then having those peers recruit, and so on until a target sample size is met. RDS has become increasingly popular not just for its potential to minimize biases of convenience sampling, but also for its logistical efficiency (although there are several examples demonstrating its inefficiencies [4–6]). Study procedures typically occur at fixed geographic locations that respondents and their recruits visit [7].

Recruitment network characteristics (network size, or degree; and assortative mixing, or homophily) are collected in RDS studies to adjust estimates for biases of common to network-based recruitment: groups with high degree and homophily tend to be oversampled [1]. Once

weighted, study estimates may be theoretically unbiased, although statistical uncertainty is larger than in traditional sampling. Variance is especially inflated when there is a "bottleneck" of recruitment in one part of the social network [8]; this occurs when there are sparse connections between groups hindering recruitment. There may also be systematic non-random sampling within the social network, in which all respondents tend to recruit peers with a particular trait (e.g., lower income level). This phenomenon violates a theoretical RDS assumption of random recruitment, and current methods used to adjust for recruitment biases do not address this one [9]. Bottlenecks and non-random recruitment may be inherently connected in cases where patterns of convenience sampling exacerbate already sparse connections between groups.

Geography is an important dimension to consider for RDS recruitment when outcomes are spatially clustered. Directly transmitted infectious diseases like HIV are often spatially clustered, insofar as the geography represents some latent properties of sexual networks [10]. A feasibility question for RDS is whether recruitment chains penetrate into the geographic distribution of networks sufficiently enough to estimate parameters efficiently. In Brazil, an RDS study of drug users found spatial diffusion of recruitment across the target geography, and while some areas were possibly underrepresented, no data on the spatial distribution of the underlying target population were available to confirm this [11]. A recent study in rural Uganda that compared an RDS sample to a known population cohort found biased estimates of socioeconomic status (among other outcomes) due to systematic under-recruitment of wealthier population members [12]. A subsequent spatial analysis of this study found variation in recruitment by location, but minimal bias in parameter estimates by geography mostly due to small spatial heterogeneity in the population [13].

In this study, we extend this line of inquiry to investigate the spatial bias of parameter estimates from an RDS study of high-risk heterosexually active adults in New York City (NYC), an urban area with considerable geographic and demographic heterogeneity [14]. The goal is to quantify any spatial bias in recruitment that may impact estimation of HIV prevalence and risk. Our hypothesis is that the recruitment network is a geographically constrained subset of the larger social network, and this constraint will spatially bias estimates to a greater degree than seen in rural Uganda given the unique heterogeneity in urban NYC. Further, we extend the primarily descriptive statistical methods used heretofore, and we also provide a framework for inferring the spatial distribution of a social network in lieu of a network census by using egocentric network data.

Methods

Procedures

Our current study is part of the National HIV Behavioral Surveillance (NHBS) project, a cross-sectional study of HIV prevalence and risk among three groups in 20 U.S. cities with elevated AIDS prevalence [15]. This analysis uses NHBS data collected in NYC in 2010, during the study cycle for high-risk heterosexuals, the methods for which have been described in detail [16]. Since the U.S. heterosexual HIV epidemic is concentrated in impoverished areas [17], we conducted background research with local HIV surveillance and U.S. Census data to determine the areas within NYC with the highest rates of HIV and household poverty. Two clusters were found (Harlem/South Bronx and Central Brooklyn), and initial data collection efforts focused on these areas.

Study ethnographers selected a small number of seeds ($n = 8$) through community-based outreach with the goal of diversity with respect to race/ethnicity, age, and gender (they were balanced across these traits). Seeds participated in the study and were then asked to recruit up to 3 peers (defined only as sexually-active adults they knew) into the study. These recruits were given the same recruitment opportunity, and recruitment continued until the target sample size ($n \approx 500$) was achieved. Unique study identification numbers were used to link the recruiters and recruits.

To be eligible, respondents had to report heterosexual vaginal or anal intercourse in the past year, be age 18–60, reside in NYC, and comprehend English or Spanish. Residence within the geographic cluster where we recruited seeds was not an eligibility criterion, but most recruits lived in these areas. The sample accordingly reflects the demography of these specific target neighborhoods and not NYC overall. Informed consent was obtained from all eligible respondents, who were compensated for their participation ($30) and also successful peer recruitment ($10 for each recruit).

Measures

In a structured interviewer-administered survey, respondents provided the closest street intersection to where they currently resided. In a sexual history module, they were asked about their last six sexual partnerships in the past year, including where (again to the closest street intersection) that partnership last occurred. Also considered in this analysis are gender, age, race/ethnicity, past-year income, past-year homelessness, past-year sexual risk factors (unprotected sex with causal partners and having multiple partners), and disease outcomes (diagnoses of non-HIV

sexually transmitted diseases in the past year, reported HIV diagnosis ever, and HIV infection). HIV infection was determined through whole-blood testing procedures. Blood collected through venipuncture was tested on HIV-1/2 enzyme-linked immunosorbent assays and confirmed on HIV-1 Western blot platforms (Bio-Rad Laboratories).

Geospatial Data Processing

Residential and sexual partnership locations were geocoded as latitude-longitude coordinates. Spatially oriented network edges (i.e., links) were established between paired spatial points (i.e., dyads), with up to nine edges possible: one edge for each of a respondent's recruits (up to three) and one edge for each of a respondent's sexual partners (up to six). Euclidean distances were calculated for all edges, and distance-based edges were mapped, using the sp and rgdal packages in R [18].

Statistical Analysis

Three analyses were conducted. First, descriptive statistics on the overall sample composition were run. The overall sample was compared to three analytic subsamples: respondents with any partnership geodata, respondents with any recruiter geodata, and respondents with both partnership and recruiter geodata. This final group is used to address the main research questions in the distance analyses as follows.

Second, the geographic distribution of the recruitment network was compared to the geographic distribution of the sexual partnership network, which we use as a proxy for the respondent's social network. Total recruitment network distance and total sexual partnership network distance were calculated by summing the distances of a respondent's unique recruitment and partner edges, respectively. Mean recruitment network and mean sexual partnership network distance are calculated by dividing these sums by the number of recruits and the number of partners, respectively. Overall sample means are thus the means of individual mean distances. We hypothesized that the mean sexual partnership distance, as a proxy for the social network distance, would be significantly greater than the mean recruitment distance; if so, the recruitment network may be a geographically constrained subset of the social network, suggesting spatial recruitment bias. Linear regression was used first to determine whether the difference of mean distances was greater than zero, and second whether the difference of mean distance varied by demographic and other covariates.

The sexual partnership distance distribution may be an inaccurate proxy for the social network distribution, because the sexual contact space differs from the larger social contact space. Thus, separate summary measures were calculated for mean recruitment network and mean sexual partnership network distance *away from the respondent's residence* by dividing the total distances above by the number of recruits and partners at a different spatial location than the respondent. We call these "truncated differences," since we remove some spatial locations.

Finally, the spatial dependence of two outcomes (multiple sexual partnerships in the past year and prevalent HIV infection) were investigated to determine whether spatial recruitment bias could influence these parameter estimates. The hypothesis was that any spatially related confounding in these associations could reflect spatial recruitment bias in RDS. Because this hypothesis concerned the overall sample independent of linked dyads, the full analytic sample was used. We used generalized additive logistic regression models parameterized as

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \mathbf{X}_i\beta + g(\mathbf{s}_i),$$

where the log odds of the outcome is modeled as a linear function of covariates, $\mathbf{X}_i$, and a spatially structured spline subfunction of the spatial coordinates, $\mathbf{s}_i$. The spatial term captures the residual spatial variation in association above the non-spatial covariates in order to assess overall spatial non-stationarity. Three demographic covariates (age, female gender, and black race) were chosen for all models as the basis for comparing non-spatial models (no spatial term added) and spatial models (as above). A difference in the beta coefficient after adding the spatial term is interpreted as spatially induced confounding (although any differences are evaluated qualitatively because there are no purely statistical methods to evaluate confounding [19]). These models were fit using the mgcv package in R [20].

Results

From the seeds (whom we dropped from the analysis), a total of 523 eligible respondents completed the study. Of these, 521 had their own complete residential geodata (a successfully geocoded coordinate). Of the 521, 442 were given RDS coupons for further peer recruitment (reasons for not providing coupons to all respondents were high socioeconomic status or current injection drug use, and also end-of-study recruitment tapering). Of the 442 given coupons, 237 had at least one recruit who successfully completed the study. With one of the 237 missing geodata, there were 236 respondents with at least one linked recruitment pair. Of the 521 with geodata, 494 reported one sexual partner whose location could be successfully

geocoded. In summary, 504 had either linked recruiter data or partner data, 494 had linked partner data only, 236 had linked recruiter data only, and 226 had both linked partner and recruiter data.

Table 1 shows the characteristics of the full sample and these analytic subsamples. Overall, we recruited a heavily African-American, impoverished, and high-risk sample of heterosexually active adults. Statistical testing was conducted to investigate whether there were any systematic differences between the full sample and analytic subsamples, with the p values representing pairwise tests with the full sample. Respondents with partner geodata had higher incomes and were less likely to be homeless. Those with recruiter geodata or both partner and recruiter geodata were older and less likely to have multiple sexual partners. Overall, the analytic subsample with complete geodata ($n = 226$) was similar to the full sample.

Figure 1 illustrates the spatial edges in the recruitment (left) and partnership dyads (right). The partnership edge distribution is denser and exhibits more cross-geography edges than the recruitment distribution. However, one caveat is that the number of recruitment edges is artificially constrained to half that of the partnership edges (three vs six).

Table 2 provides the results of the summary statistics of the distance measurements for the analytic subsample with both recruitment and partnership geodata ($n = 226$). The mean distance between a recruiter and a recruit was 1.44 miles (95 % CI 1.18–1.70 miles). Overall, 8 % of recruits were of sex partners. There was small variation around this mean for by-group comparisons, with no statistically significant differences by group. The mean distance between a respondent and a sexual partnership was 1.38 miles (95 % CI 1.08–1.69). Men had a significantly higher mean partnership distance compared to women, as did the homeless compared to the non-homeless, and those with multiple partners compared to those with only one partner. In fact, the median partner distance of the most recent sexual partnership was zero, suggesting that partnership tended to occur at the respondent's home. Overall, 48 % of respondents had a median partnership distance of zero miles (alternatively, 50 % of all partnerships were at a zero distance). There were no significant differences across the traits in Table 2 for respondents with a zero-distance median to those with a median greater than zero.

The mean difference in distances between the recruitment network and sexual partnership network proxy was

Table 1 Sample characteristics overall and by geodata completeness

	Total ($n = 521$)		Partner geodata ($n = 494$)			Recruiter geodata ($n = 236$)			Both geodata ($n = 226$)		
	n	%	n	%	p	n	%	p	n	%	p
Gender					0.38			0.63			0.52
Male	311	59.5	296	59.9		143	60.6		138	61.1	
Female	212	40.5	198	40.1		93	39.4		88	38.9	
Age					0.47			0.06			0.04
18–29	184	35.2	174	35.2		71	30.1		68	30.1	
30–39	71	13.6	68	13.8		36	15.3		35	15.5	
40–49	138	26.4	127	25.7		60	25.4		56	24.8	
50–60	130	24.9	125	25.3		69	29.2		67	29.6	
Race/ethnicity					0.97			0.07			0.11
Black	416	79.5	393	79.6		196	83.1		187	82.7	
Non-Black	107	20.5	101	20.4		40	16.9		39	17.3	
Sociodemographics[a]											
Income >$10,000	223	42.9	215	43.8	0.07	92	39.3	0.14	90	40.2	0.28
Homeless	204	39.0	187	37.9	0.03	86	36.4	0.28	80	35.4	0.14
Sexual risks[a]											
Casual unprot. sex	271	51.8	259	52.4	0.25	118	50.0	0.45	113	50.0	0.47
Multiple partners	414	79.2	394	79.8	0.16	178	75.4	0.06	171	75.7	0.09
Disease outcomes											
Reported STD diagnosis[a]	44	8.4	41	8.3	0.70	25	10.6	0.10	23	10.2	0.20
Reported HIV diagnosis	16	3.1	15	3.0	0.90	9	3.8	0.36	9	4.0	0.28
Tested HIV infection[b]	49	9.6	46	9.6	0.84	24	10.6	0.51	23	10.6	0.54

[a] In the past year

[b] HIV tested $n = 507$

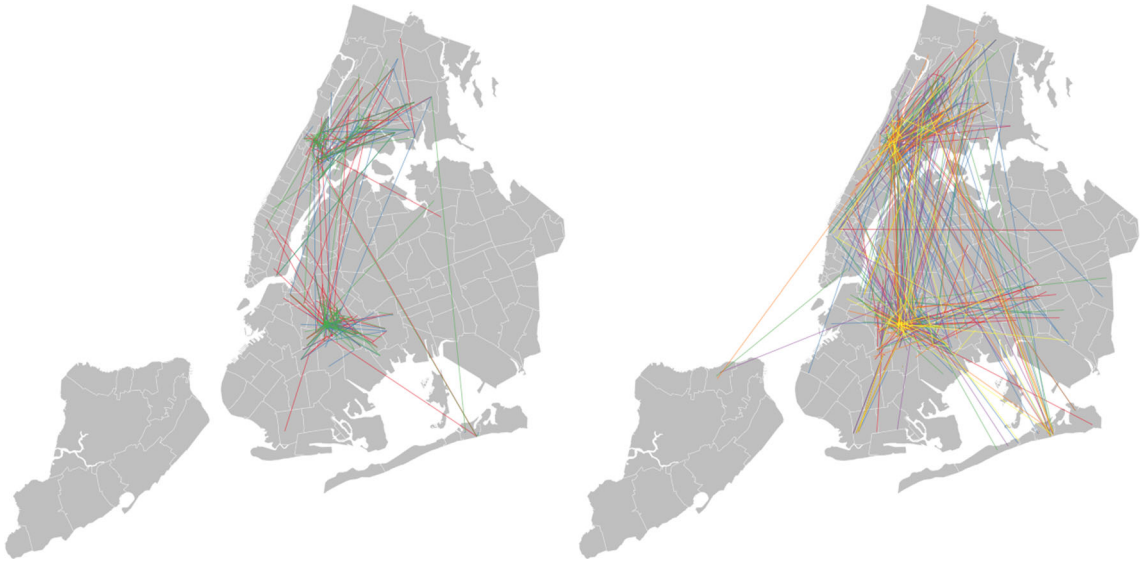

Fig. 1 The left map displays the recruitment edges of respondents with complete recruitment and partnership geodata (n = 226), with the edge color corresponding to the recruitment number (1–3). The right map displays the partnership (social) edges of the same respondents, with edge color corresponding to partner number (1–6). The partnership map is denser partially by design because respondents were given up to 3 recruitment opportunities but ask about up to 6 partners (Color figure online)

−0.06 miles, meaning that partnership dyads were on average 324 feet closer than recruiter dyads. However, after truncating the difference in mean distances to exclude those partnership dyads that occurred at zero distances, those dyads were on average 1.34 miles farther away than recruitment dyads. Men, the homeless, and the HIV-infected were all significantly more likely to have higher truncated differences.

In Table 3, we show the results of the regression models for the impact of spatial confounding on parameter estimates of the associations between three predictor variables (age, female gender, and black race) and two outcome variables (multiple sexual partnerships and HIV infection). Model coefficients are in log odds. In the non-spatial model, age and female gender were significantly associated with having multiple partners; and age, female gender, and black race were marginally or significantly associated with HIV infection. In the spatial models, the statistical significance of these associations did not substantially vary. Only the coefficient for black race on multiple partnerships showed a non-trivial change with the incorporation of the spatial function. Yet, the difference only represented a 4 % change in the adjusted probability of the outcome.

Discussion

In our study of high-risk heterosexual adults in NYC, we found evidence of spatial bias in recruitment, but minimal impact of spatial structure on two key epidemiologic outcomes. The quantitative criterion used to determine bias was a comparison of the mean distances between the recruitment dyads and the mean distances between the sexual partnership dyads. Our main limitation was the use of sexual partnership locations as a proxy for the geography of social peer interactions; we assumed that the social network spatial distribution represents a common distribution from which both sexual partnerships and recruitments arose.

Previous research has suggested that a core assumption of RDS, random recruitment within respondents' eligible social network [21], is not always met [9, 12]. The problem is not that recruitment is assortative, wherein groups on a given trait preferentially recruit among themselves, since the method includes techniques to adjust for that phenomenon [8]. Instead, uncontrolled bias occurs when *all* respondents preferentially recruit one group on that trait. An example from the Uganda study is income: all RDS recruiters tended to preferentially recruit lower income peers rather than a random selection of peers independent of income, based on the notion that higher income peers would not participate [12]. Some of this bias may be addressed in protocol development, by investigating and mitigating barriers to recruitment prior to and during study implementation [22]. Yet, some natural recruitment tendencies may be immutable.

The weighted RDS estimators have been shown to be unbiased, even with violations of the random recruitment assumption [8]. But violations have the effect of amplifying the variance. The design effect, which is the ratio of variance of a given study design to that of a simple random sample, may be as high as 10 for some parameters for

Table 2 Comparison of average Euclidean distances for recruits and partnerships, among respondents with complete geodata (n = 226)

	Recruiter distance[c]		Partner distance[d]		Distance difference[e]		Truncated difference[f]	
	Mean	95 % CI	Mean	95 % CI	Mean	95 % CI	Mean	95 % CI
Total	1.44	1.18, 1.70	1.38	1.08, 1.69	−0.06	−0.38, 0.26	1.34	0.82, 1.86
Gender								
Male	1.51	1.12, 1.89	1.48*	1.06, 1.89	−0.03*	−0.49, 0.43	1.47*	0.84, 2.10
Female	1.34	1.02, 1.66	1.23	0.76, 1.70	−0.11	−0.54, 0.32	1.12	0.15, 2.08
Age								
18–29	1.33	0.82, 1.85	1.29	0.80, 1.79	−0.04	−0.69, 0.62	1.15	0.13, 2.18
30–39	0.94	0.55, 1.33	1.06	0.38, 1.75	0.12	−0.53, 0.78	1.37	0.17, 2.57
40–49	1.71	1.07, 2.36	1.25	0.64, 1.86	−0.47	−1.12, 0.19	1.24	−0.12, 2.60
50–60	1.58	1.13, 2.04	1.74	1.03, 2.45	0.16	−0.45, 0.77	1.61	0.81, 2.41
Race/ethnicity								
Black	1.37	1.10, 1.64	1.36	1.02, 1.70	−0.01	−0.34, 0.32	1.40	0.87, 1.92
Non-Black	1.79	0.96, 2.62	1.47	0.66, 2.28	−0.32	−1.37, 0.74	1.07	−0.78, 2.91
Sociodemographics[a]								
Income >$10,000	1.31	0.86, 1.75	1.45	0.93, 1.97	0.14	−0.41, 0.69	1.74	0.81, 2.67
Homeless	1.37	0.89, 1.85	2.26*	1.57, 2.95	0.89	0.26, 1.52	2.02*	1.10, 2.94
Sexual risks[a]								
Casual unprot. sex	1.35	0.97, 1.74	1.47	1.06, 1.87	0.11	−0.24, 0.47	1.01	0.38, 1.64
Multiple partners	1.48	1.17, 1.80	1.62*	1.24, 1.99	0.13	−0.24, 0.51	1.28	0.74, 1.83
Disease outcomes								
Reported STD diagnosis[a]	0.90	0.47, 1.34	1.52	0.70, 2.35	0.62	−0.29, 1.53	0.82	−0.41, 2.06
Reported HIV diagnosis	1.39	0.30, 2.49	1.18	−0.30, 2.65	−0.22	−2.01, 1.58	4.33	−5.13, 14.93
Tested HIV infection[b]	1.76	0.95, 2.58	1.78	0.67, 2.88	0.01	−1.17, 1.19	3.17*	0.92, 5.41

* $p < 0.05$, representing by group comparisons with each distance metric

[a] In the past year

[b] HIV tested n = 507

[c] Mean distance (miles) between study subject and successfully completed recruits with geodata

[d] Mean distance (miles) between study subject and sexual partnerships with geodata

[e] Difference between sexual partnership mean distance and recruit mean distance

[f] Distance difference corrected for sexual partnerships at the same location as the study subject

Table 3 Generalized additive logistic regression model of risk and disease outcomes with and without spatial splines (n = 523)

	Non-spatial model			Spatial model		
	Coef.	s.e.	p	Coef.	s.e.	p
Multiple partners						
Age (continuous)	−0.023	0.009	0.010	−0.023	0.009	0.009
Female gender	−0.537	0.221	0.015	−0.542	0.222	0.015
Black race	0.166	0.281	0.555	0.038	0.289	0.896
HIV infection						
Age (continuous)	0.108	0.020	<0.001	0.109	0.020	<0.001
Female gender	0.531	0.324	0.102	0.530	0.328	0.107
Black race	1.375	0.749	0.066	1.400	0.757	0.064

which there is heavy bottlenecking of recruitment [23]. A design effect of this size effectively reduces a sample size of 1,000 to a size of 100. Thus, there is a significant loss of efficiency in estimation. Many RDS studies simply ignore this and analyze data as if collected in a simple random sample.

Distance and more complex spatial structures are theoretically important aspects of this recruitment bias. Since recruitment is incentivized, recruiters make choices about whom, within their social network, to recruit who will participate in the study in a reasonable time frame [13]. As recruiters have been shown to non-randomly recruit lower income persons based on this psychological mechanism [12], it is plausible that they would also recruit peers at convenience with respect to proximity too.

The main challenge to evaluating this methodological question concerns the reference group: from where do the

RDS recruiters' full network of peers within the target population originate [7]? One may compare the spatial distribution of recruits to that of some expected spatial density that is derived qualitatively [11], but this has limited scientific rigor. Alternatively, the underlying social network may be enumerated [12], but this is an expensive and lengthy process. It may be impossible in a dynamic open population, especially in dense, heterogeneous environments like cities where many RDS studies occur.

The generative process for spatial recruitment bias is ecological, where the probability of recruitment dyad formation is influenced by local population density and transportation infrastructure. This process may differ considerably by area, limiting the generalizability of any findings from small, rural settings to large, urban settings. In this respect, our substantive findings may not be generalizable to cities other than NYC, which has unique urban features. The broader issue addressed here, therefore, is methodological: how to go about evaluating potential spatial bias that is unique and local.

Our approach was to use egocentric network data on sexual partnerships to infer characteristics about the parent social network to serve as the reference group. There are clear limitations with this proxy, as indicated by the median zero distance for the most recent partnership: many of these measurements for partner location represent only the location of same-residence partnerships (e.g., cohabiting partners). The sexual partnership network is not a good approximation of the social network with respect to all scientific questions, but for geography it may be useful. This is because the sexual network is by definition a subset of the social network, and thus the spatial distribution of that sexual network conservatively estimates that of the social network. With respect to space, any biases in using the proxy would be conservative.

However, one strong conservative bias that we do account for is the overrepresentation of sexual partnerships at home, which will artificially narrow our estimate of the spatial distribution of the social network. For that reason, we use the truncated distance measure as our primary metric for evaluation here. Overall, 50 % of sexual partnerships occurred at this null distance, compared to 20 % of the recruitments; only 8 % of recruitments were of a sexual partner. So the networks are not perfectly aligned, and some of that disjoint concerns our research question (true differences between the social and RDS recruitment networks) and some is residual, conservative bias within our proxy.

As the science of RDS and related network methods evolve, it is critical to consider generally what geographic measure of the underlying social network should be used. The ideal is to measure all the places where people come into contact socially, since those are the spatial points at which RDS coupons could be distributed. One might propose the location of social network members' residences to be the gold standard, but these may not be social spaces at all, especially for populations who mainly engage in public spaces like bars, parks, street corners, and churches. Although our proxy is limited, a strength is that it measures a point where contact actually occurred, which may not be the case for network members' residences. A promising model-based approach that would infer the spatial distribution of the social network from the observed recruitment network may overcome some of our limitations [24], yet both this and our approach capitalize on egocentric network data to make inference on full networks that are difficult or impossible to survey directly.

In the end, our study suggests some degree of spatial recruitment bias, particularly among respondents who were male, homeless, or HIV-infected. But our outcome modeling, which investigated whether biased spatial recruitment could itself impact the parameter estimates for epidemiologically relevant outcomes, suggested no major spatially induced confounding. Overall, the potential for biased outcome estimates appears minimal even in the face of spatial recruitment bias. This overall pattern of results is similar to those in Uganda [13], and counter to our hypothesis that we would see a greater effect in an dense, urban environment like NYC. One reason may be that our target population was actually a rather homogenous, clustered subset of the larger heterogeneous population.

Limitations

As noted throughout, the key limitation of this analysis is the use of sexual partners as a proxy for network peers. This choice was made based on the data available. Future research on spatial RDS bias should consider alternative measures that better reflect the underlying social network from which RDS recruits are drawn. While network members' residence is one choice, we reiterate that it is also a proxy itself. There is a great potential for misreporting of locations, even with survey aides like maps; another strength to our proxy is that respondents may report on the locations of the sexual partnerships more accurately than peers' residences. In the end, ours is the first study to attempt to quantify the geography of RDS recruitment within the U.S., and we hope this work encourages other investigations into using better measures in different settings. Another limitation of our paper concerns our outcome regression modeling, in which estimation depends upon spatially neutral sampling [20]. Since this is violated by RDS, any lack of spatial dependence should not be generalized to the larger population and these results should be interpreted as in-sample characteristics only.

Conclusions

RDS is an innovative and popular method to study hidden populations, which often play an epidemiologically relevant role in infectious disease transmission. Despite its poor statistical efficiency and potential for biased estimation because of its functional assumptions about recruitment, it may be the best tool given the alternatives [25]. Continued research on the source and magnitude of these biases will help investigators address three important questions: (1) whether to use RDS versus a competing sampling method; (2) how to design a study protocol that is more robust to recruitment bias; (3) when additional statistical methods are needed to adjust RDS estimates for the effects of any remaining recruitment bias. Our study suggests that spatial recruitment bias should be considered for this type of RDS study, but that it may have minimal impact on estimation of HIV prevalence and risk.

Acknowledgments This work was supported by a cooperative agreement between the New York City Department of Health and the Centers for Disease Control and Prevention (#U62/CCU223595-03-1). We would like to thank the NYC NHBS field staff for all their efforts on the study. We also greatly appreciate the insightful comments on an earlier version of this manuscript from three anonymous reviewers.

References

1. Heckathorn D. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. Sociol Methodol. 2007;37(1):151–207.
2. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS. 2005;19(Suppl 2):S67–72.
3. Des Jarlais DC, Arasteh K, Hagan H, McKnight C, Perlman DC, Friedman SR. Persistence and change in disparities in HIV infection among injection drug users in New York City after large-scale syringe exchange programs. Am J Public Health. 2009;99(Suppl 2):S445–51.
4. Hathaway AD, Hyshka E, Erickson PG, et al. Whither RDS? An investigation of respondent driven sampling as a method of recruiting mainstream marijuana users. Harm Reduct J. 2010;7:15.
5. Johnston LG, Malekinejad M, Kendall C, Iuppa IM, Rutherford GW. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: field experiences in international settings. AIDS Behav. 2008;12(4 Suppl):S131–41.
6. Simic M, Johnston LG, Platt L, et al. Exploring barriers to 'respondent driven sampling' in sex worker and drug-injecting sex worker populations in Eastern Europe. J Urban Health. 2006;83(6 Suppl):i6–15.
7. Malekinejad M, Johnston LG, Kendall C, Kerr LR, Rifkin MR, Rutherford GW. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. AIDS Behav. 2008;12(4 Suppl): S105–30.
8. Goel S, Salganik MJ. Assessing respondent-driven sampling. Proc Natl Acad Sci USA. 2010;107(15):6743.
9. Wejnert C, Heckathorn DD. Web-based network sampling efficiency and efficacy of respondent-driven sampling for online research. Sociol Methods Res. 2008;37(1):105–34.
10. Rothenberg R, Muth SQ, Malone S, Potterat JJ, Woodhouse DE. Social and geographic distance in HIV risk. Sex Transm Dis. 2005;32(8):506–12.
11. Toledo L, Codeco CT, Bertoni N, Albuquerque E, Malta M, Bastos FI. Putting respondent-driven sampling on the map: insights from Rio de Janeiro, Brazil. J Acquir Immune Defic Syndr. 2011;57(Suppl 3):S136–43.
12. McCreesh N, Frost SDW, Seeley J, et al. Evaluation of respondent-driven sampling. Epidemiology. 2012;23(1):138–47.
13. McCreesh N, Johnston LG, Copas A, et al. Evaluation of the role of location and distance in recruitment in respondent-driven sampling. Int J Health Geogr. 2011;10:56.
14. Shepard CW, Gortakowski HW, Nasrallah H, Cutler BH, Begier EM. Using GIS-based density maps of HIV surveillance data to identify previously unrecognized geographic foci of HIV burden in an urban epidemic. Public Health Rep. 2011;126(5):741–9.
15. Centers for Disease Control and Prevention. HIV infection among heterosexuals at increased risk—United States, 2010. MMWR Morb Mortal Wkly Rep. 2013;62(10):183–8.
16. Reilly KH, Neaigus A, Jenness SM, Hagan H, Wendel T, Gelpi-Acosta C. High HIV prevalence among low-income, Black Women in New York city with self-reported HIV negative and unknown status. J Womens Health. 2013;22(9):745–54.
17. Jenness SM, Neaigus A, Murrill CS, Wendel T, Forgione L, Hagan H. Estimated HIV incidence among high-risk heterosexuals in New York City, 2007. J Acquir Immune Defic Syndr. 2011;56(2):193–7.
18. Bivand RS, Pebesma EJ, Gomez-Rubio V. Applied spatial data analysis with R. New York: Springer; 2008.
19. Pearl J. Causality: models, reasoning and inference. Cambridge: Cambridge University Press; 2000.
20. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Ser B. 2011;73(1):3–36.
21. Heckathorn D. Respondent-driven sampling: a new approach to the study of hidden populations. Soc Probl. 1997;44:174–99.
22. Johnston LG, Whitehead S, Simic-Lawson M, Kendall C. Formative research to optimize respondent-driven sampling surveys among hard-to-reach populations in HIV behavioral and biological surveillance: lessons learned from four case studies. AIDS Care. 2010;22(6):784–92.
23. Salganik MJ. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. J Urban Health. 2006;83(6 Suppl):i98–112.
24. Ott M, Gile KJ, Uuskula A, Johnston LG. Spatial dependencies in respondent-driven sampling data. Redondo Beach: Sunbelt XXXII; 2012.
25. Kendall C, Kerr LR, Gondim RC, et al. An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil. AIDS Behav. 2008;12(4 Suppl):S97–104.