A Dual-Frame Design for the National Immunization Survey

Hee-Choon Shin,[1] Noelle-Angelique Molinari,[2] and Kirk Wolter [1,3]

[1] NORC, 55 East Monroe St., Suite 2008, Chicago, IL 60603
[2] Centers for Disease Control and Prevention/ National Center for Immunization and Respiratory Diseases,
1600 Clifton Rd, NE, MS E-62, Atlanta, GA 30333
[3] Department of Statistics, University of Chicago, 5734 S. University Avenue, Chicago, IL 60637

Abstract

The National Immunization Survey (NIS)—a nationwide, list-assisted random digit-dialing (RDD) survey conducted by the National Opinion Research Center (NORC) for the Centers for Disease Control and Prevention (CDC)—monitors the vaccination rates of children between the ages of 19 and 35 months. Each year, the NIS conducts interviews with approximately 24,000 households across the United States. We analyzed 2006 NIS data (children aged 19-35 months) and NIS-Teen data (children aged 13-17 years) data to determine the effect of directory-listed status on the immunization coverage rates. We considered alternative sampling designs by varying the proportion of directory-listed households, estimated population variances of the coverage rates for directory-listed and for unlisted households, and the unit cost of each completed interview from each frame. We confirmed that the proposed dual-frame design were more cost effective and not likely to introduce unacceptable new bias in the estimation. We have demonstrated the equivalence of UTD rates between children in listed and unlisted households, and the availability of efficiency gains by utilizing the proposed optimal stratified design. Therefore, other studies would benefit from this approach.

Key Words: Dual-frame, Multi-frame, Hartley, Stratified Design

1. Introduction

The National Immunization Survey (NIS)—a nationwide, list-assisted random digit-dialing (RDD) survey conducted by the National Opinion Research Center (NORC) for the Centers for Disease Control and Prevention (CDC)—monitors the vaccination rates of children between the ages of 19 and 35 months. Each year, the NIS conducts interviews with approximately 24,000 households across the United States.

The current sampling frame for the NIS consists of listed and unlisted land-line telephone numbers in 1+ banks. Of all the possible telephone numbers in the sampling frame, fewer than 25 percent are associated with households. About 68 percent of all working residential telephone numbers are listed. The rationale for a random-digit-dial (RDD) sampling approach is to obtain unbiased estimators by including the unlisted telephone numbers in the sampling frame. In implementing the RDD method, however, disproportionate resources are dedicated to activities related to determining the status of each telephone number.

We hypothesized that a dual-frame sampling design might increase the cost-effectiveness and statistical efficiency of the NIS without introducing unacceptable new bias into estimators of immunization coverage. One frame was the standard RDD sampling frame, and the second frame was a list dense in households with an eligible child. We pursued an alternative dual-frame sampling strategy confirming or refuting our hypothesis.

In this article, we explore the strategy of oversampling directory-listed telephone numbers while retaining a RDD sample of listed and unlisted numbers. In theory, because this approach would maintain the current coverage of the eligible populations, it should introduce no new biases into estimators of immunization coverage.

We analyzed 2006 NIS data (children aged 19-35 months) and NIS-Teen data (children aged 13-17 years) data to determine the effect of directory-listed status on the immunization coverage rates. We considered alternative sampling designs by varying the proportion of directory-listed households, estimated population variances of the coverage rates for directory-listed and for unlisted households, and the unit cost of each completed interview from each frame.

2. Background: Immunization Coverage Rates by Directory-Listed Status

We analyzed 2006 NIS and NIS-Teen data to investigate the effects of directory-listed status on the coverage rates. If there were significant differences in coverage rates between listed and unlisted households, it would be difficult to rationalize adopting a new sample design even if the new design promised better estimators in terms of efficiency. First, we looked at coverage rates of all vaccines by directory-listed status and tested for differences between the two rates. Among the 19 vaccines, only 2 (Influenza 1 and Influenza 2) show statistically different up-to-date (UTD) rates of children between listed and unlisted households. For teens, there are 3 significant differences (1:3:2:1:2 Series, 1+ Influenza, and 1+ Meningococcal) between listed and unlisted households out of 14 vaccines.

Racial/ethnic difference in UTD rates is a major concern of the CDC (Keppel et al. 2005). To evaluate differences for each racial/ethnic group, separate tests were performed within each racial/ethnic group. For non-Hispanic Blacks and non-Hispanic Asians, there were no significant differences between children who live in listed and unlisted households. However, there were six significant differences for non-Hispanic Whites and seven for non-Hispanic American Indians and Alaskan Natives. There was one significant difference for Hispanics. For non-Hispanic American Indians and Alaskan Natives, UTD rates of children in unlisted households were higher than those of listed households. For other racial/ethnic groups, UTD rates of listed households were higher than those of unlisted households.

To evaluate regional or state-specific differences in UTD rates between listed and unlisted households, we examined UTD rates of three vaccines (4:3:1:3:3 Series[1], 1+ Measles, and 3+ DTaP) for each estimation area. The rates were rarely different between listed and unlisted households. Rates for Vermont were exceptional; UTD rates of unlisted households were higher than those of listed households for the three vaccines.

To evaluate the independent effect of directory-listed status on the UTD rates while controlling for the effects of other covariates, logistic regression models were employed. In addition to directory-listed status, additional control variables (child's age, mother's age, household income, mother's education, child's race, and respondent's relationship to children) were included in the models. Let $p(x)$ be the probability of being UTD for a defined vaccine given values of explanatory variables, X . The proposed logistic regression model is

$$\log\left(\frac{p(X)}{1-p(X)}\right) = X\gamma ,$$

where γ is a vector of regression coefficients and X is a matrix of explanatory variables. Each γ indicates the expected change due to the explanatory variable in log-odds ratio of being up to date, all other things being equal, and e^{γ} is the effect in terms of the odds ratio.

Regression coefficients indicated differences between listed and unlisted households in the log-odds of being up-to-date. For children, the effects of listed status were not significant. There was no difference in UTD rates between those in listed and unlisted households after controlling for the effects of demographic variables. Even the two significant differences in the bivariate models shown disappeared. For teens, the differences were not significant except for the 1+ Human papilloma virus vaccine. The positive significant effect (1.38) of unlisted status on 1+ Human papilloma virus vaccine indicated that UTD rate was higher among teens from unlisted households, other things being equal.

On balance, UTD rates of children did not differ by listed and unlisted households. Sampling variability led some results to be significant. The absence of significant differences implied that we could proceed to disproportionately select more of either listed or unlisted telephone numbers as long as it decreased the sampling variances of estimators and was cost effective. However, we note that even if there were no statistical difference in UTD rates between children who live in listed and unlisted numbers, observed UTD rates of listed households tended to be higher than those of unlisted households.

[1] Four or more doses of Diphtheria/tetanus/pertussis (DTP), three or more doses of poliovirus vaccine, one or more doses of any measles containing vaccine (MCV), three or more doses of heamophilus influenzae type B (Hib), and three or more doses of Hepatitis B.

3. Optimal Dual-Frame Design (Hartley 1962)

Let *U* be a population of interest, and *A* and *B* be sampling frames of *U*. We assume $U = A \cup B$. In the dual-frame survey, samples are independently drawn from the two frames, *A* and *B*. Following the notation of Hartley (1962), let $a = A \cap B^c$, $b = A^c \cap B$, and $ab = A \cap B$, where *c* denotes the complement of a set. Population sizes are N_a, N_{ab}, and N_b. Sample sizes are $n_A = n_a + n_{ab}$ and $n_B = n_b + n_{ba}$. Now consider the total *Y* of a survey item *y*. The population total *Y* is $Y = Y_a + Y_{ab} + Y_b$, the sum of domain totals. Corresponding totals in the two frames are $Y_A = Y_a + Y_{ab}$ and $Y_B = Y_{ab} + Y_b$. The single frame estimators $\hat{Y}(y_A, \alpha)$ from a sample from frame *A* with design parameter vector α and $\tilde{Y}(y_B, \beta)$ from a sample from frame B with design parameter vector β can be expressed as:

$$\hat{Y}(y_A, \alpha) = \hat{Y}(y_a, \alpha) + \hat{Y}(y_{ab}, \alpha),\ \tilde{Y}(y_B, \beta) = \tilde{Y}(y_{ba}, \beta) + \tilde{Y}(y_b, \beta)$$

The ingenious approach of Hartley is the way the two frame estimators are combined. The estimator $\dot{Y}$ of Y in U is $\dot{Y} = \hat{Y}(y_a, \alpha) + \tilde{Y}(y_b, \beta) + p\hat{Y}(y_{ab}, \alpha) + (1-p)\tilde{Y}(y_{ba}, \beta)$

where *p,* the sampling allocation between the two frames, should be optimized along with α and β.

By allowing σ_a^2, σ_b^2, and σ_{ab}^2 to be the domain population variances and ignoring finite population correction factors, the variance of the total can be approximated as

$$\operatorname{var}\dot{Y} \approx \frac{N_A^2}{n_A}\left\{\sigma_a^2\left(1 - \frac{N_{ab}}{N_A}\right) + p^2\sigma_{ab}^2\frac{N_{ab}}{N_A}\right\} + \frac{N_B^2}{n_B}\left\{\sigma_b^2\left(1 - \frac{N_{ab}}{N_B}\right) + (1-p)^2\sigma_{ab}^2\frac{N_{ab}}{N_B}\right\}$$

Our objective for an optimal dual-frame design is to find the optimal values of p, n_A, n_B that minimize $\operatorname{var}\dot{Y}$ given the budget constraint C.

Consider a linear cost function $C = c_A n_A + c_B n_B$

where c_A is the unit cost to obtain a completed interview from the frame *A* sample and c_B is the unit cost to obtain a completed interview from the frame *B* sample. The optimal value of *p* is given by a solution of the following bi-quadratic equation:

$$\frac{c_A p^2}{c_B(1-p)^2} = \frac{\sigma_a^2\left(1 - \frac{N_{ab}}{N_A}\right) + \frac{N_{ab}}{N_A}p^2\sigma_{ab}^2}{\sigma_b^2\left(1 - \frac{N_{ab}}{N_B}\right) + \frac{N_{ab}}{N_B}(1-p)^2\sigma_{ab}^2}$$

With *p* from the above, the optimal sampling rates are

$$\frac{n_A}{N_A} = C'\sqrt{\frac{\sigma_a^2\left(1-\frac{N_{ab}}{N_A}\right)+\frac{N_{ab}}{N_A}p^2\sigma_{ab}^2}{c_A}}\ ,\ \frac{n_B}{N_B} = C'\sqrt{\frac{\sigma_b^2\left(1-\frac{N_{ab}}{N_B}\right)+\frac{N_{ab}}{N_B}(1-p)^2\sigma_{ab}^2}{c_B}}$$

where C' is to be determined to meet the budget constraint, C.

When frame A covers all elements of the population of interest, $N_{ab} = N_B$, and $\sigma_{ab}^2 = \sigma_B^2$. The above formula simplifies to:

$$p^2 = \frac{\frac{\sigma_a^2}{\sigma_{ab}^2}\left(1-\frac{N_B}{N_A}\right)}{\frac{c_A}{c_B}-\frac{N_B}{N_A}}.$$

Sampling variances from the optimal design can be compared to the variance of the estimator from an equal cost frame A-only design with simple random sampling. Table 1 shows the expected relative efficiency, in terms of variance reduction, of the dual-frame design for various scenarios compared to the frame A-only design. When population variances were the same and the overlap was 70 percent, the variance was 87.7 percent of the frame A-only design, given a unit cost for frame B that was one-half that of frame A.

4. Application to NIS

For purposes of the current investigation, the RDD frame was regarded as frame A, and the list frame of listed residential telephone numbers was frame B.

Optimal Sample Allocation

We considered two scenarios: First, we used the proportion of listed households among possible telephone numbers in 1+ banks. The 1+ bank is 100 consecutive telephone numbers that contains at least one directory-listed number. We assumed the ratio of listed households to possible telephone numbers in 1+ banks was about 27 percent. Second, we used the proportion of listed households among the total number of households. It was estimated that the proportion of directory-listed households was 68 percent.

Table 2 shows two specific cases of the proportion of frame B (listed telephone numbers). When the proportion was 27 percent, the variance of dual-frame design was 90.6 percent that of the frame A-only design given the unit cost of frame B was one-fifth that of frame A's. When the proportion was 68 percent, the variance of dual-frame design was 68.9 percent that of the frame A-only design given the unit cost of frame B was one-fifth that of frame A's.

As discussed earlier, we saught to optimally allocate the sample elements to each frame by optimizing *p*. We assumed the total required number of completed household interviews was 24,000. Thus, for a typical NIS design, the total cost was the product of 24,000 and the average unit cost from a RDD sample of telephone numbers.

It was estimated that the population variances in the listed and unlisted groups were equal and we assumed the ratio $\sigma_B^2 / \sigma_a^2 = 1$. Four distinct values (0.67, 0.50, .0.33, 0.25) were considered.

Table 2 reports the optimal value of *p* and allocated numbers of completed interviews for the RDD and list samples. The last three columns contain the numbers of completed interviews from the RDD frame, from the list frame, and the total numbers of combined completed interviews for various scenarios and the expected relative efficiency in terms of completed interviews. For example, the last row shows that with an optimal allocation, we could obtain 43,756 interviews (instead of 24,000) with the budget constraint if the overlap were 68 percent and the cost ratio was 0.25.

5. An Alternative Cost Function

In section 3, we assumed a constant c_A for all n_A units from frame *A*. In practice, c_A is the average of the costs from each of the n_A units. A sampled unit or telephone number goes through many steps to be ready for actual interview. Before the respondent is interviewed, the telephone number should be resolved, and working residential number (WRN) status should be determined. After the status of the household is ascertained, eligibility screening and subsequent interviewing should be attempted. Let

$\pi_B =$ eligibility rate in n_B samples from list stratum
$\pi_a =$ eligibility rate in n_a samples from unlisted stratum
$m_B =$ number of completed interviews from n_B
$m_a =$ number of completed interviews from n_a
$c^I =$ unit cost of resolution, screening and interviewing
$c^S =$ unit cost of resolution and screening

Now consider the following cost function:

$$C = c^0 + \left\{c^I m_a + c^S\left(n_a - m_a\right)\right\} + \left\{c^I m_B + c^S\left(n_B - m_B\right)\right\}$$

where c^0 is a constant cost. Because m_B and m_a are not known in advance, the expected cost function is

$$E(C) = c^0 + \left\{c^I \pi_a n_a + c^S\left(n_a - \pi_a n_a\right)\right\} + \left\{c^I \pi_B n_B + c^S\left(n_B - \pi_B n_B\right)\right\}$$

Letting $r = {c^I}\big/{c^S}$, the cost function is

$$E(C) = c^0 + \left\{c^S n_a\left(\pi_a r - \pi_a + 1\right)\right\} + \left\{c^S n_B\left(\pi_B r - \pi_B + 1\right)\right\}$$

With this cost function and applying the Neyman allocation, the optimal sampling fraction or allocation for a stratified sample design is

$$f_a = k\sqrt{\frac{\pi_a}{\pi_a(r-1)+1}} \text{ , and } f_B = k\sqrt{\frac{\pi_B}{\pi_B(r-1)+1}} \text{ .}$$

where *k* is determined to meet the budget constraint *C* (Kalton 1993, 2003; Kish 1965; Waksberg, Judkins, & Massey 1997). Let the estimator and variance from the frame A-only sample with stratification be $\hat{Y}_{A.st} = \hat{Y}\left(y_{A.st}, \alpha'_{st}\right)$ and $\operatorname{var}\hat{Y}_{A.st} = V_{A.st}\left(y_{A.st}, \alpha'_{st}\right)$ where α'_{st} implies optimally allocated sample sizes between the two strata with the same budget constraint. The gain in precision from the use of disproportionate sampling with optimal sampling fractions over the use of simple proportional allocation can be found in Tables 1 and 2 of Kalton (2003).

6. Specific Application to NIS

Based on the estimated cost ratio (1.087) and eligibility rates (3.9% for the unlisted and 3.1% for the listed), the estimated optimal sampling rates for the stratified design are 0.00039 for the list stratum and 0.00012 for the unlisted stratum.

Table 3 reports relative efficiency in terms of the required number of telephone samples for each design based on optimal allocation. With traditional RDD design, we would need to select 77,358 telephone numbers to obtain 180 effective completed interviews. With optimal stratified design, we would need to select 55,529 telephone numbers to obtain 180 effective completed interviews. Table 4 displays the assumed rates for the sample size calculation.

7. Summary and Conclusion

Recent developments in communication technology jeopardize the efficiency and the effectiveness of the RDD method. The ever-increasing cell-only population raises serious questions about the adequacy of coverage in RDD surveys. The cost increase due to low WRN rates and low response rates lessens the reliability and validity of survey results by limiting necessary resources for quality surveys. The current study investigates the possibility of applying a multi-frame design to the NIS. We confirmed that the proposed dual-frame design were more cost effective and not likely to introduce unacceptable new bias in the estimation. We have demonstrated the equivalence of UTD rates between children in listed and unlisted households, and the availability of efficiency gains by utilizing the proposed optimal stratified design. Therefore, other studies would benefit from this approach.

References

Hartley, H. O. 1962. “Multiple Frame Surveys.” *Proceedings of the Social Statistics Section* 203-206. American Statistical Association.

Kalton, G. 1993. *Sampling Rare and Elusive Populations*. New York: Department of Economic and Social Information and Policy Analysis Statistics Division, United Nations.

Kalton, G. 2003. “Practical Methods for Sampling Rare and Mobile Populations.” *Statistics in Transition 6, 491-501*.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Keppel, K, E. Pamuk, J. Lynch, O. Carter-Pokras, I. Kim, V. Mays, J. Pearcy,V. Schoenbach, and J. S. Weissman. 2005. *Methodological Issues in Measuring Health Disparities. Vital and Health Statistics, Series 2*, 141, National Center for Health Statistics.

Waksberg, J.W., D. Judkins, and J. Massey. 1997. “Geographic-based oversampling in demographic surveys of the United States.” *Survey Methodology* 23, 61-7.

Table 1. Variance reduction in dual frame when $\sigma_B^2 / \sigma_a^2 = 1$.

		Proportion of population in cheap frame (N_B / N_A)								
		0.10	0.27	0.50	0.60	0.68	0.70	0.80	0.90	1.00
Cost ratio (c_B / c_A)	0.01	0.918	0.775	0.571	0.477	0.399	0.379	0.276	0.164	0.010
	0.05	0.938	0.826	0.656	0.573	0.501	0.482	0.381	0.260	0.050
	0.10	0.952	0.861	0.718	0.645	0.579	0.562	0.465	0.344	0.100
	0.20	0.968	0.906	0.800	0.742	0.689	0.674	0.589	0.475	0.200
	0.30	0.978	0.935	0.857	0.812	0.769	0.757	0.686	0.582	0.300
	0.40	0.986	0.956	0.900	0.866	0.833	0.824	0.765	0.676	0.400
	0.50	0.991	0.971	0.933	0.909	0.884	0.877	0.832	0.758	0.500
	0.60	0.994	0.983	0.958	0.942	0.925	0.920	0.888	0.831	0.600
	0.70	0.997	0.991	0.977	0.968	0.957	0.954	0.933	0.894	0.700
	0.80	0.999	0.996	0.990	0.986	0.981	0.979	0.968	0.945	0.800
	0.90	1.000	0.999	0.997	0.996	0.995	0.994	0.991	0.983	0.900
	1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: Some of the entries are also shown in Table 3 of Hartley (1962).

Table 2. Optimal sample allocation in terms of completed interviews for dual frame design with given the total cost for 24,000 completes from Frame A (RDD).

Proportion of Population in Cheap Frame (N_B / N_A)	Cost ratio (c_B / c_A)	Optimal value of *p*	Frame		
			Frame A (RDD)	Frame B (List)	Total
0.27 (Proportion of Listed Telephone Numbers)	0.67	0.7704	22,768	1,849	24,616
	0.50	0.6496	22,357	3,285	25,643
	0.33	0.5171	22,124	5,628	27,752
	0.25	0.4424	22,103	7,589	29,692
0.68 (Proportion of Listed Households)	0.67	0.6247	18,857	7,715	26,572
	0.50	0.4924	17,764	12,473	30,236
	0.33	0.3714	17,338	19,986	37,324
	0.25	0.3105	17,415	26,341	43,756

Table 3. Efficiency of optimal stratified design

	Optimal Stratified Design					One Sample Design	
	Unlisted		Listed		Total		
	Rate	Number	Rate	Number	Number	Rate	Number
Required Sample of Telephone Numbers		25,083		30,446	55,529		77,358
Resolved Telephone Numbers	79.40%	19,916	79.40%	24,174	44,089	79.40%	61,421
Households Identified	5.40%	1,075	60.30%	14,577	15,652	26.08%	16,022
Households Successfully Screened	87.66%	943	87.66%	12,777	13,720	87.66%	14,044
Households with Eligible Children	3.90%	37	3.10%	396	433	3.08%	433
Completed Interviews	83.92%	31	83.92%	332	363	83.92%	363
Completed Children	1.04	32	1.04	344	376	1.04	376
Provider Consent	80.97%	26	80.97%	279	305	80.97%	305
Conditional Adequacy	86.85%	22	86.85%	242	265	86.85%	265
Effective Completes	1.47	15	1.47	165	180	1.47	180

Table 4. Assumed Rates

	Unlisted	Listed
Resolved telephone numbers* – *Resolution rate*	79.40%	79.40%
Households identified	5.40%	60.30%
Households successfully screened for presence of age-eligible children – *Screening completion rate*	87.66%	87.66%
Households with no age-eligible children	3.90%	3.10%
Households with age-eligible children – *Eligibility rate*	83.92%	83.92%
Households with age-eligible children with completed household interviews – *Interview completion rate*	1.04	1.04
Children with consent to contact vaccination providers	80.97%	80.97%
Children with adequate provider data	86.85%	86.85%
Design Effect	1.47	1.47