Survey Methodology, Sample Representativeness, and Accurate Reporting of Population Health Statistics

R. Elsbett-Koeppen, D. Northrup, A. Noack, K. Moran
Toronto, Canada

RRFSS
Rapid Risk Factor Surveillance System
www.cehip.org/rrfss www.isr.yorku.ca

Institute for Social Research

Rationale

The collection of survey data is the core of all risk factor surveillance systems. Given the preeminent role of survey methods in conducting these studies, it is critical to assess the integrity of the data that are collected.

Declining response rates have been a bane to researchers who rely on survey data to measure behaviour, attitudes and awareness of health-related issues. In order to mitigate against declining response rates as well as increase sample representativeness and the accuracy of population health statistics, better survey research organizations now make a larger number of call attempts in order to reach people. This contrasts with some commercial survey providers who routinely make only four call attempts.

Method

Data Sources

The Rapid Risk Factor Surveillance System (RRFSS) Survey is carried out by the Institute for Social Research (ISR) at York University, Toronto, Canada, for a consortium of 21 Health Units and Health Departments (HUs) in the Province of Ontario. The RRFSS questionnaire is comprised of groups of questions or "modules". Core modules are required for all Health Units and optional modules can be selected by one or more Health Units. Each month, approximately 100 interviews are conducted in the service area of each HU. Data sets for all modules and all HUs are provided to each HU after each wave of data collection.

A sample is drawn for each of the health units for the whole duration (12 waves=1 year). Each health unit sample is randomly divided into 12 waves. Each wave's sample is then divided into six replicates that are released as needed. Consequently, each replicate is a random sub-set of the complete sample for the 12 waves.

Survey respondents are selected using a two-stage probability selection process. Random digit dialing (RDD) procedures are utilized to produce a sample of households in each HU region. Within each household, the interview is completed with a randomly selected adult (18 years of age or older). Interviews are not conducted if the respondent is not able to speak English well enough to complete the interview, or if potential respondents are identified as disabled and incapable of completing the questionnaire.

To maximize the chances of getting a completed interview from each sample telephone number, a minimum number of 14 calls are made to each number, of which at least eight are made during evening and weekend hours. For the 2002 RRFSS, 22,853 interviews were completed (a response rate of 60%) over the twelve waves: one-fifth of the interviews were realized the first time the interviewer called and, in total, three-fifths of the interviews took four or fewer call attempts to complete. In comparison, about fifteen percent of the completed interviews required ten or more calls (See Table 1 and Figure 1).

Data Analysis

The 2002 RRFSS data were used to identify 'easy to reach' (1 to 4 calls), 'not so easy to reach' (5 to 10 call attempts) and 'hard to reach' (11 or more call attempts) respondents. The distributions of the data were analyzed, as well as some associations (using cross- tabulation and Chi-Square statistics). Frequencies and cross-tabulations were used to identify 'easy to reach', 'not so easy to reach' and 'hard to reach' respondents, as well as their health indicators. A Negative Binomial Regression model was used to predict the cumulative probability of survey completion for a specific respondent profile (characteristics of an ideal respondent). SPSS was employed for cross tabulation analysis and Stata for the regression analyses.

Results

Cross-tabulations showed no significant differences between the released replicates, indicating a random distribution in the sub-samples as to respondent demographic characteristics such as age, gender, education and income (data not shown). The sampling process does not affect these results.

No significant differences were found in a comparison of the age distributions of the Ontario population (Statistics Canada, 2001) and the 2002 RRFSS age groups (Figure 2) in general, or by call attempts.

(continued on opposite panel)

Table 1: Call Attempts for the 2002 RRFSS

Call Attempts	Number	Percent
Easy to Reach (1-4 calls)	13,627	59.6
Not so Easy to Reach (5-10 calls)	6,424	28.1
Hard to Reach (11 or more calls)	2,802	12.3
Total	22,853	100.0

Figure 1: Call Attempts for the 2002 RRFSS (n=22,853)

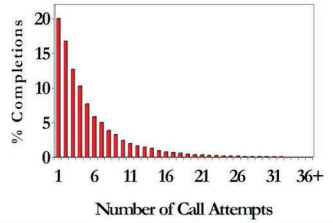

Figure 2: Ontario (Census Population) and RRFSS 2002 by Age Group and Number of Call Attempts (weighted data, n=22,346)

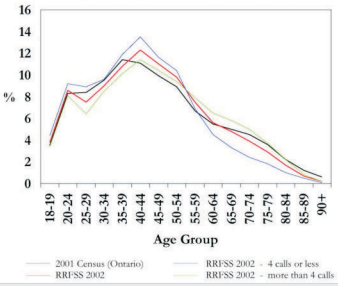

Figure 3: Completions (%) by Call Attempts for Fully Employed Respondents (n=22,752)

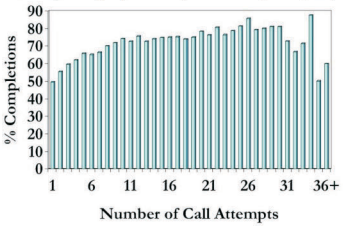

Table 2: Negative Binomial Regression for Number of Call Attempts (n=16,573)

Variable	Coefficient	Odds Ratio	p
Age (years)	-0.008	0.992	p<.001
Women	-0.114	0.893	p<.001
Education (ref: University graduate)			
High School Graduate	-0.009	0.991	p=.652
Less than High School	-0.038	0.963	p=.180
Annual Household Income (thousands)	0.001	1.001	p<.001
Number of Adults in Household	-0.027	0.974	p=.020
Has Children Living in Home	-0.005	0.996	p=.819
Fully Employed	0.262	1.299	p<.001
Constant	*1.695*	--	*p<.001*

Figure 4: Cumulative Predicted Probability of Survey Completion for Men, Fully Employed, University Degree, Earning $ 65 K, Living Alone (n=22,752)

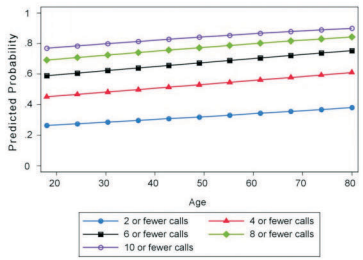

Table 3: Selected Differences Between Easy to Reach, Not so Easy to Reach and Hard to Reach Respondents (n~22,853)

Indicator	% Easy	% Not so Easy	% Hard	% Total	p
Health Status					
Reports 'excellent' or 'very good' health	56	60	60	58	<.001
Has high blood pressure	23	18	15	21	<.001
Has asthma	11	10	11	11	0.138
Has diabetes	7	5	4	6	<.001
Has other chronic disease	18	14	11	16	<.001
Health Knowledge					
Knows at least one cancer risk factor	57	59	61	58	<.001
Preventative Health Behaviour					
Ever had a pneumonia vaccination	17	13	11	15	<.001
Had a tetanus shot in past 10 years	60	63	63	61	<.001
Had a pap smear in past 2 years (women only)	72	80	81	75	<.001
Risk Behaviours					
Sunburned in last year	33	37	40	35	<.001
Consumed alcohol in last year	81	84	86	82	<.001
Consumed 2+ drinks in the hour before driving a car in past year	85	89	90	87	<.001

Results (cont'd)

For the fully employed respondents (of the total completions, n=22,752) the proportion of completed interviews by call attempts prior to completion is depicted in Figure 3. To be considered as fully employed, a respondent had to work for wages or had to be self-employed. The distribution indicates the value of increasing the number of call attempts for the fully employed.

In order to develop a model that would best predict the number of call attempts necessary to reach a specific respondent, a decision was made to use the Negative Binomial Regression model.

A base model for this regression analysis of the number of call attempts included gender, age, education, income and being fully employed. All independent variables, except education, were significant. The results indicate that respondents that are hard to reach tend to be male, younger, have a higher household income and are fully employed. For example, being a female respondent decreases the number of call attempts by a factor of 0.893 when holding all other variables constant (data not shown). That is, it takes 10% fewer call attempts for a female respondent holding all other variables constant. Inclusion of the number of adults and whether there are any children living in the household resulted in the model shown in Table 2.

The cumulative predicted probability of completing the survey (Figure 4) is represented for an 'ideal' respondent in terms of being male, being fully employed, with a university degree, earning C$ 65,000 or more (top 10%) and living alone (one adult household, no children). Again, the value of increasing the number of call attempts is demonstrated.

Selected health indicators, representing health status as well as some health behaviours, health risks and knowledge of cancer risk factors are presented in Table 3. The value of increasing the number of call attempts is again demonstrated.

Discussion

Studies have shown that single person households with employed members tend to be hard to reach, since they are at home less frequently in comparison to elderly household members. Other studies hypothesize that the 'not so easy to reach' and 'hard to reach' respondents differ significantly compared to those that are completing an interview after one, two, three or four call attempts ('easy to reach'). The results are supportive of the general population but are not consistent.

Our results stress the importance of sample representativeness such that it may affect our understanding of health issues. To collect data from 'easy to reach' respondents will underestimate, for example, good health. Good survey practice translates into better understanding of the health issues that are critical to health planning and policy development.

Cost implications are the main reason why commercial data collection agencies limit the number of call attempts to a maximum of four calls. Some commercial survey providers, who routinely make only four call attempts, are good data collection agencies for some special surveys; in general, however, four call attempts are not enough.

Limitations

The 2002 RRFSS data are subject to any of the biases associated with self-report data. The data were not weighted since we did not attempt to calculate population estimates. The regression models were used to predict whether there are any significant differences in knowledge of health behaviour between 'easy to reach' respondents, 'not so easy to reach' and 'hard to reach' respondents, rather than predicting a respondent's health behaviour and/or risk behaviour.

Conclusion

Data collection agencies that limit call attempts are more likely to ignore a part of the population that is different from the 'easy to reach' respondents in certain health risk behaviours. Consequently, users of surveillance system data must be cognizant of the efforts made by data collection agencies in order to obtain representative samples.