Health research among hard-to-reach people: six degrees of sampling

Mary Aglipay MSc, John L. Wylie PhD, Ann M. Jolly PhD

CMAJ Podcasts: author interview at https://soundcloud.com/cmajpodcasts/141076-analy

Many marginalized populations face severe health inequities in developed countries. These include people who are poor, homeless, sell sex or use alcohol or illegal drugs. In Toronto, an estimated 19% of homeless people who had a diagnosis of active tuberculosis during 1998–2007 died within 12 months after diagnosis.[1] Studies of bloodborne infections in at-risk populations showed that street youth in Montréal were 15 times as likely as the general Canadian population to have hepatitis C infection (12.6% v. 0.8%),[2] and street youth in Vancouver were 10 times as likely to have HIV infection (2% v. 0.2%).[3]

Addressing the severity and high prevalence of disease in these groups requires skillful, applied research to produce accurate results. However, research involving hard-to-reach populations is challenging. Often lacking is a list of all members (sampling frame) from which a representative study sample can be selected. Some people may be reluctant to present themselves for fear of persecution, or they may not want to or be able to provide contact information such as a phone number. Consequently, many health studies fail to include some of society's most vulnerable members. We discuss methods for recruitment of hard-to-reach people in research. We focus mainly on respondent-driven sampling and give examples of implementation in the Canadian setting.

Types of sampling

The most common method of sampling whole populations (probability sampling) begins with creating a sampling frame from which participants are selected with equal probability to represent the whole population (Figure 1). For example, if the whole population owning phones is sufficiently large and includes most, if not all, of the people with the condition under study, random-digit dialing may be an appropriate choice.[4] Because probability sampling is often not feasible in vulnerable populations, researchers will use nonprobability methods, whereby they intentionally select participants with the condition or behaviour under study. These methods include time–space sampling, targeted sampling,[5] key-informant sampling and respondent-driven sampling.

Time–space sampling requires a list of diverse and sometimes sparsely attended places and times where members of the hidden population meet. Enumeration of some of the people at the venues is conducted if they appear to belong to the group under study, after which another subsample of venues is selected and people are asked to answer brief questions to confirm their eligibility.[6] Only a subsample of those eligible are systematically selected for interview, which allows estimations to be made based on sampling the venues and the people at them with a known probability.

Targeted or outreach sampling involves the use of both probability and nonprobability methods to identify lists of specified populations within geographic areas. Sampling is iterative, with continuing re-evaluation of sampling strategies to include all members. An important feature of targeted sampling involves sending outreach field workers to various venues and locations to observe people in the community and to interact with them and with key informants.[5] Key-informant sampling involves drawing "knowledgeable" members from the hidden

Key points

- Many marginalized populations, such as injection drug users, face severe health challenges compared with the general Canadian population.
- Health research involving these hard-to-reach groups is challenging because of stigma, access and lack of resources.
- Respondent-driven sampling takes advantage of the connections between people in these groups, who recruit each other into a study in a chain-referral (friend-of-a-friend) manner.
- Analytical methods based on such nonrandom, nonindependent samples are challenging, but they can provide important measures of the relative extent of illness in vulnerable groups.

Competing interests: None declared.

This article has been peer reviewed.

Correspondence to:
Ann Jolly,
ajolly@uottawa.ca

***CMAJ* 2015. DOI:10.1503/cmaj.141076**

population, and asking them about behaviours of other members, excluding themselves.[7] Although initially intended to preserve transmission of sensitive information, the validity of proxy report is questionable.

What is the theory behind respondent-driven sampling?

The concept that participants have valuable knowledge of both their own networks and other unknown ones — known as the six degrees of separation — was proved in early research on the small-world problem.[8] In 1997, Heckathorn[9] introduced respondent-driven sampling, derived from snowball sampling, whereby future participants are recruited in a chain-referral (friend-of-a-friend) manner (Figure 2). An important assumption is that the target population is highly connected: although characteristics of the participants may diverge from those of the original leaders, each wave of participants generally resembles those in the previous wave and eventually those of the whole group.[10]

Initial formative research — assessments of the marginalized population, and the building of trust between the researchers and the community — are important prerequisites before the selection of initial participants can begin. First, the investigators identify leaders (known in sociology as "seeds"), who are well connected to the target communities under study. The leaders are interviewed at a location that is easily accessible to them; they are paid and supplied with a set number of uniquely coded, documented referral coupons (usually three) that they are asked to give to acquaintances in their own network who also have the characteristics under study. Eligible recruits present their referral coupon to the researcher, documenting the link between the recruiter and the recruited, and are asked to participate in the study. If they agree, they are interviewed, paid and given three coupons for distribution to the next "wave." This continues until the desired sample size is reached. Using three coupons allows the recruitment chain to continue even if some participants do not recruit anyone, and it allows for many sampling waves, which ensures that all members of the hidden population have a nonzero probability of being selected.[4] Participants are paid again (as in Heckathorn's original design) whenever one of their recruits is interviewed.[9]

How well does respondent-driven sampling work?

Evaluating the sucess of respondent-driven sampling is difficult, given the different types of populations included in studies, the vastly different cultures, and the rarity and complexity of comparing different sampling methods within the same population.[10]

In a systematic review of 123 studies that used respondent-driven sampling to evaluate HIV risk behaviours, participants included men who have sex with men, injection drug users and sex workers from 28 countries, excluding the United States, and the recruitment periods varied from 2 to 48 weeks.[10]

We identified studies in which respondent-driven sampling was compared with another

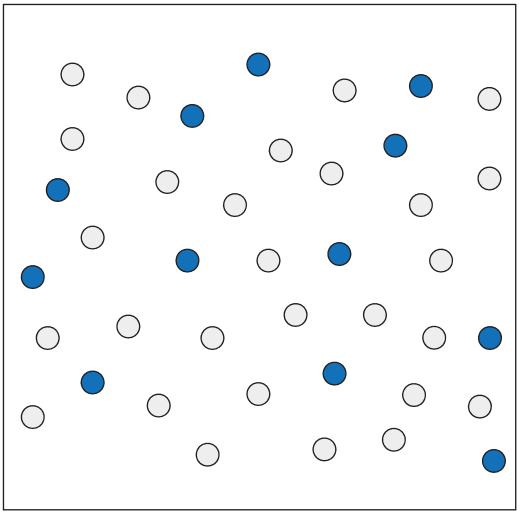

Figure 1: A group of individuals represented by circles, some of whom have been randomly selected as a representative sample with equal probability from a known list of the population.

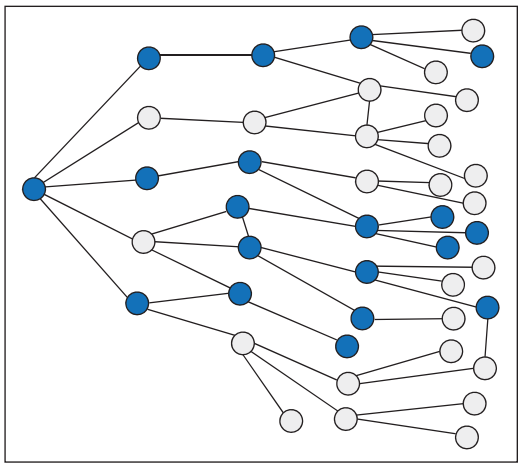

Figure 2: A network of friends represented by circles stemming from one individual with five friends, and their friends, up to four generations. Blue circles represent those selected and recruited by each other in a study using respondent-driven sampling methods.

method simultaneously. In one study, respondent-driven sampling was conducted in an open cohort of 2402 consenting male heads of households in adjacent villages in Uganda; the study began with 10 leaders, but the large number of people arriving for interviews within 9–32 days became too large to be manageable.[11] However, recruitment using respondent-driven sampling can be slow if members of the population of interest are only weakly connected.[12] In a study comparing respondent-driven sampling and targeted sampling methods in three cities in the US, respondent-driven sampling resulted in a higher proportion of eligible participants among those screened and required less staff time per recruit, with comparable cost-effectiveness.[13] In another study, the cost of sampling for cancer screening among unmarried, heterosexual and gay middle-aged and older women was higher, and participation lower, with respondent-driven sampling than with the use of print media, but both methods were superior when compared with targeted sampling at health fairs and community events.[14] In a study involving injection drug users in Estonia and Russia, respondent-driven sampling delivered results similar to those in which indigenous field workers recruited participants, though with lower salary costs.[15]

Respondent-driven sampling reached more hidden and vulnerable female sex workers (whose clients were solicited via agent ["pimp"], telephone or Internet) than time–space sampling in two Vietnamese cities.[16] In Guatemala City, respondent-driven sampling of men who have sex with men resulted in a higher population size estimate than that obtained by time–space sampling; the difference was probably because the latter method targeted men who attended venues known to be frequented by men who have sex with men, whereas respondent-driven sampling was able to reach participants not attending the venues as well.[17] Respondent-driven sampling was also better than other methods in recruiting more vulnerable participants of lower socioeconomic status among men who have sex with men in Fortaleza, Brazil,[18] and in Guangzhou, China.[19] However, the reverse appeared to be the case in a study involving injection drug users in San Francisco,[20] and no significant differences were found between the type of recruits from respondent-driven sampling and the type from targeted sampling in three US cities.[13] In the Ugandan study described earlier,[11] respondent-driven sampling produced a representative sample in most respects but an underrepresentation in others compared with the whole study population, which was difficult to quantify.

These studies suggest that respondent-driven sampling is similar or slightly better than other methods in recruiting difficult-to-reach individuals. However, studies may vary in duration of the recruitment period, and the cost-effectiveness of the sampling method and access to most marginalized groups are highly dependent on the relationships between the researchers and the participant communities under study.

How have Canadian researchers sampled hard-to-reach populations?

Respondent-driven sampling in Canada has been successful for the most part in recruiting hard-to-reach people, and most successful where health care and social counselling have been offered along with surveys.

Although the statistical theory behind respondent-driven sampling posits that initial selection of leaders[4] is independent of final estimates given sufficient waves of recruitment, Canadian researchers have carefully chosen leaders to reflect the diversity of their target populations (Appendix 1, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.141076/-/DC1) in order to minimize the number of waves until the sample reached independence from the leaders. For example, in a social network study in Winnipeg, leaders were selected from street youth, men who have sex with men, and injection drug users.[21]

The dual-incentive system[9] presents some challenges, including possible breach of the recruiter's confidentiality and coercion of recruits by recruiters.[22] In some instances, researchers in Canada noted that recruiters were being informally paid "in kind" by being helped to move house or obtain goods, eliminating the need for a second incentive. For these reasons, certain investigators have not provided second incentives, yet they have found that interest surpasses study capacity (see studies by Wylie and colleagues in Appendix 1).

Another challenge is that of excluding potential participants who do not have referral cards. This may prompt negative reactions from the community under study and jeopardize recruitment. In studies by Wylie and colleagues described in Appendix 1, two methods of leader selection were compared to determine the influence on recruitment (unpublished data). In one group, study staff selected a small number of leaders to begin recruitment chains, as per standard respondent-driven sampling. The second group consisted of individuals who self-presented to study staff without referral cards. The authors found that the self-presenters were more likely than those selected by the standard approach to be sex workers, to have less education and to rely more on gov-

ernment and other support income. This suggests that self-presenters who proactively sought out the study staff may have needed the compensation more and were more vulnerable than those who received the referral card with study contact information.

In the studies by Wylie and colleagues in Appendix 1, respondent-driven recruitment of men who have sex with men may have been less successful than similar recruitment methods among street youth and injection drug users because laboratory test results were not returned to them, or they did not receive health interventions, social system referrals or advice from an experienced street nurse, unlike street youth and injection drug users. Also, there was confusion over recruitment by coupons or response to an advertisement. Unusually cold weather coincided with decreased recruitment of men who have sex with men, which was not observed among street youth, who may have had more support services available to them. Finally, and unrelated to the study design, the networks of men who have sex with men and women who have sex with women may have been smaller and had fewer interconnections than the networks of street youth and injection drug users, which is consistent with lower recruitment.

How are data from studies that use respondent-driven sampling analyzed?

Most common statistical analyses rest on the assumptions that observations are independent of each other[9] or that participants are chosen randomly, with equal probabilities of selection. In respondent-driven sampling, participants nominate each other, so participants are not randomly selected and are not independent of one another. The simplest analytical method used in respondent-driven sampling (naive estimator) ignores the interdependency and lack of randomness and produces estimates using traditional statistical techniques. The Salganik–Heckathorn estimator adjusts estimates according to the total number of people each participant knows, thereby reducing the bias between those who know many people and those who know fewer. It also accounts for "bottlenecking," which occurs when progressive referrals result in concentrated sampling in only a small portion of the population of interest. For example, HIV-positive injection drug users may be more likely to recruit others who are also HIV positive, which results in an overestimation of HIV prevalence among injection drug users.[23] The Salganik–Heckathorn estimator mitigates this by monitoring and accounting for the changes in proportions of recruit characteristics from wave to wave. However, researchers analyzing real networks have found that Salganik–Heckathorn point and variance estimates do not differ significantly from those produced by traditional statistical techniques.[24]

Although the absolute values estimated with respondent-driving sampling methods (e.g., HIV prevalence) may be unreliable, the magnitude of relative odds within a sample is less questionable. For example, findings of a higher relative odds of HIV infection in certain networks of injection drug users than in others is important epidemiologically and remains unchallenged.[25]

The future of respondent-driven sampling

Results obtained by studies using respondent-driven sampling methods provide insight into social networks, those powerful structures of interactions between individuals through which infection is transmitted. Statistical analysis of samples is challenging, but analytical methods are rapidly developing to account for nonrandom selection and recruitment biases.[26,27] In the interim, guidelines for the use of different analytical methods under different sampling conditions have been developed.[24]

The strength of respondent-driven sampling is its ease of use in defining basic minimums of risk for infectious disease among people who are disproportionately affected. In addition, the interaction between the community and researchers serves to spread awareness to people. Finally, the same networks used to obtain information may be used to establish networks of prevention within marginalized communities in a culturally appropriate manner.

References

1. Khan K, Rea E, McDermaid C, et al. Active tuberculosis among homeless persons, Toronto, Ontario, Canada, 1998–2007. *Emerg Infect Dis* 2011;17:357-65.
2. Roy E, Haley N, Leclerc P, et al. Risk factors for hepatitis C virus infection among street youths. *CMAJ* 2001;165:557-60.
3. Summary: estimates of HIV prevalence and incidence in Canada, 2011. Ottawa: Public Health Agency of Canada; 2012. Available: http://phac-aspc.gc.ca/aids-sida/publication/survreport/estimat2011-eng.php (archived file; accessed 2015 May 27).
4. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using Respondent-Driven Sampling. *Sociol Methodol* 2004;34:193-240.
5. Watters JK, Biernacki P. Targeted sampling: options for the study of hidden populations. *Soc Probl* 1989;36:416-30.
6. Muhib FB, Lin LS, Stueve A, et al. A venue-based method for sampling hard-to-reach populations. *Public Health Rep* 2001; 116(Suppl 1):216-22.
7. Deaux E, Callaghan JW. Estimating statewide health-risk behavior: a comparison of telephone and key informant survey approaches. *Eval Rev* 1984;8:467-92.
8. Milgram S. The small-world problem. *Psychol Today* 1967;2:60-7.
9. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl* 1997;44:174-99.

10. Malekinejad M, Johnston LG, Kendall C, et al. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. *AIDS Behav* 2008;12(Suppl 4):S105-30.
11. McCreesh N, Frost SDW, Seeley J, et al. Evaluation of respondent-driven sampling. *Epidemiology* 2012;23:138-47.
12. Uusküla A, Johnston LG, Raag M, et al. Evaluating recruitment among female sex workers and injecting drug users at risk for HIV using respondent-driven sampling in Estonia. *J Urban Health* 2010;87:304-17.
13. Robinson WT, Risser JM, McGoy S, et al. Recruiting IDUs: a three-site comparison of results and experiences with respondent-driven and targeted sampling procedures. *J Urban Health* 2006;83(Suppl 6):i29-38.
14. Clark MA, Neighbors CJ, Wasserman MR, et al. Strategies and cost of recruitment of middle-aged and older unmarried women in a cancer screening study. *Cancer Epidemiol Biomarkers Prev* 2007;16:2605-14.
15. Platt L, Wall M, Rhodes T, et al. Methods to recruit hard-to-reach groups: comparing two chain referral sampling methods of recruiting injecting drug users across nine studies in Russia and Estonia. *J Urban Health* 2006;83(Suppl 6):i39-53.
16. Johnston LG, Sabin K, Mai TH, et al. Assessment of RDS for recruiting female sex workers in two Vietnamese cities: reaching the unseen sex worker. *J Urban Health* 2006;83(Suppl):i16-28.
17. Paz-Bailey G, Alvarez B, Miller W, et al. Population size estimates for MSM in Guatemala City using time location sampling and RDS [abstract P1-S4.08]. *Sex Transm Infect* 2011;87(Suppl 1):A163.
18. Kendall C, Kerr LRFS, Gondim RC, et al. An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in MSM, Fortaleza, Brazil. *AIDS Behav* 2008;12(Suppl 4):S97-104.
19. He Q, Wang Y, Li Y, et al. Accessing MSM through long-chain referral recruitment, Guangzhou, China. *AIDS Behav* 2008;12:2006-9.
20. Kral AH, Malekinejad M, Vaudrey J, et al. Comparing respondent-driven sampling and targeted sampling methods of recruiting IDUs in San Francisco. *J Urban Health* 2010;87:839-50.
21. Wylie JL, Jolly AM. Understanding recruitment: outcomes associated with alternate methods for seed selection in RDS. *BMC Med Res Methodol* 2013;13:93.
22. Scott G. "They got their program, and I got mine": a cautionary tale concerning the ethical implications of using respondent-driven sampling to study IDUs. *Int J Drug Policy* 2008;19:42-51.
23. Goel S, Salganik MJ. Respondent-driven sampling as Markov chain Monte Carlo. *Stat Med* 2009;28:2202-29.
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. Shaw SY, Shah L, Jolly AM, et al. Identifying heterogeneity among IDUs: a cluster analysis approach. *Am J Public Health* 2008;98:1430-7.
26. Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J Am Stat Assoc* 2011;106:135-46.
27. Poon AF, Brouwer KC, Strathdee SA, et al. Parsing social network survey data from hidden populations using stochastic context-free grammars. *PLoS ONE* 2009;4:e6777.

Affiliations: Department of Epidemiology and Community Medicine (Aglipay, Jolly), University of Ottawa, Ottawa, Ont.; Cadham Provincial Laboratory (Wylie), Government of Manitoba, Winnipeg, Man.; Department of Community Health (Wylie), University of Manitoba. Winnipeg, Man.

Contributors: All of the authors conceived the idea of the review and contributed to the search for unpublished studies using respondent-driven sampling in Canada. Mary Aglipay drafted the manuscript, and John Wylie and Ann Jolly revised the content for important intellectual content. All of the authors approved the final version to be published and agreed to act as guarantors of the work.

Funding: This work was funded by the Canadian Institutes of Health Research.