Social Networks xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Social Networks

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

Assessing respondent-driven sampling: A simulation study across different networks

Sandro Sperandei[a,b,*], Leonardo Soares Bastos[b], Marcelo Ribeiro-Alves[c], Francisco Inácio Bastos[a]

[a] Institute of Scientific and Technological Communication & Information in Health, Oswaldo Cruz Foundation, Brazil
[b] Scientific Computational Program, Oswaldo Cruz Foundation (FIOCRUZ), Brazil
[c] National Institute of Infectious Diseases Evandro Chagas, Oswaldo Cruz Foundation (FIOCRUZ), Brazil

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Respondent-driven sampling
Hidden population
Hard-to-reach population
Simulation method
Random graph
Epidemiology

ABSTRACT

The purpose was to assess RDS estimators in populations simulated with diverse connectivity characteristics, incorporating the putative influence of misreported degrees and transmission processes. Four populations were simulated using different random graph models. Each population was "infected" using four different transmission processes. From each combination of population x transmission, one thousand samples were obtained using a RDS-like sampling strategy. Three estimators were used to predict the population-level prevalence of the "infection". Several types of misreported degrees were simulated. Also, samples were generated using the standard random sampling method and the respective prevalence estimates, using the classical frequentist estimator. Estimation biases in relation to population parameters were assessed, as well as the variance. Variability was associated with the connectivity characteristics of each simulated population. Clustered populations yield greater variability and no RDS-based strategy could address the estimation biases. Misreporting degrees had modest effects, especially when RDS estimators were used. The best results for RDS-based samples were observed when the "infection" was randomly attributed, without any relation with the underlying network structure.

1. Introduction

Most hard-to-reach populations are marginalized, stigmatized and—depending on mores and laws—may be criminalized. Men who have sex with men (MSM), drug users, migrants belonging to ethnic/linguistic/religious minorities, people living with HIV/AIDS are some examples of these populations. Even when their members are relatively numerous in a given setting (for instance, neighborhoods where migrants from a given ethnicity cluster), it is difficult or rather impossible to use traditional sampling methods to assess them (Johnston et al., 2016; Montealegre et al., 2012a).

Such populations/groups are not easily identifiable, tend to conceal their status to protect them from actual or perceived prejudice and to avoid interactions with institutions and/or people who may be viewed as sources of additional difficulties and stigma (but see Montealegre et al., 2012b; respecting successfully HIV testing strategies for undocumented immigrant in Houston, Texas, USA, despite relevant differential rates according to education, country of origin, etc.).

Low frequencies of a given characteristic behavior and/or geographic dispersal worsens the problem because even if individuals may be candid and prone to reveal their status and habits, a large sample size and complex, costly logistics would be required to find a reasonable number of individuals (Heckathorn, 1997; Salganik and Heckathorn, 2004). Examples of such difficulties (having as a key consequence the violation of basic assumptions of random selection, an essential feature of any unbiased sampling strategy) have been documented by studies targeting rural populations, even in high-income countries (e.g., USA) where good transportation and sound infrastructure partially alleviate such hurdles and caveats (Young et al., 2014).

Currently, one of the most popular sampling technique used to assess hard-to-reach populations is respondent-driven sampling (RDS) (Heckathorn, 1997). Since the late 1990′s, its application have mushroomed and it has already proven to be efficient in finding members of several hard-to-reach populations. The recommendation and adoption of RDS by major agencies such as the Centers for the Disease Control and Prevention (CDC) (Lansky et al., 2007) and the World Health Organization (WHO) (Johnston et al., 2013)

* Corresponding author at: Rua Ferreira de Andrade, 583-202, Rio de Janeiro, RJ 20780-200, Brazil.
E-mail address: ssperandei@gmail.com (S. Sperandei).

http://dx.doi.org/10.1016/j.socnet.2017.05.004
0378-8733/

have fostered its acceptance and widespread use (Salganik and Heckathorn, 2004).

However, although RDS is able to recruit members from a hard-to-reach population, estimates based on RDS studies remain a matter of concern and debate. Clearly, RDS is a chain-referral, non-probabilistic, sampling method, similar to snowballing (Goodman, 1961; Heckathorn, 2011), and prevalence estimates based on RDS data may be biased (Goel and Salganik, 2010). As a chain-referral method, sampling results are intrinsically dependent on the underlying network structure of the population under analysis, as well as on several other factors, such as the differential recruitment of specific subgroups, geographic heterogeneities, structural bottlenecks secondary to violence or lack of transportation, less-than-optimal bridging between different segments etc. (see, for instance, Burt and Thiede, 2014; Rudolph et al., 2015; Toledo et al., 2011).

The assessment of the accuracy and validity of RDS estimates remains a challenge, since it is very difficult (or rather impossible) to know the actual contact network of each individual. Usually, the reported number of contacts is used to weight the individual information when calculating prevalence of a given characteristic or medical condition (Gile et al., 2015; Goel and Salganik, 2010).

Since the actual contact network of each individual is unknown, simulating connected populations seems to be a valid strategy to evaluate assumptions which are key to the method, as well as their putative violations when estimators are based on studies carried out in real-life situations. Some studies have assessed the accuracy and validity of standard estimators using simulated data, profiting from actual information on degree distributions (Goel and Salganik, 2010; McCreesh et al., 2012; Mills et al., 2014; Wejnert, 2009). However, one must be keep in mind that the number of contacts in common between any two individuals is hard to assess or is unknown, and there is little, if any, information about it. Even assuming that information from two individuals about their total number of contacts are precise, the extent such contacts may overlap is usually hard or impossible to estimate in real-life conditions.

Another possible relevant source of estimation error from RDS sampling is due to the dependency between the putative transmission of a given pathogen (or any other transmissible element) and the underlying network structure of the population. For instance, the transmission of some pathogens depend on close and prolonged contact between infected and at-risk individuals (e.g., HIV/AIDS and other sexually transmitted infections/diseases), whereas other conditions are less dependent on the network structure and can be transmitted even if individuals' interaction is incidental, such as in the spread of influenza virus via the shared use of public transportation.

To the best of our knowledge, a single study has addressed the impact of information error about the number of contacts on RDS estimators. Mills et al. (2014) have shown information error may determine relevant estimation biases on RDS studies.

In the present paper, we assessed RDS estimators' performance under varying conditions of network structure, misreporting degrees, and transmission dependency.

2. Material and methods

2.1. Simulated populations

Four different populations (N = 10,000) were simulated, each using a different approach based on different families of random graph models: Erdös-Renyi (ER – Erdös and Rényi, 1959), Watts-Strogatz (WS – Watts and Strogatz, 1998), Barabasi-Albert (BA – Barabasi and Albert, 1999) and Interconnected Islands (II). For the sake of the present study, only static network have been considered. Some information about the connectivity characteristics of each model used is provided as follows:

- Erdös-Renyi (ER): the connection between two individuals is established in a completely random fashion and any two individuals will be connected with a fixed probability. The only parameter needed is the probability (P) of a link between two individuals, set at 0.001;
- Watts-Strogatz (WS): starting from a regular ring lattice, an individual will be linked to a fixed number of neighbors at each side. Here, we set this parameter to five neighbors to each side. Then, each link has a certain probability to be broken and reattached to any other individual in the population, creating "shortcuts" between groups of individuals, which was set at 0.1 in our model. This model is usually known as the *small-word* model;
- Barabasi-Albert (BA): known as the *preferential attachment* model, this model starts with one individual and adds other individuals, one by one. Each entering individual will be preferentially attached to individuals with a higher number of contacts (usually mentioned as a "rich get richer" attachment strategy). The parameter to this model is the number of connections each new member of the population will add when created and was set at five in the simulation;
- Interconected-Islands (II): the original population is initially split into a number of subpopulations (five, in our simulations). Within each subpopulation, the connectivity is determined as in the ER model and a random set of individuals in each subpopulation is chosen connecting individuals from other subpopulations. In our simulation, we set five connecting individuals, which represents a highly clustered population. The third parameter needed is the probability of a random connection between individuals, as in the ER model before and was set at 0.005.

All model parameters were set to obtain a mean degree of 10 connections, irrespectively of the model used.

2.2. Disease transmission process

Each population was challenged by four transmission processes, all of them dependent on the underlying network connections. Different numbers of infection seedings (10, 100, 500, and 1500 seeds) were randomly selected and launched to transmit the condition (to "infect") to their contacts. Following a Susceptible-Infected (SI) model (which does not consider recovery as a plausible outcome), and taking HIV/AIDS as our key example, infection was spread in the population step by step. In each step, an individual connected to an infected contact had a probability of 0.05 to become infected. The infection process follows until a theoretical prevalence of ~15% "infected" individuals, which is defined here as a theoretical ¨ ceiling value". Clearly, the greater the number of infection seedings, the lesser the dependency between infection dynamic and network connections (i.e., infections spread by a huge number of infection seedings could not be distinguished from a simple "mass effect" dissemination process, where the underlying network structure is not taken into consideration). In our simulation, the ceiling value was a 1,500 seeds infectious process, where no relationship between the condition and the underlying network of contacts was made evident, and the infection can be approximately described as "randomly assigned" (data not shown). Information about individual degree was purposely "biased" in several ways, to simulate different types of misreported degree. Besides "no bias", i.e., a hypothetically perfectly accurate degree information, which corresponds to actual population data, we defined the alternatives as follows: "random misreporting", where the degree information was extracted from normally distributed data, with coefficients of variation either equal to 0.2 or 0.6; and, "systematic misreporting",

Please cite this article in press as: Sperandei, S., et al., Assessing respondent-driven sampling: A simulation study across different networks. Soc. Netw. (2017), http://dx.doi.org/10.1016/j.socnet.2017.05.004