Title

Estimating the Size of Hidden Populations Using Respondent-driven Sampling Data

Permalink
https://escholarship.org/uc/item/4hj042kt

Journal
Epidemiology, 26(6)

ISSN
1044-3983

Authors
Johnston, Lisa G
McLaughlin, Katherine R
Rhilani, Houssine El
et al.

Publication Date
2015-11-01

DOI
10.1097/ede.0000000000000362

Peer reviewed

Estimating the Size of Hidden Populations Using Respondent-driven Sampling Data

Case Examples from Morocco

Lisa G. Johnston,[a] *Katherine R. McLaughlin,*[b] *Houssine El Rhilani,*[c] *Amina Latifi,*[d] *Abdalla Toufik,*[a] *Aziza Bennani,*[d] *Kamal Alami,*[c] *Boutaina Elomari,*[e] *and Mark S. Handcock*[b]

Background: Respondent-driven sampling is used worldwide to estimate the population prevalence of characteristics, such as HIV/AIDS and associated risk factors in hard-to-reach populations. Estimating the total size of these populations is of great interest to national and international organizations; however, reliable measures of population size often do not exist.

Methods: Successive sampling-population size estimation (SS-PSE) along with network size imputation allows population size estimates to be made without relying on separate studies or additional data (as in network scale-up, multiplier, and capture–recapture methods), which may be biased.

Results: Ten population size estimates were calculated for people who inject drugs, female sex workers, men who have sex with other men, and migrants from sub-Saharan Africa in six different cities in Morocco. SS-PSE estimates fell within or very close to the likely values provided by experts and the estimates from previous studies using other methods.

Conclusions: SS-PSE is an effective method for estimating the size of hard-to-reach populations that leverages important information within respondent-driven sampling studies. The addition of a network size imputation method helps to smooth network sizes allowing for more accurate results. However, caution should be used particularly when there is reason to believe that clustered subgroups may exist within the population of interest or when the sample size is small in relation to the population.

(*Epidemiology* 2015;26: 846–852)

Submitted 25 January 2015; accepted 8 July 2015.

From the [a]Independent Consultant to UNAIDS, Rabat, Morocco; [b]University of California, Los Angeles, CA; [c]UNAIDS, Rabat, Morocco; [d]National Programme to Fight HIV/AIDS, Direction d'Epidémiologie et de Lutte Contre les Maladies, Rabat, Morocco; and [e]Global Fund to Fight AIDS, Tuberculosis, and Malaria, Rabat, Morocco.

Acknowledgment: This material is based on work supported by the National Science Foundation (NSF) Graduate Research Fellowship (DGE-1144087), National Institute of Child Health and Human Development (NICHD; R21HD063000 and R21HD075714), NSF (MMS-0851555 and MMS-1357619), and the Office of Naval Research (N00014-08-1-1015). We are grateful to the California Center for Population Research (CCPR) at UCLA for general support. CCPR receives population research infrastructure funding (R24-HD041022) from the Eunice Kennedy Shriver NICHD. Its contents are solely the responsibility of the authors and do not represent the official views of the funding institutions.

Correspondence: Lisa G. Johnston, 64F Nieuwezijds Voorburgwal, Amsterdam, The Netherlands. E-mail: lsjohnston.global@gmail.com.

ISSN: 1044-3983/15/2606-0846
DOI: 10.1097/EDE.0000000000000362

To measure the magnitude of HIV/AIDS and other infections among hard-to-reach populations, countries are required by international donors (UNAIDS, WHO, Global Fund to Fight AIDS, Tuberculosis, and Malaria, etc.) to provide population size estimates of key populations at higher risk of HIV exposure, including people who inject drugs, female sex workers, men who have sex with men, and migrants. Knowing the true population size of key populations is essential for disease transmission modeling and projections, understanding the burden of disease, gauging service coverage, guiding resource allocation, and advocating for programs to reach key populations. To meet this requirement, size estimates of key populations are currently collected in conjunction with studies using respondent-driven sampling.

Respondent-driven sampling is widely used to obtain samples of populations in which members are connected to each other via a network of social ties, and for which no sampling frame is available.[1,2] Respondent-driven sampling uses participants' recruitment data and social network sizes to derive estimates about the target population. The recruitment process begins with a convenience sample of population members ("seeds"), who use a fixed number of coupons to recruit members of their social network. Coupons, which contain a unique number to manage peer-to-peer recruitment, provide a means for anonymity, making it especially successful among populations that are stigmatized or practice illegal behaviors. The objective is to generate long recruitment chains whereby the final sample is not biased by the initial convenience sample of seeds. Data collected with respondent-driven sampling methods are adjusted based on each participant's social network size or *degree* (e.g., the number of people they know and have seen in the past 2 weeks who fulfill the study eligibility criteria) and other covariates.[1–3]

Currently, many countries estimate population sizes using service and unique object multiplier methods which require two overlapping sources of data, one of which is representative of the population being sampled (i.e., a respondent-driven sampling study).[4–7] Specifically, the unique object multiplier method involves distributing a unique, nonvaluable item to population members 1 week before the initiation of a respondent-driven sampling study. The service multiplier involves obtaining unique counts of population members who received a service during a specific time period before the initiation of a respondent-driven sampling study. During the study, participants are asked if they received the object or accessed the service during the specific time period. The number of population members to whom the object was distributed or service provided is divided by the proportion in the study who reported receiving the object and/or service. Multiplier methods, used in conjunction with respondent-driven sampling studies, are popular because they are practical and do not require much additional cost or effort in addition to the study itself. However, these methods are subject to numerous immeasurable biases (i.e., nonindependence of both data sources resulting in underestimations), require data that may be unavailable (e.g., services do not have unique counts of clients, especially when no personal identification is required to receive the service) or unreliable (e.g., services do not have records specific to the population being sampled), and often provide excessively high or low estimates.[6] A new method called successive sampling-population size estimation (SS-PSE) provides a promising alternative to other methods commonly used to estimate the size of hard-to-reach populations.[8] SS-PSE relies only on data already collected in a respondent-driven sampling study: each participant's degree, recruitment patterns, and date of enrollment.

This article uses 10 datasets from studies conducted in Morocco between 2010 and 2013 to develop population size estimates using SS-PSE.[9–12] We develop a method for imputing each person's visibility in the network and use this in place of degree, which is typically used in SS-PSE. This builds on the SS-PSE framework and accounts for typical errata we may observe from self-reported study data. We compare and evaluate the accuracy of our population size estimates to those found for the same populations through other unpublished methods. In addition, we suggest questions useful for eliciting information needed to compute SS-PSE and provide guidelines and caveats to improve the implementation of SS-PSE for real data.

METHODS

Selection of Studies and Eligibility Criteria

We use data from 10 studies conducted in Morocco (Table 1)[9–12] using respondent-driven sampling. All participants were ages 18 years or older and resided in the respective study location. In addition, eligible people who inject drugs in Nador and Tanger reported injecting drugs in the past 6 months; female sex workers in Agadir, Fez, Rabat, and Tanger were females who exchanged penetrative (vaginal/anal) sex for money or gifts with more than one male client in the past 6 months, held Moroccan nationality, and worked in the respective study location[9]; men who have sex with men in Agadir and Marrakesh were males who had anal sex with a male in the last 6 months[10]; and francophone and anglophone migrants in Rabat were persons who originated from sub-Saharan African countries (e.g., Senegal, Cameroun, Mali, Cote d'Ivoire, Democratic Republic of the Congo, Guinea, etc. for francophone participants; Nigeria, Gambia, Ghana, Liberia, Sierra Leone, etc. for anglophone participants), had resided 3 months or more in Morocco, and spoke either French or English depending on from which migrant group they originated.[11,12]

Respondent-driven Sampling Methods

Each respondent-driven sampling study presented here was conducted with the involvement of all co-authors with the exception of the second and last author with strict adherence to methodological requirements. Seeds were identified through local contacts and were well connected to and trusted by the target population. Other essential considerations included the use of a quota of no more than three recruitment coupons per participant, incentives for participation and peer recruitment, collection of each participant's social network size data, tracking who recruited whom, facilitation of long recruitment chains and consequent attainment of equilibria for key variables of interest, and inclusion of a design effect of at least two in the sample size calculation.[13] The interviewing methods, questionnaires, incentives, and specimen collection varied by population. All studies were approved through

TABLE 1. Description of Surveys Conducted Among PWID, FSW,[9] MSM,[10] and Migrants in Morocco[11,12]

Target Population	City	Sample Size	Number of Seeds	Maximum Waves	Study Period	Data Collection Period
PWID	Nador	277	7	11	November 2010–February 2011	6 weeks
PWID	Tanger	268	7	18	October 2010–February 2011	6 weeks
FSW	Agadir	372	10	8	December 2011–January 2012	8 weeks
FSW	Fez	359	4	7	December 2011–January 2012	8 weeks
FSW	Rabat	392	5	10	December 2011–January 2012	8 weeks
FSW	Tanger	324	4	10	December 2011–January 2012	8 weeks
MSM	Agadir	323	10	12	November 2010–March 2011	25 weeks
MSM	Marrakesh	346	8	23	November 2010–March 2011	23 weeks
Anglophone migrants	Rabat	277	5	14	March–April 2013	32 days
Francophone migrants	Rabat	410	6	7	March–April 2013	21 days

PWID indicates people who inject drugs; FSW, female sex workers; MSM, men who have sex with men.

the Faculty of Medicine, Casablanca, Morocco, and fulfilled a consent process before recruits could enroll.

Population Size Estimation Methods

Eliciting Prior Information

In addition to data on each participant's degree and recruitment patterns, prior beliefs about population sizes are needed for the SS-PSE. These values will not be precise, but are intended to provide a rough idea of what the true population size value should be. Efforts were made to simulate a real life situation whereby only a few key persons, with limited time, would have knowledge about each study group's population size. Questions were developed in a table to elicit information about the mode, mean, median, first quartile, third quartile, minimum, and maximum of the likely population size distribution. The table was completed by key personnel (n = 4) from UNAIDS and the Ministry of Health AIDS program, identified by the first author through her involvement with the studies. Prior information for people who inject drugs was completed by one resource person working in the country and were similar to those provided by UNAIDS and Ministry of Health personnel. The specific phrasing of the questions and the instructions provided to personnel were:

(1) What is the most likely population size? (i.e., the number that is more likely to be correct than any other single number).
(2) What is the average population size from all knowledgeable people? (i.e., if you were to ask all knowledgeable people their guess of the population size, what would the average of all these numbers be?).
(3) What is the value that the population size is just as likely to be above as below?
(4) What is the value that the population size has a one chance in four of being less than? (i.e., if you were to bet that the population size would be less than this number, you would lose the bet about one time out of four).
(5) What is the value that the population size has a one chance in four of being greater than? (i.e., if you were to bet that the population size would be higher than this number, you would lose the bet about one time out of four).
(6) What is the lowest the population size could reasonably be? By this we mean the value that has about one chance in 20 of being less than.
(7) What is the highest the population size could reasonably be? By this we mean the value that has about has one chance in 20 of being greater than.

Additional instructions provided when answering the above questions were: (1) We know that you do not know exactly what the population size is, so please do not feel like this is a test; (2) please work among yourselves to complete the table. We do not intend for you to get other opinions or data; (3) do not use outside information to come up with your estimates of population sizes (these are your best guesses based on your experience); (4) This exercise is to make sure we have no extreme numbers (your best guesses are used to avoid numbers that are unreasonably small or unreasonably large). Results of prior information elicitation are displayed in Table 2.[14–17]

Statistical Methodology (SS-PSE)

The SS-PSE assumes that individuals with higher visibility are more likely to be recruited earlier in the respondent-driven sampling process.[8] Thus, if there are fewer high reported degrees in later waves, this reflects the depletion of the members of the population with higher visibility (i.e., the sample represents a substantial portion of the population). The SS-PSE assumes that visibility and reported degree are positively associated. However, if the reported degrees stay roughly the same across recruitment waves, the sample size is likely a smaller portion of the population. If the reported degrees increase notably across waves, this may be an indication that the respondent-driven sampling recruitment process is not operating as expected and would merit caution when interpreting the results of various estimators. Thus, we leverage information about the sequential nature of data collection.

SS-PSE uses a Bayesian framework, allowing for the incorporation of information about prior knowledge and educated approximations of the population size. This allows the incorporation of expert beliefs or population size estimates from previous sources, such as enumeration through mapping, network scale-up, multipliers, or capture–recapture, if they exist. In Bayesian statistics, information about unknowns is expressed through probability distributions over their possible values. Thus, the resulting estimates take the form of a distribution with properties, such as the mean, median, and probability intervals, rather than a point estimate and confidence intervals. In SS-PSE, we estimate the posterior distribution for the population size N given our prior belief about the population size and observed data. These data include individuals' degrees and the order in which they were sampled.

We follow the Gile successive sampling estimator which assumes that sampling proceeds according to a successive sampling procedure in which each subsequent sample is selected from among the remaining units with probability proportional to size.[18] This is the likelihood that recruitment occurs in the order in which we observe it, given our understanding of the network structure. Our understanding of network structure comes from participant's reported degrees, their recruitment patterns and our prior belief about the population size. For the respondent-driven sampling context, this condition is satisfied if we consider network structures sampled from a "configuration model," in which network ties form completely at random among a population of people with fixed and observable visibilities. This approach assumes that at any point in time a recruiter is capable of recruiting any person in

TABLE 2. Data for Population Size Estimates Based on Best Guesses (Prior Estimates), Literature Sources, and Multipliers and Final Estimates Based on SS-PSE for PWID, FSW, MSM, and Migrants in Morocco

	Best Guess Population Size Estimates from Experts			Population Size Estimates Based on Literature[14–17]			Population Sizes Based on Multipliers			Population Size Estimates SS-PSE Method		
Target Population and City	**Median[a]**	**Q1[b]**	**Q3[c]**	**% of Population**	**Adult Population[d]**	**Point Estimate/ Range**	**Point/Mean**	**Q1**	**Q3**	**Point Estimate/ Median**	**Lower Bound**	**Upper Bound**
PWID, Nador	400	500	800	0.3	764,400	1,995	-	-	-	870	450	1,600
PWID, Tanger	600	400	600	0.3	1,014,770	2,625	-	-	-	680	420	1,100
FSW, Agadir	10,000	6,000	15,000	0.1–12.0	1,106,600	410–49,500	3,639–4,333	-	-	5,700	2,500	11,600
FSW, Fez	8,000	5,000	10,000	0.1–12.0	1,123,350	410–49,500	6,028	-	-	7,000	3,200	13,900
FSW, Rabat	8,000	5,000	10,000	0.1–12.0	1,014,770	790–52,560	5,683	-	-	4,100	2,300	7,000
FSW, Tanger	8,000	4,000	8,000	0.1–12.0	2,120,360	370–43,800	3,956	-	-	3,200	1,800	5,600
MSM, Agadir	10,000	4,000	10,000	2.0–5.0	1,106,600	8,250–41,250	1,633	1,029	2,426	3,700	1,700	7,400
MSM, Marrakesh	10,000	4,000	10,000	2.0–5.0	1,238,530	9,000–20,625	1,392	742	2,045	5,900	2,300	13,200
Anglophone migrants, Rabat	4,000	1,000	6,000	-	2,120,360	-	4,145	-	-	2,900	720	8,900
Francophone migrants, Rabat	6,000	1,500	10,000	-	2,120,360	-	4,427	-	-	6,400	1,200	19,200

[a] Value that the population size is just as likely to be above as below.
[b] Value that the population size has a one chance in four of being less than.
[c] Value that the population size has a one chance in four of being greater than.
[d] Adults 18 years and older, adjusted for sex by dividing total population by 0.50 (population accessed on October 3, 2014 at http://worldpopulationreview.com/countries/morocco-population/).
SS-PSE indicates successive sampling population size estimations; PWID, people who inject drugs; FSW, female sex workers; MSM, men who have sex with men.

the population who has not previously been sampled. In practice, this assumption is likely violated because a person does not know all members of the population, but it is a standard respondent-driven sampling assumption and a requirement of the successive sampling estimator.

In this approach, we specify a prior for N, our prior belief about the population size. In this article, we use the first and third quartiles provided by key personnel in Morocco. This distribution takes the form of a heavy-right-tailed distribution, with shape and scale parameters determined by the experts' guesses and feasible values for the maximum degree and population size.[8] This allows incorporation of knowledge about the sample proportion (n/N) as well as previous or concomitant sources of information about the population size.[19]

The population unit sizes are treated as independent and identically distributed samples from a super-population model based on some (unknown) distribution. Thus every person in the population has a fixed unit size (this is typically equated with their degree, although other variables could be used instead). We observe a subset n < N of people in the population in our sample, so n of the N degrees are observed, and the other N–n are unobserved. From these observed degrees, we impute unit size using a model that leverages each person's self-reported degree and efficacy as a recruiter given their time from enrollment until the end of the study. We treat these unit sizes as draws from a Conway-Maxwell-Poisson class of distributions, following Handcock et al.[19]

In addition, we know the order of observations (i.e., the order in which people are sampled) based on participants' enrollment date. The specific structure of the recruitment process is not needed; only the order in which people was sampled.

This formulation of the joint posterior distribution allows for easy computation via a four-component Gibbs sampler to produce a posterior predictive distribution of the population size.[8] Because we have a distribution, we can use a desired measure of center (mean, median, or mode) as a population size estimate, as well as place this estimate within a probability interval that gives a likely range of values.

Inference for the Visibility Distribution

Using an individual's self-reported degree as their true degree assumes there are no errors or biases in self-reports. However, self-reported data, including degrees, can be biased for many reasons, including barrier effects and transmission and recall bias.[14,15] There are also many possible discrepancies, such as extremely large outliers (possibly due to different interpretations of the degree question) and over- or under-dispersion due to groupings at "round" values (e.g., people report 60 or 70, but not 63 or 67). Furthermore, using only a person's degree (whether self-reported or true) to calculate their probability of inclusion in the sample overlooks heterogeneities of the population. For example, some people may be more geographically isolated and less able to recruit than others, or people may feel they would face differential levels of social stigma as a result of recruiting. Thus, our focus is on a measure of a person's *visibility*, rather than their degree.

We impute this visibility (u_i) and use it in place of degree (d_i) in the SS-PSE framework to get population size estimates. Rather than letting $u_i = d_i$, we instead impute u_i based on d_i, r_i (the number of eligible recruits each person enrolls in the study, capped at the maximum number of coupons each recruiter could distribute, usually three), and the time from enrollment until the end of the study. Thus, we leverage another piece of network information from participants—their efficacy as a recruiter. This accounts for problems both with self-reports and with using degree instead of visibility. We assume that recruiters make their best effort to distribute their maximum number of coupons, as has been found in many respondent-driven sampling studies,[1] and that these coupons are given to people most likely to participate. A person who reported a very large degree but was not successful at recruiting was perhaps overly optimistic about their visibility. Even though they may know many people in the population, they may be more hidden and thus less accessible for the purpose of the sample. This approach smooths degree and puts a penalty on an individual's visibility if they were unsuccessful at recruiting. Individuals that are likely to be recruited by others should also be able to recruit. Hence the number of recruits and visibility should be positively associated. We create a statistical model that captures both of these features, and use the model to estimate the visibility of each respondent. We refer to this estimate as an *imputation.*

An important feature of imputation is that researchers no longer need to specify the maximum possible degree, as in SS-PSE, because specification errors can have a large influence on the population size estimates. In addition, imputation provides an easy method to handle cases where participants report zero or have missing degree by imputing their visibility from a starting value of the number of people they recruit plus one (for the person who recruited them, as long as they are not a seed). The respondent-driven sampling recruitment structure provides a minimum network size for these people, from which we can impute their visibility. Imputing visibility and then using the new values within the SS-PSE framework leads to improved estimates.

RESULTS AND DISCUSSION

SS-PSE was implemented for 10 datasets in Morocco, covering two populations of people who inject drugs, four populations of female sex workers, two populations of men who have sex with men, and two populations of migrants. The first and third quartiles, as provided by experts' approximations, were used as prior information. Unit sizes were inferred using imputed degrees. Table 3 provides the mean, median, mode, and probability intervals of the posterior distribution for the estimates of the population size N.

We can compare these results from SS-PSE to population size estimates obtained using other methods. Table 2 compares the means and probability intervals obtained from SS-PSE to those previously published in the literature[16,17,20,21] and from multiplier methods. In general, the SS-PSE values are slightly higher than the multiplier methods, and slightly lower than the results from the literature. However, sometimes all that is available for comparison are very broad ranges. Using SS-PSE allows us to more precisely evaluate likely values of the population size and can be used in triangulation with other population size estimates to develop more precise estimates.

We had small sample fractions (less than 10%) in eight of the 10 datasets. This is not atypical for respondent-driven sampling data.[22] The method seems to slightly underestimate the population size in these cases, indicative of the fact that there may not be enough information in the respondent-driven sampling data about degree. This informs a general principle we would like to posit: *Use caution when interpreting SS-PSE results when you suspect that the sample fraction (n/N) is less than 10%.* Smaller sample fractions may indicate that there is not enough information in the unit size variables to provide informative estimates. This effect is particularly pronounced when no form of degree imputation is used.

One guideline for SS-PSE is the expectation that at least a good portion of the probability interval from the posterior distribution will fall within the minimum and maximum provided a priori by experts. If much of the mass of the posterior distribution falls outside of these values given by experts, the respondent-driven sampling data and expert values tell a different story and further investigation may be warranted. Similarly, we would expect the mean and median to fall within the first and third quartiles provided by the experts. These guidelines need not always be true, but warrant further investigation if they are not. This gives the experts an opportunity to reconcile their prior beliefs to the population size estimates given the data, and determine whether discrepancies are the result of the sampling procedure or if their prior beliefs need to be updated.

SS-PSE relies on numerous assumptions, some of which may not always hold in practice. It is important to assess how deviations from these assumptions may impact the SS-PSE. We assume that people have observable network sizes, which, for the SS-PSE, are self-reported degrees.[6] However, for this analysis, we found that some self-reported degrees were invalid and, therefore, added an innovation by imputing degrees based on inference for self-reported degree and recruiter efficacy. This new imputation method, in place of reported degrees, may also be useful in improving prevalence estimates in respondent-driven sampling data.[23]

The SS estimator, necessary for calculating the SS-PSE, assumes that each recruiter is capable of recruiting any member of the network who has not yet been recruited, which may not be true in practice. Rather, we may expect people are capable of recruiting only those in their personal networks, or, more strictly, from an *effective personal network*

TABLE 3. Results from SS-PSE Method for PWID, FSW, MSM, and Migrants in Morocco

	Estimated Sample Fraction	Prior	Degree Imputation				
			Posterior				
Target Population and City	n Divided by Median from Expert Best Guess	Q1 and Q3 Based on Best Guess from Experts	Mean	Median	Mode	Lower Bound (5%)	Upper Bound (95%)
PWID, Nador	0.69	500, 800	870	780	660	450	1,600
PWID, Tanger	0.45	400, 600	680	610	520	420	1,100
FSW, Agadir	0.04	6,000, 15,000	5,700	4,900	3,700	2,500	11,600
FSW, Fez	0.05	5,000, 10,000	7,000	6,100	4,800	3,200	13,900
FSW, Rabat	0.05	5,000, 10,000	4,100	3,800	3,400	2,300	7,000
FSW, Tanger	0.04	4,000, 8,000	3,200	3,000	2,600	1,800	5,600
MSM, Agadir	0.03	4,000, 10,000	3,700	3,300	2,700	1,700	7,400
MSM, Marrakesh	0.04	4,000, 10,000	5,900	4,800	3,500	2,300	13,200
Anglophone migrants[a]	0.07	1,000, 6,000	2,900	1,900	1,100	720	8,900
Francophone migrants	0.07	1,500, 10,000	6,400	4,300	1,900	1,200	19,200

For each set of priors, the mean, median, and probability interval are given. All simulations were performed with a burn-in period of 10,000 iterations and an interval of 1,000 iterations, attaining a sample of size 600; SS-PSE results rounded.

[a]The authors believe that a "clan-like" structure that exists among anglophone migrants resulted in two separate subgroups being formed, corresponding to the two study centers that were used. Therefore, we calculate population size estimates separately for participants at each of the two study centers, and aggregate the results. The values in Table 3 for anglophone migrants are the aggregated results.

PWID indicates people who inject drugs; FSW, female sex workers; MSM, men who have sex with men.

that may be smaller than the value they report. We might suspect that a level of *optimism* is present in personal networks, where people try to think of anyone they know who satisfies the eligibility criteria, even if they would not be able to recruit them. Using visibility instead of self-reported degree, as in SS-PSE with imputation, is a way to incorporate the idea of effective personal network size into our calculations.

When the network structure deviates substantially from the assumed sampling structure, the method may not be valid. This can occur with highly structured populations, such as populations with distinct and unconnected subgroup clusters. Respondent-driven sampling relies on the assumption that, regardless of the initial seeds selected, each person has a nonzero chance of being included in the sample. However, this assumption may be violated if small subgroup clusters of people are socially isolated from the rest of the population. As long as the number of isolated people is relatively small, SS-PSE should produce reliable estimates. However, if a large number of people are separated from the connected component of the population, estimates may be unreliable. This can happen when, for example, a region with two similarly sized but geographically separated cities is sampled. In this case, we recommend estimating the population size separately for each disjoint subgroup, and adding the results. Imputation for visibility also helps account for differential levels of isolation within a population.

When conducting respondent-driven sampling studies, SS-PSE has great potential to estimate sizes of hard-to-reach populations without the additional data collection burden required by other methods. However, the guidelines and caveats presented above should be considered when interpreting the estimates.

CONCLUSION

Many countries currently required to estimate the sizes of hard-to-reach populations are frustrated by the limitations of current methods. Given the importance of obtaining population sizes that are as accurate as possible, novel methods are desperately needed. SS-PSE is a promising new method that leverages respondent-driven sampling study data about respondents' degree to make population size estimates that do not depend on a second data source. Other common methods, such as multiplier methods, require secondary data sources, which can be unreliable, unavailable, and costly to collect.

Using 10 datasets from Morocco, we explored SS-PSE by simulating a real world situation to gather prior population size estimates from people working with these populations. Where study and elicitation procedures are similar to those presented here, it is feasible that this method will work well in other settings. In addition, we advanced SS-PSE accuracy by developing an imputation method that uses the concept of visibility in place of self-reported degree. Finally, we provide general guidelines and caveats to help researchers using SS-PSE. The SS-PSE does not perform well in all scenarios, so it is essential to understand when it is appropriate to implement, and how to discern cases where performance may be poor.

All code for SS-PSE is available through the R programming language (R Core Team, 2012) and are provided for easy use through the open-source software *RDS Analyst* found at www.hpmrg.org.[24]

REFERENCES

1. Heckathorn DD. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Soc Probl*. 2002;49:11–34.
2. Heckathorn DD. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. *Sociol Methodol*. 2007;37:151–207.
3. Gile KJ, Handcock MS. Respondent-driven sampling: an assessment of current methodology. *Sociol Methodol*. 2010;40:285–327.
4. Johnston L, Saumtally A, Corceal S, Mahadoo I, Oodally F. High HIV and hepatitis C prevalence amongst injecting drug users in Mauritius: findings from a population size estimation and respondent driven sampling survey. *Int J Drug Policy*. 2011;22:252–258.
5. Paz-Bailey G, Jacobson JO, Guardado ME, et al. How many men who have sex with men and female sex workers live in El Salvador? Using respondent-driven sampling and capture-recapture to estimate population sizes. *Sex Transm Infect*. 2011;87:279–282.
6. Johnston LG, Prybylski D, Raymond HF, Mirzazadeh A, Manopaiboon C, McFarland W. Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: case studies from around the world. *Sex Transm Dis*. 2013;40:304–310.
7. UNAIDS. *Guidelines on Estimating the Size of Populations Most at Risk to HIV*. Geneva, Switzerland; 2010.
8. Handcock MS, Gile KJ, Mar CM. Estimating hidden population size using respondent-driven sampling data. *Electron J Stat*. 2014;8:1491–1521.
9. Johnston L, Bennani A, Latifi A, et al. Using respondent-driven sampling to estimate HIV and syphilis prevalence among female sex workers in Agadir, Fes, Rabat and Tangier, Morocco. *Sex Transm Infect*. 2013;89(Suppl 1):A180.
10. Johnston LG, Alami K, El Rhilani MH, et al. HIV, syphilis and sexual risk behaviours among men who have sex with men in Agadir and Marrakesh, Morocco. *Sex Transm Infect*. 2013;89(Suppl 3):iii45–iii48.
11. Johnston LG, Oumzil H, El Rhilani H, Latifi A, Bennani A, Alami K. Sex Differences in HIV prevalence, behavioral risks and prevention needs among anglophone and francophone sub-Saharan African migrants living in Rabat, Morocco. *AIDS Behav*. 2015.
12. Tyldum G, Johnston L. *Applying Respondent Driven Sampling to Migrant Populations: Lessons from the Field*. London, UK: Palgrave Macmillan; 2014.
13. Johnston LG. *Introduction to Respondent-Driven Sampling*. Geneva, Switzerland: World Health Organization; 2013.
14. Killworth PD. Investigating the variation of personal network size under unknown error conditions. *Sociol Methods Res*. 2006;35:84–112.
15. McCormick TH, Salganik MJ, Zheng T. How many people do you know?: efficiently estimating personal network size. *J Am Stat Assoc*. 2010;105:59–70.
16. Cáceres C, Konda K, Pecheny M, Chatterjee A, Lyerla R. Estimating the number of men who have sex with men in low and middle income countries. *Sex Transm Infect*. 2006;82(Suppl 3):iii3–i9.
17. Vandepitte J, Lyerla R, Dallabetta G, Crabbé F, Alary M, Buvé A. Estimates of the number of female sex workers in different regions of the world. *Sex Transm Infect*. 2006;82(Suppl 3):iii18–i25.
18. Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J Am Stat Assoc*. 2011;106:135–146.
19. Handcock MS, Gile KJ, Mar CM. Estimating the size of populations at high risk for HIV using respondent-driven sampling data. *Biometrics*. 2015;71:258–266.
20. Mathers BM, Degenhardt L, Phillips B, et al.; 2007 Reference Group to the UN on HIV and Injecting Drug Use. Global epidemiology of injecting drug use and HIV among people who inject drugs: a systematic review. *Lancet*. 2008;372:1733–1745.
21. Aceijas C, Friedman SR, Cooper HLF, Wiessing L, Stimson G V, Hickman M. Estimates of injecting drug users at the national and local level in developing and transitional countries, and gender and age distribution. *Sex Transm Infect*. 2006;82(Suppl 3):iii10–i17.
22. Malekinejad M, Johnston LG, Kendall C, Kerr LR, 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(4 Suppl):S105–S130.
23. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *J Royal Stat Soc*. 2015;178:241–269.
24. Handcock MS, Fellows IE, Giles KJ. RDS Analyst. 2014. Available at: www.hpmrg.org. Accessed on January 26, 2015.