Respondent-driven Sampling in the Real World

Matthew J. Salganik

In many countries, HIV/AIDS infections are concentrated in 3 high-risk groups: injection drug users, sex workers, and men who have sex with men.[1] Accurate information about disease prevalence and risk behaviors in these groups is therefore critical for designing and evaluating HIV prevention programs. Respondent-driven sampling (RDS) is a relatively new method that allows researchers to collect such information about "hidden" or "hard-to-reach" groups, and the work of McCreesh and colleagues[2] provides an important contribution to our understanding of RDS. An RDS sample is collected through a peer-to-peer recruitment process, akin to "snowball sampling" and "link-tracing sampling." Once these data are collected, they might not be directly representative of the target population, and thus RDS also provides researchers a set of statistical procedures to adjust the observed data, with the hope that these adjustments will result in estimates that are more reflective of the target population.

Given the importance of the public health problem and the limitations of available alternatives, respondent-driven sampling has been rapidly adopted by the international public health community.[3–6] Despite its widespread use, however, little is currently known about the actual (as opposed to the theoretical) performance of RDS, and some recent research suggests cause for concern.[7–12] Given this, many researchers may be left to wonder: How well does RDS actually work in the real world? Our limited ability to address this question is not for lack of effort; in the past, researchers have taken several approaches to the problem, and the paper of McCreesh et al[2] is a new and important contribution to this stream of research.

RDS as a method of data collection—what I will call "RDS sampling"—and RDS as a method of data analysis—what I will call "RDS inference"—were both introduced by Heckathorn in 1997.[13] Although RDS sampling has remained largely unchanged, RDS inference has been an area of active research resulting in the RDS-I estimator,[14] the RDS-II estimator,[15] the RDS-MR estimator,[16] the RDS-SS estimator,[17] estimators currently in development,[18] and several approaches to variance estimation.[19,20]

Previous efforts to assess the performance of RDS generally fall into 3 categories: (1) analytic results, (2) simulation studies, and (3) studies using data from hidden populations. Each of these approaches has strengths and weaknesses, but it can be helpful to consider them in terms of the trade-off between precision and relevance. Some approaches—analytic results and simulation studies—allow for precise, definitive conclusions (eg, under these 5 conditions, this estimator has these 3 properties), but these conclusions may be irrelevant to actual RDS studies because they could depend on assumptions that bear little relationship to what actually happens in practice. On the other hand, studies involving data from hidden populations, while certainly relevant, rarely yield precise, unequivocal results because the underlying truth is not known. The work of McCreesh et al,[2] building on innovative work by Wejnert and Heckathorn,[21,22] attempts

From the Department of Sociology and Office of Population Research, Princeton University, Princeton, NJ.

This research was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (R01-HD062366 and R24-HD047879) and the U.S. National Science Foundation (CNS-0905086). The authors reported no other financial interests related to this research.

Opinions expressed here are the author's and not the funders.

Correspondence: Matthew J. Salganik, Department of Sociology and Office of Population Research, Princeton University, 145 Wallace Hall, Princeton, NJ 08544. E-mail: mjs3@princeton.edu.

Copyright © 2011 by Lippincott Williams & Wilkins

ISSN: 1044-3983/12/2301-0148
DOI: 10.1097/EDE.0b013e31823b6979

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

to combine precision and relevance by performing an RDS study on a population with known characteristics: a group of villages in rural Uganda. Given this design, McCreesh et al[2] can make precise statements about the performance of RDS (ie, the relationship between estimates and true population values) in a setting similar to those where RDS is typically used.

McCreesh et al[2] found that the data collected during RDS sampling was, without any statistical adjustments, roughly reflective of the population. However, performing RDS inference with both the RDS-I and RDS-II estimators—the 2 most commonly used estimators—tended to make the estimates worse, not better. This is a surprising and troubling result, which shows that more research is needed for RDS sampling and RDS inference.

Fortunately, the work of McCreesh et al[2] also suggests ways forward, by providing insights about the RDS sampling process—from both quantitative analysis of the sample and qualitative interviews conducted with members of the community. These insights explain the poor performance of the estimators and suggest ways that RDS inference might be improved. For example, McCreesh et al found that men, >50 years of age were overrepresented in the sample and that neither RDS-inference procedure corrected this problem. Through interviews with community members they were able to discover why this occurred: community members tended not to consider many younger men as a “head of household,” even if these younger men met the formal study inclusion criteria of the researchers. Neither RDS-inference procedure used was designed to handle this kind of problem, nor was effective at remedying it. Unfortunately, this mismatch between researcher and respondent conceptualization could be quite common because RDS is often used on populations, such as sex workers and men who have sex with men, whose boundaries may be more clear in the minds of researchers than in the minds of respondents.

More generally, this example shows just some of the complexities that are introduced by the “respondent-driven” nature of RDS sampling. In traditional sampling methods, researchers select respondents according to a specified design and then collect data through a process that can be monitored and controlled. In hidden populations, however, such researcher-selected samples are likely to be biased and are often logistically infeasible. RDS sampling transfers the sampling work that is normally done by the researchers to the respondents, relying on a system of coupons to track recruitment and a dual-incentive system to encourage participation (respondents are paid for participating and for recruiting others).[13] Although RDS sampling has proven effective for collecting large and diverse samples in a wide range of settings,[4] involving respondents in the sampling process means that the RDS data-generating process is largely outside of the control, and even the view, of researchers.

The work of McCreesh et al[2] has a number of characteristics that I hope and expect we will see more of in future RDS research. First, the study integrates the collection and analysis of qualitative data to produce insights about the likely sources of bias in the quantitative estimates. This integration of qualitative work both before[23] and during RDS sampling is something that would strengthen many future RDS studies. Second, the study explicitly considers several procedures for RDS inference, including the sample mean. In the minds of many researchers, there is a large difference between the sample mean and the more complicated RDS estimators. These complex estimators, however, may be similar to[11,24] or worse than the simple estimator. As more procedures for RDS inference are developed, more work will need to be done comparing their performance in a range of situations,[24] and it will be important for researchers to clearly specify which estimators they are using. Third, the study of McCreesh et al explicitly makes use of geographic data. RDS sampling is affected by physical geography,[25] and efforts to understand and then statistically model this aspect of the sampling are important areas for future work.

McCreesh et al[2] have provided a valuable contribution to the RDS literature. As a final step, I hope that the authors can release these data for analysis by other. Despite the huge number of RDS studies, there are very few publicly available data sets, and this lack of available data has hindered the development of RDS. Given the importance of RDS to global public health policy, this is quite unfortunate. Undoubtedly, data release raises concerns about the protection of human subjects, but these challenges can and must be overcome.[26] One aspect of the McCreesh et al study that should make data release easier is that, unlike most RDS studies, the population under study is not defined by illegal or stigmatized behavior. If these data were released, they would provide a test bed for future RDS-inference procedures by allowing researchers to make precise statements about the performance of new estimators using real RDS data. Thus, the same aspects of the study design that make the results of McCreesh et al so interesting, would make these data incredibly valuable for future RDS research.

ABOUT THE AUTHOR

MATTHEW SALGANIK is an Assistant Professor at Princeton University in the Department of Sociology and the Office of Population Research. He has developed and implemented network-based methods for studying hidden populations, especially those most at risk for HIV/AIDS. In addition to numerous academic venues, he has presented the results of this research to public health officials in the governments of Brazil, the Dominican Republic, Rwanda, and the United States.

© 2011 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

ACKNOWLEDGMENTS

I thank Dennis Feehan, Krista Gile, and Lisa Johnston for helpful conversations.

REFERENCES

1. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*. 2005; 19:S67.
2. McCreesh N, Frost S, Seeley J, et al. Evaluation of respondent-driven sampling. *Epidemiology*. 2012;23:138–147.
3. Malekinejad M, Johnston LG, Kendall C, Kerr LRF, Rifkin MR, Rutherford GW. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. *AIDS Behav*. 2008;12:105–130.
4. Johnston LG, Malekinejad M, Kendall C, Iuppa IM, Rutherford GW. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: field experiences in international settings. *AIDS Behav*. 2008;12(S1):131–141.
5. Lansky A, Abdul-Quader AS, Cribbin M, et al. Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System. *Public Health Rep*. 2007; 122(suppl 1):48.
6. Barbosa Júnior A, Pati Pascom AR, Szwarcwald CL, Kendall C, McFarland W. Transfer of sampling methods for studies on most-at-risk populations (MARPs) in Brazil. *Cadernos de Saúde Pública*. 2011; 27(S1):S36–S44.
7. Heimer R. Critical issues and further questions about respondent-driven sampling: comment on Ramirez-Valles et al (2005). *AIDS Behav*. 2005;9:403–408.
8. Scott G. "They got their program, and I got mine": a cautionary tale concerning the ethical implications of using respondent-driven sampling to study injection drug users. *Int J Drug Policy*. 2008;19:42–51.
9. Gile KJ, Handcock MS. Respondent-driven sampling: an assessment of current methodology. *Sociol Methodol*. 2010;40:285–327.
10. Goel S, Salganik MJ. Respondent-driven sampling as Markov chain Monte Carlo. *Stat Med*. 2009;28:2202–2229.
11. Goel S, Salganik MJ. Assessing respondent-driven sampling. *Proc Natl Acad Sci USA*. 2010;107:6743–6747.
12. Lu X, Bengtsson L, Britton T, et al. The sensitivity of respondent-driven sampling. *J Royal Stat Soc Series A Stat Soc*. In press.
13. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl*. 1997;44:174–199.
14. Salganik MJ, Heckathorn DD. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociol Methodol*. 2004;34:193–240.
15. Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *J Off Stat*. 2008;24:79.
16. Heckathorn DD. Extensions of Respondent-Driven Sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociol Methodol*. 2007;37:151–207.
17. Gile KJ. Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. *J Am Stat Assoc*. 2011;106:135–146.
18. Gile KJ, Handcock MS. Network model-assisted inference from respondent-driven sampling data. arXiv:1108.0298. 2011. Available at: http://arxiv.org/abs/1108.0298. Accessed September 27, 2011.
19. Salganik MJ. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J Urban Health*. 2006;83: 98–112.
20. Szwarcwald CL, de Souza Júnior PRB, Damacena GN, Junior AB, Kendall C. Analysis of data collected by RDS among sex workers in 10 Brazilian cities, 2009: estimation of the prevalence of HIV, variance, and design effect. *JAIDS*. 2011;57:S129–S135.
21. Wejnert C, Heckathorn DD. Web-based network sampling: efficiency and efficacy of respondent-driven sampling for online research. *Sociol Methods Res*. 2008;37:105–134.
22. 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.
23. Johnston LG, Whitehead S, Simic-Lawson M, Kendall C. Formative research to optimize respondent-driven sampling surveys among hard-to-reach populations in HIV behavioral and biological surveillance: lessons learned from four case studies. *AIDS Care*. 2010;22:784–792.
24. Tomas A, Gile KJ. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron J Stat*. 2011;5:899–934.
25. Toledo L, Codeço CT, Bertoni N, Albuquerque E, Malta M, Bastos FI. Putting respondent-driven sampling on the map: insights from Rio de Janeiro, Brazil. *JAIDS*. 2011;57:S136–S143.
26. Reiter JP, Kinney SK. Commentary: sharing confidential data for research purposes. *Epidemiology*. 2011;22:632–635.

© 2011 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.