Drug and Alcohol Dependence 142 (2014) 120–126

Contents lists available at ScienceDirect

Drug and Alcohol Dependence

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

Errors in reported degrees and respondent driven sampling: Implications for bias[☆]

Harriet L. Mills[a,*], Samuel Johnson[b], Matthew Hickman[c], Nick S. Jones[b], Caroline Colijn[b]

[a] MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Dynamics, Imperial College London, St Mary's Hospital, Norfolk Place, London W2 1PG, UK
[b] Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[c] School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK

ARTICLE INFO

Article history:
Received 11 April 2014
Received in revised form 5 June 2014
Accepted 5 June 2014
Available online 21 June 2014

Keywords:
Respondent driven sampling
At-risk populations
Contact network size

ABSTRACT

Background: Respondent Driven Sampling (RDS) is a network or chain sampling method designed to access individuals from hard-to-reach populations such as people who inject drugs (PWID). RDS surveys are used to monitor behaviour and infection occurence over time; these estimations require adjusting to account for over-sampling of individuals with many contacts. Adjustment is done based on individuals' reported total number of contacts, assuming these are correct.
Methods: Data on the number of contacts (degrees) of individuals sampled in two RDS surveys in Bristol, UK, show larger numbers of individuals reporting numbers of contacts in multiples of 5 and 10 than would be expected at random. To mimic these patterns we generate contact networks and explore different methods of mis-reporting degrees. We simulate RDS surveys and explore the sensitivity of adjusted estimates to these different methods.
Results: We find that inaccurate reporting of degrees can cause large and variable bias in estimates of prevalence or incidence. Our simulations imply that paired RDS surveys could over- or under-estimate any change in prevalence by as much as 25%. These are particularly sensitive to inaccuracies in the degree estimates of individuals with who have low degree.
Conclusions: There is a substantial risk of bias in estimates from RDS if degrees are not correctly reported. This is particularly important when analysing consecutive RDS samples to assess trends in population prevalence and behaviour. RDS questionnaires should be refined to obtain high resolution degree information, particularly from low-degree individuals. Additionally, larger sample sizes can reduce uncertainty in estimates.

1. Introduction

Respondent Driven Sampling (RDS) is a network or chain sampling method designed to access populations of individuals that are "hard-to-reach." For example, people who inject drugs (PWID) or commercial sex workers (CSW) are "hidden populations," without a recognised sampling frame and often unwilling to be identified. RDS is commonly used to deliver health education as well as to sample these populations to understand the spread of disease, the community's behavioural patterns, use of interventions, and individuals' responses to risk (Abdul-Quader et al., 2006; Broadhead et al., 2002, 1998; Des Jarlais et al., 2007; Johnston et al., 2008; Malekinejad et al., 2008; Robinson et al., 2006). RDS works as follows: a number of individuals (*seeds*) are recruited at random from the population. (We note that in reality, seeds are preferentially selected to optimise recruitment and to increase the diversity in the sample.) These individuals are interviewed and given a set number of tokens to recruit their contacts. Successfully recruited contacts are interviewed and given tokens to recruit the next *wave* of individuals. The process continues until either recruitment fails or the target number of recruits is reached. RDS carries the significant advantage that no-one is asked to name contacts directly; participants are invited through their contacts and can choose whether to participate. As such, it is the current method of choice for accessing hard-to-reach populations, not only to deliver public health interventions but to gather data to estimate the prevalence and incidence of infections such as HCV and increasingly HIV (for example, Hope et al., 2010; Iguchi et al., 2009; Sypsa et al., 2014).

[☆] Supplementary material can be found by accessing the online version of this paper. Please see Appendix A for more information.
[*] Corresponding author. Tel.: +44 20 7594 1379.
E-mail address: harriet.l.mills@gmail.com (H.L. Mills).

http://dx.doi.org/10.1016/j.drugalcdep.2014.06.015
0376-8716/© 2014 The Authors. Published by Elsevier Ireland Ltd. All rights reserved.

Accordingly, understanding sources of variability and bias in RDS estimates is increasingly important.

Inevitably, individuals with a high number of contacts will be over-sampled in RDS studies, as these individuals know more people in the target population and therefore are more likely to be recruited. (For those who may doubt the severity of this over-sampling, it can be demonstrated in simulations with minimal assumptions, and is more severe in networks with greater variability in the numbers of contacts; see Supplementary Text S1 and Fig. S1.) In addition, as individuals with high numbers of contacts may be at greater risk of becoming infected (through contact with a larger network of injectors) and also may have a greater infecting risk (e.g., being homeless; Friedman et al., 2000), the prevalence in the sample is expected to be higher than the prevalence in the at-risk community. It is therefore necessary to adjust for this bias when estimating an infection's prevalence or incidence using RDS data (Gile and Handcock, 2010; Goel and Salganik, 2010; Heckathorn, 2007; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008). The estimate $\hat{\mu}$ is [40]:

$$\hat{\mu} = \frac{\sum_{i=1}^{n}(f_i/d_i)}{\sum_{i=1}^{n}(1/d_i)} \quad (1)$$

where n is the sample size, f_i is the trait (e.g., $f_i = 1$ if the individual is infected and 0 if not) and d_i is the estimated number of contacts, or degree, of individual i (see Supplementary Text S2). Naturally, if infection were not correlated with degree, then this adjustment would not have any effect on the estimate.

An individual's degree is generally their own estimate of the number of other individuals they know by name that they have seen in a set time period, who also belong to the population being sampled (e.g., who are also PWID or CSW or other target population). This number is therefore an estimate of the number of individuals they may recruit, and also of the number of contacts relevant for the transmission of disease. However, degree may be difficult to estimate accurately as well as being dynamic in time (Brewer, 2000; Rudolph et al., 2013). Individuals may only roughly know their degree, may only recall or count close contacts or may intentionally give an inaccurate estimate, for example to hide how at risk they are or to boost their apparent popularity (desirability bias; Fisher, 1993). Degree bias or digital preference is particularly relevant in the reporting of sexual or drug use behaviours, where individuals may be uncertain or wish to avoid association with illegal or undesirable activities (Fenton et al., 2001; Schroder et al., 2003). One of the assumptions underpinning RDS and the adjustment methods is that respondents accurately report their degree. As noted by several authors, inaccuracy in degree constitutes a source of sampling bias in the adjustment procedure (Goel and Salganik, 2009; Johnston et al., 2008; Rudolph et al., 2013; Salganik and Heckathorn, 2004; Wejnert, 2009), yet to the best of our knowledge there has been no study examining the extent to which this might be important in the interpretation of RDS surveys.

There have been several other concerns about the extent to which real RDS studies match the idealised assumptions underlying the statistical estimators. Heckathorn showed that under ideal conditions, RDS samples are Markov chains whose stationary distribution is independent of the choice of seeds (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004). However, there have been concerns that preferential referral behaviour of respondents (Bengtsson and Thorson, 2010), short recruitment chains compared to the length needed for the Markov chain to reach equilibrium, and the difference between with-replacement random walk models and without-replacement real-world samples could lead to bias in RDS estimates (Gile and Handcock, 2010).

Here, we explore how reported degree data might arise from a true underlying distribution due to individuals rounding their numbers of contacts up or down to multiples of 5, 10 and 100. We use simulations of RDS to investigate the potential bias caused by inaccurately reporting degrees and compare it to other issues researchers have raised about RDS (including the difference between with- and without- replacement sampling, multiple seed individuals and multiple recruits per individual).

2. Methods

2.1. Data

We base our methodological work on two cross-sectional RDS studies of PWID in Bristol, UK, in 2006 ($n = 299$) and 2009 ($n = 292$), described elsewhere (Hickman et al., 2009; Hope et al., 2011, 2013; Mills et al., 2012). They used the same questionnaire and recruited individuals who injected in the last 4 weeks. The results were used to estimate trends in HCV prevalence and incidence in this population. We analyse the reported contact numbers (degrees) from both surveys.

2.2. RDS simulations

We generate contact networks of individuals with a defined degree distribution using the configuration model (Newman, 2003). The contact number distribution in the Bristol data is approximately long-tailed in that reported numbers vary by several orders of magnitude, so we used a long-tailed degree distribution (power law with an exponential cut off, mean degree of 10) in the simulations. We simulate the transmission of a pathogen (SIS) across the network and after a set time we simulate an RDS survey. Details of the network and transmission model are in the Supplementary Text For comparison we present results for a network with a Poisson degree distribution, where there is much less variation in degrees (Supplementary Text S3).

We determine the impact of inaccurate degrees on the prevalence estimate by re-computing the estimate in Eq. (1) using $d_i = \hat{d}_i + \Delta d_i$, where $\hat{d}_i$ are the individuals' correct degrees in the network, and Δd_i correspond to inaccuracies in these degrees. We consider five different rounding schemes to mimic patterns seen in data: (1) round all degrees up to the nearest 5, (2) round all degrees up to the nearest 10, (3) increase every degree by 5, and finally two methods to directly mimic patterns seen in the Bristol data (Fig. 1). These are (4) round all degrees between 10 and 100 to the nearest 10, and degrees greater than 100 to the nearest 100; and (5) similar, but individuals with degrees less than 10 are given a different degree between 1 and 10, chosen according to the distribution seen in the Bristol data.

We simulate a number of variations of RDS. First, we take a *standard "real world" RDS* sample: individuals recruit a number of their contacts to the sample, where this number is chosen from a Poisson distribution, mean 1.5 and limited to between [0,3] (and cannot be larger than their total number of contacts). Individuals cannot be sampled more than once. We compare this to idealised RDS, or Markov process RDS: there are *multiple seeds*, seeds recruit one individual only at random from their contacts and sampling is with replacement. We also use variants of this method, allowing *multiple tokens* (recruits), and *without replacement*. In all of our variants, seeds are chosen at random.

We simulate samples of size approximately 350 for each of these RDS variants, in a population of 10,000 individuals. We calculate the percentage difference between the prevalence estimates (both raw and using the Volz–Heckathorn estimator (Volz and Heckathorn, 2008)) and the actual population prevalence to determine which assumptions most impact error in RDS. We take two RDS surveys separated by two years, over a time when prevalence is increasing (from about 20% to 30%, see Fig. S4) and determine how accurately consecutive samples can identify changes in prevalence. We compare the true simulated population prevalence (prevalence in the modelled population) to the raw RDS sample prevalence and the prevalence after adjustment with the Volz–Heckathorn estimator.

3. Results

3.1. Reported contact numbers

Data describing the reported degrees in the Bristol surveys illustrate a pronounced preference of individuals to report their numbers of contacts to the nearest 10, 20, 30... and 100, 200, 300 (Fig. 1). However, it is likely that the true distribution of the numbers of relevant contacts has nearly as many 21s as 20s, nearly as many 31s and 30s and so on. The reported degree distribution is highly unlikely.

Since we only have the reported degrees, we cannot know what the true distribution is nor the details of how individuals modify this information. However, if we can generate degrees with a smooth distribution and show that, by applying a given rounding