Indian J Med Res 144, September 2016, pp 447-459
DOI: 10.4103/0971-5916.198665

Quick Response Code:

A single weighting approach to analyze respondent-driven sampling data

Vadivoo Selvaraj[1], Kangusamy Boopathi[1], Ramesh Paranjape[2] & Sanjay Mehendale[1]

[1]National Institute of Epidemiology, Indian Council of Medical Research, TNHB, Ayapakkam, Chennai &
[2]National AIDS Research Institute, Bhosari, Pune, India

Received January 27, 2015

***Background and objectives*: Respondent-driven sampling (RDS) is widely used to sample hidden populations and RDS data are analyzed using specially designed RDS analysis tool (RDSAT). RDSAT estimates parameters such as proportions. Analysis with RDSAT requires separate weight assignment for individual variables even in a single individual; hence, regression analysis is a problem. RDS-analyst is another advanced software that can perform three methods of estimates, namely, successive sampling method, RDS I and RDS II. All of these are in the process of refinement and need special skill to perform analysis. We propose a simple approach to analyze RDS data for comprehensive statistical analysis using any standard statistical software.**

***Methods*: We proposed an approach (RDS-MOD - respondent driven sampling-modified) that determines a single normalized weight (similar to RDS II of Volz-Heckathorn) for each participant. This approach converts the RDS data into clustered data to account the pre-existing relationship between recruits and the recruiters. Further, Taylor's linearization method was proposed for calculating confidence intervals for the estimates. Generalized estimating equation approach was used for regression analysis and parameter estimates of different software were compared.**

***Results*: The parameter estimates such as proportions obtained by our approach were matched with those from currently available special software for RDS data.**

***Interpretation & conclusions*: The proposed weight was comparable to different weights generated by RDSAT. The estimates were comparable to that by RDS II approach. RDS-MOD provided an efficient and easy-to-use method of estimation and regression accounting inter-individual recruits' dependence.**

Key words New approach - regression - respondent-driven sampling - data analysis

It is difficult to map and develop sampling frames for hard-to-reach and hidden populations such as injecting drug users (IDUs) or men having sex with men (MSM) due to privacy concerns and closed knit nature of the groups. Hence, conventional probability-based sampling methods cannot be applied to sample them. Snowball sampling, key informants sampling and targeted sampling are some of the previously described sampling methods for this population[1-3]. However, all these methods have their own limitations and known biases[4].

Slightly modified form of snowball method has been used to count the rare events such as maternal mortality[5].

Respondent-driven sampling (RDS) was introduced by Heckathorn to sample hidden and hard-to-reach populations. RDS is a modified form of snowball sampling, with a system for assigning weights to compensate for the unequal selection probability[4]. RDS starts with identifying prototype individuals known to represent a specific hidden or hard-to-reach population termed as 'seeds'. In turn, seeds recruit the first wave of respondents and then the first-wave respondents recruit the second wave of respondents and such successive 'waves' help recruiting respondents until the desired sample size is reached. Although respondents recruit those with whom they have a pre-existing relationship, the primary expectation is that the respondents recruit randomly from their personal network[6]. The probability of inclusion is derived from the extension of Markov Chain (MC) theory and random walk on the network connecting the target population[7-10]. This theoretical framework forms the basis for calculating unbiased estimates. As the selection of seeds is non-random, the RDS data lack external validity[11]. However, with attainment of more than six waves, the sample composition is expected to stabilize and become independent of seeds[4,12].

The existing software to analyze RDS data is RDS analysis tool (RDSAT), and it can generate only estimators[13]. Currently, RDSAT is in the stage of refinement and evolution. RDSAT uses bootstrapping to obtain confidence intervals (CIs) for estimates[8,14]. Goel and Salganik introduced an MC argument for population mixing[10]. They proposed an estimator by weighting the variable obtained from the size of the participants' network and the network pattern focusing on relationships within the network. However, individualized weights have to be obtained for each variable and incorporated in the estimation procedure. Therefore, only one estimate can be made at a time and hence it consumes more time for data analysis. In addition, regression analysis is not possible with RDSAT. Efforts were made to adopt RDS data to regression analysis for adjusting estimates to reflect the targeted population[12,15]. Exporting of individualized weights of a chosen variable from RDSAT for conducting univariate regression analysis was attempted. Also, multivariate-weighted regression using the weights generated by RDSAT was attempted[16]. However, RDSAT produces as many weights as the number of variables for each participant and this is the problem in applying multivariate regression to RDS data.

Volz and Heckathorn[6] generalized Horvitz-Thompson estimator to adopt RDS estimation to survey sampling (RDS II) and this was found to outperform the MC method. This was a single weight per participant approach unlike RDSAT's multiple weights per participant. Their approach made it possible to do regression analysis of RDS data. Calculation of variance analytically was made possible, but the problem of calculating CIs for a smaller group of respondents remained unresolved. Other approaches that have been proposed for analysis of RDS data are RDS-MR estimate (for analyzing continuous variables controlling for differential recruitment), RDS-SS estimates (for eliminating the condition of selection with replacement) and variance estimation[6,17-19]. All are currently in various stages of development[20]. RDS analyst *(*RDS-A) is the currently available most advanced software, but the problem with small samples and calculation of CIs for the estimation of cross-classified data remain a problem[21] and it adopts bootstrap approach for the calculation of CIs. Thus, we need an approach or interface that allows use of RDS data in standard statistical applications and software.

The objective was to propose and validate a new two-step approach termed as RDS-MOD for analyzing RDS data. We hypothesized that determining a single normalized weight for each participant (irrespective of number of variables) and transforming RDS data into clustered data without affecting the recruitment pattern (sequence and equilibrium) would enable calculating CIs of the estimates including regression coefficients analytically in case of RDS data.

Material & Methods

The RDS-MOD was applied to a real dataset from India as well as four datasets available in public domain to estimate various parameters and their CIs. In addition, three real datasets were obtained and estimates were presented. STATA SVY module was used for analysis. Taylor method of linearization was applied to calculate standard errors (SEs) of estimates[22]. The precision of the proposed estimates was assessed based on the length of the CIs.

Data sources

Indian dataset: We used the first round data of Integrated Behavioural and Biological Assessment (IBBA) conducted in Churachandpur District of Manipur State, India, during the first quarter of 2006 on 419 IDUs recruited using RDS[23,24]. Three more datasets of the same survey were also obtained and HIV prevalence estimates were presented.

The datasets available in public domain were: (*i*) 1-3: Three simulated datasets of 'RDS-A' module. The datasets were faux (RDS); fauxsycamore (RDS) and fauxmadrona (RDS)[21]. (*ii*) 4: Jazz musicians dataset of RDSAT 7.1.46[25].

Statistical analysis: Bland–Altman method was used to compare the single weight generated by RDS-MOD and the individualized weights generated by RDSAT for different variables of Churachandpur data[26]. The parameters and their 95 per cent CIs were estimated. The CIs were obtained as an output of SVY module of STATA using linearization and replication method for the calculation of SEs[27]. For multiple regression analysis, weighted generalized estimation equation (WGEE) was used.

For analysis, RDSAT 6.0.1(Cornell University Ithaca, NY), RDS-A, STATA 10 (Stata Corp, Texas USA); and SAS (Enterprise edition 4.3, SAS Institute, Cary, North Carolina, USA) were used. Drawing of network was performed by NetDraw 2.090. (Borgatti S.P, NetDraw Software for Network Visualization, Lexington, NY).

Data analysis by new approach (RDS-MOD)

Derivation of single weight and estimation method: Under the assumption in this chain referral sampling, the selection of a subject by a recruiter from his network is independent and is probability proportional to his degree (the number of men he knows and they know him) (d_i)[8]. A unique sampling weight *Wi* was derived for i^{th} participant. With these new weights and survey sampling module (SVY) of STATA, population parameters estimates were calculated[27].

Formation of clusters

Formation of clusters in Churachandpur dataset (real dataset): The recruitment pattern is depicted as recruitment network diagram using 'NetDraw' in Fig. 1A. RDS data were converted into clusters by discarding all the seeds from network chains (Fig. 1B). All participants of the branch in a network chain after discarding a seed were considered as members of that cluster. An assumption was made that the clusters thus formed were independent though some traits (characteristic affiliation) of respective seeds would prevail upon the members of the clusters thus formed. However, this correlation would minimize or vanish with expanding waves and widening of gap between recruiters and recruits, thus diluting the trait of the seed. Further, the recruits within the clusters would have intra-cluster correlations that need to be addressed in any type of analysis. Had the number of seeds been more and independent, all recruits under a seed might have to be considered as independent clusters. In addition, with more number of clusters, the estimates could be better.

Formation of clusters in other example datasets: To perform RDS-MOD on other datasets (datasets in public domain) all the recruits under a seed were considered as cluster. Hence, the clusters would be independent if the seeds were independent. Thus, if there were ten seeds, ten clusters would be formed. For example, in fauxsycamore dataset of RDS-A, there were ten seeds. Therefore, ten clusters were formed for this analysis. For the dataset 'faux (RDS)', there was only one seed, and therefore, all the recruits were assumed to be from a single cluster. This was done to study the performance of RDS-MOD in situations where all the recruits under only one seed constituted a cluster.

Formation of clusters in yet other datasets (real datasets, *viz*. Bishnupur, Phek and Wokha): In Bishnupur dataset all seeds (nine) were removed. In the process, one cluster with single respondent was not considered. Thus, there were ten exclusions (nine seeds and one first-wave respondent).

In Phek and Wokha datasets, all recruits under a seed were considered as a cluster (nine clusters each).

Data analysis by RDSAT: The analysis tool RDSAT (6.0.1) was set to use average network size by adjusted mean values method. The number of re-samplings to determine bootstrap 95 per cent CI was set to 2500. The enhanced smoothing algorithm type was employed. Homophily (Hx) for each variable was obtained to understand the magnitude of the characteristic affiliation of recruiters with their recruits. The number of waves required for attaining equilibrium was also estimated by choosing a convergence radius of 0.02. We assumed a median base population as 10,000 and set 500 as bootstrap replication to analyze additional three real datasets (*viz*. Bishnupur, Phekh and Wokha).

Comparison of weights by RDS-MOD and the individualized weights of RDSAT for different variables under analysis: Single weight per study participant derived for RDS-MOD and the individualized weights

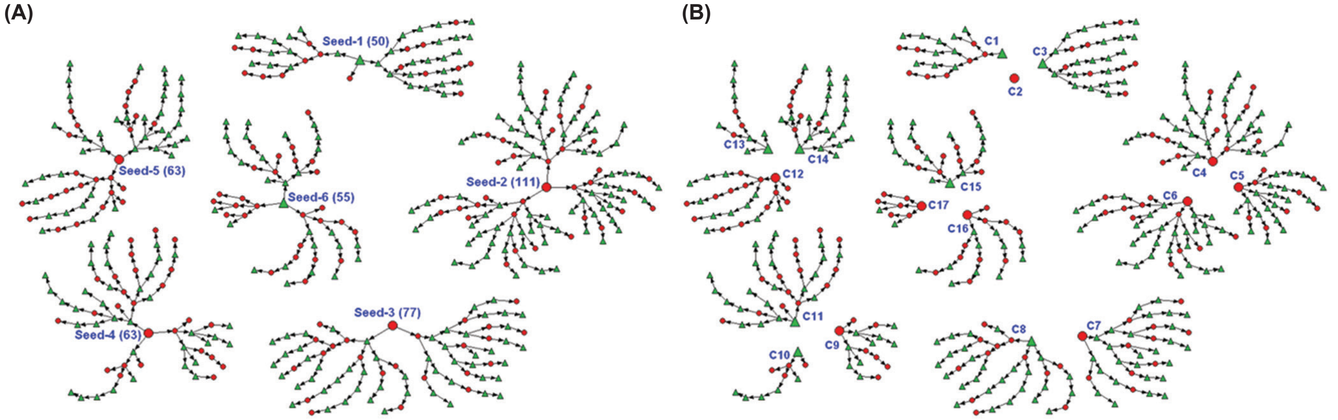

Fig. 1. (**A**) RDS Network recruitment diagram of injecting drug users (IDUs) in Churachandpur district of Manipur State of India, 2006. (**B**) Network recruitment diagram of clusters created from the IDUs in Churachandpur district of Manipur State of India, 2006. All red circles: HIV +ve; All green triangles: HIV –ve. C1 to C17 are clusters formed from networks. Seeds are in larger size. Arrow-marks represent direction of recruitment chain. (*NetDraw 2.090 software*, Data *Source*: IBBA Round 1).

generated by RDSAT for different variables were compared by Bland–Altman method[26]. The difference in weights by RDSAT and the calculated weight for RDS-MOD were plotted against the mean of the weights by these two approaches for a variable (Fig. 2). If the points on the Bland–Altman plot were uniformly scattered between the limits of agreement, it would suggest good agreement between the two weights by two different approaches. This analysis was performed to compare the individualized weights generated for each of variables by RDSAT and the single weight for each individual by RDS-MOD.

Comparison of estimates of RDS-MOD and RDS-A on example datasets: As our weights were similar to RDS II of Volz and Heckathorn[6], RDS-MOD comparison of parameters by RDS-A (RDS II) would yield similar estimates but not the CIs. The comparison of estimates by different approaches using Churachandpur data is presented in Tables I and II. The results of comparison using the datasets, *viz*. faux, fauxmadrona, fauxsycamore of RDS-A and Jazz musicians' dataset of RDSAT 7.1.46 are presented in Tables III and IV. In addition, HIV prevalence estimates were calculated both by RDS-A and RDS-MOD of yet another three real datasets (data not shown).

Regression analyses: Similarity induced by clusters would violate the standard assumption of independent observations from each individual. In the regression set-up, generalized estimating equation approach accounts for intra-cluster correlation[28]. This approach was used to study the affiliation factors for HIV positivity (factors associated with HIV) among the IDUs recruited in Churachandpur, India. Regression analysis was performed using WGEE approach with auto-regressive-1 (AR-1) correlation structure as logical ordering of recruitment was inevitably present in the RDS selection process. Furthermore, the AR-1 structure is appropriate when the correlation between various sample units is expected to decrease with the increasing distance within the recruitment chain. The CIs were obtained for the parameter estimates of regression equation by SAS software[29]. The results are presented in Table V. To understand the nature of linkages in the recruitments among IDUs with HIV status, WGEE with exchangeable correlation structure was also performed. If Hx is high for a variable, the parameter estimates with these two assumed correlation (AR-1 and exchangeable) structures would vary in the regression.

Results

The proposed weights were similar to RDS II of Volz and Heckathorn[13] but for a constant term (harmonic mean of the network size) in the numerator as normalizing factor.

Comparison of estimates of parameters of different datasets by RDS-MOD with RDSAT and RDS-A

Churachandpur, India RDS data: All the seeds grew and reached up to seven waves. The recruitment per seed ranged from 50 to 111. A random mixing pattern of recruitment was observed among HIV positives and negatives in the network recruitment diagram (Fig. 1A) and also in the demographically adjusted recruitment matrices of RDSAT (symmetry, not shown). Estimated waves required to reach equilibrium was 2-3 for all variables, except for only one subset variable. All the variables considered for this analysis attained equilibrium with more than six waves. RDSAT and RDS-A used all 419 recruits for analysis.

For our new approach (RDS-MOD), 17 clusters were formed after discarding six seeds (Fig. 1B). For example, when seed 2 was removed, three clusters were formed with assigned cluster numbers C4, C5 and C6 of size 41, 28 and 41, respectively. Only one cluster (C2) of size one from seed 1 was not considered for cluster analysis. Thus, 16 clusters of 412 recruits were available for analysis by our new approach with a loss of seven participants including, six seeds and one first-wave recruit.

Bland–Altman plots indicated that single weights of RDS-MOD and weights of RDSAT for different variables were within the acceptance limits (*i.e.* within the mean of differences of weights between these two methods ±1.96 of standard deviation of these differences) for all the four variables considered (Fig. 2). Further, the mean of the differences was nearly zero signalling that the single weight per participant was similar to multiple weights of RDSAT. However, the trends within them indicated that the differences varied with magnitude. Thus, the calculated weights by both methods depended strongly on each other. This indicated the reasonability between the weights calculated by our approach for an individual and several of RDSAT's weights for different variables of the same individual for this dataset.

The estimates of proportions by RDS-MOD across variables were similar to RDS-A and RDSAT (Table I). However, the CIs by RDS-MOD were wider compared to other two methods. The Hx was highest for the character 'sharing of injection needle' (Hx = 0.381). The other affiliation characters for HIV were 'daily injection

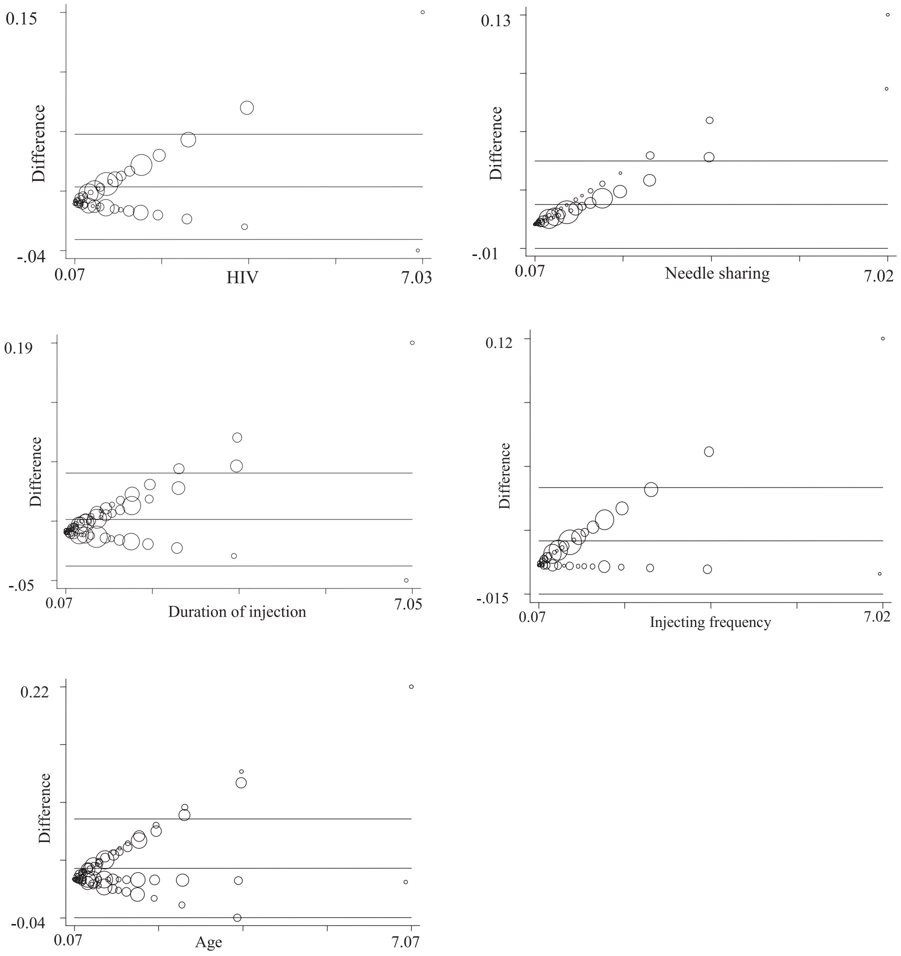

Fig. 2. Agreement between the weights generated by RDSAT and RDS-MOD by Bland - Altman method. *Source*: Paul Seed, 2014. "BAPLOT: Stata module to produce Bland-Altman plots," Statistical Software Components S457853, Boston College Department of Economics.

Table I. Estimates of proportions of population parameters by respondent-driven sampling (RDS)-MOD (modified), RDS-A (analyst) and respondent-driven sampling analysis tool (RDSAT)- Churachandpur data

Variable	RDS-MOD (n*=412)		RDS-A (RDS-II, n=419)		RDSAT (RDS-I)** (n*=419)	
	Number observed	Estimate (95% CI)	Number observed	Estimate (95% CI)	Number observed	Estimate (95% CI)
Age group (yr)						
18-20	42	10.5 (7.0-15.5)	42	10.5 (5.4-15.5)	42	10.7 (7.2-14.4)
21-25	144	36.5 (27.2-47.0)	144	36.4 (29.8-43.0)	144	37.0 (30.0-43.9)
26-30	136	32.6 (22.9-43.9)	140	32.7 (28.6-36.8)	140	32.3 (26.1-38.8)
>30	90	20.4 (16.5-25.0)	93	20.4 (14.2-26.7)	93	20.1 (15.7-25.5)
Duration of injecting drug (yr)						
1-2	91	27.1 (21.4-33.8)	91	27.0 (20.5-33.6)	91	27.2 (21.0-33.3)
3-5	156	39.3 (30.1-49.4)	157	39.2 (33.2-45.3)	157	39.3 (33.0-45.8)
>5	165	33.5 (24.8-43.5)	171	33.8 (29.3-38.3)	171	33.4 (27.7-40.7)
Frequency of injecting drug						
Daily	324	75.6 (68.0-82.0)	328	75.5 (69.3-81.6)	328	75.6 (69.3-81.8)
Others	88	24.4 (18.0-32.0)	91	24.5 (18.4-30.7)	91	24.4 (18.2-30.7)
Current needle sharing practice						
Not sharing	44	16.0 (11.2-22.4)	44	15.9 (13.1-18.8)	44	16.0 (10.7-21.7)
Sharing	368	84.0 (77.6-88.8)	375	84.1 (81.2-86.9)	375	84.0 (78.3-89.3)
HIV status						
Positive	147	32.6 (25.7-40.4)	152	32.8 (26.5-39.2)	152	32.5 (26.8-38.6)
Negative	265	67.4 (59.6-74.3)	267	67.2 (60.8-73.6)	267	67.5 (61.4-73.2)
HSV-2***						
Positive	7	32.6 (11.3-64.7)	8	-	8	0.0 (-)†
Negative	26	64.4 (34.4-86.4)	27	-	27	62.9 (-)†
Inconclusive	2	3.0 (0.7-10.6)	2	-	2	37.1 (-)†

*Observed unweighted count; **Bootstrap - 2500; enhanced data smoothening; average network size - dual component; ***HSV-2 done only on a random sample of 37; †RDSAT did not generate CI for this subgroup data. HSV-2, herpes simplex virus type 2; CI, confidence interval

of drugs' (Hx = 0.156) and 'duration of injecting drugs more than five years' (Hx = 0.143). Negative affiliation in the recruitment was noticed among those with duration of injecting drugs less than two years (Hx = −0.149). For a sub-sample, RDSAT produced estimates not in tune with the observed frequencies and RDSAT could not produce CI. For example, a random sample of 37 specimens was tested for herpes simplex virus type 2 (HSV- 2). Among them eight were 'positive', two were 'inconclusive' and the remaining 'negative'. RDSAT estimated the proportions of positives (8 out of 37) as 0 per cent and inconclusive (2 out of 37) as 37.1 per cent. RDSAT showed that the estimated mean number of waves to attain equilibrium for this variable was as high as 1960 (Table I).

RDS-MOD yielded similar and comparable parameter estimates of cross-classified variables as well (Table II). However, slightly wider CI was noticed in many of the estimates by RDS-MOD both in Tables I and II. RDS-A did not produce CI for cross-classified data (Table II).

By RDS-MOD method, all parameters were re-estimated using individualized weights per variable generated by RDSAT (data not shown). The estimates were almost similar to that of single weighting procedure of RDS-MOD implying that single weight per individual was sufficient rather than multiple weights per individual. Similar exercise was performed on cross-classified data of parameters. The estimates by

Table II. Estimates of proportions of HIV status cross-classified by factors using respondent-driven sampling (RDS)-MOD (modified), RDS-A (analyst) and respondent-driven sampling analysis tool (RDSAT)- Churachandpur data

Factors	HIV status	RDS-MOD†		n*	RDS-A‡ (RDS-II)	RDSAT (RDS-I) estimate (95% CI)
		n*	Estimate (95% CI)			
Age group (yr)						
18-20	HIV positive	3	8.6 (2.6-24.6)	3	8.6	8.4 (0.0-18.7)
	HIV negative	39	91.4 (75.4-97.4)	39	91.4	91.6 (81.3-100.0)
21-25	HIV positive	22	14.5 (9.2-22.4)	22	14.6	14.7 (8.4-22.8)
	HIV negative	122	85.4 (77.6-90.8)	122	85.4	85.3 (77.2-91.6)
26-30	HIV positive	60	42.8 (31.8-54.6)	63	43.2	42.3 (31.1-53.1)
	HIV negative	76	57.2 (45.4-68.2)	77	56.8	57.7 (46.9-68.9)
≥31	HIV positive	62	61.1 (47.8-72.8)	64	61.1	61.3 (45.6-74.8)
	HIV negative	28	38.9 (27.2-52.2)	29	38.9	38.7 (25.2-54.4)
Duration of injecting drug (yr)						
1-2	HIV positive	10	11.4 (5.9-20.8)	10	11.4	10.7 (3.9-19.7)
	HIV negative	81	88.6 (79.2-94.1)	81	88.6	89.3 (80.3-96.1)
3-5	HIV positive	42	26.1 (16.7-38.2)	42	26.0	26.9 (18.6-35.2)
	HIV negative	114	73.9 (61.8-83.3)	115	74.0	73.1 (64.8-81.4)
≥6	HIV positive	95	57.5 (47.5-67.0)	100	57.9	57.0 (46.1-67.4)
	HIV negative	70	42.5 (33.0-52.5)	71	42.1	43.0 (32.6-53.9)
Frequency of injecting drug						
Daily	HIV positive	116	35.5 (26.5-45.8)	118	35.6	35.5 (29.2-42.5)
	HIV negative	208	64.5 (54.2-73.5)	210	64.4	64.5 (57.5-70.8)
Not daily	HIV positive	31	23.5 (12.5-39.8)	34	24.5	23.9 (13.9-36.7)
	HIV negative	57	76.5 (60.2-87.5)	57	75.5	76.1 (63.3-86.4)
Current needle sharing practice						
Sharing	HIV positive	130	34.3 (26.9-42.5)	135	34.5	34.0 (27.6-41.0)
	HIV negative	238	65.7 (57.5-73.1)	240	65.5	66.0 (59.0-72.4)
Not sharing	HIV positive	17	23.8 (14.1-37.3)	17	23.8	24.8 (11.9-40.8)
	HIV negative	27	76.2 (62.7-85.9)	27	76.2	75.2 (59.2-88.1)

*Unweighted count; †SEs were calculated by the method of linearization-STATA survey module. Age group and duration of injection were significantly associated with HIV status at 5% level (χ^2). ‡RDS-A does not produce CI for cross-classified data. SEs, standard errors; CI, confidence interval

single weight approach and those with individualized weights of RDSAT were identical (data not shown). This also indicated that single weighting approach worked well for cross-classified estimations.

Comparison of results of RDS-MOD, RDSAT and RDS-A with other example datasets: The estimates of parameters and their CIs of the other example RDS datasets are presented in Tables III and IV. This exercise was done for not drawing any inference but to compare the estimates.

Dataset 1 – faux: From the faux dataset, the variable 'A', the estimates of population parameters by RDS-A (RDS II), RDS-MOD and RDSAT (RDS I) were almost similar including their CIs. However, for variable 'B' of the same datasets, population estimates by RDS-A and RDS-MOD were identical and RDSAT (RDS I) yielded varied results for all the three parameters.

Dataset 2 – fauxsycamore: For all the three variables (C, D, E), the estimates by RDS-MOD and RDS-A were comparable. The CIs produced by RDS-MOD were almost equal or narrower than RDS-A. RDSAT (RDS I) produced results differently for all these variables.

Dataset 3 – fauxmadrona: This dataset contains three variables (F, G, H). The variable 'G' has 14 groups

Table III. Estimates of proportions [given as (estimates and 95% CI)] of various subgroup of variables using respondent-driven sampling (RDS)-MOD (modified), RDS-A (analyst) and respondent-driven sampling analysis tool (RDSAT) on different datasets, *viz*. faux, fauxsycamore, fauxmadrona

Dataset	Variable	RDS-MOD	RDS-A (RDS-II)	RDSAT (RDS-I)*
Faux (RDS)				
A	Blue	31.0 (26.0-36.1)	31.1 (26.1-36.1)	31.1 (26.1-36.3)
	Red	69.0 (63.9-74.0)	69.0 (63.9-73.4)	68.9 (63.7-73.9)
B	Blue	40.7 (35.4-46.0)	40.7 (26.6-54.7)	47.6 (32.9-62.8)
	Green	37.9 (32.6-43.2)	37.9 (22.7-53.1)	24.6 (10.3-40.3)
	Black	21.4 (16.5-26.4)	21.4 (10.7-32.1)	27.8 (17.9-39.9)
Fauxsycamore (RDS)				
C	Disease			
	No	85.5 (82.9-88.1)	85.5 (77.7-93.2)	88.9 (85.6-91.9)
	Yes	14.5 (11.9-17.2)	14.6 (6.8-22.3)	11.1 (8.1-14.4)
D	To diseased (three groups)			
	0	85.9 (83.0-88.9)	85.9 (81.3-90.5)	90.7 (87.0-92.8)
	1	11.7 (9.0-14.5)	11.8 (6.8-16.7)	8.2 (6.2-11.4)
	2	2.3 (1.2-3.4)	2.3 (1.1-3.5)	1.1 (0.4-2.2)
E	To non-diseased (three groups)			
	0	57.5 (52.4-62.5)	57.5 (52.3-62.6)	60.1 (52.3-62.4)
	1	19.8 (15.8-23.8)	19.8 (14.3-25.3)	25.6 (21.1-29.2)
	2	22.8 (18.9-26.7)	22.8 (18.9-26.6)	14.4 (13.4-21.8)
Fauxmadrona (RDS)				
F	Disease			
	No	83.6 (80.7-86.5)	83.6 (76.8-90.4)	84.2 (79.6-88.0)
	Yes	16.4 (13.5-19.3)	16.4 (9.6-23.2)	15.8 (12-20.4)
G	To diseased (14 groups)			
	0	26.4 (21.7-31.2)	26.4 (22.6-30.2)	26.9 (21.5-32.8)
	1	26.4 (22-30.8)	26.4 (20.6-32.2)	26.4 (21.2-31.2)
	2	20.9 (17.1-24.7)	20.9 (17.1-24.8)	20.9 (17.5-24.4)
	3	8.9 (6.4-11.4)	8.9 (7.5-10.4)	8.8 (6.3-11.2)
	4	5.4 (3.5-7.3)	5.4 (3.8-7)	5.7 (3.7-7.6)
	5	2.5 (1.3-3.7)	2.5 (1.2-3.8)	2.5 (1.4-4)
	6	3.3 (2-4.6)	3.3 (1.7-4.9)	3.1 (1.7-4.8)
	7	2.6 (1.5-3.7)	2.6 (1.6-3.6)	2.3 (1.2-3.7)
	8	1.2 (0.5-1.8)	1.2 (0.5-1.8)	1.1 (0.4-1.9)
	9	1.2 (0.5-1.9)	1.2 (0.2-2.2)	1.2 (0.5-2)
	10	0.6 (0.1-1.1)	0.6 (−0.3-1.5)†	0.6 (0.2-1.2)
	11	0.3 (0.0-0.6)	0.3 (−1.2-1.9)†	0.3 (0-0.6)
	12	0.1 (−0.1-0.2)†	0.1 (−0.5-0.7)†	0.1 (0-0.2)
	13	0.1 (−0.1-0.3)†	0.1 (−0.5-0.8)†	0.1 (0-0.3)

Contd...

Dataset	Variable	RDS-MOD	RDS-A (RDS-II)	RDSAT (RDS-I)*
H	To non-diseased (13 groups)			
	0	0.7 (−0.7-2.0)†	0.7 (0.3-1.1)	1 (0-2.7)
	1	1.3 (−0.3-2.9)†	1.3 (−0.1-2.6)†	1.6 (0-3.5)
	2	8.9 (5.2-12.5)	8.9 (6.8-11)	8.1 (4.9-11.8)
	3	19.7 (15.4-23.9)	19.7 (17.2-22.1)	19.5 (14.9-23.9)
	4	16.3 (12.6-19.9)	16.3 (14.3-18.3)	17.1 (13.3-20.9)
	5	17.4 (13.9-20.9)	17.4 (14.2-20.5)	18 (14.6-21.2)
	6	15.9 (12.7-19)	15.9 (10.2-21.5)	15.5 (12.7-18.7)
	7	9 (6.7-11.3)	9 (7.6-10.4)	8.9 (6.6-11.6)
	8	5.8 (4.1-7.6)	5.8 (4.5-7.2)	5.9 (4.3-7.7)
	9	3.1 (1.8-4.3)	3.1 (−2.4-8.6)†	2.4 (1.2-3.7)
	10	1.4 (0.6-2.2)	1.4 (0.4-2.5)	1.4 (0.7-2.3)
	11	0.4 (0-0.9)	0.4 (−0.6-1.5)†	0.5 (0.1-0.9)
	12	0.2 (−0.1-0.5)†	0.2 (−0.1-0.5)†	0.1 (0-0.4)

*Bootstrap - 2500, Enhanced data smoothening; Average network size - Dual component; †Confidence limits were negatives

Table IV. Estimates of proportions [estimates (95% CI)] using respondent-driven sampling (RDS)-MOD (modified), RDS-A (analyst) and respondent-driven sampling analysis tool (RDSAT) on Jazz musician datasets

Variable	RDS-MOD	RDS-A (RDS-II)	RDSAT (RDS-I)*
Gender			
1	72.1 (63.5-80.7)	72.1 (62.8-81.5)	76.2 (66.1-84.2)
2	27.9 (19.3-36.5)	27.9 (18.5-37.2)	23.8 (15.8-33.9)
Race			
1	55.5 (46.5-64.6)	55.5 (44.1-67.0)	53.1 (43.1-63.6)
2	33.0 (24.3-41.7)	33.0 (22.3-43.7)	36.0 (26.0-46.8)
3	11.5 (6.3-16.6)	11.5 (6.2-16.7)	10.9 (6.2-15.5)
Air play			
1	75.1 (66.1-84.0)	75.1 (65.8-84.4)	75.1 (66.2-84.8)
2	24.9 (16.0-33.9)	24.9 (15.6-34.2)	24.9 (15.2-33.8)
Union			
1	24.1 (17.8-30.3)	24.1 (16.6-31.6)	25.0 (18.2-32.6)
2	75.9 (69.7-82.2)	75.9 (68.4-83.4)	75.0 (67.4-81.8)

*Bootstrap - 2500, Enhanced data smoothening; Average network size - Dual component

and 'H' has 13 groups. The estimates by RDS-MOD and RDS-A were identical and the estimates by all the three methods were comparable. For 'G', the CIs were wider for RDS-A. RDS-MOD gave narrow CIs. RDS-MOD and RDS-A produced negative limits for some CIs. However, RDSAT did not produce any of that type.

Dataset 4 – Jazz musician: The estimates of parameters by RDS-MOD and RDS-A were identical. RDS-MOD yielded narrow CI. RDSAT also produced comparable results for all the four variables.

Dataset 5 – Bishnupur, Phek and Wokha: HIV prevalence estimate and CI was similar to RDS-A in Wokha and CI was slightly wider in Bishnupur data. However, CI of Phek data by RDS-MOD was wider (data not shown).

Results of regression analysis: WGEE both with AR-1 and exchangeable correlation structures showed that older age groups (≥25 yr) and longer period of injecting drug use (≥6 yr) were associated with HIV positivity (Table V). Sharing of needles daily was not associated with HIV positivity. The similarity of regression coefficients both by AR-1 and exchangeable correlation

Table V. Factors associated with HIV for different correlation structure of weighted generalized estimation equation

Factors	ORs			
	Estimates with exchangeable correlation	95% CI	Estimates with AR (1)-correlation	95% CI
Age group (yr)				
18-20	1.00		1.00	
21-25	1.37	0.35-5.39	1.42	0.36-5.63
26-30	3.94	1.42-13.62†	4.06	1.20-13.78†
≥31	7.89	2.09-29.84†	7.97	2.12-29.90†
Duration of injecting drug (yr)				
1-2	1.00		1.00	
3-5	2.01	0.75-5.39	2.03	0.75-5.51
≥6	4.72	2.69-8.24†	4.65	2.60-8.28†
Frequency of injecting drug				
Daily	1.00		1.00	
Not daily	1.33	0.53-3.32	1.30	0.51-3.31
Needle sharing				
Not sharing	1.00		1.00	
Sharing	1.40	0.62-3.16	1.30	0.58-2.89

†Significant at 5% level. WGEE, weighted generalized estimating equation (SAS Enterprise Guide 4.3) using single weight; ORs, odds ratios; CI, confidence interval

structure indicated the random mixing nature of HIV status of recruiter and recruits in Churachandpur data (Table V).

Discussion

As there is no method available to make the true estimate of the parameter of RDS data, all our comparisons were mainly with RDS II (this method calculates SEs of estimates analytically) estimates, and hence, the comparison of the RDS-MOD estimates with other methods such as successive sampling method was not possible. The proposed new approach for analysis of RDS data was simple and less time-consuming. Additionally, this approach was able to generate population estimates comparable to those derived by RDS-A (RDS II), the currently available most advanced level software to analyze RDS data. Precision of estimates by our approach appeared to be superior to RDS-A in the example datasets [*viz*. faux (RDS); fauxsycamore (RDS); fauxmadrona (RDS) and Jazz musicians dataset].

In the new approach, clusters were formed without affecting sequential and natural ordering of selection. In this process, though information on all seeds (non-random) was lost, RDS data were robust and not likely to be affected by the inclusion or exclusion of out of equilibrium data, *i.e*. data collected before reaching equilibrium[30]. Discarding of earlier waves has also been recommended in previous reports[10,15]. Thus, our result might not be affected by discarding the six seeds from the analysis. Formation of clusters paved a way to account for the related characters in the recruitment process and provided ways to other statistical methods and analysis by routine statistical software.

It has been suggested that the tendency towards Hx varies among groups[31]. Hence, it is important to measure the tendency towards Hx with respect to different respondent's characteristics and to use this information to weight the sample to compensate for any biases[31]. It appears that in general, the network's composition with respect to personal attributes may exhibit Hx with respect to only a particular trait or with respect to a few characters. In our sample, 'sharing of needle' was the most prominent trait (knowing each other) irrespective of the HIV status. Therefore, technically, a single unique weight would be sufficient to compensate the Hx of different respondent characteristics. In addition, as sampling weights were used only to compensate the unequal probability of inclusion into the sample, a common weight for each individual was sufficient rather than individualized weights for each variable and the results were comparable to those of RDSAT. The similarity of the parameter estimates by our approach using RDSAT weights in one-way and two-way tables indicated that a single weight per individual was sufficient.

Our RDS-MOD approach was similar to RDS II of Volz and Heckathorn[6]. The only difference was that the weights by RDS-MOD had an additional constant multiplier compared to Volz and Heckathorn. Although this constant does not affect the estimates, it is needed for the physical comparison of weights with that of RDSAT. The very basic assumption made was that a recruiter recruited the subjects independently with probability proportional to network size of the recruiter. This assumption was based on the work of Salganik and Heckthorn[8], who showed that a random walk on network was a Markov Process, in which equilibrium occupied a node with probability proportional to degree. The applicability of Hansen-Hurwitz estimator with these assumptions provides theoretical and conceptual foundation to our approach of deriving unique weights[32].

The proposed approach necessitated the need for incorporating clustering effect in the regression model as clustering results in lack of independence among the errors in regression. Generalized estimating equation (GEE) approach resolves this problem by appropriately accounting the correlation structure of a variable of interest between recruits and recruiter[28]. The process of fitting a model should incorporate sample weights as well as information about correlation between sample units. Weighted estimating equations are the most popular methods for obtaining consistent estimates of regression coefficient with sample survey data[33,34]. Therefore, we used WGEE approach to study the affiliation factors for the HIV positivity. AR-1 correlation structure accounted for logical ordering of recruitment and the exchangeable correlation assumed equality of correlation between any two recruited individuals within a cluster. The similarity of results due to different correlation structures (AR-1 and exchangeable) suggested that HIV status was not an indicative factor for recruitment preferences in recruiting HIV-positive/negative IDUs in Churachandpur dataset. Slight variations found in model coefficients by these two procedures (AR-1 and exchangeable) could be due to possible omission of a variable from the model that had a strong interaction with the independent variables and was highly correlated with the weights[35].

The advantage of our approach is that it allows estimations using standard software such as STATA or any other software that accommodates survey sampling method. Also, the problem of subgroup analysis of RDS data could be overcome.

It was assumed that the clusters formed were independent after discarding a seed although some traits of that seed might prevail upon clusters of that seed. However, this limitation can be overcome by selecting more seeds at the stage of data collection (preferably independent) and considering each seed with its recruits as a separate cluster as has been done with other example datasets. The new approach resulted in slightly wider CIs in Churachandpur data compared to RDS-A and RDSAT. The possible reasons could be that RDS-MOD employed the analytical method to calculate these. It could also happen if the intra-cluster correlation was high. Volz and Heckathorn[6] have reported that wider CI is expected when the variances are calculated analytically. Empirical tests of RDS have indicated that the analytical method overestimates the CIs[30]. As we have also accounted for the intra-cluster correlation, still wider CIs are expected[36]. Clustering and weighting normally result in decreased precision[37]. Volz and Heckathorn[6] suggested that the loss of precision might happen due to the single weights used for all variables as was done in the present study. In contrast, RDS-MOD yielded estimates with the same or higher precision in other simulated example datasets, *viz*., faux (RDS); fauxsycamore (RDS) and fauxmadrona (RDS) and Jazz musicians dataset used for comparisons. This indicated the possibility that intra-cluster correlation in these example datasets (simulated data) was not high. As the inferences based on RDS data require many strong assumptions, Gile *et al*[38] have suggested some diagnostic tools to empower researchers to understand their RDS data better and encourage future statistical research on RDS sampling and inference.

Our study had certain limitations. The weight calculation was solely dependent on the network size (degree reported by the respondent). Thus, inaccuracies in the self-reported degree might have introduced biases in the estimates. Hanzen-Hurwitz method is applicable only when the sample elements are selected independently with replacement and that may not be true in the real sense of RDS. We assumed that the clusters formed were independent though some traits of corresponding seed would have prevailed. A few of the lower limits of our CIs in an example dataset for proportions were negative.

In conclusion, the proposed alternative approach of using single weight and converting RDS data into clusters before analysis can be recommended as it generates analytical CIs and allows for estimates for smaller groups as well. RDS data can thus be analyzed faster using commonly used statistical software that also permits wider range of statistical analysis including analysis of continuous variables.

***Conflicts of Interest*:** None.

References

1. Goodman L. Snowball sampling. *Ann Math Stat* 1961; *32* : 148-70.
2. Deaux E, Callaghan JW. Key informant versus self-report estimates of health behavior. *Eval Rev* 1985; *9* : 365-8.
3. Watters JK, Biernacki P. Targeted sampling: options for the study of hidden populations. *Soc Probl* 1989; *36* : 416-30.
4. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl* 1997; *44* : 174-99.
5. Singh P, Pandey A, Aggarwal A. House-to-house survey vs. snowball technique for capturing maternal deaths in India: a search for a cost-effective method. *Indian J Med Res* 2007; *125* : 550-6.
6. Volz E, Heckathorn DD. Probability based estimation theory for respondent-driven sampling. *J Off Stat* 2008; *24* : 79-97.
7. Fararo TJ, Skvoretz J. Biased networks and social structure theorems. *Soc Networks* 1984; *6* : 223-58.
8. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol Methodol* 2004; *34* : 193-239.
9. Gile KJ, Handcock MS. Respondent-driven sampling: an assessment of current methodology. *Sociol Methodol* 2010; *40* : 285-327.
10. Goel S, Salganik MJ. Respondent-driven sampling as Markov chain Monte Carlo. *Stat Med* 2009; *28* : 2202-29.
11. Griffiths P, Gossop M, Powis B, Strang J. Reaching hidden populations of drug users by privileged access interviewers: methodological and practical issues. *Addiction* 1993; *88* : 1617-26.
12. Johnston LG, Malekinejad M, Kendall C, Iuppa IM, Rutherford GW. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: field experiences in international settings. *AIDS Behav* 2008; *12* : S131-41.
13. Volz E, Wejnert C, Degani I, Heckathorn DD. *Respondent-driven sampling analysis*. *RDSAT 6.0.1* Ithaca, NY: Cornell University; 2007.
14. Salganik MJ. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J Urban Health* 2006; *83* : i98-112.
15. Burt RD, Thiede H. Evaluating consistency in repeat surveys of injection drug users recruited by respondent-driven sampling in the Seattle area: results from the NHBS-IDU1 and NHBS-IDU2 surveys. *Ann Epidemiol* 2012; *22* : 354-63.
16. Taran YS, Johnston LG, Pohorila NB, Saliuk TO. Correlates of HIV risk among injecting drug users in sixteen Ukrainian cities. *AIDS Behav* 2011; *15* : 65-74.
17. Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J Am Stat Assoc* 2011; *106* : 135-46.
18. Heckathorn DD. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. *Sociol Methodol* 2007; *37* : 151-207.
19. Szwarcwald CL, de Souza Júnior PRB, Damacena GN, Junior AB, Kendall C. Analysis of data collected by RDS among sex workers in 10 Brazilian cities, 2009: estimation of the prevalence of HIV, variance, and design effect. *J Acquir Immune Defic Syndr* 2011; *57* : S129-35.
20. Salganik MJ. Respondent-driven sampling in the real world. *Epidemiology* 2012; *23* : 148-50.
21. Handcock MS, Fellows IE, Gile KJ. *Software for the analysis of respondent-driven sampling data*. Version 0.42. Los Angeles, CA: Hard to Reach Population Methods Research Group . 2014.
22. Shah VB. Linearization methods of variance estimation. In: Armitage P, Colton T, editors. *Encyclopedia of Biostatistics*. New York: John Wiley & Sons, Inc.; 1998. p. 2276 - 9.
23. Mahanta J, Medhi GK, Paranjape RS, Roy N, Kohli A, Akoijam BS, *et al*. Injecting and sexual risk behaviours, sexually transmitted infections and HIV prevalence in injecting drug users in three states in India. *AIDS* 2008; *22* (Suppl 5) : S59-68.
24. Chandrasekaran P, Dallabetta G, Loo V, Mills S, Saidel T, Adhikary R, *et al*. Evaluation design for large-scale HIV prevention programmes: the case of Avahan, the India AIDS initiative. *AIDS* 2008; *22* (Suppl 5) : S1-15.
25. Volz E, Wejnert C, Cameron C, Spiller M, Barash V, Degani I, *et al*. *Respondent-driven sampling analysis tool* (*RDSAT*). Version 7.1. Ithaca, NY: Cornell University; 2012.
26. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; *32* : 307-17.
27. StataCorp. *Stata statistical software*. *Release 10*. 10th ed. TX: StataCorp LP; 2007.
28. Liang KY, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; *73* : 13-22.
29. SAS®. *SAS-administering SAS® enterprise guide®*. 4.3. Cary, NC: SAS Institute Inc.; 2010.
30. Wejnert C. An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol Methodol* 2009; *39* : 73-116.
31. Heckathorn DD. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Soc Probl* 2002; *49* : 11-34.

32. Hansen M, Hurwitz W. On the theory of sampling from finite populations. *Ann Math Stat* 1943; *14*: 333-62.

33. Binder DA. On the variances of asymptotically Normal Estimators from Complex Surveys. *Int Stat Rev* 1983; *51* : 279-92.

34. Pfeffermann D. The role of sampling weights when modeling survey data. *Int Stat Rev* 1993; *61* : 317-37.

35. Korn EL, Graubard BI. Examples of differing weighted and unweighted estimates from a sample survey. *Am Stat* 1995; *49* : 291-5.

36. Pfeffermann D. The use of sampling weights for survey data analysis. *Stat Methods Med Res* 1996; *5* : 239-61.

37. Dowd AC, Duggan MB. *Computing variances from data with complex sampling designs*. *Comparison of STATA and SPSS*. Boston: North American Stata Users Group. 2001.

38. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *J R Stat Soc Ser A Stat Soc* 2015; *178* : 241-69.

Reprint requests: Dr Sanjay Mehendale, Indian Council of Medical Research,
V. Ramalingaswami Bhawan, Ansari Nagar, New Delhi 110 029, India
e-mail: sanjaymehendale@icmr.org.in