

On the issue of homophily in respondent-driven sampling: notes based on the case of men who have sex with men in Belo Horizonte, Brazil

Carla Jorge Machado*

Mark Drew Crosland Guimarães**

Introduction

Respondent-driven sampling (RDS) is a process used to collect data from *hard-to-reach* populations, such as men who have sex with men (MSM). This is a process of accessing a hidden population of interest via links in the network of acquaintances belonging to the population, and it can be a useful epidemiological tool for estimating HIV prevalence in high-risk populations. However, a typical sample arising from an RDS process on a network contains a certain degree of homophily, and in case the quantity surveyed is a given characteristic, this can generate biased results¹ (PAJAK, 2011; ACHARYA, 2007).

In the present note we aim to assess whether there is homophily in a sample of MSM, by selected characteristics of recruits in a study conducted in Belo Horizonte, Minas Gerais, Brazil, 2008-2009.

Overview of the RDS Reasoning

RDS is based upon the idea of "snowball sampling" (getting individuals to refer those they know, who, in turn refer those they know and so on). Therefore, the sample is collected in a non-random way. Even then, RDS represents an advance in sampling methodology since it allows researching groups that are relevant to public health interventions, such as injector drug users, prostitutes, gay men, street youth, and the homeless. For this specific population groups, although standard probability sampling methods could be used, coverage of the target population would be limited, which makes RDS a more feasible alternative² (PAJAK, 2011; ACHARYA, 2007).

RDS is a network-based method that starts with a set of initial respondents (recruiters), who refer their peers; these in turn refer their peers (recruitees), and so on, as the sample expands from wave to wave. However, this approach can yield selection biases, since there is a tendency for most people to recruit those whom they resemble in race, ethnicity, education, income, and religion. Well-connected individuals tend to be over-sampled because many recruitment paths are conducive to them³ (ACHARYA, 2007).

To produce valid estimates of a given outcome, RDS proposes that respondents recruit their peers, as in network-based samples, and researchers keep track of who recruited whom and their number of social contacts. A mathematical model of the recruitment process then weights the sample to compensate for non-random recruitment patterns. The resulting statistical theory, termed RDS, enables researchers to

* Ph.D in Demography, associate professor of the Federal University of Minas Gerais.

** Sc.D in Epidemiology, associate professor of the Federal University of Minas Gerais.

¹ Available in: <<http://www.respondentdrivensampling.org/reports/RDSsummary.htm>>. Access in: 24 Jun. 2012.

² Available in: <<http://www.respondentdrivensampling.org/reports/RDSsummary.htm>>. Access in: 24 Jun. 2012.

³ Available in: <<http://www.respondentdrivensampling.org/reports/RDSsummary.htm>>. Access in: 24 Jun. 2012.

provide both unbiased population estimates and measures of the precision of those estimates⁴ (PAJAK, 2011; ACHARYA, 2007). Even then, it is important to know the size of social networks based on specific characteristics, which is difficult. Therefore, homophily should be avoided or, at least, checked for.

Data and Procedures

This is a cross-sectional study, part of the main project "*Comportamento, atitudes, práticas e prevalência de HIV e sífilis entre homens que fazem sexo com homens (HSH) em 10 cidades brasileiras*"⁵. RDS (Respondent Driven Sampling) was used to recruit individuals in their social network relations. The eligibility criteria for the Project were: to be a resident in the city of Belo Horizonte; not having participated in the research previously, having had at least one sexual intercourse (oral or anal sex) with a man in the last 12 months, presenting a valid coupon for study participation; accepting the conditions for participation, including answering a structured questionnaire and to be willing to invite peers to participate in the study; to accept signing the consent form, and not being under the influence of drugs, including alcohol, at the time of the visit⁶.

We assessed whether the HIV status of the recruiter was associated with the HIV status of the recruitee. Analyses were also undertaken for syphilis, age group, socioeconomic status (as measured by Brazilian Association of Research Companies – ABEP) and membership to a NGO in order to ascertain the degree of homophily in the sample. Statistical analyses were carried out by McNemar's Chi-square test, McNemar exact test or McNemar-Bowker test, using SPSS 12.0 statistical package. We accepted that there existed a dependency in case the significance level of the test was less than 0.05.

McNemar's test is appropriate when analyzing data from matched pairs of

subjects with a dichotomous response. It tests the null hypothesis of marginal homogeneity. Under the null hypothesis, it has an asymptotic chi-square distribution with one degree of freedom. The exact-value for McNemar's test was also calculated using binomial distribution. For $n \times n$ tables (higher level than 2×2 table), Bowker's test of symmetry needs to be used, and the null hypothesis is that the cell proportions are symmetric (DURKALSKI; PALESCH; LIPSITZ; RUST, 2003).

The study was approved by the Research Ethics Board (CONEP – registry 14494), by the Research Ethics Board of the Universidade Federal do Ceará – COMEPE (number 202/07) and by Secretaria Municipal de Saúde de Belo Horizonte (number 062/2007).

Results

We analyzed 220 to 238 recruiter-recruitee pairs among the 273 MSM included in the study. Table 1 indicates the percentage of recruiter-recruitee pairs of same characteristic, as related to the total of recruitees. In parenthesis we find the total percentage of recruitees, regardless of the recruiter's status.

For HIV+, therefore, the probability of recruiting an HIV positive individual was 28.6% when the recruiter was also HIV positive, but the total of HIV positive individuals recruited when the recruiter was either positive or negative was 7.3%. The McNemar exact test indicated no dependency on HIV status during the recruitment process ($p = 0.084$). For syphilis serology, these figures were, respectively, 16.7% and 8.5% ($p = 0.500$, indicating no recruiter-recruitee dependency).

In regard to age groups, for ages 18 to 24, the percentage of recruiter-recruitee pairs of the same age, as related to recruitees of same age-group was 65.5%, while the total percentage of recruitees aged 18 to 24,

⁴ Available in: <<http://www.respondentdrivensampling.org/reports/RDSsummary.htm>>. Access in: 24 Jun. 2012.

⁵ Available in: <<http://www.respondentdrivensampling.org/reports/RDSsummary.htm>>. Access in: 24 Jun. 2012.

⁶ Available in: <<http://www.respondentdrivensampling.org/reports/RDSsummary.htm>>. Access in: 24 Jun. 2012.

TABLE 1
Pairs of recruitee-recruiter (RDS sampling) among MSM
Belo Horizonte, Brazil – 2008-2009

	Recruiter Percentage of same characteristic for recruiter-recruitee (percentage of recruiters with characteristic in the sample)						
	HIV+	Syphilis+	Age 18-24	Age 25-34	Age 35+	Social Classes (C/D/E)	NGO+
Recruitee							
HIV +	28.6 (7.3)a						
Syphilis +		16.7 (8.5)b					
Age 18-24			65.5 (44.1)c				
Age 25-34				48.4 (37.4)c			
Age 35+					42.9 (18.5)c		
Social Classes (C/D/E)						40.5 (29.5)d	
NGO +							43.5 (15.1)b

a (n=220); exact McNemar Fisher test: p=0.084

b (n=223); exact McNemar Fisher test: p=0.500

c (n=238) ; McNemar-Bowker test p=0.023

d (n=234); McNemar test: p=0.128

e (n=238); McNemar test: p<0.001

regardless of the recruiter's age was 44.1; for ages 25 to 34 and 35+, this figures were, respectively: 48.4 and 37.4%, and 42.9 and 18.5%, indicating a significant dependency of recruitment by recruiters of similar ages of recruitees (p=0.023).

For social class groups, the results revealed no significance (p=0.128).

For NGO membership, however, the results were highly significant (p<0.001), given that the percentage of recruiter-recruitee pairs that were both members of NGOs were 43.5% as compared to the percentage of recruitees regardless of NGO membership, which was 15.1%.

Conclusions

The RDS method is still under development and data analysis is yet quite limited. Therefore, assessment of possible biases is very important (PASCOM; SZWARCOWALD; BARBOSA JR., 2010).

Estimates using information yielded by the RDS sample may have a tendency for in-group recruitment for age and NGO

membership, where participants tend to recruit others of a similar status.

The results lead us to the following thoughts: (1) indeed, it is the intention of RDS sampling to recruit elements of the population with similar characteristics. A central insight is that homophily can facilitate finding members of a population that could not be reached otherwise. Therefore, in this sense, homophily may be seen as a 'good thing' (that would be the case if MSM intercommunicate more often in NGOs); (2) However, homophily can have conflicting effects (JACKSON; LOPEZ-PINTADO, 2012): although it can facilitate initial recruitment, it might be that an increase in homophily can also lead to a decrease in the overall recruitment rate (that would be the case if MSM communicate in NGOs or in any other place at the same rate). Therefore, the eventual fraction of individuals with similar characteristics in the steady state may depend on homophily in a complicated number of ways, and researchers must be aware that homophily may exist and should explain their possible effects in the results.

References

ACHARYA, A. K. A methodological approach to study hidden populations. The case of trafficked women in Mexico. **City. Revista Internacional de Ciencias Sociales y Humanidades**, v. XVII, n. 001, p. 9-23, enero-junio 2007. Available in: <<http://www.redalyc.uaemex.mx/pdf/654/65417101.pdf>>. Access in: 02 jan. 2011.

DURKALSKI, V. L.; PALESCH, Y. Y.; LIPSITZ, S. R.; RUST, P. F. Analysis of clustered matched-pair data. **Statistics in Medicine**, v. 22, n. 15, p. 2417-28, 2003.

JACKSON, M. O.; LOPEZ-PINTADO, D. **Diffusion and contagion in networks with heterogeneous agents and homophily**. Ecore – International Association for Research, Econometrics and Statistics, March 2012 (Ecore discussion paper, 2012/31). Available in: <http://www.ecore.be/DPs/dp_1333545419.pdf>.

Access in: 29 jan. 2012.

PAJAK, D. The shadow of hierarchy – how to sample a hidden population of former employees? Jun. 2011. Available in: <<http://www.slideshare.net/dannypajak/the-shadow-of-hierarchy-how-to-sample-a-hidden-population-of-former-employees>>. Access in: 13 ago. 2011

PASCOM, A. R. P.; SZWARCOWALD, C. L.; BARBOSA JR., A. Sampling studies to estimate the HIV prevalence rate in female commercial sex workers. **Braz J Infect Dis** [online], v. 14, n. 4, p. 385-397, 2010. Available in: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-86702010000400014&lng=en&nrm=iso> Access in: 18 jan. 2011.

Recebido para publicação em 20/02/2012

Aceito para publicação em 31/08/2012