THEORY AND METHODS

Estimating the lesbian population: a capture-recapture approach

D J Aaron, Y-F Chang, N Markovic, R E LaPorte

J Epidemiol Community Health 2003;**57**:207–209

See end of article for authors' affiliations

Correspondence to: Dr D J Aaron, Department of HPRE, 155 Trees Hall, University of Pittsburgh, Pittsburgh, PA 15261 USA; debaaron@pitt.edu

Accepted for publication 15 July 2002

Study objective: Little is known about the number of women who identify as lesbian. Estimates from the US range from 1% to nearly 10%. Accurate estimates are critical in order to meet lesbian's health-care needs and to address health problems that may be more prevalent among them. This study used capture-recapture methods to estimate the lesbian population of Allegheny County, Pennsylvania.
Design: Mailing lists from four sources were used to identify lesbians. The capture-recapture method and log-linear modelling were used to estimate the number of lesbians in the defined geographical area, and the percentage of the female population they comprised there was determined through census data.
Setting: Allegheny County, Pennsylvania, USA.
Results: A total of 2185 unique names were identified. The capture-recapture method estimated that the total lesbian population of Allegheny County was 7031 (95% CI 5850 to 8576). Therefore, based on the 1990 census figures, the county's adult lesbian population was estimated to be 1.87% (95% CI 1.56% to 2.28%) of the adult female population.
Conclusions: An estimate of the lesbian population is fundamental for addressing lesbian's health needs and for developing appropriate research programmes. Capture-recapture methods have the potential to provide accurate and reliable estimates of this population in any location.

Virtually nothing is known about how many women are lesbian in the United States and worldwide. Clearly, cost effective methodologies must be developed for monitoring the lesbian population as a precondition for effectively providing it with health information and health care. Wide variations have been reported regarding the proportion of the female population that is thought to identify as a lesbian, with estimates ranging from as low as 1.3% to as high as 8.6% of the US population.[1] One problem in estimating the number of lesbians is that they are often uncountable due, in part, to social stigma associated with identifying as a lesbian. However, wildlife research has been using a technique for more than 100 years to estimate the size of rare and elusive populations, which are difficult to find and count or are highly mobile and cannot be counted at one time.[2] The technique, known as capture-recapture has also been applied to human populations that are difficult to count, such as the homeless, children on medical support, and female prostitutes.[3–5] Furthermore, capture-recapture has been used successfully to estimate rates of chronic diseases, dog bites, injuries, and other conditions in human populations.[6–11] The main objective of this study was to use capture-recapture methods to estimate the lesbian population of Allegheny County, Pennsylvania.

METHODS

Lesbians were identified with the help of four organisations that serve the lesbian and gay population of Allegheny County, Pennsylvania and maintain large mailing lists. These organisations provided a paper copy of their mailing lists for this project. All of the four lists had been updated immediately before the study. To maintain the confidentiality of names appearing on the lists, representatives of the organisations were instructed on how to match the names, and they supervised use of the lists. After the study's completion, the mailing lists that had been provided were destroyed. The University of Pittsburgh's Institutional Review Board approved the study protocol in advance.

Table 1 Number of lesbians identified by combinations of sources

	Number	Percentage
One source	*1612*	*73.8*
A only	589	26.9
B only	534	24.4
C only	281	12.9
D only	208	9.5
Two sources	*436*	*19.9*
A and B	104	4.8
A and C	44	2.0
A and D	143	6.5
B and C	64	2.9
B and D	53	2.4
C and D	28	1.3
Three sources	*114*	*5.2*
A and B and C	27	1.2
A and B and D	48	2.2
A and C and D	19	0.9
B and C and D	20	0.9
Four sources A and B and C and D	*23*	*1.1*
Total	*2185*	*100*

A=community centre, B=event promoter, C=foundation, D=research study.

To limit the study's results to Allegheny County, women with zip codes outside this area were excluded. The names on the mailing lists were manually cross referenced and linking the names and addresses identified those appearing more than once. Three of the mailing lists were obtained from a community centre, an event promoter, and a foundation, all of which primarily serve the local lesbian and gay population. There are no data available to estimate the age range or other characteristics of the individuals included on these three lists.

Table 2 Log-linear model with four sources

Model	Deviance	df	Estimate (95% CI)	AIC*
Independent Model	174.81	10	3739 (3578 to 3919)	154.81
All two way interactions	6.72	4	7126 (5908 to 8734)	–1.28
All two way interactions + ABD	2.77	3	6127 (4956 to 7819)	–3.23
Heterogeneity model	98.13	9	6964 (5846 to 8407)	80.13
Heterogeneity + AD + BD + CD†	7.24	6	7031 (5850 to 8576)	–4.76

*AIC=Akaike information criterion. †Best model.

The fourth list was provided by a lesbian health research project at the University of Pittsburgh.[12] Surveys for the project were widely distributed throughout the lesbian community by a number of methods including; social, political, and religious organisations, media outlets, community events, mailing lists, and social networks. Participants had the option of providing their name and address on a postcard, indicating their interest in future research projects, that was mailed independent of the anonymous survey. Thus, the list from the research project included the names of women who returned the postcard. The age range of women completing the anonymous survey was 18–74 years. However, as the survey was anonymous and we had no way of linking the return postcards with completed surveys, we cannot estimate the age range of women that returned a post card and were included in the study.

Capture-recapture method and log-linear modelling[13] were used to estimate the total number of lesbians in Allegheny County based on these four lists. Source dependence was modelled by adding the corresponding interaction term to the model. The significance of the interaction was assessed using likelihood ratio statistics, and the goodness of fit of the model was measured by deviance. To model heterogeneity among the population, a method suggested by Agresti[14] and Darroch *et al*[15] was used. A confidence interval (CI) for the estimated number of cases was computed via the profile likelikhood method and Akaike Information Criterion (AIC) was used for the model selection. For the log-linear model the value for this criterion is AIC = Deviance B 2×df, where df is the number of degrees of freedom of the model, and the model with the smallest AIC is selected as the best model.[16] The proportion of the county's lesbian population was determined by dividing the number of lesbians estimated through the capture-recapture method by the number of adult women residing in the county according to the 1990 census. The statistical software SAS with GENMOD procedure was used for the analyses.

RESULTS

A total of 2185 unique women were identified from the four lists as follows: 947—community centre (list A), 873—event promoter (list B), 506—foundation (list C), and 542—research project (list D). However, we know that by their nature these lists are incomplete and that the total number of lesbians is larger than 2185 (0.6% of the adult female population of Allegheny County). But how much larger is it? Capture-recapture utilised the overlap between the lists (table 1) to determine the degree of underascertainment for the raw count of 2185 and thus provided an estimate of the total lesbian population. Most (74%) of the names appeared on only one list (n=1612). Nearly 20% of them appeared on two lists (n=436), 5% appeared on three (n=114), and only 1% appeared on all four (n=23). The log-linear model that best fit the data included both heterogeneity and dependence among the sources (table 2). On the basis of this model, the number of lesbians in Allegheny County was estimated at 7031 (95% CI 5850 to 8576). The number of adult women in Allegheny County in 1990 was 375 901. Therefore, the county's adult lesbian population constituted 1.87% (95% CI 1.56% to 2.28%) of its entire adult female population.

DISCUSSION

This study shows the usefulness of the capture-recapture method in determining the population of lesbians in a defined geographical area. That usefulness must nevertheless be qualified. The study's sources of data depended on a woman's willingness to have her name included on a mailing list. Of course, the voluntary nature of being on a mailing list entails that some members of the lesbian population will have a low probability of capture. This may apply particularly to women who do not want to be identified as lesbian and whom our approach thus could not count. Yet a principle of the capture-recapture method is that its estimate is valid only for a population that can be captured without selectivity. Given this stricture, we must qualify our results because all four mailing lists used in the current study capture only those women who are visible or active in the lesbian community. In effect then, we have estimated the number of women in Allegheny County who identify with that community. It is also possible that a woman may be included on a list and not be a lesbian. We have no way of estimating this number but expect that it would be extremely small. In addition, while the matching of the names was done manually, there could have been problems with this process because of different spellings of names, incomplete names (providing only first or last initial), or the possible use of an alias on one of the lists). All of these factors could contribute to potential mismatches.

There are advantages to studying the population of lesbians that are out and identify with and participate in the lesbian community. Social learning theory postulates that a health behaviour is influenced by interaction among a given person's characteristics, the person's environment, and the behaviour itself.[17] Therefore, possible differences between the health behaviours of lesbian and heterosexual women may be attributable in part to the shared environment of the lesbian community. Hertzman *et al*[18] define special populations as those having shared characteristics. They further maintain that the social environment within such populations may condition certain health habits, producing a pattern of negative or positive behaviours that ultimately influence health. Thus, to understand the complexity of health behaviours within the lesbian population it seems necessary to enumerate and then study those women that identify with the community.

The capture-recapture method with log-linear modelling suggested that there was heterogeneity in capture probability among the lesbian population, that is, different individuals would have a different chance of being listed by the sources. In addition, the model also indicated the existence of source dependence in the lists. It is likely that some positive dependence exists among the four sources used for this study. In other words, appearing on one mailing list may increase a woman's chance of being on the other three as women who are willing to have their names on a lesbian related mailing list are more likely to have their names on other such lists. In addition, the women participating in the research project (list D) were recruited from many sources in the lesbian community including publications and advertisements that would have likely been read by women on lists A, B, and C.

Using capture-recapture has the potential to provide accurate and reliable estimates of the lesbian population in

any location. The method's basic tenet is that one need not identify every member of the population to estimate its size accurately, provided that appropriate sources are used for case identification and an adequate model is selected for the estimation. Consequently, under proper conditions, capture-recapture is incredibly cost effective in counting any elusive population and could easily be used to enumerate the lesbian population locally, nationally, and globally.

Furthermore, studies such as this have important public health implications. After all, estimating the lesbian population is critical in planning for healthcare needs and developing relevant research programmes within the population. For example, if the proportion of lesbians thought to be overweight is 29%, then the number of lesbians in Allegheny County that could benefit from a weight reduction programme would be 2039. Statistics such as this would certainly be worth knowing. In summary, the results of this study in combination with epidemiological studies on health behaviours, will allow us to estimate not only the prevalence of risk factors for disease among lesbians residing in Allegheny County but how many women are in need of health promotion programmes and health care services.

ACKNOWLEDGEMENTS

The authors extend their appreciation to the Gay and Lesbian Community Center of Pittsburgh, The Lambda Foundation, Triangle Productions, and the ESTHER Project for allowing us to use their mailing lists for this project and to Mr Jerry Heverly for his editorial assistance.

Authors' affiliations

D J Aaron, Y-F Chang, N Markovic, R E LaPorte, Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, USA

Funding: this research was supported by a grant from the Lesbian Health Fund of the Gay and Lesbian Medical Association.

Conflicts of interest: none.

REFERENCES

1 **Laumann EO**, Gagon JH, Michael RT, *et al*. *The social organizaiton of sexuality: sexual practices in the United States*. Chicago, IL: University of Chicago Press, 1994.
2 **Sudan S**, Sirken MG, Cowan CD. Sampling rare and elusive populations. *Science* 1988;**240**:991–6.
3 **Fisher N**, Turner SW, Pugh R, *et al*. Estimating numbers of homeless and homeless mentally ill people in northeast Westminster by using capture-recapture analysis. *BMJ* 1994;**308**:27–9.
4 **Palfrey JS**, Haynie M, Porter S, *et al*. Prevalence of medical technology assistance among children in Massachusetts in 1987 and 1990. *Public Health Rep* 1994;**109**:226–33.
5 **Bloor M**, Leyland A, Barnard M, *et al*. Estimating hidden populations: a new method of calculating the prevalence of drug-injecting and non-injecting female prostitution. *Br J Addiction* 1991;**86**:1477–83.
6 **LaPorte RE**, Tull ES, McCarty D. Monitoring the incidence of myocardial infarctions: applications of capture-mark-recapture technology. *Int J Epidemiol* 1992;**21**:258–63.
7 **Robels SC**, Marrett LD, Clarke EA, *et al*. An application of capture-recapture methods to the estimation of completeness of cancer registry. *J Clin Epidemiol* 1988;**41**:495–501.
8 **Chang YF**, McMahon JE, Hennon DL, *et al*. Dog bite incidence in the city of Pittsburgh: a capture-recapture approach. *Am J Public Health* 1997;**87**:1703–5.
9 **LaPorte RE**, Dearwater SR, Chang YF, *et al*. Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring. *Am J Epidemiol* 1995;**142**:1069–77.
10 **Chiu W**, Dearwater SR, McCarty DJ, *et al*. Establishment of accurate incidence rates for head and spinal cord injuries in developing and developed countries: capture-recapture approach. *J Trauma* 1993;**35**:206–11.
11 **International Working Group for Disease Monitoring and Forecasting**. Capture-recapture and multiple record systems estimation II: application in human disease. *Am J Epidemiol* 1995;**142**:1059–68.
12 **Aaron DJ**, Markovic N, Danielson ME, *et al*. Behavioral risk factors for disease and preventive health practices among lesbians. *Am J Public Health* 2001;**91**:972–5.
13 **Cormack RM**. Interval estimation for mark-recapture studies of closed populations. *Biometrics* 1992;**48**:567–76.
14 **Agresti A**. Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* 1994;**50**:494–500.
15 **Darroch JN**, Fienberg SE, Glonek GFV, *et al*. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J American Statistical Association* 1993;**15**:1137–48.
16 **Hook EB**, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995;**17**:243–64.
17 **Bandura A**. *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
18 **Hertzman C**, Frank J, Evans RG. Heterogeneities in health status and the determinants of population health. In: Evans RG, Barer ML, Marmor TR, eds. *Why are some people healthy and others not? The determinants of health of populations*. New York, NY; Aldine DeGruyter, 1994.