Social Science & Medicine xxx (2014) 1–15

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Challenges to recruiting population representative samples of female sex workers in China using Respondent Driven Sampling

M. Giovanna Merli [a,b,*], James Moody [b], Jeffrey Smith [c], Jing Li [e], Sharon Weir [d], Xiangsheng Chen [e]

[a] Sanford School of Public Policy & Duke Global Health Institute, Duke Population Research Institute, Duke University, Box 90312, Durham, NC 27708, USA
[b] Department of Sociology, Duke University, Durham, NC 27708, USA
[c] Department of Sociology, University of Nebraska, Lincoln, NE 68508, USA
[d] The Carolina Population Center and the Department of Epidemiology, Gillings School of Global Public Health, Campus Box 8120, University of North Carolina at Chapel Hill, Chapel Hill, NC 27546, USA
[e] National Center for STD Control, 12 Jiangwangmiao Street, Nanjing 210042, China

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Respondent Driven Sampling
HIV
Female sex workers
Network simulation
Social networks
China

ABSTRACT

We explore the network coverage of a sample of female sex workers (FSWs) in China recruited through Respondent Drive Sampling (RDS) as part of an effort to evaluate the claim of RDS of population representation with empirical data. We take advantage of unique information on the social networks of FSWs obtained from two overlapping studies – RDS and a venue-based sampling approach (PLACE) – and use an exponential random graph modeling (ERGM) framework from local networks to construct a likely network from which our observed RDS sample is drawn. We then run recruitment chains over this simulated network to assess the assumption that the RDS chain referral process samples participants in proportion to their degree and the extent to which RDS satisfactorily covers certain parts of the network. We find evidence that, contrary to assumptions, RDS oversamples low degree nodes and geographically central areas of the network. Unlike previous evaluations of RDS which have explored the performance of RDS sampling chains on a non-hidden population, or the performance of simulated chains over previously mapped realistic social networks, our study provides a robust, empirically grounded evaluation of the performance of RDS chains on a real-world hidden population.

1. Introduction

With 780,000 people who were living with HIV by the end of 2011 [range: 620,000–940,000] (China Ministry of Health, 2011), corresponding to .058 percent of the adult population, China is still a low HIV prevalence setting, with infections mostly concentrated in subpopulations such as female sex workers (FSWs), their clients and injecting drug users (IDUs) (China Ministry of Health, 2011). A recent review of 15 studies of HIV infection among small ($n < 400$) samples of FSWs recruited mainly by convenience sampling revealed a median prevalence rate of .6% and prevalence as high as 10% in samples of FSWs who inject drugs (Yan et al., 2011). Despite still relatively low HIV prevalence, FSWs in China are believed to display widespread risky behaviors such as large numbers of clients, low rates of condom use and lack of knowledge of HIV/STD risk behaviors (Chen et al., 2005; Hesketh et al., 2005). The role of FSWs in the progression of HIV and other STDs in China is crucial, especially at this early stage of the epidemic when sex workers may act as bridges of infection to the general population (Pirkle et al., 2007; Merli et al., 2006). These concerns are motivated by new evidence that the number of infections due to heterosexual transmission is growing (China Ministry of Health, 2011) and by a very rapid increase in the incidence of other STDs, especially syphilis, both among female sex workers and the general population (Chen et al., 2007; Tucker et al., 2010; Tucker and Cohen, 2011; Van Den Hoek et al., 2001; Wang et al., 2009a,b).

The recruitment of samples for representation of the population of FSWs in China is significant for understanding the distribution of health behaviors and health status of its members, for the identification of the most vulnerable population segments in need of interventions and for their impact on the health dynamics of the population at large. Beyond China, the importance of deriving valid estimates of population characteristics is of interest to an ongoing

* Corresponding author. Sanford School of Public Policy & Duke Global Health Institute, Duke University, Box 90312, Durham, NC 27708, USA.
E-mail address: giovanna.merli@duke.edu (M.G. Merli).

http://dx.doi.org/10.1016/j.socscimed.2014.04.022
0277-9536/

global debate regarding sampling methodologies of hidden and marginalized populations at risk of HIV/STDs (UNAIDS, 2008).

The study of FSWs is complicated because of the difficulty to recruit samples from which inference to the population can be drawn. FSWs are a hidden population. Because of social stigma and the illegal status of sex work, FSWs are characterized by the absence of complete sampling frames which prevents the use of standard probability sampling methods. In addition to sampling challenges, the organization of sex work in China is structured around a semi-rigid hierarchy of tiers of sex work, ranging from high to low according to place where sex work is solicited, price charged per sexual transaction and socioeconomic background of clients (Yamanis et al., 2013; Chen et al., 2012; Hershatter, 1997; Huang et al., 2004; Lim, 1998; Parish and Pan, 2006; Rogers et al., 2002; Xia and Yang, 2005). High-tier sex workers solicit clients by phoning rooms in star hotels; or are the female staff of karaoke halls who accompany customers in singing, dancing and drinking and provide sexual services for additional compensation. Middle-tier sex workers include the female staff of establishments which offer commercial sex services under the guise of personal services such as bathing and massage, hair washing, beauty services, and foot cleaning. Low-tier sex workers are women who solicit clients on the streets, in parks or other public spaces, or on construction sites. This variability in the organization of sex work complicates researchers' efforts to equally access the various strata of the population and poses significant challenges for intervention as programs sponsored by the Chinese Ministry of Health have been mainly directed at middle tier sex workers, while the hardest to reach population pockets in need of interventions remain underserved.

In recent years, Respondent Driven Sampling (RDS) (Heckathorn, 1997, 2002) has become a popular sample recruitment tool for hidden and hard to reach populations. RDS uses a link tracing sampling design, similar to earlier snowball sampling (Goodman, 1961) and random walk (Klovdahl, 1989) approaches, but seeks to provide a probability-based inferential structure for studying hidden populations that allows one to generate unbiased estimates of population characteristics and behaviors. Relative to other methods to recruit samples of hard to reach populations, RDS has quickly, efficiently and cost-effectively recruited large, diverse samples (Robinson et al., 2006; Kendall et al., 2008; Johnston and Sabin, 2006; Carballo-Dieguez et al., 2011). The public health community has made significant investments in this sampling approach (Lansky et al., 2007) with a large number of applications among populations at risk of acquiring and transmitting HIV/STDs such as injecting drug users, FSWs, and men who have sex with men (Malekinejad et al., 2008). However, the ability of RDS to faithfully represent hidden populations such as FSWs relies on strong, largely empirically untested assumptions regarding the unobserved participant referral process and the unobserved structure of the social network over which the RDS sampling process is running.

Here, we explore the ability of RDS to demographically and spatially cover the underlying network and to meet crucial assumptions about how participants refer other participants into the RDS sample in the context of FSWs in China. Our goal is to evaluate the performance of RDS chains on a hidden population network of the type RDS was developed to study with state-of-the-art simulations of empirical networks from sampled data. We take advantage of two concurrently implemented surveys of FSWs in Liuzhou (Guangxi Province, China), an RDS survey and a venue-based survey, to identify: (a) features of the RDS sampling process which produce imprecise descriptions of the risk profile of the hidden population; and (b) features of the network over which the RDS sampling process is running that interfere with the ability of RDS to adequately cover the network and reveal the most vulnerable population segments in need of health interventions.

2. Respondent Driven Sampling

In RDS, the target for representation is typically the hidden population of a well-defined geographic area or community (e.g. a city, a school, etc). RDS tracks links between members of the population who recruit other members to the sample. The sample recruitment process is initiated by the researcher administering the study through the selection of a small convenience sample of "seed" respondents known to the researchers from the population of interest. Seeds are interviewed and given a limited number of uniquely numbered coupons which they are asked to distribute to their immediate social contacts in the target population as a means of recruiting other participants from among their social networks. Members of the seeds' social circles who receive coupons and then choose to participate in the study form the first "wave" of the sample. This process advances recursively through multiple waves until a desired sample size is reached.

The relative ease of recruiting a sample with RDS comes at the cost of stringent assumptions about the RDS sampling process and the network over which this process is running. These assumptions are required to generate unbiased estimates of population characteristics, a feature that distinguishes the RDS approach from more familiar snowball sampling methods. In link tracing designs, one can calculate inclusion probabilities (and thus sampling weights) if network structure and position of the seeds are known (Thompson, 1990; Thompson and Seber, 1995). Population averages are then computed by applying sampling weights to subgroup members with a higher or lower known probability of being sampled. In RDS, the full network over which the referral process is running is not known and only a fraction of the network links of a sample unit are traced. Since nodes have variable numbers of in-ties, they have variable – but generally unknown – probability of being sampled (Frank, 1977) which makes it impossible to calculate inclusion probabilities. RDS solves this problem by using the logic of a random walk on a graph as a model for the RDS referral process. If the random walk proceeds long enough, it settles into an equilibrium state (or stationary distribution) where sampling probabilities are inversely proportional to a respondent's degree (Lovasz, 1993; Lawler and Coyle, 1999).

Based on the stationary distribution of a random walk and the assumption of a non-repeating, non-branching random walk, conventional RDS theory and corresponding inference (Volz and Heckathorn, 2008; Salganik and Heckathorn, 2004) assumes that as long as all members of the target population can be reached through the chain-referral process, an individual's sample inclusion probability is *exactly* proportional to the number of reciprocal ties she has with other members of the target population (her degree), which we refer to as the sampling with probability proportional to degree (SPPD) assumption (Neely, 2009). The two most commonly used RDS estimators rely on the SPPD assumption to make population proportion estimates. The S-H estimator presented in Salganik and Heckathorn (2004) is based on both the Sampling Probability Assumption and a Markov Model Maximum Likelihood Estimate (MLE) of the sampling proportions. It is estimated using cross-group recruitments and the relative probability of sampling group members, represented by an estimate of the harmonic mean degrees of each group, based on the rationale that this is exactly the Horvitz-Thompson estimator for mean degree when the probability of sampling is proportional to degree (see Thompson, 2002). The V-H estimator (Volz and Heckathorn, 2008) has the same form as the S-H estimator, except that the sampling proportions are estimated using the observed group proportions from the data rather than as Markov model MLE, hence the V-H only relies on the SPPD assumption. As discussed below, this estimator directly connects to conventional sampling estimation and was found to consistently outperform the S-H estimator in terms of bias and variance except when there is

Please cite this article in press as: Merli, M.G., et al., Challenges to recruiting population representative samples of female sex workers in China using Respondent Driven Sampling, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.04.022