Undercoverage Analysis in Surveys Using Telephone Exchange Information

Ting Yan[1], Wei Zeng[1], Zhen Zhao[2]
[1]NORC at the University of Chicago, 55 East Monroe Street, Chicago, IL 60615
[2]Centers for Disease Control and Prevention, National Center for Immunizations and Respiratory Diseases, 1600 Clifton Road, Atlanta, GA 30333

Abstract

Surveys targeting particular segments of the general population are prone to undercoverage where the prevalence of the subpopulation found through the survey is less than the prevalence expected based on external sources (such as Census estimates). Both the noncoverage of the eligible population on the sampling frame and the nonresponse from the eligible population on the sampling frame contribute to the undercoverage problem experienced by surveys targeting a special population.

This paper examines the undercoverage of children aged 19 to 35 months in the National Immunization Survey (NIS). Census tract information tied to telephone exchanges is employed in the analysis. For the NIS, we identified variables that are highly correlated with the undercoverage of children aged 19 to 35 months old. Characterizing sampled persons that are most susceptible to being undercovered in surveys is essential for survey organizations to better target and recruit them on the one hand, and to search for poststratification variables that are most effective in reducing bias caused by nonresponse and noncoverage on the other.

Key words: undercoverage, NIS, telephone-exchange information

1. Introduction

For surveys targeting a particular segment of the general population, sample households are usually screened first to see if an eligible person lives in the household. If an eligible person is found, a main interview will be continued. The interview is terminated if no one living in the household is eligible for the survey. In general, screening processes tend to find fewer eligible sample persons than expected based on external sources (such as Census estimates), resulting in undercoverage of eligible population.

Many factors contribute to the undercoverage problem experienced by surveys targeting special populations. First, certain eligible sample persons are simply not covered by the sampling frame. In landline RDD surveys, for instance, households with only wireless telephone service and households with no telephone services are not covered on the sampling frame of telephone numbers. Consequently, eligible persons living in these households will not have a chance to be contacted in RDD surveys. Second, sampled households covered on the sampling frame may not agree to participate in the screening interview even if reached and contacted regardless of whether they have eligible persons or not. Third, for various reasons, the respondents of sampled households may not be willing to disclose that an eligible person lives in the household even if they agree to participate in the screening interview. Finally, identified eligible sample persons may refuse to participate in the main interview.

Noncoverage of eligible population on the sampling frame could be minimized or reduced if a different sampling frame is adopted. However nonresponse from sampled households and nonresponse from

eligible sample persons are experienced by all surveys to a different extent, regardless of the types of sample frame and the modes of data collection (Groves, 2006; Groves and Couper, 1998). In addition, noncoverage of the eligible population on the sampling frame and the nonresponse from the eligible population on the sampling frame are not unique to surveys targeting special populations; general population surveys are also prone to these two problems (Groves, 1989). The additional screening step exacerbates the extent of undercoverage of eligible population.

Undercoverage of the eligible population is problematic and undesirable. It increases the cost of the data collection since a larger sample would be needed in order to find enough eligible sample persons. In addition, there exists the potential for bias in survey estimates. Coverage bias can result if eligible persons not covered on the sampling frame and/or eligible persons not identified in the screening process differ systematically from those in the full population of interest on key statistics of interest. Nonresponse from each step of the screening process could produce additional bias on top of coverage bias when those who choose not to respond to the survey request are systematically different from those who agree to participate in the survey (Groves, 1989). Neither bias is desired.

Various operational procedures and statistical techniques are employed to reduce the extent of bias. Poststratification might help reduce the bias resulting from noncoverage and nonresponse. However, it is only efficient when variables used in poststratification are highly correlated with the variable of interest and with the sample person's propensity to respond to the survey (Little and Vartivarian, 2003, 2005). Therefore, if variables related to undercoverage are included in poststratification, the efficiency of poststratification might be improved.

This paper examines the undercoverage experienced by the National Immunization Survey (NIS), which targets households with children aged 19 to 35 months old. The NIS tends to find fewer children in the eligible age range than what is indicated by the American Community Survey (ACS). Since NIS is a RDD survey, little information is available about those who are not on the sample frame and those who do not respond. This paper attempts to identify geographical variables (such as census region and state) and telephone-exchange level variables that are highly correlated with households most susceptible to being undercovered in the NIS. Characterizing households at a greater risk of being undercovered in the NIS will help the NIS choose survey design features that allow the NIS to better target and recruit those households so as to improve the overall coverage of households with eligible children. Furthermore, these characterization variables could be used in poststratification to reduce bias caused by nonresponse and noncoverage of eligible children.

2. Data and Analytic Methods

The 2007 National Immunization Survey (NIS) data were used in this study. The NIS is a nationwide, list-assisted random digit-dialing (RDD) survey sponsored by the Centers for Disease Control and Prevention. It monitors the vaccination coverage rates among children aged 19 to 35 months. It uses a nationally representative quarterly data collection cycle and produces biannual vaccination coverage estimates at the national, state, and local area levels.

Each year, the NIS dials approximately 4.5 million telephone numbers and conducts interviews with approximately 24,000 households across the United States. The NIS purchases samples from the Genesys Sampling Systems (http://www.m-s-g.com), a commercial provider of RDD samples. Genesys attaches census tract information to every telephone exchange during the sample drawing process. Examples of these information include percent of people aged between 0 to 17 years living in the census tract, percent

of whites in the census tract, percent of people in the census tract whose income is less than $10,000, number of college graduates in the tract, and percent of home owners in the tract. For the NIS sample, every telephone number is associated with census tract information at the exchange level as mentioned above.

We limited our analyses to the 2007 first quarter data. In this quarter, 1,792,877 telephone numbers were sampled. 43.52% of them were screened out as business phone numbers, cell phone numbers, modem/fax numbers, or disconnected numbers. The remaining 1,009,869 numbers were sent to the phone center for dialing, resulting in 12,075 screened as an eligible household and 10,386 households with completed interviews. The final eligibility rate (the percent of screened households having at least one child in the eligible age range) is 3.26%. The CASRO response rate is 64.7%.

The goal of the analyses is to compare the percent of households with eligible children observed in the NIS to the expected percent of households with eligible children as indicated in the American Community Survey (ACS). The NIS call record data were used to compute the observed eligibility rate and the three year (2005-2007) ACS data were used to compute the expected eligibility rate. Since every telephone number belongs to a state and a census region, the comparison of observed and expected eligibility rates is straightforward at the geographic level. We computed eligibility rates for each state and census region in both the NIS and the ACS data and calculated an eligibility rate difference as the difference between the observed and the expected eligibility rates for each state and census region.

The comparison at the telephone exchange level is less straightforward. This is because the smallest geographical area in the three year ACS data is public use microdata area (PUMA). PUMAs are areas, typically of population 100,000 or more, that the Census Bureau defines for the purpose of inclusion in public-use microdata computer files, too protect the respondent's confidential data. Information on census tracts is only included in the 5-year ACS data, which will not be released until Fall 2010. Therefore, we have to extrapolate the expected eligibility rates at the telephone exchange level in order to make comparisons feasible.

Since every telephone number in the NIS sample is associated with census tract information at the telephone exchange level, we are able to segment our sample based on these census tract information. For example, based on the "percent of home owners" census information attached to each telephone number, we grouped our sample into four groups: exchanges that have 0% to less than 25% of home owners in census tracts; exchanges that have 25% to less than 50% of home owners in census tracts; exchanges that have 50% to less than 75% of home owners in tracts; and exchanges that have 75% or more of home owners in census tracts. For each segment, we are able to calculate the number of NIS-eligible households and the number of NIS-ineligible households, deriving the observed eligibility rate for that segment. We next compute benchmark eligibility rates from the ACS data. We can get the benchmark eligibility rate for home owners and for non-home owners separately. Using these two eligibility rates, we can calculate expected NIS eligibility rates for exchanges with a certain proportion of home owners, using the formula below:

> Expected NIS Eligibility Rate = % home owners * eligibility rate among home owners + (1 - % home owners) * eligibility rate among non-home owners.

3. Results

We first present the eligibility differences by two geographical variables (region and state) and then show the same analysis by three social-economic variables (home ownership, income, and race).

3.1. Eligibility Rate Differences by Geographies

Genesys split Alaska and Hawaii from the west as a separate region category before attaching census region to each telephone number. As shown in Table 1, we display observed and expected eligibility rates in five geographies -- Northeast, South, Midwest, West (excluding Alaska and Hawaii), and Alaska and Hawaii. Overall, the NIS observed eligibility rate is smaller than what is indicated by the ACS by 1.41 percentage points. The observed NIS eligibility rate is closest to that expected in the Midwest (1.23 percentage points) and the farthest in Alaska and Hawaii (1.59 percentage points).

Table 1: Eligibility Rate Differences by Region

	2007Q1 Sample Size (Percent)	(Observed) NIS Eligibility Rates	Expected NIS Eligibility Rates Using 3-year ACS data	Eligibility Rate Differences
Northeast	324,657 (18.1%)	2.92%	4.20%	1.28%
Midwest	381,124 (21.3%)	3.37%	4.60%	1.23%
South	732,327 (16.9%)	3.10%	4.66%	1.56%
West (Excluding Alaska and Hawaii)	302,861 (16.9%)	3.78%	5.14%	1.36%
Alaska and Hawaii	51,998 (2.9%)	3.55%	5.14%	1.59%
Total	1,792,877	3.26%	4.67%	1.41%

We conducted further analyses and compared observed and expected eligibility rates at the state level. The differences range from 0.39 percentage points in Nebraska to 2.05 percentage points in Texas.

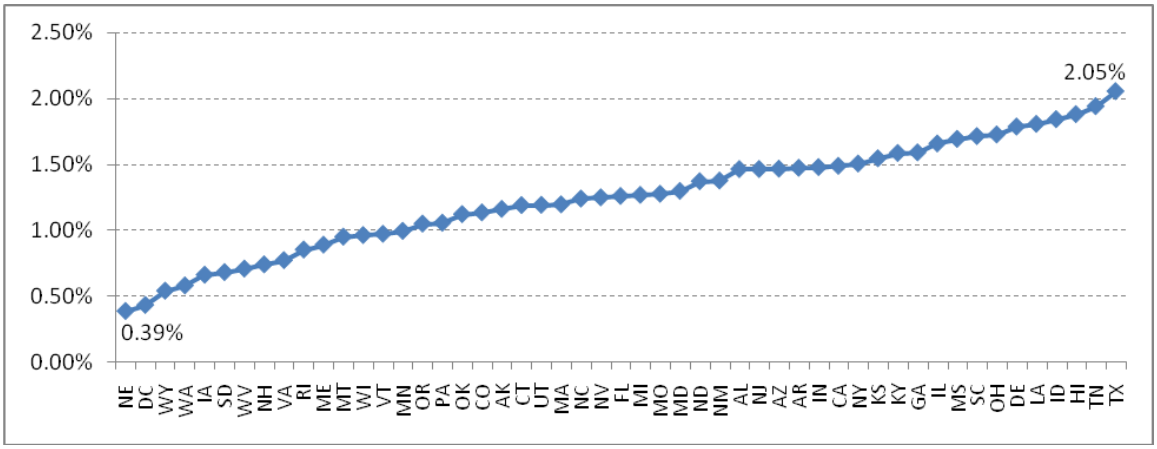

Figure 1: Eligibility Rate Differences by State

3.2 Eligibility Rate Differences by Social-Economic Characteristics

We identified three socio-economic dimensions available both in the sample file (as exchange-level information) and in ACS data; home ownership; income; and race. To facilitate our analysis, we characterized exchanges by the prevalence of homeowners in tracts, the prevalence of households whose

income is less than $50,000 in tracts, and the prevalence of whites in census tracts. Table 2 displays observed and expected rate of eligible households by home ownership. The eligibility rate difference is the largest in exchanges with less than 25% of home owners in tracts. The difference becomes smaller as the proportion of home owners increases. It seems that the eligibility rate difference is related to the prevalence of homeowners in tracts.

Table 2: Eligibility Rate Difference by Prevalence of Home Owners

Exchanges with	2007Q1 sample size (Percent)	Observed NIS Eligibility Rates	Expected NIS Eligibility Rates Using 3-year ACS data	Eligibility Rate Difference
0-less than 25% of home owners	79491 (4.43%)	3.38%	5.30 - 5.68%	1.92 - 2.30%
25% to less than 50% of home owners	311622 (17.38%)	3.27%	4.93 - 5.30%	1.66 - 2.03%
50% to less than 75% of home owners	875873 (48.85%)	3.29%	4.55 - 4.93%	1.26 - 1.64%
75% to 100% of home owners	525891 (29.33%)	3.19%	4.17 - 4.55%	0.98 - 1.35%

Table 3 shows the result by income. Although not a consistent finding as with home ownership, the difference is greater for exchanges with at least 25% of households whose income is less than $50,000 than for exchanges with less than 25% of households whose income is less than $50,000.

Table 3: Eligibility Rate Difference by Prevalence of Households with Income Less than $50,000

Exchanges with	2007Q1 sample size (Percent)	Observed NIS Eligibility Rates	Expected NIS Eligibility Rates Using 3-year ACS data	Eligibility Rate Difference
0-less than 25% of households whose income is less than $50,000	73757 (4.11%)	3.97%	4.86 - 5.04%	0.90 – 1.07%
25% to less than 50% of households whose income is less than $50,000	607115 (33.87%)	3.28%	4.68 - 4.86%	1.40 - 1.58%
50% to less than 75% of households whose income is less than $50,000	1025507 (57.20%)	3.18%	4.51 - 4.68%	1.32 – 1.50%
75% to 100% of households whose income is less than $50,000	86498 (4.82%)	3.23%	4.33 - 4.51%	1.10 – 1.28%

Eligibility rate difference by prevalence of whites in tracts is presented in Table 4. It appears that difference in eligibility rate is related to the prevalence of whites in tracts - exchanges with less than 25% of whites in tracts have the largest difference while exchanges with 75% and more whites have the smallest difference.

Table 4: Eligibility Rate Difference by Prevalence of Whites

Exchanges with	2007Q1 sample size (Percent)	Observed NIS Eligibility Rates	Expected NIS Eligibility Rates Using 3-year ACS data	Eligibility Rate Difference
0-less than 25% of whites	261441 (14.58%)	3.52%	6.01 - 6.68%	2.49 – 3.17%
25% to less than 50% of whites	291192 (16.24%)	3.32%	5.33 - 6.01%	2.01 – 2.68%
50% to less than 75% of whites	416032 (23.20%)	3.29%	4.65 - 5.33%	1.36 – 2.04%
75% to 100% of whites	824212 (45.97%)	3.16%	3.97 - 4.65%	0.82 – 1.49%

4. Discussion

This paper explored differences in households with eligible children aged 19-35 months in the 2007 Q1 NIS data. The analyses took advantage of exchange level information to identify geographical and social-economic characteristics of exchanges with large differences in terms of eligibility rate. On average, the eligibility rate observed in the NIS is smaller than what is indicated in the 3-year ACS data by 1.41%. Our analyses indicate that exchanges most prone to undercoverage are mostly in the south, have fewer home owners, income less than $50,000, and fewer whites. The results suggest that eligibility rate differences are related to the prevalence of home owners, prevalence of whites, and income.

Eligible households can be missed at several different steps. The eligible households may not be covered by the sampling frame; the household may decline to the survey request; the household may not want to reveal that they have children in the eligible age range even after they agreed to participate in the NIS.

This paper did not attempt to identify which step(s) is the major cause of the eligibility rate differences observed in the NIS. Nor did we attempt to quantify how much impact each step contributes to the difference experienced by the NIS. However, our results suggest that the presence of households with only wireless service and households without any telephone services play a role in the eligibility rate differences. Low-income, renters, and minorities are important predictors of whether or not a household is a wireless-only household or no-phone household. In addition, recent government statistics on the prevalence of wireless phone-only households show that most states in the south have at least 15% of households with only wireless telephones (Blumberg et al., 2009). These characteristics also apply to areas where the NIS experiences large eligibility rate differences, as suggested by our results. Even though our study does not provide conclusive data on the effect of wireless-only households and no-phone households on the eligibility rate differences, two additional studies launched by the NIS (NIS cell-phone study and NIS Address-based Sampling study) could show the impact of wireless-only and no-phone households on NIS.

This paper identified several exchange level variables related to eligibility rate differences. The efficiency of poststratification might be improved if these exchange variables are incorporated. We did not test this speculation in this paper. However, Copeland and co-authors (2009) did find that while the NIS must adjust for undercoverage of non-landline telephone households, it is not necessary to use population controls based on telephone characteristics. Rather, use of appropriate socio-demographic variables should be sufficient, which are already part of NIS weighting methodology.

References

1. Groves, R. M. (2006). Nonresponse rates and nonresponse error in household surveys. *Public Opinion Quarterly*, 70, 646-675.

2. Groves, R. and M. Couper (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

3. Groves, R. M. (1989). *Survey Errors and Survey Cost*. John Wiley and Sons.

4. Little, R. J. and S. Vartivarian (2003). On weighting the rates in non-response weights. *Statistics In Medicine*, 22, 1589-1599.

5. Little, R. and S. Vartivarian (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 161-168.

6. Blumberg, S. J., Luke, J. V., Davidson, G., Davern, M.E., Yu, T-C., and Soderberg, K. (2009). Wireless Substitution: State-level Estimates From the National Health Interview Survey, January–December 2007, National Health Statistics Report, Number 14, March 11, 2009. Available at http://www.cdc.gov/nchs/data/nhsr/nhsr014.pdf.

7. Copeland, K. R., Khare, M., Ganesh, N., Zhao, Z., and Wouhib, A. (2009). An evaluation of weighting methodology in an RDD survey with multiple population controls. Paper presented at the Joint Statistical Meeting, 2009, Washington, DC.