Comparison of Methods for Identifying a Population-Based Sample of Filipina Women for a Health Survey

Katherine J. Hoggatt, MPH; Susan L. Stewart, PhD; Rena Pasick, DrPH; Fabio Sabogal, PhD; Noe Tuason, MA; Melen R. McBride, PhD, RN; Grace Villarin Duenas, MPH; Pamela L. Horn-Ross, PhD

This paper describes and compares 2 random-digit dialing (RDD) methods that have been used to select minority subjects for population-based research. These methods encompass the census-based method, which draws its primary sampling units from census tracts with a high proportion of minority persons, and the registry-based method, which derives its primary sampling units from a population-based cancer registry. Our study targeted Filipinos living in 10 Northern California counties, where they constitute 4% of the total population. Eligible participants (Filipina women, at least aged 20, who spoke 1 of 4 interview languages) were asked to complete a short telephone interview. Both the census and registry methods located Filipino households with comparable efficiency and with a higher yield than would be expected in a non-targeted population survey, such as the Mitofsky-Waksberg RDD method. No systematic pattern of responses was evident that would indicate that either method sampled women who were systematically less acculturated or less likely to use cancer screening tests. Although both methods offer substantial gains in efficiency, their utility is limited by generating samples that tend to over-represent high-density areas. The degree to which these methods are considered viable depends on further refinement to limit, or eliminate, their inherent selection biases without sacrificing their increased efficiency to locate minority populations. (*Ethn Dis*. 2004;14:21–25)

Key Words: Random Digit-Dialing, Minority Population, Cancer Registry, Census, Bias, Efficiency, Methods, Population-Based Study

From the Northern California Cancer Center, Union City (KJH, SLS, RP, NT, GVD, PLH), Comprehensive Cancer Center, University of California San Francisco (RP), California Medical Review Inc., San Francisco (FS), Stanford University School of Medicine, Palo Alto (MRM), California.

Address correspondence and reprint requests to Pamela L. Horn-Ross; Northern California Cancer Center; 32960 Alvarado-Niles Rd; Suite 600; Union City, CA 94587; phornros@nccc.org

Introduction

In population-based research, study participants are chosen to represent the characteristics of the underlying population.[1] To achieve a representative sample, researchers must first define the target population they wish to study, in terms of geography, age, time period, etc, and then develop a list or sampling frame from which to select the subjects for their study. When the target population is a small proportion of the total population or is dispersed over a large area, selecting a representative, unbiased sample at a reasonable cost may prove difficult. Telephone surveys can offer a means of contacting potential subjects as an alternative to in-person screening, a method that may be prohibitively expensive or logistically difficult. The overall effectiveness of a telephone survey depends on 2 factors: the efficiency with which the survey reaches eligible households and the representativeness, or the absence of bias of the sample generated.

One method that can enhance the efficiency and ensure the objectivity of the telephone survey is a process called "random-digit dialing" (RDD). Utilizing this procedure, randomly generated telephone numbers define the population to be sampled, referred to as the"sampling frame". The most common form of RDD is a 2-stage sample. The method is roughly outlined below.[2–4]

1. Telephone numbers are generated at random to cover the geographic area of interest. Many sources comprise these primary sampling units or PSUs, the first-stage sampling units that link them to geographic areas.[2]
2. The numbers are called and screened to determine if they represent a residence or non-residence (eg, business, non-working number, cell phone, etc);
   a. if the number is residential, the last 2 digits are truncated to form the PSU (identification PSUs are typical in RDD, however, a smaller or larger number may be used);
   b. if the number is non-residential, the PSU is discarded.
3. Additional random numbers, typically 99, are generated in each residential PSU and are then called and screened to find additional *k* eligible households, where *k* is selected by the researcher.

This RDD strategy is commonly referred to as the Mitofsky-Waksberg method and is considered the "standard" RDD procedure.[3] In theory, the Mitofsky-Waksberg method holds 2 key advantages. First, by screening out a large number of non-residential PSUs at the first stage, the researcher will locate households with greater efficiency than a simple random sample consisting of telephone numbers dialed at random. Second, and of equal importance, the final sample will be self-weighted, meaning each household in the target population has an equal probability of selection.[3] However, the Mitofsky-Waksberg method often fails to be either efficient or unbiased, particularly when the target population is not the majority population in the area. In this case, the Mitofsky-Waksberg method may offer only small gains in efficiency

To achieve a representative sample, researchers must first define the target population they wish to study, in terms of geography, age, time period, etc, and then develop a list or sampling frame from which to select the subjects for their study.

relative to a simple random sample (eg, a single-stage telephone survey where each randomly generated number is screened for eligibility). Even more troubling, a large number of unproductive PSUs may exist, where the number of eligible households is less than *k*. Without *k* eligible households per PSU, the sample will not be self-weighting, and the RDD procedure will introduce a selection bias into the sample.

A number of articles in the epidemiologic literature have evaluated the strengths and shortcomings of the Mitofsky-Waksberg method and have discussed the complications involved in finding subjects of a specific age or sex.[2,4–6] In contrast, the problem of locating "rare" or "minority" ethnic populations through RDD has not been widely studied.[7–8] The Mitofsky-Waksberg method was developed to locate households which are generally rare in non-residential PSUs and on average represent at least 50% of telephone numbers in a given residential PSU.[9] Therefore, the method may not be an effective means of reaching an ethnic population, such as Filipinos, that constitute only a small percentage of the general population even within a residential PSU. Due to these limitations, researchers who are interested in studying minority populations have sought alternatives to the Mitofsky-Waksberg method of RDD that will increase the efficiency of the procedure without introducing bias. This paper describes and compares two such alternative RDD methods that have been used to select minority subjects for population-based research, herein referred to as the census-based and registry-based methods.

METHODS

Between June and August 1996, we conducted a RDD pilot study as part of the Filipino Women's Health Study. The target population for both the pilot and main studies was Filipino women at least 20 years of age who resided in one of 10 Northern California counties and spoke one of 4 interview languages (ie, English, Tagalog, Ilocano, or Cebuano). We generated 2 RDD samples of approximately 3500 telephone numbers each, using the census method and the registry method, then screened and subsequently interviewed eligible women.

Generating the Samples

Census Method

The census-based method used PSUs drawn from census tracts where at least 5% of the residents were Filipino-American according to the 1990 US Census. These census tracts included 65% of the Filipino population in the 10-county area. Following the methods outlined by Marin et al to determine PSUs, a commercial service selected 3 streets at random within each census tract and identified telephone numbers corresponding to all addresses on the selected streets.[10] These telephone numbers were truncated to form the PSUs for the second stage of sampling.

A set quantity of random phone numbers was generated for each PSU. These were pre-screened using ProPhone, a proprietary list of business/non-working numbers, to eliminate non-residential phone numbers. This technique has previously been shown to be efficient in locating minority populations.[10]

Registry Method

The registry-based method generated PSUs from the phone numbers of *every* Filipino cancer patient diagnosed between 1990 and 1993 in the 10-county area who were reported to the California Cancer Registry (CCR). The CCR is a population-based registry covering the entire state of California and is estimated to be over 95% complete. This RDD method assumes that cancer patients, of all ages, both sexes, and from all sites combined, are *geographically* representative of all persons within the population of interest. If a patient did not have a phone number on record with the registry, reverse directories were used to secure the patient's or nearest neighbor's phone number.

In contrast to the census method, *k* random numbers were generated for each patient's phone number, replacing the last 2 digits with random numbers, rather than generating a set quantity of phone numbers for distinct PSUs. To avoid biasing the sample in favor of individuals with a personal or family history of cancer, patient phone numbers were excluded from the final stage.

Data Collection

We collected data in 2 phases. In the first, a screening call was placed to each of the randomly generated numbers to determine if an eligible woman resided in the household. A follow-up interview was scheduled if an eligible woman was identified. This screening interview was conducted in English. Once a household was determined to include at least one eligible woman, we selected the woman from the oldest age stratum represented for the interview—these strata were aged 65 and over, aged 50–64, and aged 20–49. Older women were therefore over-sampled to ensure adequate numbers for the analysis of this subgroup. If more than one woman within an age stratum per household was eli-

Table 1. Detailed enumeration status of samples drawn using the census- and registry-based RDD methods to locate Filipino households

	Census Method*		Registry Method	
	N	%	*N*	%
Total	3406		3526	
Non-residential numbers:	429	13	1180	33
Disconnected/not in service	346		716	
Business	33		254	
Modem/fax/beeper, etc	50		210	
Status unknown:	97	3	227	6
Residential numbers without enumeration:	284	8	270	8
No adult reached in household	45		20	
Only reached answering machine	209		210	
Residence was out of study area	3		13	
No one in household spoke English	6		5	
Refused enumeration	21		22	
Enumerated households:				
Not a Filipino household	2435	71	1710	48
Filipino household	161	5	139	4

* Business numbers screened out by a commercial service.

gible, we interviewed the older woman. Finally, we limited the number of younger women interviewed. Once we reached the appropriate number within this age stratum, households, where the only eligible woman was in the younger strata, were not interviewed.

In the second phase, trained staff members interviewed selected women over the phone. Respondents chose the language for the interview. The standardized, structured questionnaire included questions on acculturation, access to health care, attitudes toward health issues, traditional beliefs, and use of breast and cervical cancer screening tests.

RESULTS

Table 1 shows the detailed resolution of the census and registry-based RDD samples. While Filipino households represented a similar proportion of the total numbers generated in each of the samples, an important difference was implemented between the 2 methods. In the census method, we screened out business/non-working numbers prior to enumeration, as reflected in the large difference in the proportion of non-residential numbers between the registry and census samples (33% vs 13%).

As indicated in Table 2, several reasonable measures may account for the efficiency of the 2 methods. When the most conservative measure was utilized, distinguished by the proportion of numbers belonging to Filipino households out of all numbers generated, both methods yielded samples with comparable efficiency (4.7% vs 3.9% for the census and registry methods, respectively). Given that Filipinos represent only 4.1% of the residential population in the 10-county area, according to the 1990 US Census, we also looked at the yield of Filipino households out of the population of potential and confirmed households. Under these measures of efficiency, the registry method offered an approximately 50% higher yield than would be expected from a general population survey. This figure was slightly better than the yield from the census method offering an approximately 30% higher yield than would be expected from a general population survey.

As shown in Table 3, once households with eligible women were located, both samples yielded a response rate of 73%. We interviewed women to determine if the 2 sampling methods had succeeded in locating women who differed systematically with respect to acculturation or use of breast and cervical cancer screening exams. These results are summarized in Table 4. For most of the variables we measured, small differences were evident between the groups, but no clear pattern emerged. However, we did observe that women in the census method sample were more likely to report ever having a mammogram com-

Table 2. Efficiency of the census- and registry-based RDD methods in locating Filipino households

	Census Method		Registry Method	
	N	%	*N*	%
Total	3406		3526	
Potential residential numbers:	2977		2346	
Confirmed residential numbers:	2880		2119	
Efficiency in locating Filipino households:				
Filipino households/total phone numbers generated	161/3406	4.7	139/3526	3.9
Filipino households/potential households	161/2977	5.4	139/2346	5.9
Filipino households/confirmed households	161/2880	5.6	139/2119	6.6

Table 3. Participation rates among eligible Filipinas located using the census- and registry-based RDD methods

	Census Method		Registry Method	
	N	%	*N*	%
Total Filipino households:	161		139	
Filipino households without eligible women:	33		33	
Filipino households not sampled:*	29		22	
Filipino households with eligible women:	99		84	
Refused interview	8	8	13	15
Too ill to participate	4	4	1	1
Did not speak one of 4 interview languages	8	8	5	6
Never home after 8+ tries	4	4	3	4
Partial interview	3	3	1	1
Completed interview	72	73	61	73

* These households included Filipina women age 20 or older but were contacted after we had identified and selected the total number of women we needed for that age strata.

We are considering improvements to the registry method that would enhance our ability to locate minority populations efficiently, while minimizing the selection bias.

pared to women in the registry method sample. Women in the registry sample were, on the other hand, more likely to report performing a monthly breast self-exam in the past year. This difference was most pronounced among women aged 40–49.

DISCUSSION

In this pilot study, both the census and registry methods located Filipino households with comparable efficiency and with a higher yield than would be expected in a population survey without a targeted population, using either the Mitofsky-Waksberg RDD method or a simple random sampling technique. No systematic pattern of responses was evident indicating that neither method sampled women who were systematically less-acculturated or less likely to use cancer screening tests. Clearly though, our samples were small; thus, even large differences between the groups yielded *P* values that were consistent with the null hypothesis of no association. Further details from a study on the characteristics associated with cervical cancer screening in a Filipino population, for which the present study was a pilot, have been published elsewhere.[11]

For researchers with access to population-based disease registries like the cancer registry used in this study, and now available in most states, the registry-based method offers an inexpensive way to locate PSUs. Researchers may therefore find the registry-based method preferable to the census-based method. The census-based method involves substantially more staff time to identify PSUs, unless the mapping of streets to PSUs could be purchased from a commercial firm at a reasonable price. The registry method could be improved by screening out business numbers prior to enumeration, as they were screened out for the census-based sample in this study. While the PSUs generated from the registry method usually reflect residential numbers, in a separate pilot using the same method for number generation, 5.1% of numbers generated in this manner were identified as business numbers by Pro-Phone.

While the evident gains in efficiency might seem to justify using either the registry or census methods, this premise overlooks the selection bias that these methods may introduce into the samples. As we noted in the introduction, one of the important strengths of the self-weighting Mitofsky-Waksberg method is that it generates an equal probability sample of the underlying population by using second-stage selection probabilities inversely proportional to the density of the PSU. Without this crucial second-stage weighting, the probability of selection for each element is proportional to the density of the target population (eg, Filipinos) in each PSU. Thus, even in the registry-based sample, households in areas with a higher density of Filipinos will be sampled at a higher probability than will households in lower-density areas. In addition, areas with a very-low density of Filipinos may not generate any Filipino cancer cases; thus, individuals in these areas are given zero probability of selection. Drawbacks are also inherent in census-based methods, particularly in cases where numbers are drawn only from high-density census tracts, and low-density areas are excluded from the sample entirely. To our knowledge, the effects of this population-density selection bias have not been explored in the epidemiologic literature. Although we did not have the data to look at this bias in our own study, our *a priori* hypotheses would be that higher density areas contain, on average, less-acculturated households and that Filipinas living in these households may differ from more-acculturated Filipinas in a systematic way. Both of these theories suggest areas for further research. In summary, balancing efficiency while minimizing bias may be more difficult depending on

Table 4. Comparison of selected characteristics of Filipinas located using the census- and registry-based RDD methods

	Census Method		Registry Method		
	N	%	*N*	%	*P* Value*
Language respondent is most comfortable speaking					
Filipino dialect	20	28	19	31	
English	19	26	15	25	
Both	33	46	27	44	.91
Raised in Philippines first 10 years	59	83	47	77	.38
How long respondent has been in United States					
0–10 yr	25	35	19	31	
11–20 yr	23	32	20	33	
>20 yr	24	33	22	36	.90
Married	56	78	44	72	.45
Catholic	61	87	52	85	.75
Monthly breast self-exam in past year					
Age 40–49	8	36	8	67	.15
Age 50+	10	32	12	44	.42
Ever had a mammogram					
Age 40–49	19	86	7	58	.10
Age 50+	26	81	19	70	.37
Two or more PAP smears in the past 5 years					
Age 20–39	15	83	19	86	1.00
Age 40–49	20	91	11	92	1.00
Age 50+	23	72	23	85	.35

* *P* value for the chi-square test H_0: no association.

characteristics of the ethnic population of interest.

We are considering improvements to the registry method that would enhance our ability to locate minority populations efficiently, while minimizing the selection bias. One option would be to alter the methods so that they are self-weighting. In essence, rather than calling all the randomly generated telephone numbers and retaining individuals that match the study criteria, we would designate a target number. Our goal would then be to identify *k* individuals per PSU. *K* would have to be chosen both to maximize the efficiency of the method (efficiency considerations favor higher values of *k*) and to minimize the number of unproductive PSUs (the argument for a lower value of *k*). Finding the optimal *k* depends on the population under study (Filipinos vs African Americans, for instance), the degree to which population density and acculturation vary throughout the population, and other technical factors.[3]

The need to locate subgroups that are rare in the general population has led to several modifications of standard RDD procedures. We have demonstrated that 2 of these modifications, the census and registry methods, offer substantial efficiency but, their utility is limited since both generate samples that tend to over-sample high-density areas. The degree to which these methods are considered viable depends on further refinement to limit, if not eliminate, their inherent selection biases without sacrificing their increased efficiency to locate minority populations.

ACKNOWLEDGMENT

This research was supported by Grant DAMD 17-94-J-4215 from US Army Medical Research Acquisition Activity.

REFERENCES

1. Rothman KJ, Greenland S. *Modern Epidemiology.* 2nd ed. Philadelphia, Pa: Lippincott-Raven Publishers; 1998.
2. Potthoff RF. Telephone sampling in epidemiologic research: to reap the benefits, avoid the pitfalls. *Am J Epidemiol.* 1994;139:967–978.
3. Waksberg J. Sampling methods for random digit dialing. *J Am Stat Assoc.* 1978;73:40–46.
4. Hartge P, Brinton LA, Rosenthal JF, Cahill JI, Hoover RN, Waksberg J. Random digit dialing in selecting a population-based control group. *Am J Epidemiol.* 1984;120:825–833.
5. Harlow BL, Davis S. Two one-step methods for household screening and interviewing using random digit dialing. *Am J Epidemiol.* 1988;127:857–863.
6. Olson SH, Kelsey JL, Pearson TA, Levin B. Evaluation of random digit dialing as a method of control selection in case-control studies. *Am J Epidemiol.* 1992;135:210–222.
7. Blair J, Czaja R. Locating a special population using random digit dialing. *Public Opin Q.* 1982;46:585–590.
8. Waksberg J. A note on "locating a special population using random digit dialing." *Public Opin Q.* 1983;47:576–579.
9. Casady RJ, Lepkowski JM. Stratified telephone survey designs. *Survey Methodol.* 1993; 19:103–113.
10. Marin G, Vanoss Marin B, Perez-Stable EJ. Feasibility of a telephone survey to study a minority community: Hispanics in San Francisco. *Am J Public Health.* 1990;80:323–326.
11. McBride MR, Pasick RJ, Stewart S, Tuason N, Sabogal F, Duenas G. Factors associated with cervical cancer screening among Filipino Women in California. *Asian Am Pac Isl J Health.* 1998;6:358–367.

AUTHOR CONTRIBUTIONS

Design and concept of study: Stewart, Pasick, Sabogal, Tuason, McBride, Horn-Ross
Acquisition of data: Stewart, Pasick, Sabogal, Tuason, McBride, Horn-Ross
Data analysis and interpretation: Hoggatt, Stewart, Pasick, Tuason, McBride, Horn-Ross
Manuscript draft: Hoggatt, Stewart, Sabogal, McBride, Horn-Ross
Statistical expertise: Hoggatt, Stewart, Tuason
Acquisition of funding: Stewart, Pasick, McBride, Horn-Ross
Administrative, technical, or material assistance: Sabogal, McBride
Supervision: Pasick, McBride, Horn-Ross
Testing protocol: McBride