Probability Sampling in Matched Case-Control Study in Drug Abuse

Surya Raj Niraula[1,*] and Frederick A. Connell[2]

[1]*School of Public Health and Community Medicine, B.P. Koirala Institute of Health Sciences, Dharan, Nepal*

[2]*School of Public Health and Community Medicine, University of Washington, Seattle, USA*

Abstract: Although random sampling is generally considered to be the gold standard for population-based research, the majority of drug abuse research is based on non-random sampling despite the well-known limitations of this kind of sampling. We compared the statistical properties of two surveys of drug abuse in the same community: one using snowball sampling of drug users who then identified “friend controls” and the other using a random sample of non-drug users (controls) who then identified “friend cases”. Models to predict drug abuse based on risk factors were developed for each data set using conditional logistic regression. Bootstrap analysis of the random-sample data set showed less variation, and did not change the significance of the predictors when compared to the non-bootstrap analysis. Comparison of ROC curves using the model derived from the random-sample data set was similar when fitted to either data set (0.93 for random-sample data vs. 0.91 for snowball-sample data (p=0.35)); however, when the model derived from the snowball-sample data set was fitted to each of the data sets, the areas under the curve were significantly different (0.98 vs. 0.83, p<.001). The proposed method of random sampling of controls appears to be superior from a statistical perspective to snowball sampling and may represent a viable alternative to snowball sampling.

Keywords: Random sampling, bootstrapping, non-random sampling, ROC curve.

INTRODUCTION

The illicit drug use is a ‘hidden’ and often socially stigmatized activity [1]. The illegal and stigmatized behaviors of illicit drug users endow them with ‘low social visibility’ [2]. The illegality of drug usage and the heterogeneity of drug users make representative community survey difficult. Such problem does not occur in alcohol and smoking research [3,4].

A common methodological limitation in drug abuse research is that it is frequently based on the non-probability sampling methods. Commonly used non-random sampling methods include snowball sampling, convenience sampling, privileged access interviewer method, respondent driven sampling and contact tracing [5-8]. Furthermore, if one uses an institution based case-control design, there is a high likelihood of Berkson’s bias [9]. Regardless of the care with which research based on these sampling methods is conducted and the ‘adequacy’ of sample size, there is no guarantee that results from these studies will be generalizable to the population from which subjects were selected.

This paper presents a new random sampling strategy for research on ‘hidden’ populations and compares statistical properties of this method to those of a sample from the same community derived from snowball sampling.

METHODS

The data for this paper were derived from a study of risk factors for drug abuse conducted in Dharan municipality in eastern Nepal. Nepal is a landlocked country covering an area of 147,181 km^2 with a population of about 26.5 million bordered by India and China. A total of 116,181 people reside in 103.38 km^2 areas of Dharan [10]. Two matched case-control data sets were formed using 1) snowball sampling and 2) community-based random sampling methods for comparison (Figure **1**). In both samples cases (drug abusers) were persons aged between 15 and 40 years who met the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders-IV) [11] criteria for drug abuse using the CAGE screen [12]. Controls were restricted to persons with same age group, who had never taken any psychoactive drugs, except as prescribed by doctors.

Snowball Sample

Sixteen potential drug abusers were identified by interviewing five ex-drug abusers, four drop-in-center in-charges (Auxiliary Nurse Midwives) and four drug abuse outreach workers. Six of these were under severe influence of drugs and were excluded. The remaining ten agreed to participate in the interview. Each case was asked to name a friend who was a drug abuser (a new case) and a friend who had never been involved in the abuse of drugs (control). One hundred fifty case-control pairs were identified in this way (Figure **1**).

[*] Address correspondence to this author at the School of Public Health and Community Medicine, B.P. Koirala Institute of Health Sciences, Dharan, Nepal; Tel: +977 9842035218; Fax: +977 25 520251; E-mail: sniraula@yahoo.com

Random Sample of Controls

A total of 158 households were selected randomly from 19 wards of Dharan Municipality. As the number of houses is heterogeneously distributed in the wards, stratified random sampling with proportional allocation method was adopted. One person aged 15 to 40 years without history of drug abuse in each selected house was asked to be interviewed for this study as a control. If more than one eligible person was found in the household, a lottery method was used to select the household respondent. Each potential control was informed about the objectives of the study and assured of anonymity and confidentiality before the interview. After the interview, each was asked the name-list of his or her friends who were drug abusers (potential cases). One of the drug abusers from the list was randomly selected as a *friendship-matched case* and interviewed with his/her consent. In seven households, the interviewees were themselves identified as drug abusers. They were included in the study sample as 'cases' and then seven respective matched controls were randomly selected from the name-list of non-user friends provided by the drug abusers. The house numbers of four houses could not be traced in the community. Two houses were found locked even during the third visit. Nine controls could not name even a single known drug abuser and two potential cases did not meet the criteria of drug abuser in the first screening and were excluded from the study sample. In this way, 141 matched control-case pairs were included for study (Figure **1**).

Survey Instrument and Variables

Pre-testing of the questionnaire was done among ten sets of drug abusers and controls from a drug rehabilitation center and its surrounding areas. These individuals were not included in the study samples. Based on the pre-test, corrections were made in a few questions and the time required for an interview was estimated.

The instrument used in both studies consisted of standard scales: Kuppuswamy scale [13] of socio-economic status, Central for Epidemiological Studies-Depression scale (CES-D) [14], Fagerstrom scale for nicotine dependence [15], CAGE screening scale for drug abuse [12] and Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) scale [11]. It also obtained information on socio-demographic characteristics and potential risk factors for drug abuse, including peer, family, social, psychiatric, personality and educational factors. All the possible risk factors were assessed using either dichotomous category (yes/no) or Likert scale.

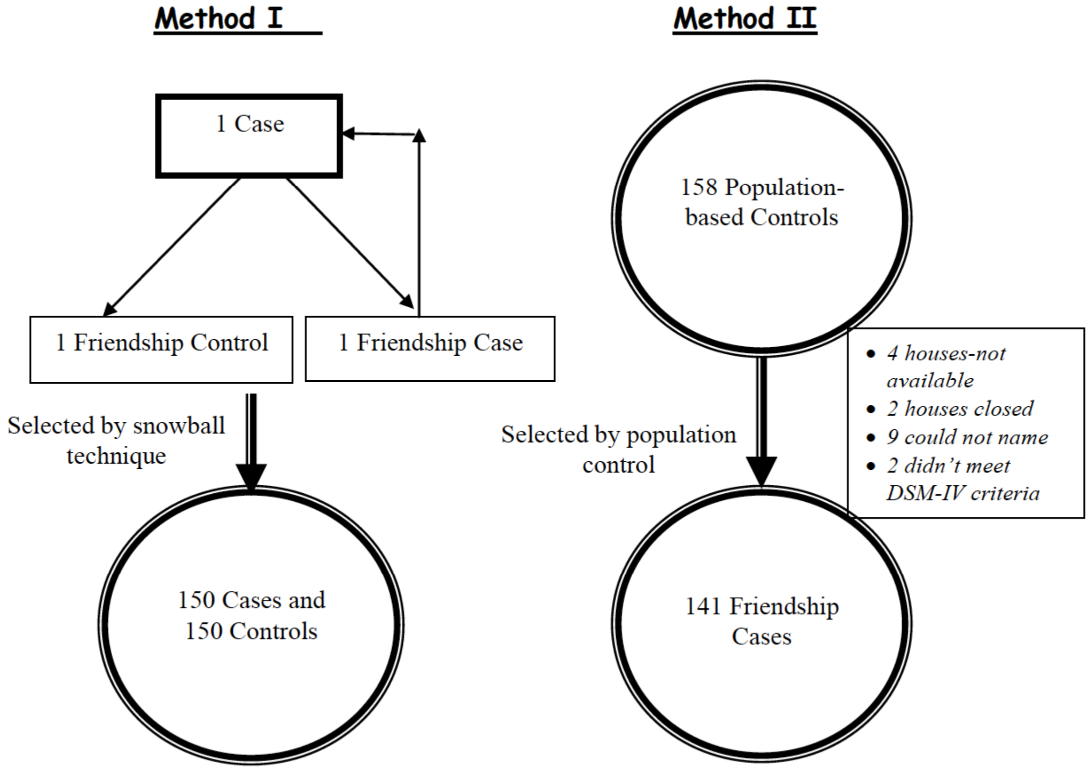
Method I
1 Case
1 Friendship Control
1 Friendship Case
Selected by snowball technique
150 Cases and 150 Controls
Method II
158 Population-based Controls
Selected by population control
• 4 houses-not available
• 2 houses closed
• 9 could not name
• 2 didn't meet DSM-IV criteria
141 Friendship Cases

Figure 1: Methodological Chart.

Ethical Consent

The Ethical Review Board of the Research Committee of B P Koirala Institute of Health Sciences (BPKIHS), Dharan, an institution authorized by the Nepal Health Research Council, approved the study. Verbal consent was taken from each potential subject after informing him or her about the purpose of the study. Subjects were also provided assurance of anonymity and confidentiality of data. Only cases (drug abusers) were compensated for their time by providing free tickets for the outpatient treatment of drug abuse/addiction.

Reliability and Validity

Measures taken to enhance validity were: use of standard scales, use of pre-tested questionnaire, giving respondents sufficient time to remember and respond, confidential (one-to-one) interviews, and multiple questions to obtain the same information.

The Spearman-Brown split-half reliability test [16] was employed for the risk factors involved in the instrument. In both the cases, the coefficients of Spearman-Brown test are more than 0.76, indicating that the individuals responded consistently to the instrument items.

Data analysis

In both studies, bivariate analysis was performed to select risk factors for inclusion in the multivariate models for drug abuse. Risk factors whose p-values were less than 0.2 were selected [17]. The conditional logistic regression with stepwise backward elimination was used to create multivariate models for predicting drug abuse using Stata (9.0) software. The details of assessment of interaction, confounding, precision, sample size, and multi-collinearity are described elsewhere [18].

Two independent multiple conditional logistic regression models identified factors associated with the risk of drug abuse for the two different data sets. Models were also developed with a bootstrap method [19,20] using the same predictors from the (non-bootstrap) multivariate models. The non-parametric bootstrapping technique [20] allowed us to estimate the sampling distribution of the standard error of the betas empirically without making assumptions about the form of the population. It also allowed estimating confidence intervals.

The goodness of fit for each of these four models was tested with the Hosmer-Lemshow statistic [21,22]. The predictive abilities of the two models were assessed using the area under ROC curves. Both were compared for the best result, where an area of 1.0 indicates perfect positive predictive ability and an area of 0.5 indicates a predictive ability no greater than chance alone.

RESULTS

Cases and controls in both samples had a broad range of socio-demographic characteristics [18]. Tables **1** and **2** compare the standard errors of the beta coefficients in the multivariate models without bootstrapping to those with bootstrapping, using 100 replications. Since Z values of the Wald test is based on estimated coefficients and its bootstrapping standard error, change in standard errors of the beta coefficients for the predictors yielded insignificant values of Z-statistic. In other words, the bootstrapping of the snowball-sample data set resulted in a large

Table 1: Change in Standard Errors of Estimation after Bootstrapping as Determined by the Snowball-Sample Data Set: A Backward Conditional Logistic Model

Factors	Coefficient[a]	Without bootstrapping		With Bootstrapping	
	β	SE	P value	BSE[+]	P value
Education (<10 yrs)	1.39	0.56	0.014	9.69	0.886
Occupation (Student)	-3.03	0.78	<0.001	3.91	0.438
Domination	1.38	0.63	0.030	7.87	0.861
Undeniable	1.36	0.55	0.013	2.02	0.500
Shyness behave	0.89	0.51	0.083	0.76	0.241
Short temper	0.96	0.48	0.045	7.71	0.901
Depression	1.76	0.57	0.002	5.82	0.763

[a]Estimated values adjusted for each of the other factors. [+]Bootstrap Standard Error.

Table 2: Change in Standard Errors of Estimation after Bootstrapping as Determined by the Random-Sample Data Set: A Backward Conditional Logistic Model

Factors	Coefficient[a]	Without bootstrapping		With Bootstrapping	
	β	SE	P value	BSE[+]	P value
Education (< 10 yrs)	1.10	0.44	0.012	0.53	0.040
Religion (Hindu)	-1.95	0.53	<0.001	0.65	0.003
Occupation (Student)	-1.86	0.54	0.001	0.74	0.012
Peer pressure	1.31	0.45	0.003	0.58	0.024
Undeniable	2.60	0.52	<0.001	0.60	<0.001
Commitment	-0.79	0.41	0.054	0.49	0.105
Depression	1.04	0.41	0.012	0.51	0.042

[a]Estimated values adjusted for each of the other factors. [+]Bootstrap Standard Error.

increase of the standard error of the beta coefficients for all seven predictors and resulted in a loss of significance (at the 5% level) of the beta coefficients. However, with the random-sample data set, the bootstrapping model had less impact on the standard errors of the beta coefficients which remained significant at the 5% level.

The model developed by the snowball sample (Model I) fitted the data well (area under ROC curve = 0.98). When this model was used to predict drug abuse using the random-sample data set, its predictive power was less (area under ROC curve = 0.83). These two statistics were significantly different (p<0.001).

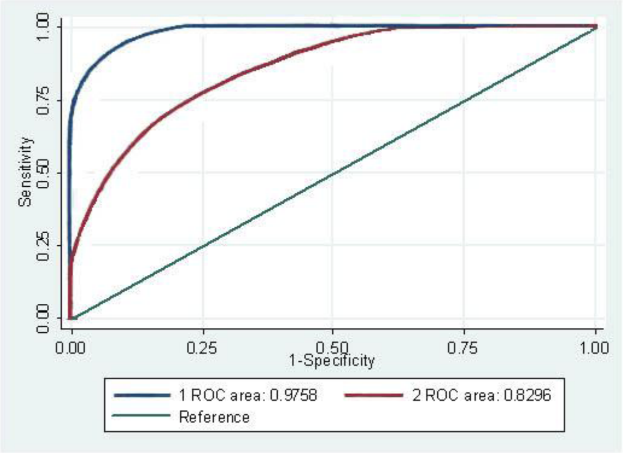

Figure 2: ROC Curves Fitted to Two Data Sets Developed by Model I.

On the other hand, the model based on the random-sample data set (Model II) was almost similar predictive power when apply to either the random-sample data set (area under the ROC curve = 0.93) or the snowball sample data set (area under the ROC curve = 0.91). There was no significant difference in these ROC areas (p=0.352).

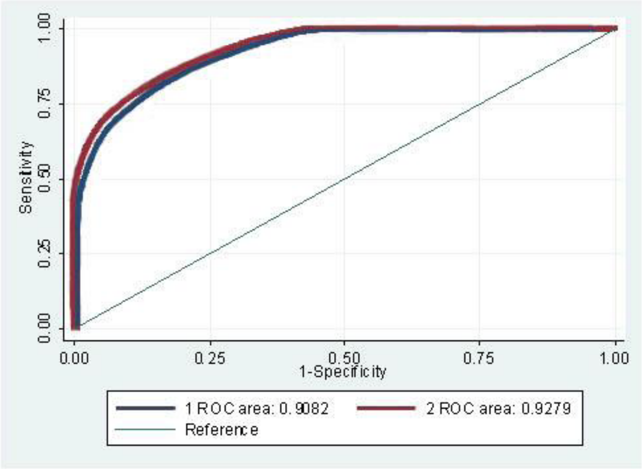

Figure 3: ROC Curves Fitted to Two Data Sets Developed by Model II.

DISCUSSION

Non-probability sampling methods, such as social network analysis [23], capture-recapture [24], contact tracing [25], or snowball sampling have been increasingly applied to the study of drug abusers or other "hidden" populations. For conditions which have a low prevalence, it is far easier to accrue large number of patients directly from specialized clinics or residential rehabilitation centers. Samples of patients from hospital treatment services or patients' data obtained from readily available sources, such as hospitals admission statistics, case registers and case notes are also used. Such approaches can be more cost-efficient than using a probability sampling and measures can be taken to try to make the samples representative, such as ensuring that non-probability samples are heterogeneous, e.g., patients can be drawn from various settings like hospitals, private clinics, and rehab centers. Nevertheless, with non-probability samples, one cannot be certain of generalizability. For random sampling selection it is necessary to have a sampling frame. Unfortunately, such frames are rarely available for “hidden populations” in community settings.

This paper demonstrates a strategy that relies on the random sampling of *controls,* which is feasible using geographic and census data. *Cases* that are identified by the randomly selected controls are more likely to be random than cases identified by traditional non-probability sampling methods.

In a matched case-control design with friendship matching, there is a possibility of selection bias because selection into the study may not be independent of the study factor within each stratum of the matching factors [26]. Lopes *et al.* have demonstrated, however, that one can conclude that there is no selection bias in a friendship-matched case control design if the proportion of controls selected by exposed cases (P_1) and the proportion of exposed controls selected by unexposed cases (P_2) are equal [27]. When this logic is applied to the random sample of controls and friendship mated cases, one would want to see that the proportion of cases selected by exposed controls (P_1) and the proportion of exposed cases selected by unexposed controls (P_2) are equal. In the present study, we found that there was no selection bias in either sample, as they both satisfied the condition of $P_1= P_2$ (Tables **3** and **4**).

While the predictive ability (as measured by the area under the ROC curve) of snowball-sample model (model I) is slightly higher than that of the random-sample model (model II), model II appeared to fit both data sets while model I did not. Weber *et al.* adopted a similar procedure for testing the fitness of their data on the model found by Feldman *et al.* to be predictive of endometrial neoplasia [20]. Thus, model II may be considered suitable to fit both data sets in our context.

Furthermore, the 100 replications of bootstrapping samples in estimating the standard error of model coefficients produced a greater increase in standard errors, such that insignificant probability values were obtained for all the beta coefficients of all factors model I, which indicated that the data did not fit well into the model (P = 0.292). But in case of the model II, the result showed a better adjustment to data (P = 0.001). There was little change in the statistical significance of beta coefficients; however, the probability values slightly differed. Bootstrapping is a method that repeatedly analyzes sub-samples of the data and calculates the standard error of estimation, which reflects sampling variation of the collected data [28,29]. Hence, these findings suggest the second sample (randomly selected from the community) had less variability than the data set drawn from a snowball sampling.

In absence of selection bias, the population-based probability technique of selecting controls -- and friend cases -- may be better than snowball sampling. This study suggests that this sampling method had superior statistical properties compared to a non-probability sample drawn from the same community. The possibility of the technique had already been described

Table 3: Depression in Friend Controls Stratified by Depression Status of Cases Sampled through the Snowball Sampling

Depression in Cases	Depression in matched controls		Total	Significance
	Depression	No depression		
Depression	58 (50.4)	57 (49.6)	115	χ^2=0.24, df=1 P=0.62
No depression	16 (45.7)	19 (54.3)	35	
Total	74 (49.3)	76 (50.7)	150	

*Note: Figures in the parentheses show percentage.

Table 4: Depression in Cases Identified by Randomly Selected Controls

Depression in Controls	Depression in matched cases		Total	Significance
	Depression	No depression		
Depression	41 (48.2)	44 (51.8)	85	χ^2=0.68, df=1 P=0.17
No depression	25 (44.6)	31 (55.4)	56	
Total	66 (46.8)	75 (53.2)	141	

*Note: Figures in the parentheses show percentage.

in the previous article [30]. As the controls are easily identifiable in the community, the problem of random sampling for "hidden" cases is lessened. After the selection of random controls, the cases can be identified through the list of drug abusers provided by the randomly selected controls (Figure **1**). The cases that are selected in this manner should also be representative of all drug abusers in the community under the assumption that drug abuse practice of friends is known to the randomly selected controls. Another advantage of this method is that it can be applied to population distributed in a large (or small) geographical area.

Whether the results of samples chosen from the population truly represent the parameters of population cannot be confirmed unless the sampling frame and the population parameters are known. As already discussed, drug abusers in the community are difficult to trace out, thus one cannot be sure whether the results of this proposed method represent the target population or not. The validation of the result of this sampling method should be tested with other studies.

One cannot conclude that the statistics obtained from this random control-case pair method of sample selection is closer to actual population parameters unless the population parameters are known. But an adequate sampling frame (or enumeration) of drug abusers is generally not feasible in a community setting. However, the experiment can be repeated in a closed population, with well-defined denominators like schools, campuses, etc. with both the proposed sampling method and a census enumeration. With these data, results of the technique described in this study can be compared to the parameters obtained from a census. Further research is suggested to examine and fortify this finding.

ACKNOWLEDGEMENTS

We thank especially to the Fulbright program and USEF Nepal to prepare this collaborative paper. This research was initially supported by Institute Advisory Authority, BPKIHS for data collection. Those who involved and provided help to this study are thankful.

CONFLICT OF INTEREST

None declared.

KEY MESSAGES

- There is hardly a community-based case-control study, which has been conducted except some prevalent studies based on non-probability sampling technique in the context of Nepal.
- There is always a problem associated in random sample selection of drug abusers from the community.
- This paper adds a technique of random sample selection in matched case-control design, which can be better than a snowballing method in a given condition.

REFERENCES

[1] 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. https://doi.org/10.1111/j.1360-0443.1993.tb02036.x

[2] Watters JK and Biernacki P. Targeted sampling: options for the study of hidden populations. Soc Problem 1989; 36: 416–30. https://doi.org/10.2307/800824

[3] Niraula SR, Shyangwa PM, Jha N, Paudel RK, Pokharel PK. Alcohol use among women in a town of Eastern Nepal. Journal of Nepal Medical Association 2004; 43(155): 244-9.

[4] Niraula SR. Tobacco use among women in Dharan, Eastern Nepal. Journal of Health, Population and Nutrition 2004; 22(1): 68-74.

[5] Patrick B and Waldorf D. Snowball sampling: problems and techniques in chain-referral sampling. Social Methods and Research 1981; 10: 141-63. https://doi.org/10.1177/004912418101000205

[6] Watiters JK and Biernacki P. Targeted Sampling: Options for the Study of Hidden Populations. Social Problems 1989; 36(4): 416-30. https://doi.org/10.2307/800824

[7] Blacker P, Tindall B, Wodak AD, Cooper D. Exposure of intravenous drug users to AIDS, Sydney, 1985. Aust N Z J Med 1986; 16: 686-90. https://doi.org/10.1111/j.1445-5994.1986.tb00013.x

[8] Acharya LB, Dhungel N, Ross JL. Factor associated to HIV prevalence among male IDUs in the Eastern Terai of Nepal. Proceedings of the 15th International AIDS Conference; 2004 July 11-16; Bangkok, Thailand. International AIDS Society; 2004.

[9] Hayden GF, Kramer MS, Horwitz RI. The Case-Control Study-A practical Review for the Clinician. JAMA 1982; 247(3): 326-31. https://doi.org/10.1001/jama.1982.03320280046028

[10] National Population and Housing Census 2011 (Village Development Committee/Municipality). Government of Nepal, National Planning Commission Secretariat, Central Bureau of Statistics, Kathmandu; 2: 2012

[11] American Psychiatric Association. Quick Reference to the Diagnostic Criteria from DSM-IV-TRTM. Washington, D.C. and London, U.K.: 2003.

[12] Potter B and Fleming MF. Obstet Gynecol Clin N Am 2003; 30: 583–99. https://doi.org/10.1016/S0889-8545(03)00081-0

[13] Kuppuswamy B. Manual of socio-economic status scale (Urban).Manasayan, 32, Netaji Subash Nagar, Delhi-110006: 2001.

[14] Lewinsohn PM, Seeley JR, Roberts RE, Allen NB. Center for Epidemiological Studies Depression Scale (CES-D) as a

screening instrument for depression among community-residing older adults. Psychol Aging 1997; 12: 277-87. https://doi.org/10.1037/0882-7974.12.2.277

[15] Health Alert. Asia-pacific Edition (Vol. 6): Health Action Information Network 2005.

[16] Singh ML, editor. Understanding Research Methodology. 1st ed. Kathmandu; Neeti and Nitendra Singh publisher; 1991.

[17] Weber AM, Belinson JL, Peidmonte MR. Risk factors for endometrial hyperplasia and cancer among women with abnormal bleeding. Obstetrics and Gynecology 1999; 93(4): 594-8.

[18] Niraula SR. Development of model to identify the risk factors for drug abuse [dissertation]. Nepal: Tribhuvan University; 2008.

[19] Efron B. Bootstrap methods: Another look at the jackknife. Annal of Statistics 1979; 7(1): 1-26. https://doi.org/10.1214/aos/1176344552

[20] Efron B. Non parametric standard errors and confidence intervals (with discussion). Canad J Statist 1981; 9: 139-72. https://doi.org/10.2307/3314608

[21] Hosmer DW, Hosmer T, Le Cessie S, Lemshow S. A comparison of goodness-of-fit-tests for the logistic regression model. Statist. Med. 1997; 16: 965-80. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O

[22] Qin J and Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. Biometrika 1997; 84: 609-18. https://doi.org/10.1093/biomet/84.3.609

[23] O'Reilly P. Methodological issues in social support and social network research. Soc Sci Med 1988; 26: 863-73. https://doi.org/10.1016/0277-9536(88)90179-7

[24] Laporte RE. Assessing the human condition: Capture-recapture techniques. BMJ 1994; 308: 5-6. https://doi.org/10.1136/bmj.308.6920.5

[25] Hall W and Dolan K. Is there a role for contact tracing in preventing HIV transmission among injecting drug users. Addiction 1996; 91: 917-9. https://doi.org/10.1111/j.1360-0443.1996.tb03585.x

[26] Flanders WD and Austin H. Possibility of selection bias in matched case-control studies using friend contacts. Am J Epidemiol 1986; 124: 150-3. https://doi.org/10.1093/oxfordjournals.aje.a114359

[27] Lopes CS, Rodrigues LC, Sichieri R. The lack of selection bias in a snowball sampled case-control study on drug abuse. Int J Epidemiol 1996; 25(6): 1267-70. https://doi.org/10.1093/ije/25.6.1267

[28] Liu H, Li G, Cumberland WG, Wu T. Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping. Journal of Data Science 2005; 3: 257-78.

[29] Efron B and Tibshirani RJ. An introduction to the bootstrap. London: Chapman & Hall, 1993. https://doi.org/10.1007/978-1-4899-4541-9

[30] Niraula SR, Chhetry DB, Singh GK, Shyangwa PM. Sampling methods and possibility of random sampling to select drug abusers from the community. Health Renaissance 2007; 4(1): 1-6.

Received on 13-11-2017 Accepted on 02-01-2018 Published on 02-02-2018

https://doi.org/10.6000/1929-6029.2018.07.01.3