INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
2023, VOL. 26, NO. 3, 319–333
https://doi.org/10.1080/13645579.2022.2031153

OPEN ACCESS

Assessing logistic regression applied to respondent-driven sampling studies: a simulation study with an application to empirical data

Sandro Sperandei[a], Leonardo Soares Bastos[b], Marcelo Ribeiro-Alves[c], Arianne Reis[d] and Francisco Inácio Bastos[e]

[a]Translational Health Research Institute, Western Sydney University, Penrith, Australia; [b]Scientific Computing Program, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil; [c]National Institute of Infectious Diseases Evandro Chagas, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil; [d]School of Health Sciences, Translational Health Research Institute, Western Sydney University, Richmond, Australia; [e]Department of Health Information, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil

ABSTRACT

The aim of this study is to investigate the impact of different logistic regression estimators applied to RDS studies via simulation and the analysis of empirical data. Four simulated populations were created with different connectivity characteristics. Each simulated individual received two attributes, one of them associated to the infection process. RDS samples with different sizes were obtained. The observed coverage of three logistic regression estimators were applied to assess the association between the attributes and the infection status. In simulated datasets, unweighted logistic regression estimators emerged as the best option, although all estimators showed a fairly good performance. In the empirical dataset, the performance of weighted estimators presented an unexpected behavior, making them a risky option. The unweighted logistic regression estimator is a reliable option to be applied to RDS samples, with a performance roughly similar to random samples and, therefore, should be the preferred option.

KEYWORDS

Respondent-driven sampling; logistic regression; simulation studies; hard-to-reach populations; statistical methods

Introduction

Respondent-driven sampling (RDS) is a chain-referral sampling method based on the key principle that the best recruiter for a hard-to-reach, marginalized or hidden population is a member of this very population (Heckathorn, 1997). It is similar to snowballing in that it begins with an initial pool of participants who then refer others from their personal networks. However, it reduces common snowballing biases by including a seed recruitment process and estimation methodology that addresses these issues (Johnston & Sabin, 2010). The method's success in recruiting individuals from hard-to-reach populations is well accepted, and major international organizations have advocated its use, including the Centers for Disease Control and Prevention (Lansky & Mastro, 2008) and the World Health Organization (Johnston et al., 2013). Its appeal for research investigating rare and elusive populations is greatly due to its ability to reduce biases in situations where probability sampling is not possible (Johnston & Sabin, 2010).

CONTACT Sandro Sperandei s.sperandei@westernsydney.edu.au Translational Health Research Institute, Western Sydney University, Building 3, Room G.05, Campbelltown Campus, Locked Bag 1797, Penrith, NSW 2571, Australia

As an estimation method, it is based on the assumption that the size of an individual's contact network is related to the probability of this individual being recruited to the sample. For this reason, the accepted procedure is to weight individuals as the inverse value of their network size, resulting in individuals with smaller networks, and therefore less likely to being recruited, receiving higher weights in prevalence studies (Gile et al., 2015).

The performance of RDS prevalence estimators has been assessed in many studies, using different methods, particularly simulations (Goel & Salganik, 2010; Mills et al., 2014), with varying results.undefined In general, studies have shown an intermediate to high performance of RDS prevalence estimators (Mills et al., 2014; Rocha et al., 2016; Sperandei et al., 2018). However, almost all currently proposed estimators for RDS-driven studies aim to estimate the prevalence of a given condition in the population of interest. The identification of factors associated with that condition has been seldom addressed. In order to address this, (Bastos et al., 2018) proposed a model-based estimator, called RDS-B, which can be used to estimate both prevalence and associated factors. Notwithstanding the capacity of RDS-B to estimate associated factors, the authors used the estimator in its simplest form to estimate prevalence and did not fully address the underlying characteristics of the modeling process.

Several researchers who have analyzed RDS-based datasets have applied simple logistic regression estimators to assess the putative association between covariates and outcomes, irrespective of the varied study designs and the very characteristics of the method, especially the underlying network structures (Do et al., 2018; Liu et al., 2018; Toro-Tobón et al., 2018). Conversely, others try to use some form of weighted logistic regression, adding weights obtained from actual networks (Hotton et al., 2018; Ndori-Mharadze et al., 2018; Szwarcwald et al., 2018). However, the influence of such sampling weights has not been assessed beyond what has been defined as the basic diagnostic tools to double-check either the sound or improper use of the standard RDS procedures (Gile et al., 2015).

The purpose of this paper is to address this gap in knowledge by assessing the performance of three logistic regression models in estimating RDS-based samples generated by simulations. These estimators were then applied to a real-life RDS sample data from an empirical study on Brazilian transgender women (2,846 participants).

Methods

Simulation

A total of four connected populations (N = 10,000) were simulated using two random graph models, with and without the simulation of nested subpopulations. The random graphs used and the main parameters for each population were as follows:

- Erdös-Rényi random networks without subpopulations (ER1): the simplest random graph structure, initially proposed by Erdös and Rényi (1959), where links between two members of the population were established at random, with a fixed probability (P). P was set at 0.0025;
- Erdös-Rényi random networks with nested subpopulations (ER2): this population is similar to the previous (ER1). However, instead of one population, five subpopulations were nested within the P set at 0.0125. Only ten individuals in each of the five subpopulations were allowed to connect with other subpopulations. They were chosen at random.
- Barabási-Albert scale-free networks without nested subpopulations (BA1): the scale-free model created by Barabási and Albert (1999), also known as the 'richer get richer', follows a power-law distribution for connectivity. In summary, the population starts with one individual and every new individual entering the population has the probability of linking with old members proportionally to the connectivity degree (i.e. number of contacts) of each individual. It generates few

individuals with extremely high connectivity degrees and the majority of the population with few connections. The parameter needed is the number of links each new individual will establish when joining the population. In this simulation, such links were set to 12.5.

- Barabási-Albert scale-free networks with nested subpopulations *(BA2)*: five subpopulations with 2,000 individuals each were generated to construct this population. Subsequently, ten individuals from each subpopulation with the highest degree were chosen to link randomly across these subpopulations.

All parameters were set in order to obtain, whatever the model, a mean connectivity degree of 20, irrespectively of the population under analysis.

Explanatory variables

To assess the performance of logistic estimators emulating actual associations, two binary explanatory variables were added as attributes of each individual, apart from the infected/not-infected status (see infection process below). For the sake of the present study they were designated: E1 and E2. Each one presents 50% randomly distributed positive and negative cases. Over time, the infection process unfolds and each individual in the population with a positive E1 variable will have the chance of being infected two times. The purpose of this mandatory rule is to impose a statistical association between E1 and the disease, whereas E2 will be excluded with any relationship with the putative disease.

Infection processes

Four infection processes were simulated to emulate the dissemination of a particular disease in each of the populations. All processes are variations of the classical Susceptible-Infected (SI) model, where the infected individual does not recover from the disease (i.e. the classic R component is absent). In the first process, individuals were selected at random and defined as 'infected'. The other three processes were dependent on network contacts. All three started with some randomly selected individuals defined as 'infected'. However, unlike the first process, from there on the infection followed through the network contacts in successive waves. In each wave, all individuals connected to the infected ones had a probability of 0.005 to be infected. This infection rate was selected to avoid a surge (i.e. an out of control increase of the infected population). Each newly infected individual could infect their contacts in subsequent waves. All infected individuals kept infecting their contacts until the desired prevalence was reached. The infection prevalence was set at 30%.

Processes started with 10, 100, and 500 infected individuals, creating infections dependent on network connectivity. In the case of 10 initially infected individuals, all those infected 'individuals' were more closely related to the network of the initial individuals, given each individual would generate, on average, an infected 'tree' (or equivalent branching process) of about 300 individuals. In the case of 500 initially infected individuals, there would be a lower network connectivity dependency, with expected 'trees' of only six individuals each. Also, the random process can be considered a particular case, where the process starts with 3,000 infected individuals (prevalence = 30% of 10,000). These processes simulate diseases that depend on interaction between susceptible and infected individuals.

Sampling process

Benchmark samples were obtained based on a simple random process, applied to each combination of population versus simulated infection pattern.

RDS samples were obtained simulating an RDS classic study. All RDS sampling processes were launched using three randomly selected individuals ('seeds'). Each seed recruited randomly from their network one to three contacts, with probabilities of 0.40, 0.40, and 0.20, respectively. These probabilities were based on empirical data from a study with drug users from Belo Horizonte, Brazil (unpublished data). Each recruited individual repeats the process, recruiting additional individuals from their network, and this pattern is repeated until the desired sample size was obtained. It is essential to highlight that, although similar to the infection process previously described, each individual in the population recruits only one to three individuals. In contrast, in the infection process, they keep infecting other individuals until the process reaches a dead end/it´s exhausted.

No homophily-related bias was explicitly incorporated into the recruitment process, although previous studies have suggested that homophily may influence the process (Gile et al., 2015). The simulated samples were designed to reproduce a 'perfect world', following the RDS pristine assumptions as originally proposed by its originator, Douglas Heckathorn. Seeds are recruited randomly, each recruiter recruits randomly among their contacts, no recruitee refuses to participate and all report their network size accurately.

In all cases, 1,000 samples with three sample sizes (100, 250, and 500 individuals) were obtained from each combination of population and infection and applied to all three logistic estimators.

Logistic estimators

Three variations of logistic regression estimators were applied to the abovementioned simulated data. For each one, a model with both variables and interaction was fitted.

The first, used on both RDS and random samples, was the logistic regression estimator (Sperandei, 2014), with the frequentist likelihood estimator. It will be named here the 'unweighted logistic', given the other two estimators are weighted.

The second type of regression, called here 'RDS-weighted logistic', takes into consideration the study design and weightings of each individual using the same form of weighting used in RDS-I and RDS-II estimators (Heckathorn, 1997, 2002; Salganik & Heckathorn, 2004). It weighs results from the simulations proportionally to the inverse of the reported degree (i.e. the number of connections) of each individual (Volz & Heckathorn, 2008).

The third type of regression estimator, called 'RDS-B' (Bastos et al., 2018), is a Bayesian version of the RDS-weighted logistic, where weakly informative priors are set to the coefficients (Gelman et al., 2008), and the weighted likelihood, called pseudo-likelihood, is combined with the prior using Bayes theorem. It yields a pseudo-posterior distribution (Savitsky & Toth, 2015). Posterior means were used in order to make a comparison among estimators, and 95% credible intervals were used to represent uncertainty.

In the case of randomly selected samples, only the unweighted logistic estimator was used, defining a benchmark performance.

Performance assessment

The performance assessment was accomplished by the observed coverage metric, also known as coverage probability (Dodge et al., 2003). This is the proportion of times the confidence interval of each estimator contains the populational parameters simulated. It means that, for the coefficient of E1, the OR confidence interval contains the parameter 2 simulated for each population. For this coefficient, the confidence interval also needs to exclude the value of 1, meaning a significant coefficient. The rationale for this second criterion is to avoid too wide confidence intervals might be improperly considered as indicators of an acceptable performance. For the coefficients of E2 and the interaction E1xE2, the OR confidence interval must contain the value of 1, meaning a non-significant interval, which is the simulated situation. For these two coefficients, the complementary probability (1 – coverage) will be used as an estimate of type-I error probability. Finally, a combination of E1, E2, and interaction results will be built to investigate the probability of a combined correct estimation from the model, meaning a significant E1 coefficient and non-significant coefficients for E2 and interaction E1 x E2. The word 'significant' here was used in a broad sense, related to the usual 95% confidence interval, although we acknowledge that in Bayesian models these definitions are not strictly adequate.

All performances were compared to the random samples' performance for each combination of population and infection.

Real-Life, empirical data

All four estimators were subsequently applied to the Divas Research dataset (Bastos et al., 201 8undefined), which is a large RDS-based study conducted across 12 cities in Brazil that collected data from 2,846 transgender women.

The entire dataset was merged and considered for the sake of the present study as one population, from where the expected parameters were estimated. Four variables were considered in this study to assess the performance of the estimators. HIV status (positive x negative) was considered the main outcome. The two explanatory variables considered were whether the person had acted as a sexual worker anytime in their life (explanatory variable 1 – E1) and whether the person had moved from their place of birth anytime during life (E2). E1 is expected to be related to HIV status, while E2 is not. The fourth variable was the reported number of contacts (network degree), which was used in RDS estimations. A total of 2,548 individuals were used to avoid missing information in any of the variables considered.

From this population, samples were extracted with sizes of 100, 250, and 500 individuals. First, 1,000 random samples of each sample size were used as benchmarks, similar to what was done in the simulation. Second, 1,000 samples were drawn following the RDS process. As the objective here is to observe the impact of real-world constraints and bottlenecks in the sampling procedure, these samples were extracted respecting the original RDS sampling from the dataset. Real seeds were randomly selected and the original recruitment trees were followed from each seed until the desired sample size was reached. By doing this, each sample used was a subsample of the original dataset, presenting all the characteristics found in real-life sampling.

Again, similarly to the process used in the simulation, the sample results were compared to the observed result from the population, and the number of correct estimations was counted.

The Divas study received ethics approval from the Escola Nacional de Saúde Publica (CAAE 49359415.9.0000.5240). All participants signed an informed consent form to take part in the study. The dataset was provided in an unidentified form and no additional approval was necessary for the current study.

All simulations and analyses used R software, version 3.4.4 (R Core Team, 2021) and its libraries *igraph* (Csardi & Nepusz, 2006)), *survey* (Lumley, 2004), and *arm* (Gelman & Su, 2018).

Results

Results of the simulated populations can be seen in Figure 1. Red dots represent infected individuals, while blue dots represent non-infected individuals. A considerably different pattern can be noted between the two random graph models used and an even more dramatic effect between clustered and non-clustered populations. Comparing ER and BA networks, it is clear that highly connected individuals, located on the external borders of the figure, have a higher chance of becoming infected in the BA model. In ER models, as the distribution of degrees does not present heavy tails, the infection is more uniformly spread. The same pattern can be observed in models with subpopulations well defined, with one additional characteristic: the clustered nature of these models resulted in parts of the population being almost untouched by infection.

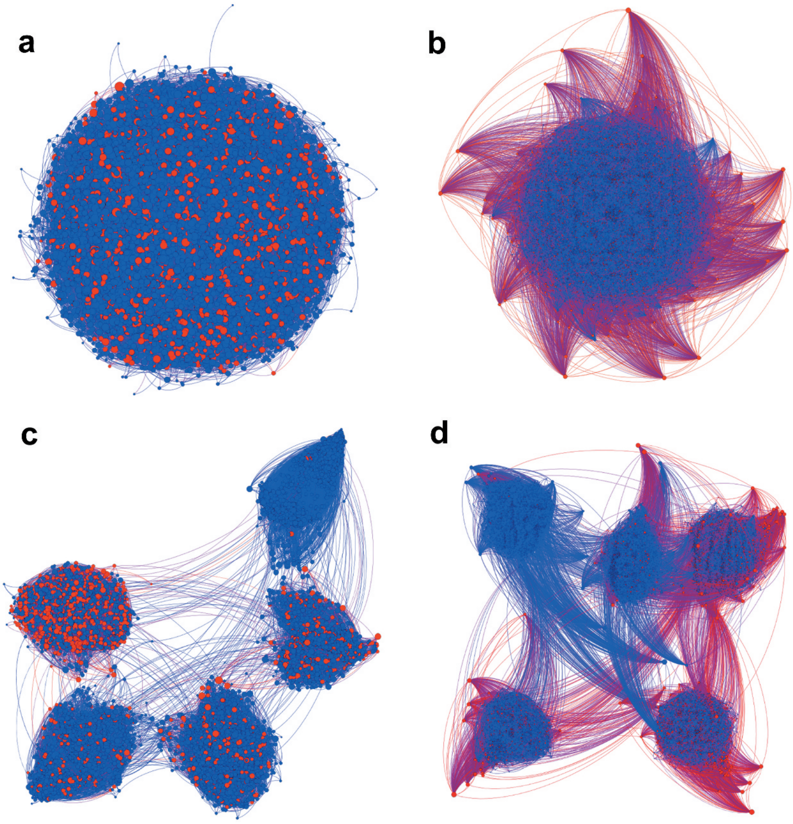

Figure 1. Populations created. Blue vertices and edges are for non-infected individuals. Red vertices and edges are for infected individuals. A: ER1 model. B: BA1 model. C: ER2 model. D: BA2 model.

Table 1 presents the main characteristics of each simulated population as well as the Divas dataset. It can be noted that all main characteristics were successfully simulated. The Barabási-Albert models showed a discrepancy between the average and the median degree due to the asymmetric nature of the model degree's distribution.

The simulated prevalence ranged from 14.6% (ER2) to 17.2% (BA1), very close to the desired value (15%). Regarding true ORs observed in the population, general logistic models fitted to the whole population (one for each population) detected significant ORs for variable E1, all between 1.95 and 2.05, after adjusting for E2 and the interaction. For variable E2, true ORs ranged from 0.81

Table 1. Main characteristics of simulated and Divas populations.

	Population				
Characteristic	ER1	ER2	BA1	BA2	Divas
Mean Degree	20.03	19.95	19.99	19.95	20.21
Median Degree	20.0	20.0	14.0	14.0	10.0
Min – Max Degree	4–37	5–39	10–541	10–247	2–100
Infection Prevalence (%)	30.2–31.6*	30.4–31.9*	30.0–33.0*	29.8–32.5*	29.98
E1 Prevalence (%)	50	50	50	50	76.4
E2 Prevalence (%)	50	50	50	50	60.9
E1 Odds Ratio	1.97–2.00*	1.98–2.05*	1.97–2.05*	1.98–2.04*	1.83
E2 Odds Ratio	0.90–1.01*	0.85–1.10*	0.76–1.09*	0.83–1.04*	1.26
E3 Odds Ratio	0.95–1.33*	0.94–1.21*	0.92–1.45*	0.98–1.40*	1.31

* Values represent the minimum and maximum range across the four types of infection

to 1.10, all of them non-significant, as expected. Lastly, for the interaction factor (E1xE2), true ORs varied from 0.90 (ER1) to 1.20 (BA2). These results confirm the simulation process was adequate. Regarding the Divas population, a pattern towards a power-law distribution of connectivity and a clustered behavior is expected, given the way the population was created, joining samples from twelve cities. This means that no individual will recruit out of their own city. Overall, the Divas dataset was most similar to the BA2 simulated population.

Figure 2 presents observed coverage probability results according to the network model, infection process, sample size, and estimators used for coefficient E1 alone. The most evident effect was related to the sample size. The bigger the sample size, the higher the coverage. Regarding estimators themselves, three of them had similar performances, with a slightly better performance obtained by the traditional logistic estimator applied to RDS samples. The estimator with the worst performance was the weighted-logistic estimator. However, even this estimator did not perform substantially below the logistic estimator applied to random samples (taken here as the benchmark) and could be considered a satisfactory estimator. In regards to the effect of network models, it can be observed that populations without heavy tails in the distribution of degrees (ER1 and ER2) present very small difference between estimators, while heavy tail distributions of degree inside the population (BA1 and BA2) seems to affect heavily the weighted estimators (RDS and Bayes) and favor the unweighted estimator applied to RDS samples. The presence of subpopulations (ER2 and BA2) had little to no effect on the estimators' performance for E1 or the analysis of the combined coefficients. Lastly, it is interesting to note that, in Barabási-Albert model-based populations, the unweighted estimator applied to RDS samples presented a better performance when the infection was not random even when compared to random samples.

The Type-I error rate for the combined coefficients shows a general trend for an addictive effect, showing a certain independence between the coefficients error (Figure 3). Irrespective of the type of infection, sample size, network model or estimator, the probability for both coefficients was close to the expected value of 5%, appearing close to 10% due to addictive effect. Only for BA networks, under random infection, with n = 500 (and to a lesser extent with n = 250), the error rate was above this threshold, especially for the unweighted estimator applied to RDS samples.

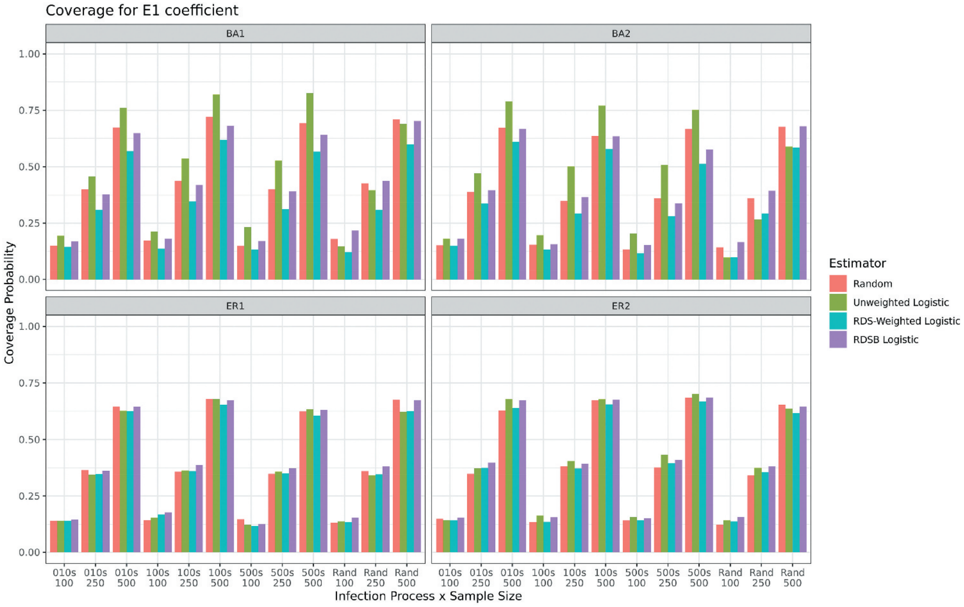

Figure 2. Observed coverage probability results according to the combination of network models (each subgraph, as labelled), sample size (100, 250, 500) and infection process (10s, 100s, 500s, Rand).

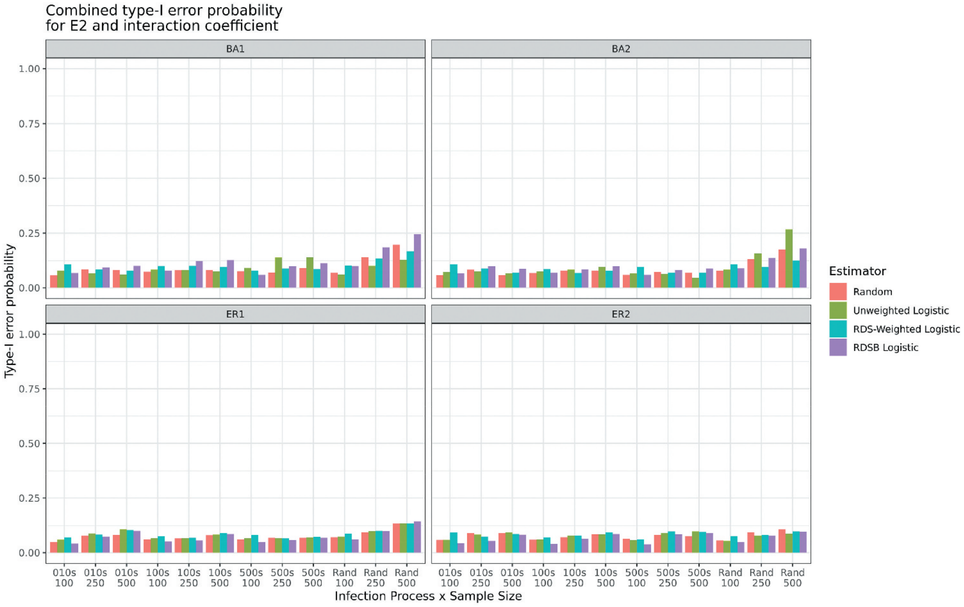

Figure 3. Type-I error rate for the E2 and the interaction coefficients according to network models (each subgraph, as labelled), sample size (100, 250, 500) and infection process (10s, 100s, 500s, Rand).

When the analyses of all three coefficients are combined, it is possible to notice the general performance of the estimators to find the 'right answer' from the samples: a significant E1 coefficient with a confidence interval containing the simulated E1 effect plus non-significant E2 and interaction coefficients. Figure 4 illustrates how results are very similar to those for the E1 coefficient, given the general stability of E2 and interaction results. The results for the random infection were the most affected, especially by the higher type-I error rate.

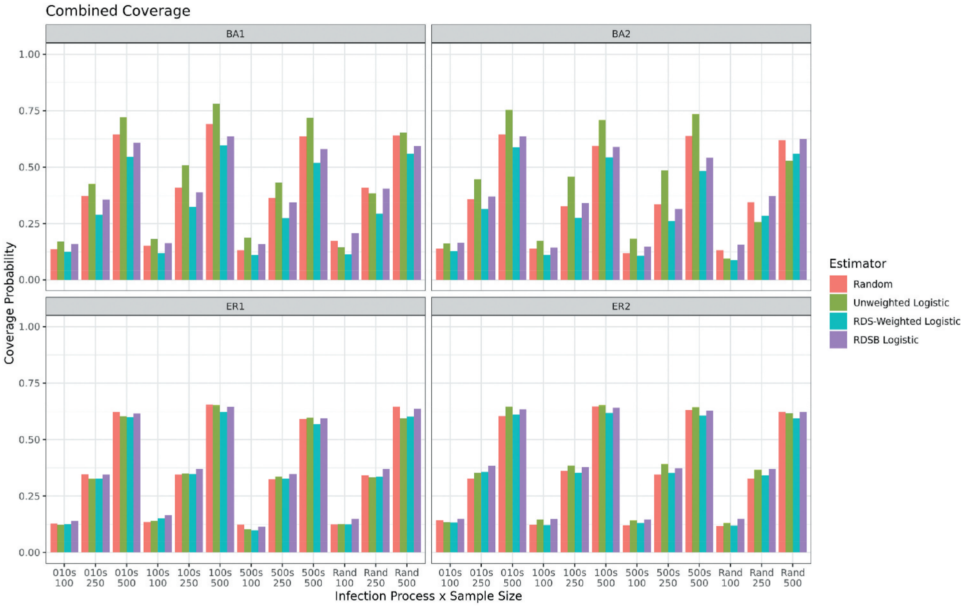

Figure 4. Observed coverage probability results for the combination of coefficients according to network models (each subgraph, as labelled), sample size (100, 250, 500) and infection process (10s, 100s, 500s, Rand).

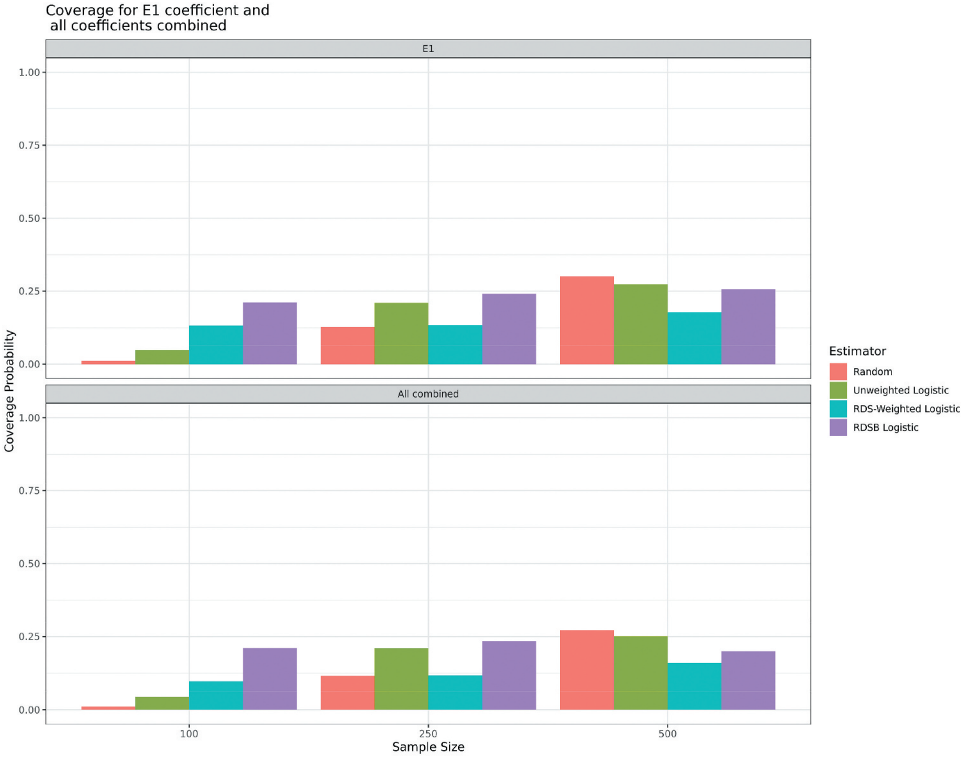

Figure 5. Observed coverage probability for E1 and all coefficients combined according to sample size and estimator.

An exciting result was observed when the estimators were applied to the Divas dataset. The random samples behaved as expected, with a proportional increase in coverage for the E1 coefficient according to the sample size (Figure 5). In the same way, the unweighted estimator presented a similar behavior when applied to RDS samples compared to random samples. However, weighted estimators presented a somewhat strange behavior, with unusual high coverage for smaller samples (compared to random), and smaller improvements with increasing size, especially the RDS-B, which demonstrated a drop when the sample reached 500 individuals. This pattern was the same for the combination of all coefficients.

When looking at the type-I error rate (Figure 6), they were well below the expected for the sample size of 100 and around 5% for the unweighted logistic estimator, either applied to random or RDS samples. The weighted estimators showed a higher error rate, especially for the RDS-weighted logistic estimator, which reached more than 40% with sample size of 100. This may correspond to a high probability of obtaining wrong results when using this estimator.

Discussion

The RDS method has been widely used and recommended as a sampling method to recruit hard-to-reach populations, such as people who use substances, sex workers, transgender individuals, among others (Marpsat & Razafindratsima, 2010). Although its ability to find and recruit members of these ‘hidden’ populations is uncontroversial, its use as an estimator method is still disputed (Sperandei et al., 2018). Moreover, the use of model-based estimators to study relationships between response and explanatory variables has been poorly assessed, especially in regard to the basic question of when to use sampling weightings (Schonlau & Liebau, 2012). These issues notwithstanding, researchers have used traditional logistic estimators or some form of weighted logistic applied to

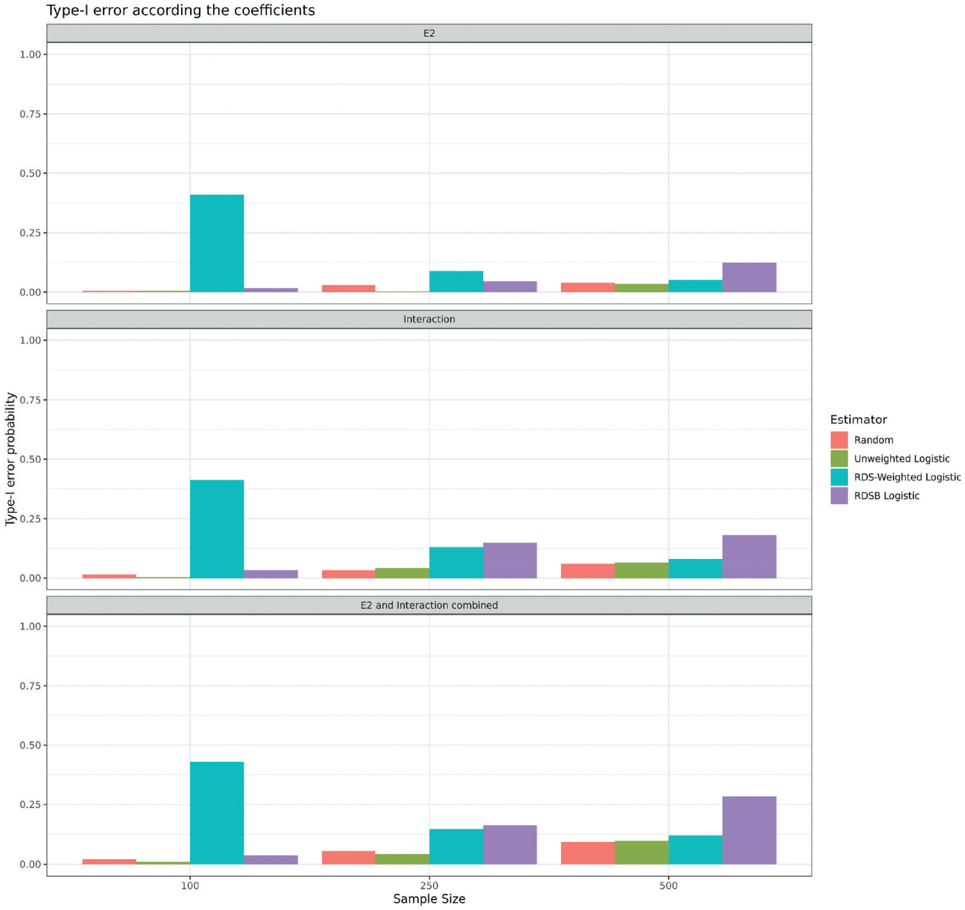

Figure 6. Type-I error rate for the E2 and the interaction coefficients according to sample size and estimator.

RDS samples. A quick survey of the Pubmed database identified 70 studies published between 2018 and 2019 applying logistic regression models to RDS samples, with 48.6% (n = 34) using unweighted estimators, 44.3% (n = 31) using some form of weighting with network degrees, and 7.1% (n = 5) presenting both weighted and unweighted models. This pattern highlights the evident lack of consensus in the current literature on which type of estimator should be used. The current study addressed this issue by testing which of the models currently being employed in studies using the RDS method is most effective in assessing associations between variables. In addition, the current study provided the degree of reliability of different models in order to equip researchers with better tools to evaluate results of RDS studies.

Our simulations have demonstrated not only the impact of data and population characteristics but also the estimator used on results of an RDS study. Here it is important to note that our results do not suggest that RDS is as good as a random sample in representing the characteristics of the population. In fact, previous studies have identified the limitations of RDS in achieving this (Sperandei et al., 2018). However, results from the simulation, respecting all assumptions of the RDS method, as well as real RDS data suggest that unweighted logistic regression is an effective method for estimating associations between variables. Clearly, the more the data moves away from the method's assumptions, the less effective the performance of the estimators will be – which would be true for any estimator used –, and this was demonstrated in the results when the models were applied to real data. And here lies an important contribution of the current study: it provides an overview of what would happen in 'the real world' where the methodological assumptions of RDS are not always followed, for the most different reasons. Although some interactions with other factors must be considered, it seems that weighted and unweighted estimators performed relatively well when compared to logistic regression applied to random samples.

To the best of our knowledge, only one study assessed the impact of weights used on RDS sample estimates from logistic regression models (or any other form of model estimates) using a simulation approach, and, similarly to our results, it concluded that unweighted estimators perform better than the weighted ones (Avery et al., 2019). However, the lack of a clear structure in the simulated network connections and the absence of real data to reflect real sampling problems, in comparison to 'perfect' simulated samples, left many issues unaddressed. First, Avery et al.'s (2019) study used only simple logistic models, with just one explanatory variable, not considering the effect of interaction between explanatory variables on the result. Second, this study confounded clustering with homophily, when they are, in fact, different concepts (Rocha et al., 2016; Sperandei et al., 2018). Clustering represents the phenomenon of individuals being more connected to their similar ones (in one or more characteristics such as age, geography, etc.), whereas homophily relates to preferential recruitment, where people choose to recruit those peers with particular characteristics (that the recruiter may also possess or with which they have a close relation), instead of recruiting randomly (Lu et al., 2012). In the present study, we addressed these limitations by creating populations based on theoretical graph models, controlling the connectivity process. From the results, comparing the two models used here, it is clear the impact of the nature of connectivity on the performance of estimators, which is reinforced by previous research on simple prevalence estimators (Rocha et al., 2016; Sperandei et al., 2018). Of course, this theoretical approach falls short of real life challenges, as demonstrated by classic ethnographic studies such as the comprehensive mapping of actual social networks of people who sell, share and use substances, in Bushwick, Brooklyn, NYC, USA (Friedman et al., 2006).

In addition, we used an adapted concept of 'coverage probability' to reflect not only the identification of correct estimation of the E1 coefficient but also the simultaneous identification of E2 and the E1xE2 interaction, representing the proportion of correct estimation for the complete hypothesis. It represents a more restrictive criterion compared to the usual coverage because it requires all three hypotheses (E1, E2, E1xE2) being true at the same time.

In this dataset, the random samples acted as a benchmark to what would be expected, given that, for any population, random samples are considered the gold standard sampling method. The results show the expected increase in coverage according to the sample size. The most exciting finding was the performance of the unweighted logistic estimator applied to RDS samples, which showed similar results compared to random samples, sometimes even better. The results with real data represent a decreased performance in comparison to simulation results, showing the effects of differences between theoretical sampling procedures and real ones; however, it still performs well and is a good alternative to be used with RDS samples, similarly to what Avery et al. (2019) found.

On the other hand, weighted estimators presented a somewhat aberrant behavior, especially the RDS-B, which presented higher coverage with smaller samples. At a lower intensity, the RDS-weighted estimator also showed an unexpectedly high power with the 100 samples, but the increase with bigger sample sizes was not so considerable. This behavior, also partially observed in the performance of unweighted logistic, is probably related to the differences in simulated and real sampling procedures. In relation to the type-I error rate, the RDS-weighted estimator showed a very high result, representing a big chance of a wrong result.

Several studies have demonstrated the advantages of weighting procedures for the simple prevalence of RDS estimators (Goel & Salganik, 2010; Mills et al., 2014; Sperandei et al., 2018). However, the present results demonstrate that weighting may not be the best option when it comes to regression coefficient estimates, making the unweighted estimator the preferable one instead.

Strengths and limitations

Simulations can only approximate the characteristics of the real world, their success being dependent on previous knowledge about the population being simulated. Models must reduce the complexity and dimensionality of phenomena under analysis in order to make them amenable to different analytic strategies and the interpretation of their findings (Weisberg, 2012). This knowledge, in the case of hard-to-reach populations, can be very restricted. The use of real data allowed us to observe what happens when RDS is applied in the real world.

In our simulated scenarios, RDS sampling followed best practices described for the method, with putative random selection of seeds (usually absent in the vast majority of empirical studies), long recruitment trees, and each recruiter 'selecting' randomly amongst their peers (Salganik & Heckathorn, 2004; Volz & Heckathorn, 2008). In practice, it is common to see 'non-generative seeds' (i.e. seeds that do not recruit any peers [Reisner et al., 2010]), recruitment trees with mixed length, and true homophily, with recruiters choosing selectively amongst their peers (Li et al., 2018). Also, time, resource, and logistical constraints are common, and their impacts on estimation are unknown (Truong et al., 2013; Valois-Santos et al., 2020). Considering a large sample as a population, and using real recruitment trees as RDS samples, is not a perfect approach. However, we argue that it is one of the best possible ways of assessing RDS estimators in real life. In addition, these limitations notwithstanding, the simulation presented here confirms the capacity of RDS sampling, with the use of unweighted estimators, to perform exceptionally well in circumstances where probability sampling is not possible, which is frequently the case among rare and elusive populations.

Conclusion

In summary, this study suggests that unweighted logistic regression is the best option to be used with RDS samples, particularly when the basic assumptions of the RDS method are duly respected (what may or may not be observed in the field). Indeed, even in real RDS samples, it may achieve a performance that is surprisingly equivalent to the random sampling performance in assessing associations. These findings suggest that the RDS method is applicable to a broader spectrum of research designs, even where true random sampling is nothing but an elusive goal. This therefore goes beyond hard-to-reach or elusive populations to include studies with other population groups where random sampling of participants may not be feasible.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

SS was the recipient of a postdoctoral scholarship from Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Coordenação de Aperfeiç oamento de Pessoal de Nível Superior (CAPES).

Notes on contributors

Dr Sandro Sperandei is a Lecturer in Epidemiology with a PhD in Computational Biology and Systems. He has expertise in statistical simulation and analysis. Dr Sperandei focus of research is in suicide prevention, mental health, and marginalised populations.

Dr Leonardo Soares Bastos is a public health researcher at Oswaldo Cruz Foundation with a PhD in Statistics. He hasexpertise in Bayesian statistical modelling and infectious disease epidemiology.

Dr Marcelo Ribeiro-Alves has a PhD in Biomedical Engineering and is a public health technologist at the National Institute of Infectious Diseases Evandro Chagas (INI/FIOCRUZ), associated with the Laboratory of Clinical Research in STD/AIDS, where he develops research related to epidemiology, pharmacogenetics/pharmacogenomics of antiretroviral treatment of people living with HIV/AIDS, with interest in Pattern Recognition in biomolecular, genetic and genomic data, and molecular epidemiology.

Dr Arianne Reis is an Associate Professor in the School of Health Sciences and has training in both quantitative and qualitative research. She has worked in the field of health sciences for more than 20 years and has led several research projects in the field of leisure studies, health promotion and public health, working particularly with vulnerable populations.

Dr Francisco Inácio Bastos is a senior researcher at the Department of Health Information (LIS-ICICT), at the Oswaldo Cruz Foundation (FIOCRUZ). His main interests have been linked with the misuse of substances and associated harms and risks. Dr Bastos has been working with different sampling/estimation methods such as respondent-driven sampling, time-location sampling and network scale-up.

ORCID

Sandro Sperandei http://orcid.org/0000-0001-5367-3397
Leonardo Soares Bastos http://orcid.org/0000-0002-1406-0122
Marcelo Ribeiro-Alves http://orcid.org/0000-0002-8663-3364
Arianne Reis http://orcid.org/0000-0002-1630-8857
Francisco Inácio Bastos http://orcid.org/0000-0001-5970-8896

References

Avery, L., Rotondi, N., McKnight, C., Firestone, M., Smylie, J., & Rotondi, M. (2019). Unweighted regression models perform better than weighted regression techniques for respondent-driven sampling data: Results from a simulation study. *BMC Medical Research Methodology*, *19*(1), 202. https://doi.org/10.1186/s12874-019-0842-5

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science (New York, N.Y.)*, *286*(5439), 509–512. http://www.ncbi.nlm.nih.gov/pubmed/10521342

Bastos, F. I., Bastos, L. S., Coutinho, C., Toledo, L., Mota, J. C., Velasco-de-castro, C. A., Sperandei, S., Brignol, S., Travassos, T. S., Dos Santos, C. M., & Malta, M. S. (2018). HIV, HCV, HBV, and syphilis among transgender women from Brazil. *Medicine*, *97*(1S Suppl 1), S16–S24. https://doi.org/10.1097/MD.0000000000009447

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Sy*. 1695(5) , 1–9 https://www.researchgate.net/profile/Gabor-Csardi/publication/221995787_The_Igraph_Software_Package_for_Complex_Network_Research/links/0c96051d301a30f265000000/The-Igraph-Software-Package-for-Complex-Network-Research.pdf.

Do, T. T. T., Le, M. D., Van Nguyen, T., Tran, B. X., Le, H. T., Nguyen, H. D., Nguyen, L. H., Nguyen, C. T., Tran, T. D., Latkin, C. A., Ho, R. C. M., & Zhang, M. W. B. (2018). Receptiveness and preferences of health-related smartphone applications among Vietnamese youth and young adults. *BMC Public Health*, *18*(1), 764. https://doi.org/10.1186/s12889-018-5641-0

Dodge, Y., Marriott, F. H. C., & International Statistical, I. (2003). *The Oxford dictionary of statistical terms*. Oxford University Press.

Erdös, P., & Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, *6*, 290–297. citeulike-article-id:4012374. http://www.renyi.hu/~p_erdos/Erdos.html#1959-11

Friedman, S. R., Curtis, R., Neaigus, A., Jose, B., & Des Jarlais, D. C. (2006). *Social networks, drug injectors' lives, and HIV/AIDS*. Springer Science & Business Media.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Gelman, A., & Su, Y.-S. (2018) arm: Data analysis using regression and multilevel/hierarchical models. http://CRAN.R-project.org/package=arm.Rpackageversion,1-3.

Gile, K. J., Johnston, L. G., & Salganik, M. J. (2015). Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 241–269. https://doi.org/10.1111/rssa.12059

Goel, S., & Salganik, M. J. (2010). Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(15), 6743–6747. https://doi.org/10.1073/pnas.1000261107

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, *44*(2), 174–199. https://doi.org/10.2307/3096941

Heckathorn, D. D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, *49*(1), 11–34. http://www.respondentdrivensampling.org/reports/RDS2.pdf

Hotton, A., Quinn, K., Schneider, J., & Voisin, D. (2018). Exposure to community violence and substance use among Black men who have sex with men: Examining the role of psychological distress and criminal justice involvement. *AIDS Care 31*(3) , 370–378. https://doi.org/10.1080/09540121.2018.1529294

Johnston, L. G., Chen, Y.-H., Silva-Santisteban, A., & Raymond, H. F. (2013). An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS and Behavior*, *17*(6), 2202–2210. https://doi.org/10.1007/s10461-012-0394-8

Johnston, L. G., & Sabin, K. (2010). Sampling hard-to-reach populations with respondent driven sampling. *Methodological Innovations Online*, *5*(2), 38–48 https://doi.org/10.4256/mio.2010.0017.

Lansky, A., & Mastro, T. D. (2008). Using respondent-driven sampling for behavioural surveillance: Response to Scott. *The International Journal on Drug Policy*, *19*(3), 241–243. discussion 246-247. https://doi.org/10.1016/j.drugpo.2008.03.004

Li, J., Valente, T. W., Shin, H.-S., Weeks, M., Zelenev, A., Moothi, G., Mosher, H., Heimer, R., Robles, E., Palmer, G., & Obidoa, C. (2018). Overlooked threats to respondent driven sampling estimators: Peer recruitment reality, degree measures, and random selection assumption. *AIDS and Behavior*, *22*(7), 2340–2359. https://doi.org/10.1007/s10461-017-1827-1

Liu, Y., Jiang, C., Li, S., Gu, Y., Zhou, Y., An, X., Zhao, L., & Pan, G. (2018). Association of recent gay-related stressful events with depressive symptoms in Chinese men who have sex with men. *BMC Psychiatry*, *18*(1), 217. https://doi.org/10.1186/s12888-018-1787-7

Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B. J., Thorson, A., & Liljeros, F. (2012). The sensitivity of respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(1), 191–216. https://doi.org/10.1111/j.1467-985X.2011.00711.x

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1–19.https://doi.org/10.18637/jss.v009.i08

Marpsat, M., & Razafindratsima, N. (2010). Survey methods for hard-to-reach populations: Introduction to the special issue. *Methodological Innovations Online*, *5*(2), 3–16 https://doi.org/10.4256/mio.2010.0014.

Mills, H. L., Johnson, S., Hickman, M., Jones, N. S., & Colijn, C. (2014). Errors in reported degrees and respondent driven sampling: Implications for bias. *Drug and Alcohol Dependence*, *142*, 120–126. https://doi.org/10.1016/j.drugalcdep.2014.06.015

Ndori-Mharadze, T., Fearon, E., Busza, J., Dirawo, J., Musemburi, S., Davey, C., Acharya, X., Mtetwa, S., Hargreaves, J. R., & Cowan, F. (2018). Changes in engagement in HIV prevention and care services among female sex workers during intensified community mobilization in 3 sites in Zimbabwe, 2011 to 2015. *Journal of the International AIDS Society*, *21*(Suppl Suppl 5), e25138–e25138. https://doi.org/10.1002/jia2.25138

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ .

Reisner, S. L., Mimiaga, M. J., Johnson, C. V., Bland, S., Case, P., Safren, S. A., & Mayer, K. H. (2010). What makes a respondent-driven sampling “seed” productive? Example of finding at-risk Massachusetts men who have sex with men. *Journal of Urban Health*, *87*(3), 467–479 https://doi.org/10.1007/s11524-010-9439-3.

Rocha, L. E. C., Thorson, A. E., Lambiotte, R., & Liljeros, F. (2016). Respondent-driven sampling bias induced by community structure and response rates in social networks. *Journal of the Royal Statistical Society 180*(1), 99–118. *Series A (Statistics in Society)*. https://doi.org/10.1111/rssa.12180

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, *34*(1), 193–240. https://doi.org/10.1111/j.0081-1750.2004.00152.x

Savitsky, T. D., & Toth, D. (2015). *Bayesian estimation under informative sampling*. http://arxiv.org/abs/1507.07050

Schonlau, M., & Liebau, E. (2012). Respondent-driven sampling. *Stata Journal*, *12*(1), 72–93. http://www.stata-journal.com/article.html?article=st0247. http://www.stata-journal.com/sjpdf.html?article=st0247

Sperandei, S., Bastos, L. S., Ribeiro-Alves, M., & Bastos, F. I. (2018). Assessing respondent-driven sampling: A simulation study across different networks. *Social Networks*, *52*, 48–55. https://doi.org/10.1016/j.socnet.2017.05.004

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, *24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Szwarcwald, C. L., Damacena, G. N., de Souza-júnior, P. R. B., Guimarães, M. D. C., de Almeida, W. D. S., de Souza Ferreira, A. P., Ferreira-Júnior, O. D. C., & Dourado, I. (2018). Factors associated with HIV infection among female sex workers in Brazil. *Medicine*, *97*(1S Suppl 1), S54–S61. https://doi.org/10.1097/MD.0000000000009013

Toro-Tobón, D., Berbesi-Fernandez, D., Mateu-Gelabert, P., Segura-Cardona, Á. M., & Montoya-Vélez, L. P. (2018). Prevalence of hepatitis C virus in young people who inject drugs in four Colombian cities: A cross-sectional study using respondent driven sampling. *The International Journal on Drug Policy*, *60*, 56–64. https://doi.org/10.1016/j.drugpo.2018.07.002

Truong, -H.-H. M., Grasso, M., Chen, Y.-H., Kellogg, T. A., Robertson, T., Curotto, A., Steward, W. T., & McFarland, W. (2013). Balancing theory and practice in respondent-driven sampling: A case study of innovations developed to overcome recruitment challenges. *PLOS ONE*, *8*(8), e70344–e70344. https://doi.org/10.1371/journal.pone.0070344

Valois-Santos, N. T., Niquini, R. P., Sperandei, S., Bastos, L. S., Bertoni, N., Brito, A. M. D., & Bastos, F. I. (2020). Reassessing geographic bottlenecks in a respondent-driven sampling based multicity study in Brazil. *Salud Colectiva*, *16*, e2524–e2524. https://doi.org/10.18294/sc.2020.2524

Volz, E., & Heckathorn, D. D. (2008). Probability based estimation theory for respondent-driven sampling. *Journal of Official Statistics*, *24*(1), 79–97.

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.