Supplementary Material: Errors in Reported Degrees and Respondent Driven Sampling: Implications for Bias

Harriet L Mills [1], Samuel Johnson[2], Matthew Hickman[3], Nick S Jones[2], and Caroline Colijn[2]

[1]*MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Dynamics, Imperial College London, UK*
[2]*Department of Mathematics, Imperial College London, UK*
[3]*School of Social and Community Medicine, University of Bristol, Bristol, UK*

This material supplements but does not replace the content of the peer-reviewed paper published in *Drug and Alcohol Dependence*.

Supplementary Text

S1 Oversampling high degree individuals in RDS

Individuals with many contacts are more likely to be recruited during RDS than those with few contacts; we demonstrate the extent of this oversampling using simulations. We ran simulations of RDS on networks of 10,000 individuals (for networks with Poisson and long tailed degree distribution, mean of 10). We compared the degree distribution of the population to the degree distribution of the sampled individuals, to see if there are any significant differences.

We plotted the degree against the probability of being selected to determine if high degree nodes have a higher probability of selection. In every simulation there was oversampling of high degree individuals, Fig. S1. The mean degree was higher in the simulations than in the underlying population (increasing from 6 to about 7.45), the variance in degree was slightly higher in the simulations than in the underlying population, Table S1.

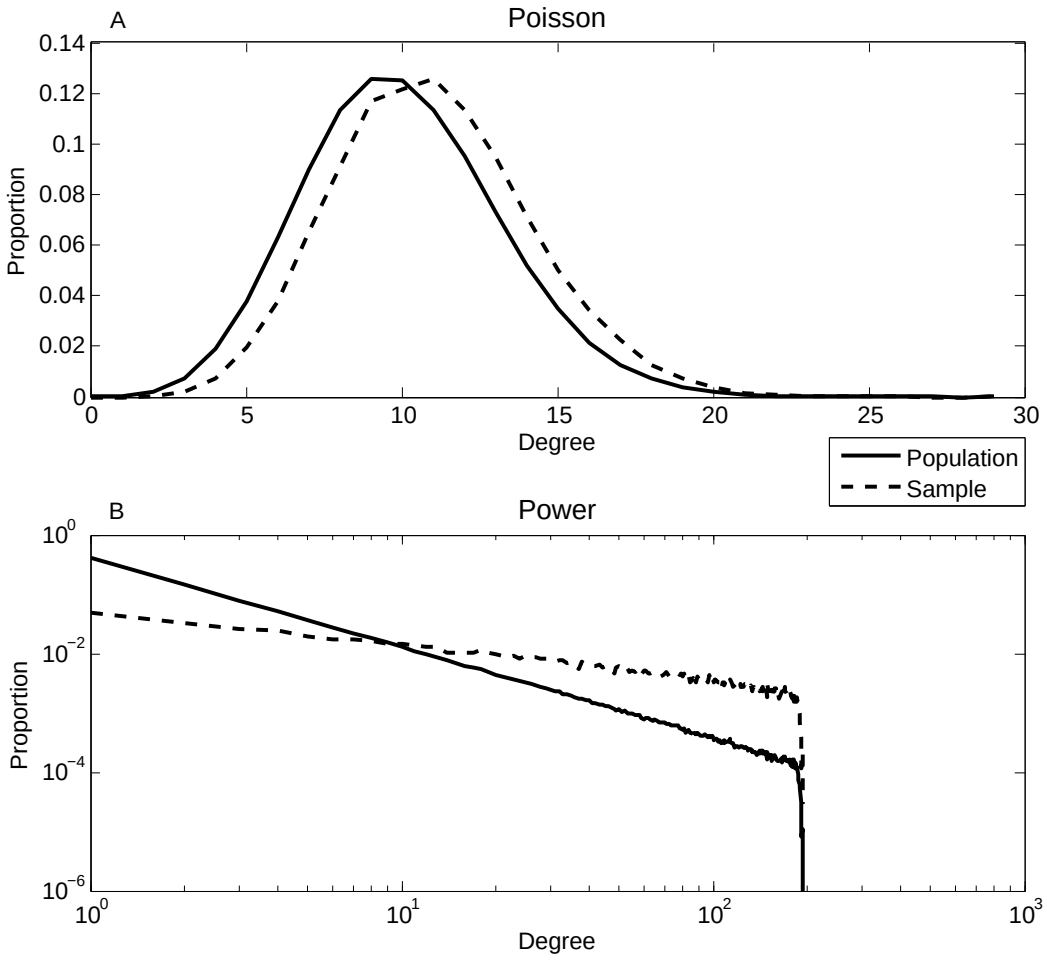

Figure S1: Oversampling of high degree individuals in RDS simulations. RDS samples were carried out on each of 100 networks of 10,000 individuals for both Poisson (A) and long tailed (B) degree distribution. Note scales on (B) are logarithmic. Oversampling of high degree individuals occurs in both networks.

Property	Poisson		Long tailed	
	Popn	**Sample**	**Popn**	**Sample**
Mean	9.99	10.97	10.03	57.72
Variance	9.98	10.04	537.00	2801.10
Minimum	0	2	1	1
Maximum	29	29	195	195

Table S1: The mean of the degree of individuals in the RDS simulations. RDS samples were carried out on each of 100 networks of 10,000 individuals, with Poisson or long tailed degree distributions (mean degree 10). Oversampling of high degree individuals occurs in both networks, but particularly so in the long tailed network.

S2 The Volz-Heckathorn estimator

The most commonly used estimator to calculate the population mean of a property using an RDS sample uses probability-theoretic methods (defined by Volz and Heckathorn (Volz and Heckathorn, 2008) and again by Goel and Salganik (Goel and Salganik, 2010)). This estimator adjusts by the given degree of the individuals sampled (the number of contacts), to remove the bias caused by the over-representation of high degree individuals. The estimator is derived as follows (Volz and Heckathorn, 2008):

A chain referral sample such as RDS is analogous to a random walk on a network; this is a Markov Process (MP) which at equilibrium occupies a node with probability proportional to degree (Salganik and Heckathorn, 2004). Therefore, a chain referral sampling procedure such as RDS will recruit individuals with probability proportional to their degree. As the random walks (RDS samples) are finite and irreducible, there is a unique equation for the MP and the MP will converge to it.

Therefore, for a network where node i has degree d_i the probability of node i choosing node j is $\sigma_{ij} = 1/d_i$. Then, the vector $x*_i = d_i / \sum_j d_j$ is an unique equilibrium for the MP.

To define their RDS estimator, Volz and Heckathorn use a Hansen-Hurwitz (HH) type estimator (Hansen and Hurwitz, 1943) with selection probabilities $p_i = d_i/N\bar{d}$, where d_u is the average degree of the population. The selection probabilities are estimated by $\hat{p}_i = d_i/N\bar{d}$, where $\bar{d}$ is an estimate of the average degree of the population and N is the whole population size.

Now by (Salganik and Heckathorn, 2004), $\bar{d}$ can be estimated using a ratio estimator of HH estimators.

$$\bar{d} = \frac{\sum_{i=1}^{n} \frac{d_i}{np_i}}{\sum_{i=1}^{n} \frac{1}{np_i}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{S1}$$

where n is the sample size. Let f_i be the variable of interest, for example age. Then let T_f be the total value of f in the population (so $T_f = \sum_{i=1}^{N} f_i$). Then the HH estimator of T_f is

$$\begin{aligned} \hat{T}_f &= \frac{1}{n} \sum_{i=1}^{n} \frac{f_i}{\hat{p}_i} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\bar{d} N f_i}{d_i} \\ &= \frac{\bar{d}N}{n} \sum_{i=1}^{n} \frac{f_i}{d_i}. \end{aligned} \tag{S2}$$

If N is unknown (which is common), then the estimate of the average f in the population (e.g. the average age) is

$$< \hat{f} > = \frac{\bar{d}}{n} \sum_{i=1}^{n} \frac{f_i}{d_i}$$
$$= \frac{\sum_{i=1}^{n} \frac{f_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{S3}$$

using Eqn. (S1).

If we wish to estimate the proportion of a population of type A, for example the proportion female, then we let f_i be an indicator function such that $f_i = I_A(i)$ which equals 1 if node $i \in A$ and 0 else. Then,

$$\hat{P}_A = \frac{\sum_{i \in A} \frac{1}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}}. \tag{S4}$$

Volz and Heckathorn showed that their estimator was accurate (had low bias compared to previous estimators), but acknowledge that it is still assuming with replacement and one recruit per respondent and is more accurate for larger sample sizes. These formulas are commonly used to estimate population properties from an RDS sample, using RDSAT, a freely available RDS analysis tool (Volz et al., 2007).

S3 Network and transmission model

We model a population of size $n = 10,000$. We model networks with Poisson and long-tailed degree distributions. The long tailed degree distribution is a power law with an exponential cut off (power $\alpha = 1.5$ and the cut-off is $\tau = 1000$). Both distributions have a mean degree of 10. We connect the network using the configuration model: nodes (individuals) are each assigned a degree chosen from a defined degree distribution. Each node then has that number of "stubs". We choose two stubs at random from a list of all stubs and connect them with an edge (with the condition to exclude double edges and self-loops).

The transmission model structure is susceptible-infected-susceptible (SIS); we do not parametrise the model for any particular pathogen, although this is a suitable framework for Hepatitis C, where recovery may occur (spontaneously or through treatment) but individuals may be re-infected. The prevalences we consider (up to 40%) are realistic in many PWID populations. The model is initialised with 1% of the population infected. Each time-step a node i will become infected with probability equal to $p = 1 - (1 - \alpha)^{k_i}$ where α is the probability of transmission, per infected contact, per time-step and k_i is the number of infected contacts of node i. Infected nodes may recover (to the susceptible state) with a fixed probability r each time step. Nodes may die in all states, with an increased probability if infected. Nodes which die are replaced one time-step later by a susceptible node (which retains all contacts of the dead node, for simplicity). We chose parameters such that on both networks the stable equilibrium was around 30% prevalence, due to the long tails in the power law distribution this stable prevalence was reached within around 5 years, on the Poisson distributed network this took around 10 years. The time scales didn't matter to this investigation, as long as the RDS samples were taken at around the same prevalences on both networks (and the consecutive samples were taken over a time when prevalence was still increasing).

S4 Analysing reported contact numbers in the Bristol datasets

We define the distance z between two degree distributions f and h as

$$z = \left\langle \frac{|f_k - h_k|}{f_k + h_k} \right\rangle, \tag{S5}$$

where the average $\langle\rangle$ is over the range of k (the term $|f_k - h_k|/(f_k + h_k)$ is interpreted as zero if $f_k = h_k = 0$). The distance between the 2006 and 2009 Bristol distributions is $z_B = 0.0214$ (assuming a range of $0 \leq k < 1000$).

The contact number distribution in the Bristol data is approximately long-tailed (Fig. 1 in the main text) 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) to represent it.

For each rounding scheme, we draw samples from the long-tailed distribution, and obtain the distance z as defined above. The values of z in themselves are not easily interpretable, but when compared to z_B they provide measures of the distance to the empirical distributions in terms of the distance between each of these.

The rounding schemes we apply are shown in Table S2, each with a label then used in the subsequent table and figure.

Rounding label	**Description**
0	As drawn from candidate distribution (no rounding)
1	Rounded up to nearest 5, $\forall k$
2	Rounded up to nearest 10, $\forall k$
3	Increased by 5, $\forall k$
4	Rounded to nearest 10 if $k \in [10, 100]$, or to nearest 100 if $k > 100$
5	From Bristol ('06+'09) if $k < 10$, else as in scheme 4 (above)

Table S2: Description of each rounding scheme.

Table S3 lists the average relative z-distance, $\overline{z} = (z_{06} + z_{09})/(2z_B)$, for each rounding scheme (where z_X is the mean distance to the Bristol data of year X), along with standard deviations for averages over 10^4 sampled distributions. Fig. S2 displays the information in the tables visually.

Rounding label	$\overline{z}$	**SD(z)**
0	5.173868	0.288580
1	2.702984	0.198006
2	2.347472	0.157871
3	5.471744	0.292656
4	1.429304	0.051137
5	1.206379	0.060924

Table S3: Mean relative distances, $\overline{z}$, to the Bristol distributions of each rounding scheme (as described in main text) applied to samples drawn from a long-tailed sitribution (power law with an exponential cut-off, described in Section S3); and standard deviations for averages over 10^4 samples.

The distance results show that adding 5 to every degree is worse than drawing directly from the fitted distribution; rounding all degrees up improves upon no rounding; and the more

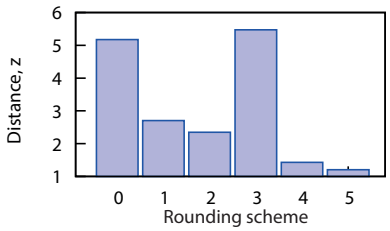

Figure S2: Mean distances $\overline{z}$ for each rounding scheme, from Table S3.

sophisticated schemes 4 and 5 yield distributions closest to the empirical ones. In other words, artificial distributions generated by schemes 4 and 5 are as close to the empirical ones as the two empirical ones are to each other. This would seem to justify a posteriori this choice of underlying distribution, as well as the hypothesis that people do indeed tend to report their perceived degrees in this way.

Supplementary Figures

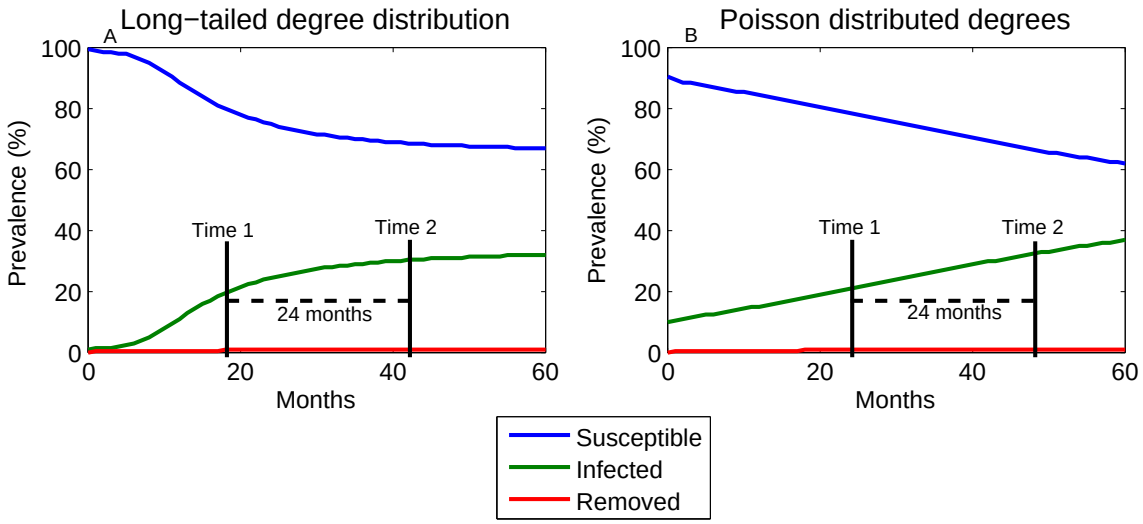

Figure S3: The prevalence in the modelled epidemics. The black lines indicate the times of the surveys, consecutive surveys were 24 months (2 years) apart. (A) Long-tailed distributed degrees. (B) Poisson distributed degrees.

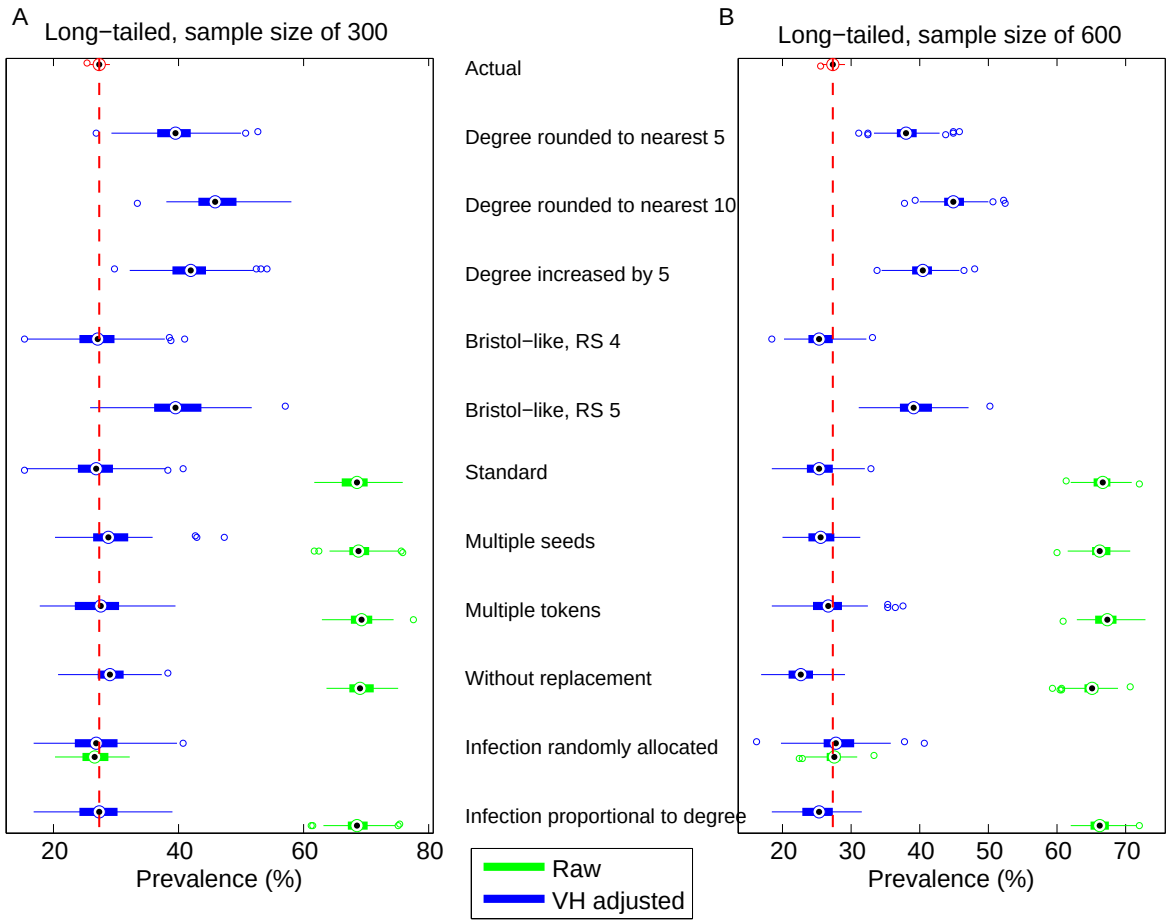

Figure S4: The estimates from RDS, for networks with a long-tailed degree distribution. The red dashed line indicates the actual median prevalence in the simulated populations. (A) Sample size 300, (B) Sample size 600. The Volz-Heckathorn estimate of prevalence is always more accurate than the raw data, however inaccuracies in the degrees can increase the error. The variance in both raw and adjusted data is large with both sample sizes, however the estimate is more accurate when the sample size is larger.

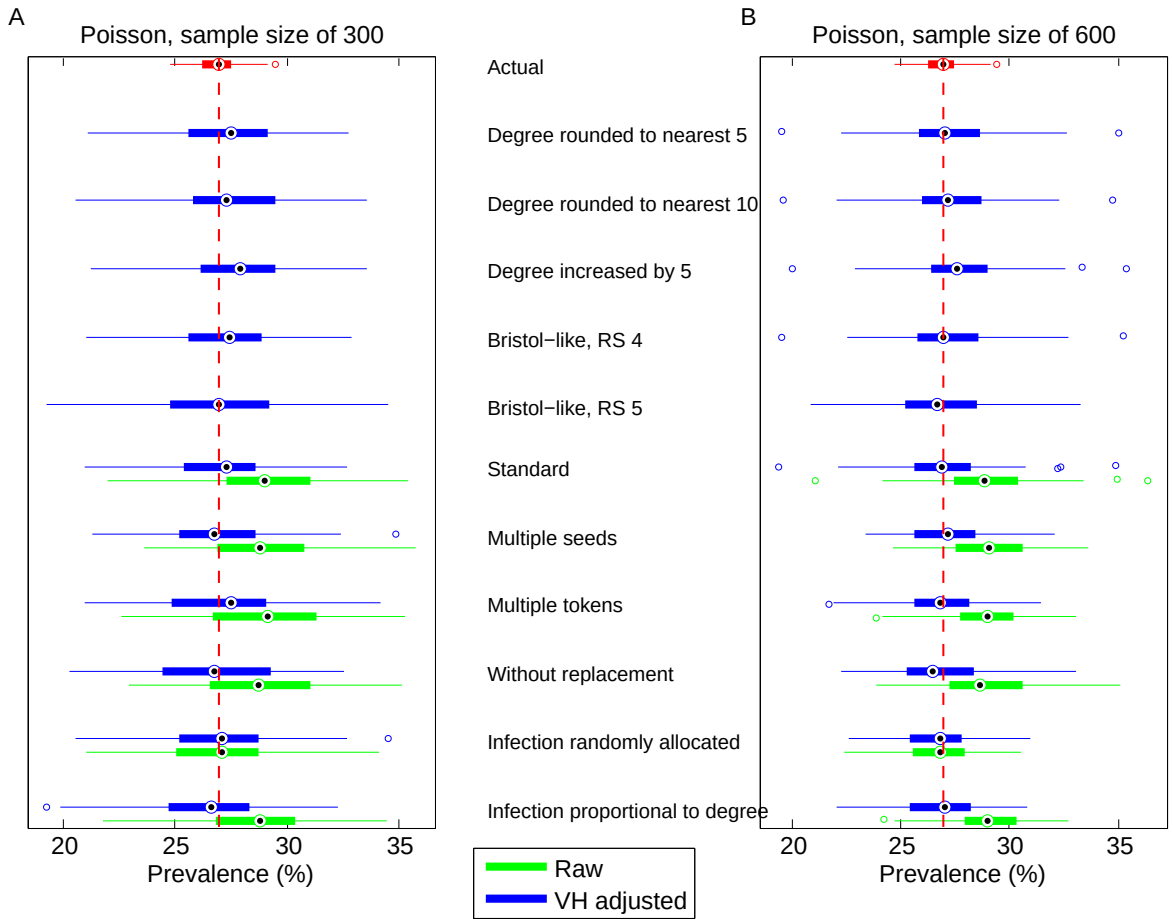

Figure S5: The estimates from RDS, for networks with a Poisson degree distribution. The red dashed line indicates the actual median prevalence in the simulated populations. (A) Sample size 300, (B) Sample size 600. The Volz-Heckathorn estimate of prevalence is always more accurate than the raw data, however inaccuracies in the degrees can increase the error. Generally, RDS estimates from the Poisson network are more accurate than from the long tailed network, see Figure S4. The Poisson degree distribution has a smaller range of degrees and consequently a lower correlation between degree and infection than the long tailed degree distribution, this leads to less biased RDS estimates from these networks compared to the long-tailed network. As with the long-tailed degree distribution, a larger sample size results in lower errors.

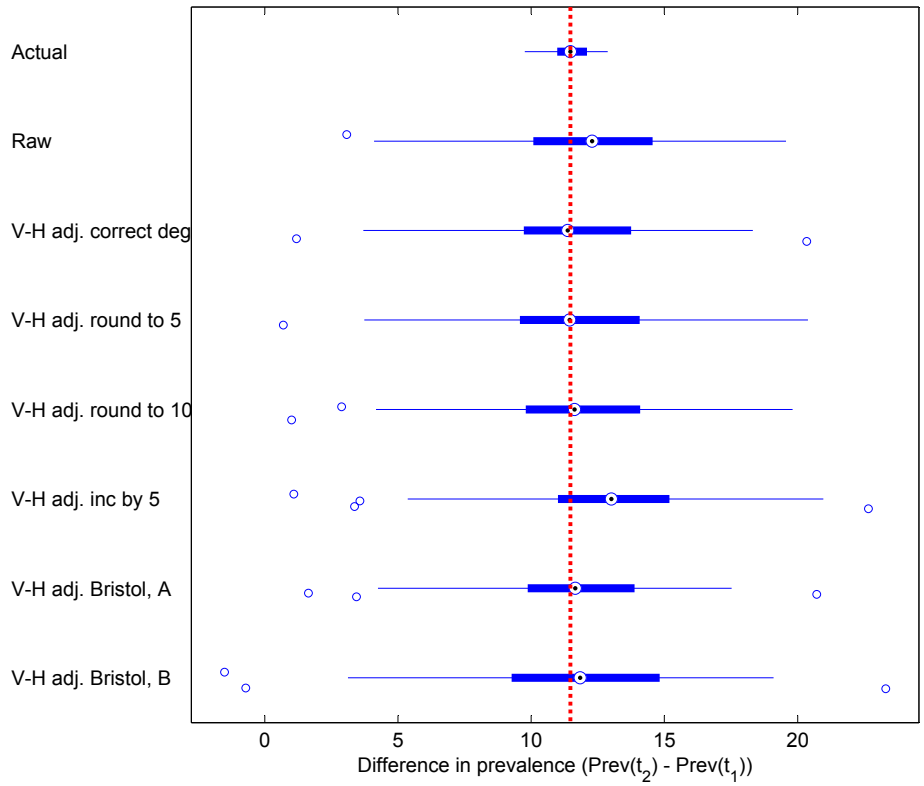

Figure S6: Box plots illustrating the extent of the errors in estimates of the increase in prevalence between two consecutive RDS surveys of the same population, in a network with a Poisson degree distribution. Consecutive samples were taken in 100 populations, the actual prevalence increased from 21% to 32% in the 2 year gap: this difference of 11% is indicated by the dashed line across the plot. The methods are fully described in the main text. Generally samples from a network with a Poisson degree distribution more accurately estimate the prevalence than those from a network with a long-tailed distribution (Figure 3 in the main text), however the variation and range in estimates is very large, as in the long-tailed simulations.

Supplementary Tables

	Mean prevalence at time 1 (%)	Mean prevalence at time 2 (%)	Mean difference [95% CI]	Proportion of samples which underestimate the increase	Mean percentage error in the underestimate [95% CI]	Proportion of samples which overestimate the increase	Mean percentage error in the overestimate [95% CI]
Actual	19.2	30.1	10.9 [9.5 – 12.6]				
Raw	61.2	69.8	8.5 [1.6 – 15.6]	0.68	39.61 [2.41 – 84.43]	0.31	17.73 [2.08 – 36.49]
V-H adjusted, real degrees	18.0	29.4	11.4 [2.4 – 20.8]	0.48	34.68 [2.31 – 83.28]	0.51	44.69 [3.15 – 96.88]
V-H adjusted, degree rounded to nearest 5	28.5	42.0	13.5 [5.5 – 22.6]	0.34	26.49 [1.50 – 60.10]	0.66	49.88 [5.55 – 120.92]
V-H adjusted, degree rounded to nearest 10	34.9	48.8	13.9 [6.2 – 22.5]	0.29	26.67 [0.48 – 70.66]	0.71	49.99 [6.28 – 107.82]
V-H adjusted, degree increased by 5	31.8	44.4	12.6 [3.6 – 21.7]	0.29	38.96 [1.97 – 76.13]	0.70	39.71 [3.60 – 93.22]
V-H adjusted, degree rounded as Bristol data, 1	18.2	29.6	11.4 [2.5 – 20.7]	0.48	34.84 [1.72 – 82.83]	0.51	44.88 [4.66 – 95.89]
V-H adjusted, degree rounded as Bristol data, 2	28.9	43.1	14.3 [2.1–24.2]	0.25	47.59 [1.82– 86.37]	0.75	56.49 [3.65– 114.07]

Table S4: *Long tailed degree distribution: the results of 100 consecutive (paired) RDS samples for a population with increasing prevalence.* Adding degree inaccuracy shifts the proportion of samples which over- and under-estimate so that more samples over-estimate the increase than when accurate degrees were used. The range of over-estimates (95% confidence intervals) is huge, indicating that using consecutive samples to evaluate the prevalence trends in a population may lead to inaccurate conclusions. Compare to Fig. 3 in the main text.

	Mean prevalence at time 1 (%)	Mean prevalence at time 2 (%)	Mean difference [95% CI]	Proportion of samples which underestimate the increase	Mean percentage error in the underestimate [95% CI]	Proportion of samples which overestimate the increase	Mean percentage error in the overestimate [95% CI]
Actual	20.6	32.2	11.5 [10.4 – 12.5]				
Raw	22.4	34.6	12.2 [5.3 – 17.6]	0.40	23.18 [1.92 – 66.05]	0.60	26.19 [1.61 – 59.70]
V-H adjusted, real degrees	20.8	32.5	11.6 [6.1 – 16.8]	0.47	23.46 [1.19 – 68.87]	0.53	23.56 [2.16 – 54.51]
V-H adjusted, degree rounded to nearest 5	21.1	32.9	11.8 [5.8 – 16.9]	0.45	24.09 [4.09 – 67.48]	0.55	24.50 [3.04 – 57.91]
V-H adjusted, degree rounded to nearest 10	21.1	32.9	11.8 [5.5 – 17.0]	0.44	24.49 [1.84 – 65.91]	0.56	23.80 [2.65 – 54.88]
V-H adjusted, degree increased by 5	23.7	36.8	13.1 [6.0 – 19.1]	0.33	21.14 [0.76 – 70.79]	0.67	31.90 [1.37 – 73.70]
V-H adjusted, degree rounded as Bristol data, 1	21.0	32.7	11.7 [6.4 – 17.1]	0.46	23.44 [3.25 – 64.96]	0.54	23.90 [0.46 – 56.24]
V-H adjusted, degree rounded as Bristol data, 2	20.4	32.4	11.9 [5.4–19.1]	0.46	25.56 [5.26– 67.26]	0.54	29.00 [2.08– 71.89]

Table S5: *Poisson distributed degrees: the results of 100 consecutive (paired) RDS samples for a population with increasing prevalence.* In the network with Poisson distributed degrees the over and under-estimates are slightly lower than in the network with long-tails, around 25% in both the raw and adjusted data and the average increase (over all 100 repeats) is close to the real increase between timepoints. In particular, increasing all the degrees by 5 increases the error the most of all the degree - biasing methods. This was also seen in Fig. S5; increasing the degree by 5 increased the percentage error from 1 to 4%. However, the errors on the network with the long-tailed degree distribution are much higher and we note that injecting networks in reality appear to follow a long tailed degree distribution (the Bristol data for example).

References

S. Goel and M.J. Salganik. Assessing respondent-driven sampling. *Proc Natl Acad Sci USA*, 107(15):6743, 2010.

M.H. Hansen and W.N. Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.

M.J. Salganik and D.D. Heckathorn. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34(1):193–240, 2004. ISSN 1467-9531.

E. Volz and D.D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics, Stockholm*, 24(1):79–97, 2008. ISSN 0282-423X.

E. Volz, C. Wejnert, I. Degani, and D.D. Heckathorn. Respondent-driven sampling analysis tool (RDSAT) version 5.6. *Ithaca, NY: Cornell University*, 2007.