Author’s response to reviews

Title: Unweighted regression models perform better than weighted regression techniques for respondent-driven sampling data: Results from a simulation study

Authors:

Lisa Avery (lavery@maths.otago.ac.nz)
Nooshin Rotondi (rotondin@smh.ca)
Constance McKnight (CMcKnight@dahac.ca)
Michelle Firestone (firestonem@smh.ca)
Janet Smylie (janet.smylie@utoronto.ca)
Michael Rotondi (mrotondi@yorku.ca)

Version: 1 Date: 05 Jun 2019

Author’s response to reviews:

Dear Dr. Salim,

Thank you for the opportunity to revise our manuscript. We appreciate your detailed and thoughtful feedback and feel that this new version has improved significantly. Our responses to your comments, and those of the reviewers are included in bold-face.

Response to Reviewer & Editor Comments

Editor Comments:

In addition to comments from the reviewers, I have several major comments that would need to be addressed:

1) I found it surprising that the unweighted analysis gives better results than weighted analysis. But perhaps it's because in the authors' simulation, the outcome variable is not correlated with the network size and hence not correlated with sampling probability. In the Discussion section, the authors mentioned that this is an interesting aspect to pursue in the future. I strongly feel that the authors need to extend their simulation to this setting. Not only because network size-dependent outcome is a real possibility in practice but it's also because others (eg., Pfeffermann 1996) have found inverse probability weighting (IPW) to be most beneficial in such cases.

Reference:


We agree that the correlation may be an important concern, but given the large number of parameters to investigate in RDS (in particular, homophily), we did not focus on it initially. Nonetheless, in light of your comments, we are happy to investigate it now and have undertaken additional simulations:

.
To examine the effect of a correlation between the outcome and network size we re-assigned the group allocations for each of our 12 simulated populations in four different ways. The prevalence level was unchanged, so, for instance, in the populations with prevalence of 10% there were still 10% x 10,000 = 1000 people with outcome = 1, but these outcomes were assigned as follows:

1) Outcomes of 1 were assigned to the individuals with the highest network size. This group was called extreme positive correlation.

2) Starting with the top decile of network size, 50% more individuals were assigned outcome = 1 than would be expected if outcome and network size were uncorrelated. This was repeated with successive deciles until enough individuals had been allocated to the outcome =1 group. This was called moderate positive correlation.

3) As with #2, but assigning groups starting with the bottom decile of network size. This was called moderate negative correlation.

4) As with #1, but outcome=1 was assigned to individuals with the lowest reported network size. This was called extreme negative correlation.

Our results (shown below) indicated that type-I error was still unacceptable for the weighted analysis. However, this procedure nullified the homophily, so a new set of eight additional populations were created, with prevalence 10%, and homophily varied at 1.25 and 1.5, for each of the four outcomes described above.

The error rate using weighted regression is still unacceptable. This is unsurprising, because degree, though related to outcome (as may be expected with HIV infection) is not associated with the relationship between the outcome and the predictor. This second analysis has been presented in the revised manuscript and we feel that it is indeed an important point to highlight. Please note that we did find that inverse probability weighting (which is what RDS-II weights are) was an effective method of computing the outcome prevalence in our study.

2) In Figure 1, it would seem that the sampling of the blue points are network-dependent but the red points are not. My understanding is that the authors setup the simulation so that the sampling for both is network-dependent? I think it is worth providing the readers with evidence that the sampling in the simulation achieves what the authors set out for and in particular whether the RDS-II weight is really good proxy of the reciprocal of the actual sampling probability here.

In Figure 1, the red dots indicate the seeds. For each sample, these seeds were randomly selected from the networked populations. The sampling of the seeds is therefore not network dependent, which is why they tend to be excluded from RDS analyses. RDS-simulated samples were then drawn from the population, so that the first wave of recruitment was from individuals networked to the seeds. The blue dots are individuals with outcome=1. We have amended the figure to clearly show the individuals with outcome=0, which are open circles. We have calculated the correlation between the network degree of each individual in the population, and the number of times the individual was sampled and added this information to Table 1. if each participant is sampled with a probability proportional to network degree, then these correlations should be near 1. In addition, we found that the RDS-II weights enabled accurate estimation of the outcome prevalence.

3) The authors mention about difficulty with large weights. Have the authors tried a truncated weight and how does that affect the results?
We have not attempted to modify the weights in this study, though we do feel that this is a useful area of future research. Our rationale for not truncating the weights in this study is that, because data were simulated, network size was both accurately and precisely known and was therefore a true indication of the sampling probability.

Reviewer reports:

Mohammad A. Mansournia (Reviewer 1):

1) I suggest that the authors use mean square error (MSE) for comparison of different models. While we initially calculated the MSE we ultimately chose to report bias and coverage probability because we felt these measures were more informative than MSE. Our primary concern was that MSE doesn’t provide information on the direction of the bias and we felt this was necessary for this stage of the investigation as the magnitude and direction of bias can be inconsistent in RDS estimators.

2) P. 8, line 24: Please clarify how the weight is related to inverse probability weights (IPW) given unequal probability of sampling. Please cite the following paper for IPW:

http://www.bmj.com/content/352/bmj.i189.long

We have clarified the text to make clear that the RDS-II weights are inverse probability weights.

3) I suggest that the authors discuss the importance of weighting and adjustment for clustering when associations are of interest using previous experiences in the analysis of complex surveys e.g., see the following paper:

ajph.aphapublications.org/doi/10.2105/AJPH.81.9.1166

Weighting is indeed important for the estimation of disease prevalence in these complex, RDS surveys and numerous papers have been written about this, including an excellent review by Giles (http://www.annualreviews.org/doi/10.1146/annurev-statistics-031017-100704 )

However, our results indicate that weighted regression, as a means of estimating relationships between outcomes and predictors is more problematic because those with low degree act as leverage points leading to inaccurate results and inflated type I error rates. Note that clustering plays a role, but this is less important in the context of regression models than in the estimation of prevalence as the impact of clustering is reduced.

Luis Rocha, PhD (Reviewer 2): This is an important study given that (I agree) the use of regression models in RDS data has increased in the medical literature however standards and routines are not well-tested and well-established. The authors consider various parameters and models, covering diverse practical scenarios, that is positive. For the same reason, I think the presentation of the paper should improve to facilitate understanding the rationale and the different analysis.

Some suggestions and comments:
I suggest that authors create for example a diagram to guide readers on the various cases (particularly for section "Data Analysis").

Thank you for your comments and feedback, we have included a flow diagram to clarify our simulation process. This is included at the end of the cover letter, and we suggest that it should be included as an online supplement.

It is a bit confusing which tables and which graphs go to main text or appendix. We have tried to better clarify this in the revision.

In p6, it would be interesting to report the homophily of the original real network for reference. We agree that it would be interesting and useful information. However, recent work by Crawford et al shows that estimates of both homophily and preferential recruitment in RDS samples to be non-identifiable and furthermore that the identification intervals are too large to be of use. Therefore, we have not provided the homophily estimate for the motivating sample.

Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling
Forrest W Crawford Peter M Aronow Li Zeng Jianghong Li

In p6, row 23-24. By more variation, do authors mean that they control the "variance" of the degree distribution, keeping the mean/median? We have clarified this in the text. By ‘variation’ we meant that, instead of using only the network sizes reported in the OHC study, which tended to be multiples of 5, we instead used a distribution containing a greater variety of network sizes.

In p6, row 34. By population, do authors mean "a fixed set of parameters"? No, we simulated populations from a fixed set of parameters. These populations were then networked, given our desired level of homophily and samples were drawn from these populations using a respondent-driven sampling strategy. Hopefully our study flow diagram helps to clarify this.

In p6, row 34. Though "data sets" are correct, nicer to use "samples" since you are generating samples from the same set of parameters. Agreed – we have changed this in the text.

In p8, row 14. Not sure what you mean by knowing "precisely the degree". One may know the degree distribution with a "1-person" resolution but this does not mean that the degree (or network size) was accurately reported in the survey. Several studies have fine-tuned numbers (in contrast to ranges, that seems to be the point of the authors).
In this study, because network data was simulated, and sample data drawn from those networks we do have access to the actual network degree and we have used this as the ‘reported degree’ in our samples.

In p12, row 8. I think there is a typo in the sentence "prevalence, as shown in Figure 2. 8 in Appendix B details the"?
Yes, corrected, thank-you – the second sentence reads Table 8 in Appendix B...

In p12, row 51. This is minor but I would rather write "suggest" or "recommend" instead of "prefer".
Agreed – we have changed this in the text.

In p16, row 46. I think here the authors could make a strong conclusion and use "recommend" or "suggest" rather than "feel".
Agreed – we have changed this in the text.

In p16, row 44. What do you mean by "very low prevalence"? 10% or less? 10% is not very low.
This was a general statement about the relationship between the odds ratio and the relative risk. We have amended our phrasing

In p16, row 42. Not sure what authors mean by saying that prevalence is more likely to be known than homophily in hard-to-reach-populations. Homophily can be estimated using RDS.
Again, we refer to the paper by Crawford et al. that indicates that homophily can not be estimated with any great confidence from RDS samples. Also, in RDS samples homophily is conflated with preferential recruitment, so can not be estimated as accurately as prevalence.

In p16, row 17. Not sure I agree with "nice" number. Rounding does happen but collecting "non-nice" numbers is common as well. I am not aware of studies pointing that rounding (to 5, 10, 20 and 100) is the standard. In your Fig3, these numbers are not always the most relevant choice.
This statement was based on the distribution of the reported network sizes in the OHC study. We have corrected our statement in the text from “whereas in actual RDS studies we know that people round the reported degree to the nearest ‘nice’ number” to “whereas in the OHC study we observed ‘heaping’, the tendency for people to report degree in clusters (such as 5, 10, 20, 100).

In Fig.4. I am not sure why this degree distribution does not resemble much Fig.3? Specially this valley between degrees 30 and 40 that looks unconventional. Could you comment on that? Have you tried any theoretical cases using for example the log-normal distribution with same mean (possibly standard deviation) as the real data?
Our goal was to try to mimic the observed distribution, but to substantially reduce the ‘heaping’, the tendency for the reported degree to be multiples of 5 or 10. The valley was simply a reflection of what we observed in the real data. We actually just started experimenting with using a log-normal distribution, which does a fairly good job of re-creating our observed population, and is also easier to
In Table 3, I am not sure I understand the systematic under-estimation of the prevalence using RDS-II (and it's worse estimation than the naive method), that is as high as 20% lower for 10% homophily. Is this a result of some particularities of your network model?

We did investigate this. It seems that the underestimation at low prevalence is caused by the degree distribution (which as you point out is essentially log-normal). The lower the prevalence, the more likely people with low reported degree (and therefore high influence) will belong to the ‘not diseased’ group and therefore, the prevalence estimate is biased toward the null value, while at prevalence of 50% high influence participants are just as likely to be diseased as healthy, and so the bias disappears.

The word "cluster" has multiple meanings in social networks, it may be related to clustering of traits (homophily) or links (either triangles or network communities). They affect RDS in different ways (see e.g. Rocha et al. (2017) Respondent-driven sampling bias induced by community structure and response rates in social networks. J Royal Statistical Society A). The authors should clarify these differences in the introduction. I understand that the study focuses on homophily but it would be interesting to include a short discussion on how clustering of links may affect the results (but I don't think new analysis should be done since this would be too much).

We have tried to clarify that our study involved a single community and did not look at network communities. We haven’t included a discussion of how clustering of links may affect the results, because we feel that our main message is that high-influence people can bias the results and we suspect that would be true regardless of the clustering mechanism. We also added a sentence to clarify this in the Introduction.

Thank you again for the detailed comments and the opportunity to revise our manuscript.

Sincerely,

Lisa Avery

Study Workflow Diagram