Respondent driven sampling—where we are and where should we be going?

Richard G White,[1] Amy Lansky,[2] Sharad Goel,[3] David Wilson,[4] Wolfgang Hladik,[5] Avi Hakim,[5] Simon DW Frost[6]

Respondent Driven Sampling (RDS) is a novel variant of link tracing sampling that has primarily been used to estimate the characteristics of hard-to-reach groups, such as the HIV prevalence of drug users.[1] 'Seeds' are selected by convenience from a population of interest (target population) and given coupons. Seeds then use these coupons to recruit other people, who themselves become recruiters. Recruits are given compensation, usually money, for taking part in the survey and also an incentive for recruiting others. This process continues in recruitment 'waves' until the survey is stopped. Estimation methods are then applied to account for the biased recruitment, for example, the presumed over-recruitment of people with more acquaintances, in an attempt to generate estimates for the underlying population. RDS has quickly become popular and relied on by major public health organisations, including the US Centers for Disease Control and Prevention and Family Health International, chiefly because it is often found to be an efficient method of recruitment in hard-to-reach groups, but also because of the availability of custom written software incorporating inference methods that are designed to generate estimates that are representative of the wider population of interest, despite the biased sampling.

As demonstrated by RDS's popularity,[1] there was a clear need for new methods of data collection on hard-to-reach groups. However, RDS has not been without its critics. Its reliance on the target population for recruitment introduced ethical[w1] and sampling concerns.[w2] If RDS estimates are overly biased or the variance is unacceptably high, then RDS will be little more than another method of convenience sampling. If these errors can be minimised however, then RDS has the potential to become a very useful survey methodology.

In this editorial we highlight that 'RDS' includes both data collection and statistical inference methods, discuss the limitations of current RDS inference methods for generating representative estimates, highlight other applications of RDS for which it may be more reliable, propose and request feedback on a draft RDS reporting checklist, and finally suggest priority areas for RDS research.

As commonly discussed, RDS is actually a collection of methods to carry out two primary tasks, a method to sample a population and a method of statistical inference to generate population estimates.[2] A custom-written computer package, *'RDSAT'*, has been released to assist with data handling, tabulation and inference.[3] The RDS method of sampling is often efficient, with samples usually accruing quickly and with minimal perceived need for intervention by project staff, and has led to the collection of a wealth of data.[1] Lessons learned from the design and implementation of RDS have been shared and coalesced into standard protocols for data collection.[w3]

However, the performance of the inference methods is far less certain. There is much disagreement and confusion about the suitability and utility of the current methods of statistical inference and therefore the ability of RDS to generate representative data. Current inference methods rely on multiple assumptions of the sampling process, most of which may not be met in practice.[4 5 w4] Hence, in retrospect, it might seem reasonable to expect that RDS estimates are likely to suffer from (perhaps large) error. Unfortunately, we really do not know if this is true or not, because there have been few robust evaluations. This is in part because such evaluations are methodologically challenging to carry out. The representative or total-population data that are required are generally unavailable for hard-to-reach groups (hence the need for RDS). The most convincing studies that do exist (see[5–7]; Goel *et al*[7] also has a useful summary of other evaluations) suggest that RDS samples (i) may indeed suffer from bias and the bias may be difficult to detect, (ii) that the current inference methods do not reduce these biases, and perhaps most importantly, (iii) estimates probably have higher variance than initially thought. The latter is important because it means sample sizes in the thousands may be required to get the levels of precision currently assumed obtainable from sample sizes in the hundreds, and would therefore make RDS studies substantially larger, longer and more expensive than current common practice. The practical implications of these findings are that when interpreting RDS surveys that make statements about the wider population beyond the sample, CIs should be assumed to be too narrow, and adjustments should not be assumed to have made the unadjusted estimates more representative. Readers should also consider the unadjusted estimates and how representative they might be of the wider population.

That said, generating representative estimates is one of the most difficult things we could ask of RDS. Other potential applications for the RDS method or data collected using RDS include risk factor identification,[w5] social network data collection,[w6] population size estimation,[w7] and implementation of interventions.[w8] These other applications require fewer (or no) sampling assumptions be met. RDS might not be a panacea, but could still be the best method to collect data on many hard-to-reach groups. However, it is critical that the concerns about statistical inference be addressed, and the current benefits and limitations of RDS be better communicated to the broader public health community.

Another concern is that currently RDS studies are not being adequately reported.[1] Ultimately his reduces

[1]Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK; [2]Division of HIV/AIDS Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; [3]Microsoft Research, New York, USA; [4]World Bank, USA; [5]Division of Global HIV/AIDS, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; [6]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

Correspondence to Dr Richard White, Centre for the Mathematical Modelling of Infectious Diseases and Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK; richard.white@lshtm.ac.uk

Table 1 Summary of proposed additional information to be reported for RDS studies (STROBE-RDS study reporting checklist)

	Proposed additional information to be reported for RDS studies
Title and abstract	Indicate the study's design (Respondent-Driven Sampling) in the title or abstract
Methods	
Study design	State why RDS is considered the most appropriate sampling method
Setting	Describe formative research methods and findings used to inform RDS study design
Participants	Give the eligibility criteria, number, sources and methods of seed selection State if additional seeds were required, and if so, when and how recruited and started State if there was any variation in study design during data collection (eg, changing numbers of coupons per recruit, or stopping chains) Give the eligibility criteria for subsequent recruits if it differs from seeds Give number, types (eg, mobile/static) and location of recruitment venue(s) Report wording of network size question(s)
Variables	State if and how recruiter-recruit relationship was tracked
Data sources/ measurement	Describe methods to assess eligibility and reduce repeat enrolment (eg, coupon manager software, biometrics, detection of commercial exchange of coupons) Quality checks (eg, were returned coupons actually distributed and redeemed only once?)
Statistical methods	Describe all statistical methods, including name and description of the analytical methods (ie, point and interval estimators) used to take account of RDS sampling strategy. Report software package/version number and settings values Report any criteria used to support statements on whether estimator conditions or assumptions were met, for example, 'RDS equilibrium reached' State if seeds included in each analysis
Results	
Participants	Report numbers of individuals at each stage of study, that is, final number of seeds, number examined for eligibility, number confirmed eligible, number included in study, number returned for incentive collection and (if applicable) re-interview, and number analysed Give reasons for non-participation at each stage, including reason for coupon rejection Report number of coupons distributed and returned Report number of recruits by seed and number of RDS recruitment waves
Main results	Report unadjusted estimates and their stated precision (eg, 95% CI) If applicable, report adjusted estimates and their stated precision (eg, 95% CI) If adjusted estimates presented, report enough information so that the reason for the magnitude of the adjustment is clear (eg, network sizes and homophily by group)
Other analyses	Report other sensitivity analyses for example, different RDS estimators, different network size definitions
Discussion	
Limitations	Consider limitations of RDS sampling method and, if used, the RDS method(s) of inference. Include comment on how representative the unadjusted sample is thought to be

Note: this is a summary of the full checklist. See web appendix table W1 or the Equator Network website[8] for full details. Guidelines development proceeding according to Moher *et al* 2010.[w12] Checklist adapted from STROBE guidelines[w9] checklist for cross-sectional studies.[w10]
RDS, Respondent Driven Sampling.

the utility of published data and hinders assessment of study quality. Development of specific RDS reporting guidelines will assist in the interpretation of estimates and findings from RDS studies and in the evaluation of the RDS method itself. To facilitate this process, we have drafted a RDS study reporting checklist (Summary in table 1; Full version in web appendix table W1 or on Equator Network website[8]) that we have adapted from the STROBE guidelines[w9] for cross-sectional studies[w10] after receiving feedback from RDS experts contacted via the RDS list server[w11] and personal contacts. We invite further comments on the full draft checklist (web appendix table W1), either directly to the corresponding author, a Rapid Response on the STI website, or via the Equator Network website.[8] These comments will feed directly into a planned guidelines-setting meeting during which the contents of this checklist and the accompanying guidelines will be discussed and hopefully a consensus reached, followed by formal publication of the resulting guidelines.

There are many current priority areas for RDS research. There needs to be a clearer distinction between the methods used for RDS sampling and the methods used for statistical inference. There is a need for more systematic reporting of RDS studies. There is a critical need for more robust empirical evaluations to measure RDS sampling errors (bias, and variance if possible) in a range of different populations as context is likely to be important. Current efforts to develop new inference methods should be intensified,[9, 10] and more focus should be given to designing diagnostics to identify when problems are occurring during RDS recruitment. If problems are detected during data collection, steps could be taken to alleviate these problems immediately by correcting the problem or collecting additional data that may allow correction during the estimation stage. As new inference methods are developed, it would be preferable if they were released as open source programmes in commonly used statistics packages (R, STATA, SAS etc) so they can be more easily evaluated and compared to existing methods.

RDS has undoubtedly generated a wealth of new data on populations that have historically been difficult to access, and RDS is here to stay. The challenge now is to improve the methods (both sampling and inference) and ensure studies are reported well enough so that we can make the most of these data to improve public health.

Acknowledgements This editorial was largely based on presentations and discussions at the symposium organised by RGW *"Respondent Driven Sampling: Where we are and where should we be going?"* at the 19th Biennial International Society for STD Research conference in Quebec, Canada in July 2011. We would like to thank many individuals for comments on the draft reporting checklist, in particular Amy Drake and Lisa Johnston.

► Additional supplementary files are published online only. To view these please visit the journal online (http://dx.doi.org/10.1136/sextrans-2012-050703).

Contributors RGW wrote the first draft of the editorial and STROBE-RDS check list. All other authors revised the editorial and checklist critically for important intellectual content and helped redraft. RGW wrote the final draft. All authors agreed with the final version of the manuscript.

Funding RGW is funded by a Medical Research Council (UK) Methodology Research Fellowship (G0802414), the Consortium to Respond Effectively to the AIDS/TB Epidemic (19790.01), and the Bill and Melinda Gates Foundation (21675). SDWF is funded by the National

Institute of Nursing Research (grant NR10961), the National Institute on Drug Abuse (grant DA24998), and by a Royal Society Wolfson Research Merit Award. The funders had no involvement in the design, collection, analysis or interpretation of the data, in writing the report or in the decision to submit.

Competing interests None.

Provenance and peer review Non commissioned, internally peer reviewed.

Accepted 8 August 2012

Sex Transm Infect 2012;**88**:397–399.
doi:10.1136/sextrans-2012-050703

REFERENCES

1. **Malekinejad M**, *et al*. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. *AIDS Behav* 2008;**12**(S1):105–30.
2. **Salganik MJ.** Respondent-driven sampling in the real world. *Epidemiology* 2012;**23**:148–50. 10.1097/EDE.0b013e31823b6979
3. **Volz E**, *et al*. *Respondent-driven sampling analysis tool (RDSAT)*. http://www.respondentdrivensampling.org. Ithaca, NY: Cornell University, 2007.
4. **Rudolph AE**, *et al*. Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: implications for research and public health practice. *Ann Epidemiol* 2011;**21**:280–9.
5. **McCreesh N**, *et al*. Evaluation of respondent-driven sampling. *Epidemiology* 2012;**23**:138–47. 10.1097/EDE.0b013e31823ac17c
6. **Wejnert C.** An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol Methodol* 2009;**39**:73–116.
7. **Goel S,** Salganik MJ. Assessing respondent-driven sampling. *Proc Natl Acad Sci U S A* 2010;**107**:6743–7.
8. **Equator Network.** Equator Network—Reporting-guidelines-under-development. 2012; http://www.equator-network.org/resource-centre/library-of-health-research-reporting/reporting-guidelines-under-development/
9. **McCreesh N**, *et al*. Exploration of determinants of recruitment and a novel point estimator for respondent driven sampling. *Epidemiology*, In review.
10. **Gile KJ.** Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J Am Stat Assoc* 2011;**106**:135–46.