Respondent-driven sampling to recruit in-country migrant workers in China: A methodological assessment

Peiyuan Qiu, Yang Yang, Xiao Ma, Fang Wu, Ping Yuan, Qiaolan Liu and Eric Caine

Scand J Public Health 2012 40: 92 originally published online 22 September 2011
DOI: 10.1177/1403494811418276

Scandinavian Journal of Public Health, 2012; 40: 92–101

ORIGINAL ARTICLE

Respondent-driven sampling to recruit in-country migrant workers in China: A methodological assessment

PEIYUAN QIU[1,*], YANG YANG[1,*], XIAO MA[1], FANG WU[2], PING YUAN[3], QIAOLAN LIU[1] & ERIC CAINE[4]

[1]*Department of Health Education, West China School of Public Health, Sichuan University, Chengdu, China,* [2]*Tiaosanta Community Health Service Center, Chengdu, Sichuan, China,* [3]*Department of Epidemiology, West China School of Public Health, Sichuan University, Chengdu, China, and* [4]*Department of Psychiatry, University of Rochester Medical Center, Rochester, NY, USA*

Abstract

Aim: Respondent-driven Sampling (RDS) is a new form of chain-referral sampling, which is superior to random sampling and traditional non-probability sampling in hard-to-reach populations. We employed RDS to recruit internal migrant workers in Chengdu, Sichuan Province, China, and examined whether it could be successfully used in this population. *Methods*: 1,270 migrant workers were recruited into the study. Social demographic information, social network size, and geographic information about participants' residential locations were collected. RDSAT software and geographic information system (GIS) technology were used to examine whether RDS was successful. *Results*: The results showed that the sample compositions converged to equilibrium very quickly. Sample representativeness testing results showed that females ($t = 3.61$, $p < 0.001$) and people aged 46 years old and above ($t = 3.222$, $p < 0.001$) were under-represented. GIS results showed that respondents were concentrated in the vicinity of the third road ring of Chengdu, especially in the areas of 3, 4 and 21, which were close to the investigation site. ***Conclusions*: The results demonstrated that RDS is a robust sampling method in the study of migrant workers. Despite its potential utility, it is also important to recognize and mitigate potential limitations, such as geographic proximity.**

Key Words: *Respondent-driven sampling, migrant workers, epidemiology, China*

Background

China has been experiencing the largest in-country migration in history, one that is characterized both by permanent relocations and seasonal shifts in the labour force. Data from the 1% national sample survey of population in 2005 showed that China had a minimum of 147 million internal migrants, and this number has been increasing yearly [1]. It is indisputable that migrant workers have made extraordinary contributions to China's rapid economic development. However, national and regional policies, and social systems and city infrastructures have not been prepared to receive this magnitude of population influx. Despite readily evident macroeconomic benefits arising from migrants' work, the attitudes of urban residents are mixed, at best.

Migrant workers are defined as people who are 16 years and older, and who leaves their *hukou* (registered residence in China) to work in another place. Most often internal migration occurs without a change in *hukou* status. Under the prevailing system, people are categorized as either residing in agricultural or non-agricultural households, holding rural *hukou* or urban *hukou*, respectively. When strictly enforced in the past, *hukou* defined where an individual could live and work [2]; even now it

*These authors contributed equally to this article.

Correspondence: Xiao Ma, West China School of Public Health, Sichuan University, No. 17, Section 3, South Renmin Road, Chengdu, Sichuan, 610041, The People's Republic of China. E-mail: huaxihe2009@gmail.com

(Accepted 6 July 2011)

DOI: 10.1177/1403494811418276

affords different privileges to migrants and local residents. Although *hukou* control has been relaxing steadily since the 1980s, migrant workers remain excluded from the benefits of city systems, resulting in greatly diminished access to job opportunities, educational opportunities, good residential surroundings, social welfare support, and medical insurance.

All of these barriers potentially serve to establish conditions that may compromise the health status of migrant workers, whose occupational choices can also add to potential health burdens. Most are involved in manufacturing or physical labour, and many, especially men, are hired for dirty or dangerous employment that local people avoid [3]. They work long hours, receive less pay, and hold lower social status [4]. In addition, migrant workers are usually geographically segregated, live in poor communities, and are perceived by urban dwellers as undesirable people who create multiple social problems, such as traffic congestion, crimes, illegal marriage, and disruption of family planning in the cities. It is in the midst of such contexts that the health issues of migrant workers have aroused attention [5].

Given their marginal status and often transient residential circumstances, researchers have frequently encountered numerous obstacles when seeking to recruit a representative sample of migrant workers. While current policy requires that migrants register at local police stations and obtain a temporary residential permit, it has been a daunting task to construct a representative sample. Despite requirements, many migrant workers do not register. Their high level of job mobility, the essential feature that makes them attractive to employers, also makes it very difficult to recruit them into a study. Thus, random sampling is not feasible.

For most recent studies, non-probability methods such as convenience and targeted sampling have been employed to recruit migrants in China. Sites selected for sample recruitment have included work sites (e.g., factories and construction sites), restaurants, and entertainment settings [6–8]. Strengths of non-probability sampling methods include the lack of a need to construct a sampling frame, which is hardly possible for migrant workers, and the convenience of readily identifying potential migrant participants, especially when going to working sites. However, there are recognized limitations, including an inability to make generalizing statistical inferences, which is possible with probability samples, and the considerable likelihood of powerful sample biases occurring when tapping specific work sites to define potential subjects.

A new method introduced by Heckathorn, called ''respondent-driven sampling'' (RDS), has been developed to recruit difficult-to-reach people, such as intravenous drug users (IDUs), men who have sex with men (MSM), and commercial sex workers (CSWs) [9–11]. In RDS, a sample is collected using a subject-referred or chain-referral procedure, and this approach is suitable for reaching people who may be members of low frequency population groups or who exist on social margins. It starts with identification of potential ''seeds,'' individuals who are then asked to recruit their peers into the study, who in turn further refer their peers, and so on. Each participant, however, is limited to referring two to four peers to avoid bias, and the process will continue until the sample achieves ''equilibrium,'' a condition where the final sample composition has stabilized and become independent of the number and types of seeds. The recruitment of RDS can be considered as a Markov process. According to the law of large numbers for Markov chains, the sample compositions should reach equilibrium as the recruitment process unfolds – irrespective of the characteristics of the initial sample [9]. Theoretical equilibrium sample composition can be calculated using the equations shown in a paper of Heckathorn [10]. Limiting the number of referrals that each participant may enlist serves to extend the recruitment chain to multiple waves before achieving equilibrium.

RDS goes beyond random sampling and is inherently more effective when there is no defining boundary of a potential target population, preventing development of a definitive sampling frame, or when it is difficult to clarify. The method is superior to other snowball sampling techniques by incorporating sampling procedures and analytical tools that allow for calculating unbiased population estimates. Consequently, RDS utilizes the reach of snowball sampling while maintaining the potential for formulating unbiased statistical inferences [12].

The effectiveness of respondent-driven sampling has been demonstrated in several studies [9,10,13]. In China, although this method has been used in the study of MSM and boys involved in prostitution [14,15], it has yet to be tested among migrant workers. Migrant workers are not as hidden as IDU, CSW, and MSM; however, they are scattered, transient, and hard to reach. RDS may provide a method to better obtain representative samples of migrant workers.

In this study, RDS was applied for the first time to recruit migrant workers in Chengdu, Sichuan Province, China. RDS depends, in essence, upon using the inherent sample bias of the group to be

studied; in this case, migrant workers are more likely than a random sample or geographically based sampling to recruit other migrants, who when the process is completed most closely represent the overall universe of these workers. However, not having an *a priori* measure of the members of this unbounded universe, it is necessary to use other tests to ascertain the representative quality of the sample and to specifically guard against unwanted internal sample biases (i.e., that members of the sample inadvertently reflect idiosyncratic characteristics of one or several of the initial ''seeds'').

Thus, in this paper we report on an examination of our sample to determine whether it was ''successful,'' specifically in terms of the criteria that others have established to define the quality of recruitment. These criteria include two elements, convergence and representativeness. 1) Convergence of the sample: A mean absolute discrepancy between the actual sample and the theoretically computed equilibrium sample composition smaller than 2% indicates convergence. When calculating mean discrepancy across more than two groups, weighted mean absolute discrepancy is used where the frequency of recruits in each group was used as the weight [11]; 2) The inferred representativeness of the sample: Once it has been established that the sample waves reach convergence, representativeness is demonstrated by non-significant discrepancies by *t*-test between actual sample compositions and estimated corresponding population compositions (e.g., once there has been convergence on gender composition in successive waves, it is possible to calculate the likely gender composition of the population from which the sample was drawn), the latter of which could be computed by RDSAT computer program that is available free at www.respondentdrivensampling.org. This is computed using the equation shown in Appendix A of Wang (2005) [16].

In addition, we used geographic information system (GIS) technology to map the expanding geographic pattern of RDS from wave to wave, an innovation that has, to the best of our knowledge, not been used in previous RDS studies. We recognized the possibility that our sample might well be representative within a limited geographic region, but yet fail to sufficiently recruit farther-away migrant workers for whom participating in the study would be very inconvenient. Mapping the sample thus allowed a means of examining whether our subjects came primarily from one region of Chengdu, or more broadly reflected the residential sites where migrants are known to live. It thus served as an external check that complemented the internal criteria that have been used previously.

Materials and methods

Respondents

The population of interest for this study included migrant workers in Chengdu, Sichuan Province, China. To be eligible for the study, respondents had to satisfy three inclusive criteria: 1) respondent did not hold a *hukou* indicative of living in central areas or near suburbs of Chengdu city; 2) respondent must have been 16 years or older; 3) respondent could not be a student. Face-to-face interviews were conducted in the Wuhou district, which is located downtown and conveniently accessible using public transport. Informed consent was obtained from all participants following a protocol that was approved by Sichuan University Medical Ethic Committee.

Implementation of RDS

The RDS was implemented so that the final sample would be independent of the number and types of seeds, irrespective of the strategy used to recruit seeds. Using a heterogeneous group of seeds, we accelerated the speed of achieving equilibrium; ethnographic research was used to identify seeds [17,18]. In this study, we used a total of 12 seeds after taking gender, age, occupation, and residential sites into consideration. It is notable that the optimal number of seeds to recruit a particular sample size is still being tested. Experience to date has suggested that between six and 20 seeds is preferable, depending on the desired sample [19].

Each seed was given three coded coupons to recruit peers. We then consented and enrolled persons who presented one of these (valid) coupons and who we deemed eligible; in turn, each new enrollee was given three coded coupons for the purpose of recruiting peers. A dual incentive system was used: A ''primary incentive'' of RMB40 (about US$5.9) was paid to each participant as compensation for the time spent completing an interview, which lasted about 1–1.5 hours. The ''second incentive'' was built using a laddered method to promote enthusiasm for referring, which included RMB10 (about US$1.5) for a single referral to the project, RMB25 (about US$3.7) for two, and RMB45 (about US$6.6) for three.

In order to control repeated participation, an ID number registration system was created. In China, every person has an identification card with a unique number and a photo; ID numbers were registered once each person provided consent. Upon enrolment, new potential participants were asked to show their ID, which was checked against our list of previous participants to assure that they were new to the study. In addition, respondents were asked to

report the number of potential participants they knew, which was used as well as part of our RDS analysis.

Analytic approach

Sample analyses were conducted following completion of the RDS process. As a first step, we described the seeds and respondents in each wave, including the productivity of the seeds, the referral chains connected to each seed, and compositional characteristics of the sample. The characteristics selected for sample analysis included gender (1 – males; 2 – females), age (1 – age 16–25; 2 – age 26–35; 3 – age 36–45; 4 – age $\geq$ 46), occupation settings (1 – construction; 2 – manufacturing; 3 – restaurant and entertainment; 4 – commercial; 5 – services; 6 – office worker; 7 – unemployed). Next, the characteristics of the actual sample composition were described and the equilibrium sample composition was calculated based upon the characteristics of the convergent waves. From this, the estimated population composition was computed. Analyses were done by RDSAT computer program.

Weighted discrepancy between actual sample and equilibrium sample was computed using frequency of recruits in each group as the weight. And differences between actual sample and estimated population were tested by using the equation shown in Appendix A of a paper of Wang et al. [16].

Geographical mapping

A geographic information system is a collection of computer hardware, software and geographic data used to analyze and display geographically referenced information. GIS provides a method by which geographically dependent data can be displayed in a readily accessible visual format.

In this study, each respondent was asked for the location of residence that in turn was mapped by a geographic information system (ArcGIS). In the first stage, a digital map of central area Chengdu was vectored in terms of main roads and boundaries, and residential location. Then vectored data were rendered and developed into thematic maps. In this way, the geographic distribution of respondents of each wave and subsequent waves could be easily visualized from the map.

Coordinates of the investigation site (Wuhou district CDC) and residential locations of respondents were recorded, which were in turn used to calculate maximum, minimum, and median distances between the investigation site and residential locations of respondents in each wave, which was intended to allow us to track the expanding response pattern of RDS.

Results

Descriptive results

The study was initiated in September 2008 and completed in June 2009. Twelve seeds were identified. Of the 12 seeds, one seed did not refer anyone into the study, and one seed referred only one participant. They were defined as ''infertile'' [16]. Therefore, a total of 1,256 respondents were recruited based upon 10 seeds: 10.5% of respondents were referred by their relatives, 9.1% by people from the same county (Lao Xiang), 29.7% by colleagues, 46.7% by friends, and 4.1% by friends of friends.

The increases in the number of recruits, as well as the sample compositions over recruitment waves in terms of gender, age, and occupation are shown in Table I. Although gender, age, and occupation were taken into consideration at the stage of seed selection, the initial sample of 10 was viewed by us as deficient in two respects. No seeds in the group were older than age 46 years, and none were working in construction sites. The former may have reflected a fundamental age skew among migrant workers, but it initially raised concerns. The latter clearly did not represent a bias among workers, because migrants compose much of the local pool of construction workers in Chengdu; rather it was an unanticipated outcome of our initial approach to sampling. Fortunately, as the referral chain grew over the time of our sampling, the composition of each wave changed and gradually stabilized. The final sample compositions are shown in the last row of Table I.

Results from RDSAT

Because seeds were obtained by a different sampling strategy, they were not included in the RDSAT analysis. The results of sample analysis are shown in Table II. The weighted mean absolute discrepancies between the actual and equilibrium sample composition were 0.3%, 0.6%, and 0.4% for gender, age, and occupation, respectively, all of which were much less than 2%, indicating equilibrium had been achieved. Sample representativeness testing results showed that males ($t = 3.61$, $p < 0.001$), people in the age group of 26–35 ($t = 3.85$, $p < 0.001$), and people working in offices were over-sampled ($t = 5.286$, $p < 0.001$). In contrast, females ($t = 3.16$, $p < 0.001$), people aged 46 years old and above ($t = 3.222$, $p < 0.001$), and

Table I. Sample size and gender, age, and occupation compositions of each wave.

Wave	*n*	Change in recruits	Gender(*n*)(%)		Age(*n*)(%)				Occupation[a] (*n*)(%)						
			Male	Female	16–25	26–35	36–45	≥46	1	2	3	4	5	6	7
0	10	-	7(70.0)	3(30.0)	3(30.0)	3(30.0)	4(40.0)	0(0.0)	0(0.0)	1(10.0)	1(10.0)	3(30.0)	3(30.0)	1(10.0)	1(10.0)
1	26	26	11(42.3)	15(57.7)	8(30.8)	16(61.5)	1(3.8)	1(3.8)	0(0.0)	4(15.4%)	1(3.8%)	13(50.0%)	4(15.4%)	4(15.4%)	0(0.0)
1–2	85	59	22(37.3)	37(62.7)	14(23.7)	31(52.5)	7(11.9)	7(11.9)	2(3.4)	5(8.5)	10(16.9)	13(22.0)	17(28.8)	5(8.5)	7(11.9)
2–3	212	127	60(47.2)	67(52.8)	39(30.7)	42(33.1)	31(24.4)	15(11.8)	8(6.3)	9(7.1)	19(15.0)	35(27.6)	37(29.1)	8(6.3)	11(8.7)
3–4	441	229	100(43.7)	129(56.3)	89(38.9)	65(28.4)	54(23.6)	21(9.2)	16(7.0)	16(7.0)	32(14.0)	44(19.2)	78(34.1)	28(12.2)	15(6.6)
4–5	902	461	236(51.2)	225(48.8)	162(35.1)	110(23.9)	129(28.0)	60(13.0)	43(9.3)	33(7.2)	67(14.5)	96(20.8)	149(32.3)	41(8.9)	32(6.9)
5–6	1,256	354	165(46.6)	189(53.4)	164(46.3)	85(24.0)	77(21.8)	28(7.9)	37(10.5)	26(7.3)	47(13.3)	63(17.8)	132(37.3)	24(6.8)	25(7.1)
Total	1,266	1,256	601(47.5)	665(52.5)	479(37.8)	352(27.8)	303(23.9)	132(10.4)	106(8.4)	94(7.4)	177(14.0)	267(21.1)	420(33.2)	111(8.8)	91(7.2)

[a]Occupation: 1 – construction; 2 – manufacturing; 3 – restaurant and entertainment; 4 – commercial; 5 – services; 6 – office work; 7 – unemployment [20].

unemployed people were under-sampled ($t = 2.333$, $p < 0.05$).

GIS results

The four circles in Figure 1 are the first to the fourth rings surrounding the centre of Chengdu city, respectively. Main roads and boundaries were used to divide central areas of Chengdu into 27 areas. The area within the third ring is usually regarded as the downtown area of Chengdu; the area between the third and the fourth ring is suburban; and the area out of the fourth ring is exurban. The number of respondents living in downtown, suburban, and exurban areas is described in Table III.

Figure 1 shows the cumulative changes in the distribution of residential locations over waves. Eight of the ten seeds were located within the third ring, as shown in Figure 1(a). Distribution of respondents in wave 1 was accordant with seeds as shown in Figure 1(b). In waves 2 and 3, distribution of respondents was obviously expanded as shown in Figure 1(c) and 1(d). In waves 4, 5 and 6, as shown in Figure 1(f)–1(g), respondents did not go farther. The five-pointed star in the map refers to the investigation site (Wuhou district CDC).

Most respondents were living within the third ring of Chengdu, and the areas of most density were 3, 4 and 21, which were the closest to the investigation site. Out of the third ring, respondents were scattered, except two clusters in area 9 and the location between area 19 and 26, respectively.

Minimum distance, maximum distance, median distance, percentiles 25 and percentiles 75 distances between residential locations of respondents and investigation site of each wave were calculated based on coordinates information, the results of which are shown in Table IV. Through the whole sampling process, the minimum distance was 251 m, and the maximum distance was 18,300 m. The median distance increased first, and then decreased since the third wave.

Discussion

Seed selection and dual incentive system

Identifying an effective seed is a critical issue in RDS. In the first stage of our study, we found it was very difficult to obtain the trust of potential seeds in a short time, no matter how we might present the study to them. Even though some showed their willingness to participate in a research interview, they refused when they learned that these were to be conducted in a specific place. Therefore, we changed

Table II. Descriptions of gender, age, and occupation compositions of the actual samples, equilibrium samples, and estimated populations and results of differences tests between actual sample and equilibrium sample, and actual sample and estimated population.

Characteristic of recruiters	Characteristics of recruits						
Gender	Gender						
	Male		Female				
Actual sample composition (p_s)	47.3%		52.7%				
Equilibrium sample composition (p_e)[d]	47.0%		53.0%				
Population composition ($\hat{p}$)(C.I)	40.8% (37.4%, 44.4%)		59.2% (55.6%, 62.6%)				
Weighted mean absolute discrepancy between p_s and p_e[c]	0.3%						
t-test for (p_s–$\hat{p}$)[d]	$t = 3.61$, $p < 0.001$		$t = 3.16$, $p < 0.001$				
	Age						
Age	16–25		26–35		36–45		≥46
Actual sample composition (p_s)	37.9%		27.8%		23.8%		10.5%
Equilibrium sample composition (p_e)	37.2%		27.3%		24.4%		11.1%
Population composition ($\hat{p}$)(C.I)	34.4% (30.1%, 38.7%)		22.8% (20.4%, 25.7%)		26.5% (22.7%, 30.8%)		16.3% (12.3%, 19.5%)
Weighted mean absolute discrepancy between p_s and p_e	0.6%						
t-test for (p_s-$\hat{p}$)[d]	$t = 1.591$, $p > 0.1$		$t = 3.85$, $p < 0.001$		$t = 1.286$, $p > 0.1$		$t = 3.222$, $p < 0.001$
	Occupation						
Occupation[a]	1	2	3	4	5	6	7
Actual sample composition (p_s)	8.4%	7.4%	14.0%	21.0%	33.2%	8.8%	7.2%
Equilibrium sample composition (p_e)	9.2%	7.4%	14.6%	20.9%	32.7%	7.8%	7.4%
Population composition ($\hat{p}$)(C.I)	10.1% (7.2%, 13.4%)	6.9% (5.1%, 8.6%)	13.5% (10.4%, 16.3%)	19.7% (16.7%, 23.0%)	34.0% (30.2%, 37.5%)	5.1% (3.9%, 6.5%)	10.7% (8.2%,14%)
Weighted mean absolute discrepancy between p_s and p_e	0.4%						
t-test for (p_s-$\hat{p}$)[d]	$t = 1.063$, $p > 0.2$	$t = 0.556$, $p > 0.5$	$t = 0.333$, $p > 0.5$	$t = 0.813$, $p > 0.4$	$t = 0.444$, $p > 0.5$	$t = 5.286$, $p < 0.001$	$t = 2.333$, $p < 0.05$

[a]Occupation: 1 – construction; 2 – manufacturing; 3 – restaurant and entertainment; 4 – commercial; 5 – services; 6 – office work; 7 – unemployment. [b]Equilibrium compositions were computed by solving the linear equations shown in equation (1) of Heckathorn [10]. [c]Weighted mean absolute difference between the actual and equilibrium sample composition. Frequency of recruits in each group was used as the weight [16]. [d]t-test was computed by equation shown in Appendix A of Wang [16].

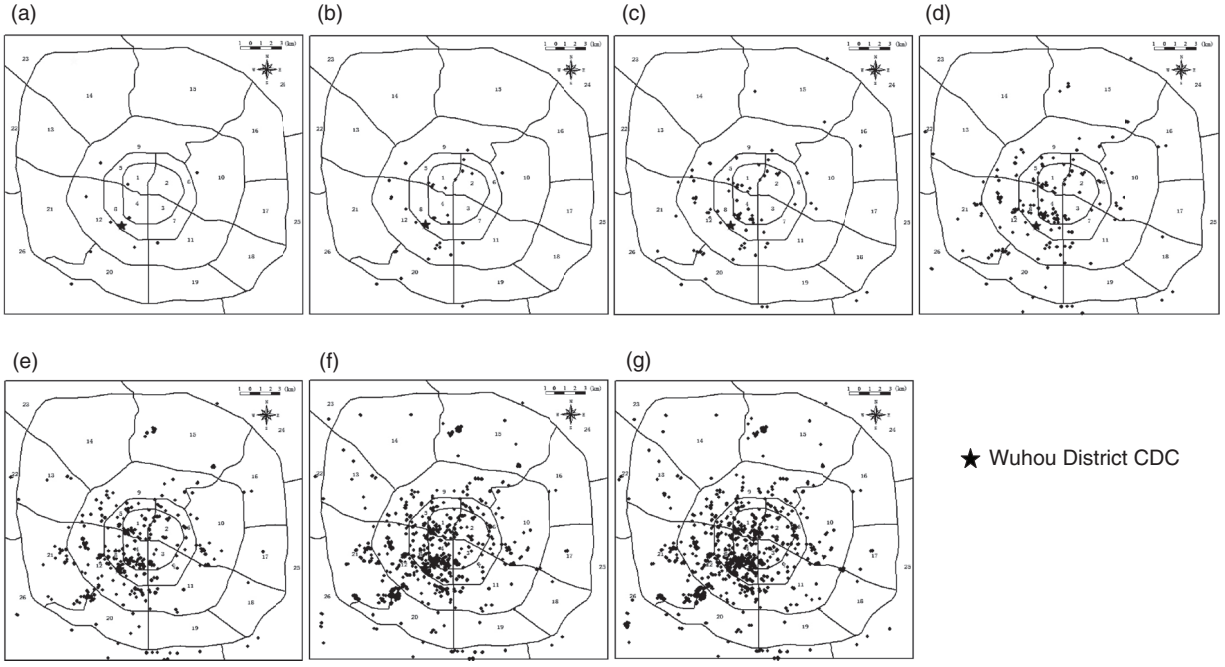

Figure 1. Cumulative changes in the distribution of residential locations over waves. (a) Zero wave; (b) first wave; (c) second wave; (d) third wave; (e) fourth wave; (f) fifth wave; (g) sixth wave.

Table III. The number of respondents and corresponding compositions in downtown, suburban, and exurban areas.

	Area	GIS code	*n* (%)
Downtown	With the first ring	1, 2, 3, 4	453 (35.8)
	Between the first and the second ring	5, 6, 7, 8	229 (18.1)
	Between the second and the third ring	9, 10, 11, 12	281(22.2)
Suburban	Between the third and the fourth ring	13, 14, 15, 16, 17, 18,19,20,21	250 (19.7)
Exurban	Out of the fourth ring	22, 23, 24, 25 26	53 (4.2)

GIS: geographic information system.

Table IV. Distances between living locations of respondents and investigation site of each wave.

Wave	Minimum distance (m)	Maximum distance (m)	Median (m)	Percentiles 25 (m)	Percentiles 75 (m)
Zero wave	990.8	8,760.8	3,911.6	2143.4	6523.0
First wave	915.0	8,519.6	3,998.1	2,498.7	6173.8
Second wave	949.67	18,300.0	4,322.2	2,239.6	6542.1
Third wave	1,075.3	17,700.0	4,284.3	2,207.9	6933.0
Fourth wave	269.1	13,900.0	3,862.1	2,143.7	5919.4
Fifth wave	273.2	17,100.0	3,897.9	2,487.9	6457.8
Sixth wave	251.0	17,800.0	3,653.8	1,585.7	5635.4

our strategy to identify seeds introduced by research team members or team members' friends, which turned out to be successful.

This strategy had two advantages. First, mutual confidence between seeds and the research team was easier to establish, and as a result, our communication was more informal and informative. The better that seeds understood the objectives and procedures of the study, the greater the possibility of their recruiting well-informed, cooperative peers. Second,

the connection between the seeds and the team were more robust – that is, more effectively sustained – which led to a greater percentage of referrals. Eight of the 12 seeds (66.7%) recruited at least three peers, which greatly exceeds the 11.8% reported in a previous study done in MDMA (3,4-Methylenedioxymethamphetamine (''Ecstasy'')) users [16].

RDS has a dual incentive system – rewarding initial participation and peer referral. An intensive ''laddering'' strategy was employed in the study, which we think would promote activity of respondents to refer their peers. Among people who did referral, 80.4% referred three peers into the study, 10.6% referred two peers, and 9.1% referred one peer. This pattern was different from the previous ''benchmark'' study, in which 44.1% had a single recruit, 31.8% had two, 16.4% had three, and 7.7% had more than three [16]. The laddering incentive system apparently had a positive motivational impact on respondents. Beyond this direct financial incentive, we observed that peer pressure played an important role. For example, several participants reported that they came to the study to support their friend and not for the direct incentive – statements which had greater credibility when we noted their relative high incomes.

Equilibrium and representativeness of the sample

The results indicated that the sample achieved equilibrium, implying that the referral chain of the RDS method of recruitment would converge quickly for migrant workers. However, an RDS sample may not necessarily be representative of the target population even though sample composition reaches equilibrium due to discrepancies between the earliest waves and subsequent waves that reached convergence. In our study, the results showed that females, people aged 46 years old and above, and unemployed migrants were under-represented (Table II).

Our data suggest that, compared with previous studies, we had a smaller overall proportion of male migrants and construction workers in our estimated population, and a larger proportion of migrants aged 46 years and above. There were several potential explanations. The targeted population for this study was different from others. We focused on recruited migrant workers in Chengdu, who included rural-to-urban migrants and urban-to-urban migrants; in contrast, many other studies dealt exclusively with rural-to-urban migrant workers.

A second explanation reflects the gender composition of migrant workers in Sichuan. According to the Year 2000 Census data for the province, there were 87 male for every 100 female workers [21], which may be different from that of the national level or other province. The ratio in the present study was 69 male workers for every 100 females, which was lower than that in the Census data. This difference may reflect geographic limitations of RDS, which we will discuss, and it might also result in part between changes in the migrant population between 2000 and June 2009.

Thirdly, different sampling methods were employed. In previous studies, non-probability sampling strategies were employed so that construction sectors and factories were the most popular study sites for researchers. These sampling strategies were more likely to recruit male and younger migrant workers working in construction and manufacturing sectors, but less likely to include dispersed migrant workers, such as those who work for themselves, babysitters, unemployed people, and so on. RDS has its unique advantages in penetrating a targeted population and recruiting scattered subjects. Because most of the construction sector and manufacturing factories are located outside the third ring of Chengdu, which is a long way from the investigation site, it could inhibit including people working there into our study. This is attributed to the geographic limitation of RDS.

These differences provided us a picture of characteristics of migrant workers in Chengdu – that is, having more female migrant workers and having more migrants working in commercial sectors. They also indicated that previous studies might have underestimated the proportion of female migrants, and migrants aged 46 or above. Since construction migrant workers, factory migrant workers, and entertainment migrant workers are easier to recruit into a study compared with migrant workers working in another area, most previous studies chose them as study subjects. RDS provides us an alternative and feasible way to obtain a better representative sample of migrant workers.

However, RDS has two limitations. The estimated compositions are asymptotically unbiased under the assumption of random selection of peers from personal networks. This is difficult to fulfill in practice. Fortunately, Heckathorn's simulation study showed that RDS was able to generate a sample that provides a good cross-section of the targeted hidden population [9]. Although the RDS sample was not a random sample of a targeted population, it could be considered the best one could expect from a target population where random sampling was impossible.

Geographic distribution of respondents by RDS

The other limitation that we uncovered related to the geographic challenges associated with a study using

RDS. We estimated that in our study the numbers of construction workers and manufacturing workers were underestimated due to the specific geographic distribution of these sectors in Chengdu. Our GIS maps visually depicted the distribution of the residential locations of respondents and how the distribution changed with additional waves. As seen in Figure 1, higher responder density was associated with residing closer to the investigation site, with most respondents residing in downtown and suburban areas of Chengdu. These findings reinforce the observations that we had less recruitment from the construction sites that were primarily near the 3rd ring. RDS, when it is tied to a face-to-face interview, appears to have an underlying spatial dimension that complements its social network assumptions. We, in retrospect, detected an apparent relationship between incentive and distance-to-interview site. By computing for each wave the minimum distance, maximum distance, median distance, and the 25th–75th percentile distances between residential locations of respondents and the interview site, we found that respondents expanded in the first three waves and then stopped going farther under current incentive strategy.

It appears that respondents computed their own cost–benefit analyses, including accessibility, and time and transportation costs. Potentially, one might expand the sampling range further by increasing the incentive. Alternatively, increasing the number of investigation sites likely enhances representativeness by promoting respondent convenience, and lessens the tendency for potentially biasing ''cost–benefit'' analyses.

Thus, even if one achieves a fully ''representative'' sample using the internal modelling criteria that have been employed in this and other studies, it may be entirely possible to miss a meaningfully different segment of a potential target population based on the location of interview sites. This finding adds to our understanding of the sampling strengths and limitations of on-the-ground RDS.

Conclusions

Given that respondents refer their peers into further waves of a study, our results reinforce the belief that RDS can be used as a strategy for sampling difficult-to-identify migrant workers, in addition to socially marginal individuals. Thus these findings have practical implications for future applications of RDS to research involving migrant workers in China, and perhaps in other countries as well. Moreover, researchers do not need to employ field workers who struggle to identify subjects in dispersed communities, saving resources and improving access to people who work is outside the scope of most researchers community ''reach.'' Since each respondent has self-selected, study participants are very likely to be cooperative, which further enhances the quality of the study.

RDS also has some disadvantages that must be recognized for future studies. First, researchers do not control when potential respondents would appear. For example, at the beginning of a study, there might be a few, but when the chain extends to the third wave, an investigational site may be inundated, overwhelming the capacity of researchers to efficiently interview new participants. To avoid this possible complication, we recruited our initial 12 seeds in a staggered fashion. We also scheduled each respondent's interview as a way to control the flow of new entries. Together these approaches proved to be effective.

RDS is built upon an intentional sample bias; that is, it depends specifically on peers recruiting peers. We used RDS successfully for recruiting migrant workers into this study. This method had the advantage of penetrating into migrant groups that lay beyond the reach of outside investigators, and enabled us to reach highly dispersed workers who would have been much more challenging to identify using conventional non-probability sampling methods. Even as we see great advantages for RDS methods for future epidemiological or sociological studies of migrant workers, it is important to recognize and mitigate potential limitations, such as geographic proximity. We suggest a multisite strategy to decrease such effects for future studies.

Funding

This work was supported, in part, by grants from the National Natural Science Foundation of China [Grant number: 70673067]; the National Doctoral Foundation of Ministry of Education of China [Grant number: 20060610087]; the Fogarty International Center, National Institutes of Health, USA [Grant number: 2 D43 TW005814]; and the China Medical Board [Grant number: 08-905].

References

[1] National Bureau of Statistics of China. National Statistical Communiqué on 1% population sampling survey in China in 2005 [in Chinese]. 2006.

[2] Chan KW. Post-Mao China: a two-class urban society in the making. Int J Urban Reg Res 1996;20:134–50.

[3] Li P. Rural-to-urban migrant workers: social and economic analysis on rural-to-urban migrant workers in China [in Chinese]. Beijing: Social Science Academic Press; 2003.

[4] Dong X, Bowles P. Segmentation and discrimination in China's emerging industrial labour market. China Econ Rev 2002;13:170–96.

[5] Lee CK. Production politics and labour identities: migrant eorkers in South China. In: Lo CK, Pepper S, Tsui KY, editors. China review. Vol. 15. Hong Kong: The Chinese University Press; 2003. pp. 1–28.

[6] Fu D, Wong K, He X, Leung G, Lau Y, Chang Y. Mental health of migrant workers in China: prevalence and correlates. Soc Psychiatry Psychiatr Epidemol 2008;43:483–9.

[7] Chen Z, Zhang X, Chen X, Chen B, Wei P, Hu H, et al. Relationship between depression and self-rated health among floating population [in Chinese]. Chinese J Health Educ 2006;22(10):747–9.

[8] Li X, Fang X, Lin D, Mao R, Wang J, Cottrell L, et al. HIV/STD risk behaviors and perceptions among rural-to-urban migrants in China. AIDS Educ Prev 2004; 16(6):538–56.

[9] Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. Soc Probl 1997;44(2):174–99.

[10] Heckathorn DD. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Soc Probl 2002;49(1):11–34.

[11] Heckathorn DD, Semaan S, Broadhead RS, Hughes JJ. Extensions of respondent-driven sampling: a new approach to the study of injection drug users aged 18–25. AIDS Behav 2002;6(1):55–67.

[12] Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent driven sampling. Sociol Methodol 2004;34:193–239.

[13] Wang J, Falck RS, Li L, Rahman A, Carlson RG. Respondent-driven sampling in the recruitment of illicit stimulant drug users in a rural setting: findings and technical issues. Addict Behav 2007;32:924–37.

[14] Liu H, Liu H, Cai Y, Rhodes AG, Hong F. Money boys, HIV risks, and the associations between norms and safer sex: a respondent-driven sampling study in Shenzhen, China. AIDS Behav 2009;13:652–62.

[15] He Q, Wang Y, Li Y, Zhang Y, Lin P, Yang F, et al. Accessing men who have sex with men through long-chain referral recruitment, Guangzhou, China. AIDS Behav 2008;12(supplement 1):93–6.

[16] Wang J, Carlson RG, Falck RS, Siegal HA, Rahman A, Li L. Respondent-driven sampling to recruit MDMA users: a methodological assessment. Drug Alcohol Depend 2005;78:147–57.

[17] Draus PJ, Siegal HA, Carlson RG, Falck RS, Wang J. Cracking the cornfields: recruiting illicit stimulant drug users in rural Ohio. Sociol Q 2005;46:165–89.

[18] Falck RS, Siegal HA, Wang J, Carlson RG, Draus PJ. Nonmedical drug use among stimulant-using adults in small towns in rural Ohio. J Subst Abuse Treat 2005;28:341–9.

[19] Malekinejad M, Johnston LG, Kendall C, Kerr LRFS, Rifkin MR, Rutherford GW. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. AIDS Behav 2008;12(Suppl 1):S105–30.

[20] Qiu P, Caine E, Yang Y, Chen Q, Li J, Ma X. Depression and associated factors in internal migrant workers in China. J Affect Disord 2011;00:0–0 (accessed 31 August 2011).

[21] Lu N. Analysis on demographic characteristics of Sichuan migrants [in Chinese]. Statistic Educ 2005;8:15–18.