

# A Study on the Capture-Recapture Method for Estimating the Population Size of Injecting Drug Users in Southwest China

Rongsheng Luan, Boheng Liang, Ping Yuan, Li Fan, Yong Huang, Gang Zeng, Linglang Wang, and Shan Wang\*

Department of Epidemiology, West China School of Public Health, Sichuan University, Chengdu, Sichuan 610041, P.R. China

(Received December 21, 2004; Accepted March 5, 2005)

We used the capture-recapture method to estimate the population size of injecting drug users (IDU). We collected and analyzed registry data from the Drug Use Rehabilitation Center in Leshan Downtown District, designed the interval between capture phase and recapture phase, explored the data model of the capture-recapture method (CR), and estimated the size of a local IDU population. The estimated size was 3455. Establishment of a consistent data model played a key role in estimating the IDU population size using CR. CR is an easy-to-use and inexpensive technique for IDU population size estimation, which generates reliable results.

**Key words** — capture-recapture method, injecting drug user, data model, population size estimation

## INTRODUCTION

In 2003, a total of three million people were killed by AIDS globally and five million became newly infected. Total of HIV infection cases were estimated to be million worldwide by the end of 2003,<sup>1)</sup> with 10–20% attributing to syringe sharing, especially in South America, Western Europe, mid-Asia, the Middle East and Southeast Asia (except for Thailand), where AIDS is primarily triggered by injecting drug use.<sup>2)</sup> The cumulative number of HIV infection cases reported was 102 thousand (85–102 thousand) in Mainland China, and cases related to injecting drug use had the largest share of 68–71%. The situation was particularly critical in Dehong and Ruili in Yunan Province, Yining in the Xingjiang Autonomous Region, and Liangshan in Sichuan Province. Reliable estimates of the population size of injecting drug users (IDUs), a population at high risk for HIV infection, are significant.

IDUs are often a hidden or hard-to-reach population.<sup>3)</sup> Since drug use is illegal, detailed information on the population is rarely accessed in a direct

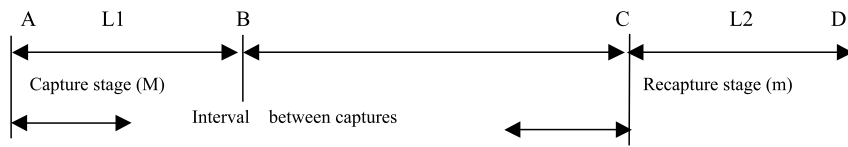
manner.<sup>4)</sup> Routine methods cannot be used in surveys of this population. In our study, we used an indirect approach, the capture-recapture method (CR), to estimate the number of IDUs. Based on the registry data from the Drug Use Rehabilitation Center in Leshan Downtown District, we designed the periods of the two captures and the interval in between and established the data model to calculate the number of the population.

## MATERIALS AND METHODS

**Rationale** — CR was originally developed to estimate the number of wild animals in closed areas. At one time as many animals as possible in an area are captured, tagged (making marks on the captured allowing them to be easily recognized in the population) and released which is the capture stage. At a later time this is repeated which is the recapture stage. The number of animals in each sample, and the number common to both, are used to estimate the number of the total population (assuming that capture and recapture are independent). CR later became popular in research on human diseases and health issues. This methodology is categorized into two types according to the number of samples: two-sample CR and multiple CR.

In this study, we used two-sample CR, the total

\*To whom correspondence should be addressed: Department of Epidemiology, West China School of Public Health, Sichuan University, Chengdu, Sichuan 610041, P.R. China. Tel.: +86-028-8550164; Fax: +86-028-85443652; E-mail: aupers100@163.com, aupers100@tom.com



**Fig. 1.** Illustration of the Rational

**Table 1.** Interval Distribution Two or More Admissions of IDU to the Drug Use Rehabilitation Center

The total number of captured	Intervals (month)										
	< 3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	> 30
412	37	47	74	42	35	27	22	23	17	20	68

of individuals in a population is N. From the population, a total of M individuals are randomly captured, tagged and released back to the population, which is regarded as the capture stage. After an interval, a recapture stage occurs and another sample of n individuals is captured, among whom are some tagged individuals (m). When the samples are independent, the relation below should hold (as Fig. 1 shows):

$$\frac{\text{Number captured in the first stage } (M)}{\text{Total number of the group } (N)} \times 100\% = \frac{\text{Number of recaptured } (m)}{\text{Number captured in second stage } (n)} \times 100\% .$$

**Formula** — We adopted the non biased Chapman formula rather than the original formula, which would have resulted in biased estimation.

$$N = [(M + 1)(n + 1)/(m + 1)] - 1$$

$$95\% \text{ confidence interval of } N: N = \pm 1.96 \sqrt{\text{Var}(N)}$$

$$\text{Var}(N) = \frac{[(M + 1)(m + 1)(M - m)(n - m)]}{[(m + 1)^2(m + 2)]}$$

The CR method was first used for surveys in demography in the 1940s<sup>5)</sup> and became widespread after Wittes *et al.*'s work in the 1970s.<sup>6)</sup> This method has been applied to research on hidden populations in recent years, such as sex workers (SWs), homeless people, and IDUs. The specific methodologies have varied. Some groups used variations similar to the original version, directly capturing, tagging and recapturing human samples.<sup>7-9)</sup> Others turned to indirect applications of CR by utilizing data collected from clinics, treatment centers, or other social services.<sup>10-12)</sup> In epidemiology, CR often estimates disease prevalence by comparing information from two

or more sources such as register records, personal status or social investigations.<sup>13-15)</sup> In our study, an individual was defined as being captured upon his or her admission to the Drug Use Rehabilitation Center.

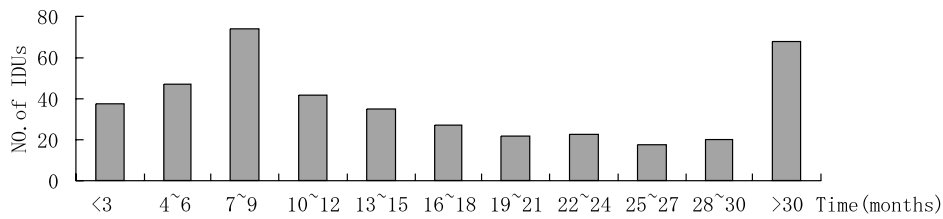
**Subjects** — All IDUs registered during the first capture period (L1) and the second capture period (L2) were included.

## RESULTS

### Data Modeling

*Collecting and Preparing Data:* The registries of all IDUs admitted to the Drug Use Rehabilitation Center, including personal identity status (*e.g.*, name, gender, and occupation), time of admission, and time of discharge. A database was established with the prepared and checked data as the basis of data modeling. Our study utilized the personal status of IDUs admitted to the Drug Use Rehabilitation Center in Leshan Downtown District from April 1, 1997, to October 31, 2002.

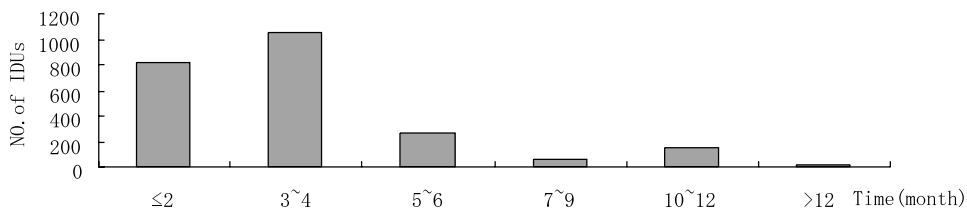
*Defining the Interval between the Capture Stage and the Recapture Stage:* We estimated the local IDU population size using the registries of IDUs admitted to the Rehabilitation Center. To meet the methodology requirements and generate reasonable estimates, we analyzed the registries and determined the central tendency of intervals between two or more admissions of an IDU to the Rehabilitation Center and used the central tendency as the interval between two captures (P) (Table 1 and Fig. 2). We also calculated the central tendency of the IDUs' stay in the Rehabilitation Center to assign the capture stage (L) (Table 2 and Fig. 3). In Table 1 shows the intervals between two or more admissions of IDUs to the Rehabilitation Center.



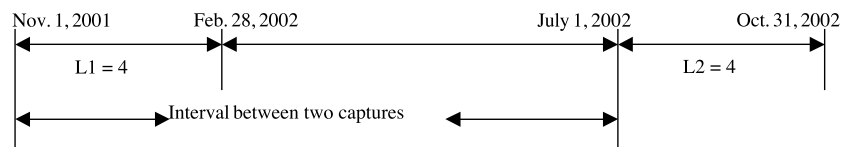
**Fig. 2.** Intervals between Two or More Admissions to the Drug Us Rehabilitation Center

**Table 2.** Time Distribution of IDUs Stay in the Drug Use Rehabilitation Center

Total	Time in Rehabilitation Center (month)						Mean (month)
	≤ 2	3-4	5-6	7-9	10-12	> 12	
2349	809	1047	264	64	151	14	3.48



**Fig. 3.** Length of Stay in the Drug Use Rehabilitation Center



**Fig. 4.** Illustration of Study Schedule

A total of 412 person-time admissions were recorded for two or more IDU admissions to the Rehabilitation Center between April 1997 and October 2002. The intervals between two or more admissions generally concentrated in a period of 7 to 9 months (although a considerable number of IDUs were admitted after an interval > 24 months, and the consistency of the captured population could not be confirmed because of the lengthy interval). Therefore we selected a period of 8 months for the interval between two captures (P).

Table 2 shows the periods of time for which IDUs stayed in the Rehabilitation Center according to the registry data. As indicated in Fig. 3, the period of time for which most IUDs stayed in the center concentrated on 3 to 4 months, for a means of 3.48 months. We therefore selected 4 months for the period of capture (L).

Assigning the starting times of the two captures (A–D): To estimate the current population size of IDUs, we chose the previous four months (July 1, 2002–October 31, 2002) for the time of the second capture (L2). By scheduling 8 months backward, we obtained the time of the first capture (L1) (November 1, 2001–February 28, 2002) (Fig. 4).

The number of IDUs registered during the first capture period (L1) was assigned for M, and that during the second capture period (L2) after the interval (P) was assigned for n. The number of IDUs common to both L1 and L2 was calculated for m:  $M = 71, n = 191, m = 3$ .

**Estimated Size**

$$N = [(71 + 1)(191 + 1)/(3 + 1)] - 1 = 3455$$

95% confidence interval: (542, 6368).

The result was close to the number acquired us-

ing the Delphi method based on data from local police authorities.

## DISCUSSION

Sentinel surveillance is predominant in the Chinese AIDS surveillance system. The target populations of the system include drug users, commercial SWs, and commercial sex clients. These people are often referred to as hidden populations, who are hard to reach, and so far there have no golden standards for their size estimation. Conventional methods like random digit dialing (RDD) are appropriate for large populations but are not capable of estimating hard-to-reach groups. CR method is an indirect approach, that has no specific requirements for the a target population.

IDUs are one major vulnerable population at high risk of contracting AIDS in China. A reliable estimate of their population size has substantial significance for estimating the population infected with AIDS and formulating health policies. In this study, we used CR for three main reasons. First, drug use is most likely to be underreported in a routine survey due to its illegal and stigmatized nature.<sup>16)</sup> Second, the drug use lifestyle drives the users away from ordinary households, which results in less participation in routine surveys. Third, drug use rehabilitation centers are our primary source of information on IDUs, whose data are often neither complete nor comprehensive.

The following assumptions must be fulfilled to use CR: the two captures take place with an interval in between, which is sufficiently long for the tagged individuals to mingle in a population; the population is closed during the period of study, a stable population results in a precise estimate; missing tags should be avoided, so that the recaptured individuals can match those who are tagged in the first capture (*i.e.*, individuals common to both samples can be identified); all individuals in the population share an equal probability of being captured; and the two samples are independent, *i.e.*, an individual's probability of being captured in one sample does not affect his or her probability of being captured in the other.

Furthermore, when CR is employed to estimate a population size based on existing data,  $M$ ,  $n$ , and  $m$  vary with the differently assigned periods of capture ( $L$ ) and interval between two captures ( $P$ ). This leads to varied estimates. Therefore, ascertainment

of  $L$  and  $P$  is the key to successful CR application. Our study offers a preliminary strategy and method for ascertaining  $L$  and  $P$ .

Assigning the interval ( $P$ ): According to the CR rationale, a reasonable interval should allow the captured individuals to mingle in the population and eliminate the influence of nonrandom factors (*e.g.*, policy factors) upon an IUD's admissions to the Drug Use Rehabilitation Center. The strategy was to use the central tendency of intervals between two or more admissions to the Rehabilitation Center as the interval between the two captures ( $P$ ).  $P$  then represented the interval between two or more admissions of most IDUs and exclude the influence of incidental factors.

Assigning the periods of captures ( $L_1$  and  $L_2$ ): To ensure the independence of the two samples and all individuals' equal probability of being captured, the period of each capture must be properly assigned. The periods should not be too short, otherwise IDUs captured in the first sample might miss the recapture stage when they were still residing in the Rehabilitation Center. Neither should the time be so long that one IDU might be present twice in one sample. Our solution was to designate the central tendency of the periods when IDUs remained in the Rehabilitation Center as the period of capture ( $L$ ). This is guaranteed that the majority of subjects would return to the population and become well mingled to be recaptured.

Fixing A–D: According to CR methodology, the study estimated the size of the target population during the period from A to D, after  $P$  and  $L$  were assigned. Accordingly, the starting and ending times of  $L_1$  and  $L_2$  were concerned with the purpose of the study, which could be either historical or prospective.

Four points must be noted when CR is applied to estimating an IDU population size: an assumption must be made to mobilize the data from a drug use rehabilitation center that all IDUs in the concerned locality have the same probability of admission to the center. The estimated population size otherwise would refer to those IDUs who had the access to the center rather than the entire local IDU population. Determination of the capture periods and their interval was the key to our successful modeling. It should be noted that the length of the periods and interval must be properly assigned based on relevant intelligence. They should not be too long (usually no longer than 2 months), or the figures might be unstable. Additionally, the starting and ending

points of P should be carefully designed, which starts at the beginning of L1 and ends as L2 finishes. The coverage of the IDU population should be clarified using the obtained data, *i.e.*, places where IDUs are from. Therefore the estimated size of a local IDU population would be most precise when all data on IDUs from all local drug use rehabilitation facilities (both voluntary and compulsory) are obtained and analyzed. Since our study was based on the existing registry records, the completeness and validity of data were the foundation of good modeling, and all data obtained must be carefully prepared and checked. If labeling marks such as names are not available in some localities, other personal information (*e.g.*, gender, occupation, or address) should be integrated and analyzed for identification.

## REFERENCES

- 1) UNAIDS/WHO (2003) UNAIDS reports for 2003: Most deaths and new infections Ever; some good news, AIDS Treatment News, <http://www.aids.org/atn/a-396-02.htm>.
- 2) Adler, M. W. (2001) ABC of Aids: Development of the epidemic [review]. *BMJ*, **322**, 1226–1229.
- 3) Lambert, E. L. (1990) The Collection and Interpretation of Data from Hidden Populations, Vol. 98, NIDA Research Monograph, US Government Printing Office, Washington, DC, pp. 4–10, <http://www.drugabuse.gov/pdf/monographs/98.pdf>.
- 4) Des Jarlais, D. C., Dehne, K. and Casabona, J. (2001) HIV surveillance among injecting drug users. *AIDS*, **15**(Suppl. 3), 13–22.
- 5) Sekar, C. C. and Deming, W. E. (1949) On a method of estimating birth and death rates and the extent of registration. *J. Am. Stat. Assoc.*, **44**, 101–115.
- 6) Wittes, J. T., Colton, T. and Sidel, V. W. (1974) Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J. Chronic. Dis.*, **27**, 25–36.
- 7) Ghys, P. D., Jenkins, C. and Pisani, E. (2001) HIV surveillance among female sex workers. *AIDS*, **15**(Suppl. 3), S33–S40.
- 8) McKeganey, N., Barnard, M., Leyland, A., Coote, I. and Follet, E. (1992) Female streetworking prostitution and HIV infection in Glasgow. *BMJ*, **305**, 801–804.
- 9) UNAIDS (1997) The Status and trends of HIV/AIDS/STD Epidemics in Asia and the Pacific. A report of the official symposium at the 4th International Conference on AIDS in Asia and the Pacific, Monitoring the AIDS Pandemic Network, Manila, Philippines, [http://www.usaid.gov/our\\_work/global\\_health/aids/Publications/docs/map2001.pdf](http://www.usaid.gov/our_work/global_health/aids/Publications/docs/map2001.pdf).
- 10) Mastro, T. D., Kitayaporn, D. and Weniger, B. G. (1994) Estimating the number of HIV-infected injection drug users in Bangkok: a capture-recapture method. *Am. J. Public Health*, **84**, 1094–1099.
- 11) Fisher, N., Turner, S. W., Pugh, R. and Taylor, C. (1994) Estimating numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. *BMJ*, **308**, 27–30.
- 12) Archibald, C. P., Jayaraman, G. C., Major, C., Patrick, D. M., Houston, S. M. and Sutherland, D. (2001) Estimating the size of hard-to-reach populations: a novel method using HIV testing data compared to other methods. *AIDS*, **15**(Suppl. 3), S41–S48.
- 13) Hsieh, Y. H., Chen, C. W. and Lee, S. M. (2000) Empirical Bayes approach to estimating the number of HIV-infected individuals in hidden and elusive populations. *Stat. Med.*, **19**, 3095–3108.
- 14) Stephen, C. (1996) Capture-recapture methods in epidemiological studies. *Infect. Control Hosp. Epidemiol.*, **17**, 262–266.
- 15) Laporte, R. E. (1994) Assessing the human condition: capture-recapture techniques. *BMJ*, **308**, 5–6.
- 16) Hall, W. D. and Ross, J. R. (2000) How many dependent heroin users are there in Australia. *Med. J. Aust.*, **173**, 528–531.