Errors in reported degrees and respondent driven sampling: Implications for bias
Mills, Harriet L; Johnson, Samuel; Hickman, Matthew; Jones, Nick S; Colijn, Caroline

DOI: 10.1016/j.drugalcdep.2014.06.015
License: Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
https://doi.org/10.1016/j.drugalcdep.2014.06.015

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

• Users may freely distribute the URL that is used to identify this publication.
• Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
• Users may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
• Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.
Errors in reported degrees and respondent driven sampling: Implications for bias

Harriet L. Mills\textsuperscript{a,}\ast, Samuel Johnson\textsuperscript{b}, Matthew Hickman\textsuperscript{c}, Nick S. Jones\textsuperscript{b}, Caroline Colijn\textsuperscript{b}

\textsuperscript{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
\textsuperscript{b} Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
\textsuperscript{c} School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK

\begin{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 occurrence 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.

\end{abstract}

© 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]:

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

(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 \( d_i = \bar{d}_i + \Delta d_i \), where \( \bar{d}_i \) are the individuals’ correct degrees in the network, and \( \Delta d_i \) 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 estimates 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
scheme, the resulting modified distribution resembles the Bristol data, we have some justification both for the choice of original distribution and the rounding scheme in question. With this objective, in Supplementary Text S4 we define a simple measure of distance between distributions. It is not immediately obvious how close two distributions should be to be considered similar. But fortunately we have two sets of reported degrees for Bristol at different times; we can therefore use the distance between the two, which we call $z_B$, as a yardstick for resemblance.

We generated long-tailed degree distributions using the power law with exponential cut-off described in Section 2, and found that the average distance to the empirical distributions was about 5.2 times $z_B$. We then applied each of the rounding schemes described in Section 2, Scheme 1 (rounding all degrees up by 5) and Scheme 2 (by 10), reduced the factor from 5.2 to 2.7 and 2.3, respectively. Scheme 3 (adding 5 to every degree) increased the distance somewhat to 5.5. However, the more sophisticated Scheme 4 (rounding to the nearest 10 for $k<100$ and to the nearest 100 for $k>100$) reduced the factor to 1.4; while Scheme 5, which is like Scheme 4 but also draws all degrees under 10 from the combined Bristol distributions, decreases this factor further still to 1.2, Table S3. In other words, these schemes produce distributions almost as close to the empirical ones as the two Bristol datasets are to each other. Note, however, that the level of interference involved in Schemes 4 and 5 should be seen as the minimum reporting error required to obtain realistic reported distributions from smooth underlying ones. If in fact it were the individuals with few contacts who nonetheless claimed to have hundreds while the highly connected reported only a small number, this would not be evident in the data. The bias introduced by inaccuracies in reported degrees which we go on to analyse should therefore be regarded as a lower bound to the potential importance of this effect.

3.2. RDS simulations

Inaccuracy in reported degrees had a large effect on the reliability of estimates of prevalence and incidence (Fig. 2). The top half of Fig. 2 shows estimates of prevalence from RDS surveys where degree was mis-reported by the 5 rounding schemes. The estimates were calculated using the Volz–Heckathorn estimator. Mis-reporting degrees caused all surveys to over-estimate prevalence (compare to the ‘Actual’ prevalence in the whole network, top). However, if degrees were correctly reported (standard RDS) the average prevalence estimate from 100 surveys was accurate, but individual variation was large. Even with inaccurate degrees, the adjusted estimates (blue bars) were still closer to the true prevalence or incidence than the point estimate from the raw data (green bars).

Two of our degree-biasing rounding schemes were based on degrees collected in Bristol, UK. Scheme 4 adjusted only those degrees larger than 10: the prevalence estimate is comparable with the estimate using correct degrees. However, the error increased when inaccuracies were added to the lower degrees ($1 \leq d \leq 10$) in Scheme 5. Those with low degree have a higher weighting in the estimator (Eq. (1)) than those with high degree; therefore mis-reporting these degrees had a larger effect on the estimate. The average prevalence for rounding Scheme 5 was 39.8% [31.1–51.4% 95% CI] compared to the actual average prevalence of 27.2% [26.1–28.4% 95% CI]. If only the high degrees were mis-reported the average estimated prevalence is closer to the actual prevalence,
however separate simulations have a large variation in estimates (average 27.1% [17.0–38.6% 95% CI]). As low-degree individuals are far less likely to be infected than high-degree individuals, error in their degrees affects the denominator in the estimator without a comparable effect on the numerator (recall Eq. (1)).

The extent of bias due to mis-reporting degrees also depends on the contact network in the population being studied. Networks with a long-tailed degree distribution have a strong correlation between degree and infection status and we showed that inaccurate degrees can lead to substantial bias in the estimation of sero-prevalence. However, in a network with a lower variance in degrees, such as a Poisson degree distribution, the correlation was much weaker and there was a correspondingly weaker effect on prevalence estimates when degrees were inaccurate (see Fig. S6).

Other differences between real RDS surveys and the idealised circumstances from which the mathematics of RDS is derived did not cause any substantial error in the estimates of incidence or prevalence in our simulations (lower half of Fig. 2). The fact that real RDS surveys use sampling without replacement, multiple original seeds rather than just one, and that individuals can recruit more than one contact into the study can be expected to lead to relatively minor differences between the theoretically unbiased estimates and estimates from data.

### 3.3. Consecutive RDS surveys

While nearly all of the simulated consecutive samples correctly identified that prevalence increased between the two time points, there were marked differences in the estimates of the sampled trend (Fig. 3 and Table S5). The true simulated prevalence on average increased from 19% to 30% in the two year gap. The raw sample data overestimated prevalence, but identified the trend correctly. When the sample was adjusted using the Volz–Heckathorn estimator with accurate degrees the average difference across all

![Fig. 2. Boxplots indicating the prevalence estimates from simulated RDS surveys of 100 populations on networks with long-tailed degree distributions. The adjusted prevalence estimates (black) are always better than the raw, unadjusted data (grey), but incorrect degrees can cause significant inaccuracies in the estimated prevalence. The true prevalence in the simulated populations is indicated by the dashed line.](image)

![Fig. 3. Box plots illustrating the extent of the over-estimates of the increase in prevalence between two consecutive RDS surveys of the same population. Consecutive samples were taken in 100 different populations, the actual prevalence increased from 19% to 30% in the 2 year gap; this difference of 11% is indicated by the dashed line across the plot. The scenarios are fully described in the text.](image)
100 repeats was very close to the actual increase (3rd boxplot from the top in Fig. 3). However, the variation between individual paired samples was very large, indicating large inaccuracies in individual runs. As repeated samples are impractical in reality, this implies that conclusions from consecutive studies have a high probability of being quite inaccurate, even if degrees are correctly given. This is the case for all of the rounding methods we compared.

All scenarios with mis-reported degrees had a large variance in the estimated trend in prevalence. When degrees were rounded up to 5, to 10 or increased by 5, the prevalence and the increase in prevalence between the two surveys was over-estimated. When only degrees larger than 10 were adjusted (rounding Scheme 4) the estimator returned very similar trends to those seen when adjusting with accurate degrees, indicating that the method did not remove enough of the information in degree to alter the effect of the weightings. Rounding Scheme 5, however, which added inaccuracy to those with degrees <10, showed a large average overestimate and variation in results. This indicated that it is particularly important to obtain correct degrees for low degree individuals as even small inaccuracies can have a large impact on results. The same simulation on networks with a Poisson degree distribution (and therefore a lower variance in degrees) showed a lower average over-estimate but still a large variation in results, Fig. S7.

4. Discussion

There is a clear indication in the reported degrees of the Bristol data that individuals round or bin their number of contacts to the nearest 5, 10 and 100. Indeed, these empirical distributions were part of the motivation for this work; high frequencies of degrees that were multiples of 5, 10, etc. suggest that individuals may be guessing or rounding their reported degree. We analysed the effect of rounding schemes on the degree distribution and showed that schemes which round degrees to the nearest order of magnitude result in degrees with a distribution close to that seen in the Bristol data.

It is well-known that the Volz–Heckathon adjustment reliably recovers prevalence and incidence estimates in the presence of over-sampling of high-degree individuals, in contrast to raw RDS data. However, we have found that the necessity of weighting individuals’ contributions by their reported degree can lead to significant bias if degrees are inaccurately reported. This source of bias is very likely greater than inaccuracies resulting from other variations in RDS (e.g., with- or without-replacement sampling, multiple or single recruitment).

Our results highlight the importance of obtaining correct degrees for accurate analysis of RDS surveys. This has been described previously, but the extent of the effect of inaccurate degrees, particularly on serial estimates using RDS, has not been determined. If inaccurate degrees can be introduced into RDS surveys, and confidence in the estimates will be low, causing uncertainty in the calculated trends from paired samples.

We note that our methods of adding inaccuracy to degrees are fairly conservative: it is likely that realistic biasing behaviour is heterogeneous across a population and may depend on factors like gender, age, behaviour, degree or disease status (Bell et al., 2007; Brewer, 2000; Marsden, 1990; Rudolph et al., 2013). For example, men usually report a far higher number of sexual partners than women, giving inconsistency in the number of sexual partnerships that could have occurred (Brown and Sinclair, 1999; Liljeros et al., 2001; Smith, 1992). Similar problems may occur among PWIDs recalling injecting partnerships. In addition, PWIDs in different countries or regions where different laws and restrictions apply may bias their answers differently. These more systematic inaccuracies will likely cause a larger error in estimates, enhanced by correlations between those factors and infection. Testing the accuracy of reported degrees would be very challenging in the “hidden populations” in which RDS is used. Because of these complications, we are uncertain as to whether developing a general method to recover accurate degrees is realistic.

Most RDS questionnaires ask a variety of questions about degree, such as how many PWIDs they know by name and have seen in the last X days, how many PWIDs have they injected with in the last Y weeks and how many of these were new partners or regular partners. A recent study asking multiple questions about degree determined that the first of these obtained the most accurate answers, though mis-reporting was common (Wejnert, 2009). However, it may be possible to alter the questions to gain a more accurate understanding of an individual’s risk (Rudolph et al., 2013), or to use some combination of answers to determine an alternative weighting for use with the Volz–Heckathorn estimator (Lu, 2013). Alternatively, self completion of the contact portion of the questionnaire may improve accuracy of answers (Schroder et al., 2003). Our simulations show that it is most crucial to obtain accurate reports of degree for low-degree individuals. If questioned in detail, this group may be more likely to remember contacts more accurately and may better be able to answer questions about
contact numbers, times and type of contact than individuals with dozens of contacts.

Not surprisingly, our simulations indicated that the variation in results decreases if the sample size is increased (see Figs. S5 and S6, also shown in Goel and Salganik, 2009; Mills et al., 2012). Additionally, taking multiple surveys of the same population can improve the estimate. However, multiple surveys are generally not practical. If instead a larger survey were taken, the error in estimates could be reduced by using a bootstrapping method, as described by Salganik (2006). Researchers would estimate \( \hat{\mu} \) many times, each time using a subset of the (larger) RDS sample. The resulting distribution of estimates would be used to construct confidence intervals, for example, and ultimately \( p \)-values for any estimated change in prevalence, incidence or other estimate.

We have shown that inaccuracy in degree can reduce the accuracy of prevalence estimates from RDS surveys and decrease the ability to identify accurately the magnitude of prevalence trends in the underlying population. We recognise that RDS is an extremely useful method to quickly access hidden populations such as PWIDs, but we urge users to consider results cautiously and to make every effort to estimate degrees carefully, particularly those of low-degree individuals, and particularly when comparing surveys.

**Role of funding source**

H.L.M. would like to acknowledge funding from Wellcome Trust University Award 093488/2/10/Z. S.J. is grateful for financial support from the European Commission under the Marie Curie Intra-European Fellowship Programme PIEF-GA-2010-276454. M.H. would like to acknowledge funding from NINQAD MRC grant G1000021 and National Institute for Health Research (NIHR)’s School for Public Health Research (SPHR). N.S.J. acknowledges support from EPSRC grant EP/I005765/1. C.C. would like to acknowledge EPSRC grant EP/I031626/1 and EP/K028003/1. The funding sources had no role in the study design or analysis.

**Contributors**

H.L.M., S.J., C.C., N.S.J. designed the study. H.L.M. ran the model and statistical analysis. C.C., N.S.J., S.J. and M.H. advised on the analysis. M.H. provided the datasets. H.L.M. and S.J. analysed the datasets. H.L.M. wrote the first draft of the manuscript. All authors contributed to and have approved the final manuscript.

**Conflict of interest**

None declared.

**Appendix A. Supplementary data**

Supplementary data associated with this article can be found, in the online version, at [http://dx.doi.org/10.1016/j.drugalcdep.2014.06.015](http://dx.doi.org/10.1016/j.drugalcdep.2014.06.015).

**References**


