White, RG; Lansky, A; Goel, S; Wilson, D; Hladik, W; Hakim, A; Frost, SD (2012) Respondent driven sampling–where we are and where should we be going? Sexually transmitted infections, 88 (6). pp. 397-9. ISSN 1368-4973

Downloaded from: http://researchonline.lshtm.ac.uk/305977/

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: Creative Commons Attribution Non-commercial No Derivatives http://creativecommons.org/licenses/by-nc-nd/2.5/
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 Frost6

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 and sampling concerns.2,3 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 has primarily been used to estimate the 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 and sampling concerns.2,3 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.4,5

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.4,5,6,7,8 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 (see5,7) also have 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,5,9 social network data collection,6,9,10 population size estimation,7,9 and implementation of interventions.8,10 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
Table 1 Summary of proposed additional information to be reported for RDS studies (STROBE-RDS study reporting checklist)

<table>
<thead>
<tr>
<th>Proposed additional information to be reported for RDS studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Methods</strong></td>
</tr>
<tr>
<td>Study design</td>
</tr>
<tr>
<td>Setting</td>
</tr>
<tr>
<td>Participants</td>
</tr>
<tr>
<td>Variables</td>
</tr>
<tr>
<td>Data sources/measurement</td>
</tr>
<tr>
<td>Statistical methods</td>
</tr>
</tbody>
</table>

**Results**

Participants

Main results

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.

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 wording of network size question(s) Give number, types (eg, mobile/static) and location of recruitment venue(s) Give reasons for non-participation at each stage, including reason for coupon rejection. Report number of recruiters by seed and number of RDS recruitment waves. Report number of recruits by seed and number of RDS recruitment waves. Report adjusted estimates and their stated precision (eg, 95% CI) Report number of recruits by seed and number of RDS recruitment waves. Report wording of network size question(s) State if seeds included in each analysis. |

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. |
REFERENCES