DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 545–557, http://dx.doi.org/10.1515/JOS-2015-0034

# Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations

*Martin Barron*[1], *Michael Davern*[1], *Robert Montgomery*[1], *Xian Tao*[1], *Kirk M. Wolter*[1], *Wei Zeng*[1], *Christina Dorell*[2], *and Carla Black*[2]

We investigate disproportionate stratified sampling as a possibly efficient method of surveying members of a rare domain in circumstances in which there is no acceptable list of members. In this work, we assume that information is available at the sampling stage to stratify the general-population sampling frame into high- and low-density strata. Under a fixed constraint on the variance of the estimator of the domain mean, we make the optimum allocation of sample size to the several strata and show that, in comparison to proportional allocation, the optimum allocation requires (a) a smaller total sample size but (b) a larger number of interviews of members of the rare domain. We illustrate the methods using information about American consumers maintained by market-research companies. Such companies are able to develop lists of households that are thought to have a defined attribute of interest, such as at least one resident in a user-specified age range. The lists are imperfect, with false positives and negatives. We apply an age-targeted list to the National Immunization Survey (NIS), conducted by the Centers for Disease Control and Prevention, which targets the relatively rare population of children age 19–35 months. The age-targeted list comprises the high-density stratum and the rest of the survey's sampling frame comprises the low-density stratum. Given the optimum allocation, we demonstrate potential cost savings for the NIS in excess of ten percent.

*Key words:* Optimum allocation; cost model; variance; disproportionate stratification; rare population; age-targeted list; telephone surveys; National Immunization Survey.

## 1. Introduction

Surveys of rare populations are common in a variety of scientific fields. For example, health surveys often target low-prevalence domains, such as people with a specific disease, a specific chronic condition, a special healthcare need, or people who have received specific healthcare services. While in general there is no universally accepted demarcation between rare and nonrare, we have in mind possible rare domains that comprise less than ten percent of the general population.

[1] NORC – University of Chicago, 55 East Monroe Street Suite 3000, IL 60603-5805, Chicago, Illinois, U.S.A. Emails: martin-barron@norc.org, davern-michael@norc.org, montgomery-robert@norc.org, tao-xian @norc.org, wolter-kirk@norc.org, and zeng-wei@norc.org.
[2] National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, U.S.A. Emails: eqw1@cdc.gov and cblack2@cdc.gov.
**Acknowledgments:** The findings and conclusions in this article are those of the author(s) and do not necessarily represent the views of Centers for Disease Control and Prevention. Kirk Wolter presented this article in October 2012 at the International Conference on Survey Methods for Hard to Reach Populations in New Orleans, Louisiana. The authors thank the associate editor and referees for helpful comments on earlier drafts of the article.

We consider the problem of sampling when two circumstances are true: (1) no acceptable sampling frame exists for the rare domain of interest, henceforth denoted by $D$, and (2) an acceptable sampling frame does exist for the general population and auxiliary information is available at the time of sampling that enables the survey statistician to partition this frame into high- and low-density strata. The former are presumed to have higher prevalence rates (also called the *eligibility rate*) of the rare population than the latter. A sample is selected from each stratum; a brief screening interview is administered to persons in the sample to ascertain membership in $D$; and then members receive the main survey interview and nonmembers are not interviewed. Practical applications of this problem may encounter a range of eligibility rates in the various strata. Throughout this article, we use the labels *high density* and *low density* simply to indicate that one or more strata have higher eligibility rates, perhaps much higher, than the other strata, not to imply any absolute level of eligibility.

One example of this sampling problem occurs when a list (possibly quite imperfect, reflecting false positives and false negatives) of members of $D$ exists and is available at the time of sampling. The list itself may be considered the high-density stratum and all persons represented on the general sampling frame and not on the list may be considered the low-density stratum. A second example occurs when the sampling frame is stratified by census variables that are thought to be associated with membership in $D$. Such examples may become increasingly important in the future as cost pressures on surveys mount.

Our main aim in this article is to develop a method of *disproportionate stratification* in which the high-density strata are sampled at higher rates than the low-density strata. We examine whether the use of different sampling rates can result in lower data-collection costs than when the same sampling rate is used across the entire sampling frame. Aspects of this sampling problem have been treated previously by Sudman (1972), Waksberg (1973), and Kalton and Anderson (1986). Kalton (2009) arrived at the general conclusion that disproportionate stratification can reduce cost only when three conditions are true: (a) the prevalence rates in the high-density strata are much higher than those in the low-density strata, (b) the high-density strata contain a substantial portion of the overall rare domain $D$, and (c) the per-unit cost of the main data collection must be high relative to the cost of screening. Valliant et al. (2014) study the use of stratification of address-based samples of households in which the strata are defined by auxiliary information from commercial sources.

Our specific aims are to give a precise definition of the method of disproportionate stratification and demonstrate the optimum design and its sample sizes within this class (Section 2), to describe certain information available from market-research companies that can be used for implementation of such stratification (Section 3), and to illustrate the optimum design and select market-research information using an age-targeted list applied to the National Immunization Survey (NIS), a project conducted on an ongoing basis by the Centers for Disease Control and Prevention to measure the vaccination status of young children (Section 4). The article closes with a brief summary and recommendations (Section 5).

## 2. Methods for an Optimum Allocation

Two notions of optimality are standard in survey sampling: first, one can fix the variance of a key survey statistic of interest and design the sample to minimize the cost of data

collection, or second, one can fix the cost of data collection and design the sample to minimize the variance of the key statistic. Both notions of optimality lead to a similar relative allocation of the sample size across the several strata (Cochran 1977). We will focus on the first notion of optimality, and comment briefly on the second notion at the end of this section.

We consider a sampling design in which there are $L$ strata indexed by $h$, and, without loss of generality, take the eligibility rate of the rare domain $D$ to be decreasing from $h = 1$ to $h = L$. Simple random samples are taken from each of the strata, resulting in the selection of some members of the rare domain and some nonmembers.

A brief screening interview is conducted to determine the members of $D$, followed by the main interview of such members. In this section, we consider the ideal circumstance of complete response, while in Section 4 we give an illustration in which nonresponse does occur. Furthermore, throughout the article, we assume that domain membership can be ascertained without error in the screening interview. This setting is in contrast with some survey applications in which reporting, coding, or definitional problems can result in erroneous classifications of sampling units as in $D$ or not in $D$.

We let $c_{scr}$ denote the cost (or hours) per screening interview and $c_{inv}$ the cost (or hours) per main interview. We let $n_h$ be the number of completed screening interviews and $m_h$ the number of completed main interviews in stratum $h$. Moreover, we let $r_h = N_{Dh}/N_h$ denote the population eligibility rate (size of the rare domain $D$ as a proportion of the size of the sampling frame) in stratum $h$ and $r = \sum_{h=1}^{1} W_h r_h = N_D/N$ the overall eligibility rate across the entire sampling frame, where $W_h = N_h/N$ is the proportion of units on the sampling frame that are classified in stratum $h$.

Total expected survey costs can be expressed by

$$T = \sum_{h=1}^{1} \left( c_{scr} n_h + c_{inv} E\{m_h\} \right) = \sum_{h=1}^{1} t_h n_h \,, \tag{1}$$

where $t_h = c_{scr} + c_{inv} r_h$ is the average combined cost per unit in the sample. On average, each unit in the sample incurs its own cost of screening plus a fractional share of the cost of the main interview, where the fraction is the eligibility rate. For simplicity, we have omitted fixed costs from the model, because they have no bearing on the optimum allocation. Also for simplicity, we have assumed that the per-unit costs are identical in the two strata. The methods extend directly to the case where the cost components vary by stratum, such as when response rates vary by stratum.

We assume the main aim of the survey is to estimate the mean of the rare domain, say $R = Y/X$, where $Y_{hi}$ is the variable of interest for members $(h, i)$ of the rare domain and is zero for nonmembers, $X_{hi}$ is 1.0 for members of the rare domain and is zero for nonmembers, and $Y$ and $X$ are the population totals of these variables. We let $\hat{R} = \hat{Y}/\hat{X}$ be the standard ratio estimator of $R$, where $\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} d_{hi} y_{hi}$ is the estimated domain total of the variable of interest, $\hat{X} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} d_{hi} x_{hi}$ is the estimated total number of members of the rare domain, and $d_{hi} = N_h/n_h$ is the design weight for all $i = 1, \ldots, n_h$ and $h = 1, \ldots, L$.

Assuming that finite population correction terms can be ignored and that the means and variance components are of similar value in the various strata, the Taylor series

approximation to the variance of the estimator is given approximately by

$$Var\{\hat{R}\} \doteq \frac{S^2}{r^2}\sum_{h=1}^{L}\frac{W_h^2 r_h}{n_h}\,, \tag{2}$$

where $S^2$ is the variance component among members of the rare domain. Kalton and Anderson (1986) give a similar expression for this variance. An alternative exact expression for the variance can be given in lieu of (2) in the event that the variance components differ from stratum to stratum.

Given the foregoing, the classical optimum allocation (Cochran 1977) of the sample to the two strata, which minimizes cost subject to a constraint on the variance, is given by

$$n_h^o = a_h n^o\,, \tag{3}$$

where

$$a_h = \frac{W_h\sqrt{r_h}/\sqrt{t_h}}{\sum_{h'=1}^{L} W_{h'}\sqrt{r_{h'}}/\sqrt{t_{h'}}}\,, \tag{4}$$

$$n^o = \frac{S^2}{V^o r^2}\sum_{h=1}^{L} W_h\sqrt{r_h}\sqrt{t_h}\sum_{h=1}^{1} W_h\sqrt{r_h}/\sqrt{t_h}\,, \tag{5}$$

and $V^o$ is the specified fixed constraint on the variance. The sample size within a stratum is proportional to the size of the stratum and to the root of the eligibility rate in the stratum, and inversely proportional to the root of the per-unit cost of data collection in the stratum. The expected number of interviews of members of the rare domain is

$$m^o = \sum_{h=1}^{1} n_h^o r_h = \frac{S^2}{V^o r^2}\left(\sum_{h=1}^{1} W_h r_h\frac{\sqrt{t_h}}{\sqrt{r_h}}\right)\left(\sum_{h=1}^{L} W_h r_h\frac{\sqrt{r_h}}{\sqrt{t_h}}\right), \tag{6a}$$

and the minimum total cost under the optimum allocation is

$$T^o = \frac{S^2}{V^o r^2}\left(\sum_{h=1}^{1} W_h\sqrt{r_h}\sqrt{t_h}\right)^2. \tag{6b}$$

An alternative sampling design that is used in many surveys involves the selection of the sample without regard to the high- and low-density strata, or effectively the selection of the sample from the sampling strata using proportional allocation. The sample sizes required to achieve the variance constraint are

$$n_h^p = W_h n^p \tag{7}$$

$$n^p = \frac{S^2}{V^o r^2}\sum_{h=1}^{1} W_h r_h\,, \tag{8}$$

the expected number of interviews of members of the rare domain is

$$m^p = \frac{S^2}{V^o r^2}\left(\sum_{h=1}^{1} W_h r_h\right)^2, \tag{9a}$$

and the total cost given this allocation is

$$T^p = \frac{S^2}{V^o r^2}\sum_{h=1}^{1} W_h r_h \sum_{h=1}^{1} W_h t_h\,. \tag{9b}$$

A measure of the cost savings associated with the optimum allocation is the ratio of total costs $T^o/T^p$, where the superscripts "$o$" and "$p$" signify optimum and proportional allocation, respectively. This ratio is guaranteed to be less than or equal to 1 by construction. If the eligibility rates are homogeneous, that is, $r_h = r$, for all $h$, then the ratio is equal to 1. Cost can be reduced relative to proportional allocation when the eligibility rates are variable and there are high-density strata of non-negligible size.

In comparing optimum and proportional allocations when variance is fixed, two inequalities are true: (i) $n^o/n^p \leq 1$ and (ii) $m^o/m^p \geq 1$. Because $T^o/T^p \leq 1$, the ratio of sample sizes is $n^o/n^p \leq c_{scr} + c_{inv}r^p/c_{scr} + c_{inv}r^o$, where $r^p = \sum_{h=1}^{1} W_h r_h = r$ and $r^o = \sum_{h=1}^{1} a_h r_h$. Inequality (i) follows from the fact that $r^o \geq r^p$. Applying the Cauchy-Schwarz inequality to (6) and (9) gives inequality (ii).

Summarizing the results for fixed variance, the optimum allocation results in cost savings relative to proportional allocation; it requires a smaller total sample size but a larger number of interviews of members of the rare domain than does proportional allocation. The optimum allocation involves disproportionate sampling, it creates a weighting effect, and it therefore requires more interviews to achieve the fixed variance.

Briefly, for fixed cost, the variance-minimizing optimum allocation is given by (3) and (4), where $n^o = T^o\left(\sum_{h=1}^{1} W_h\sqrt{r_h}/\sqrt{t_h}\right)/\left(\sum_{h=1}^{1} W_h\sqrt{r_h}\sqrt{t_h}\right)$. Consider the special case $c_{scr} = 0$, $c_{inv} = 1$, and $T^o = \sum_{h=1}^{1} n_h r_h$, which corresponds to fixing the expected sample size in the rare domain $D$. For this case, the optimum allocation is proportional allocation with $n^o = T^o/r$.

## 3. Market-Research Lists for Stratification

Market-research companies have developed proprietary databases containing demographic, behavioral, and consumer information on people and households throughout the world. These data can be used as the basis for the stratification used in Section 2. Even though the specific details of their construction are proprietary, it is known that the databases are compiled from product registrations, store loyalty programs, credit-card purchases, cable-television viewing, internet searching, smartphone applications, coupon redemptions, mobile health devices, voter registration databases, publicly available real-estate transactions, as well as many other sources. And while the data from market-research companies are not always accurate at the individual case level (Pasek et al. 2014), they may still be useful for stratifying a survey sampling frame of the general population into high- and low-density strata for households or people who have the rare characteristic of interest. Using the lists provided by market-research companies containing names, telephone numbers or addresses (depending on the sampling frame used), the sampling

statistician can divide the sampling frame into two or more strata based on whether the market-research company has associated the name, telephone number or address with a specific rare trait or characteristic of interest (domain *D*).

The general approach of stratifying the sampling frame into high- and low-density strata is not limited to lists provided by market-research companies. For example, if a team of researchers was interested in studying asthma among children using an address-based sample frame, they might be able to obtain a high-density list of addresses from administrative data of children on Medicaid (Medicaid is a government health-insurance program for needy people in the U.S.) with asthma-related prescriptions. The low-density stratum would be comprised of all remaining addresses. And there could be combinations with one high-density list coming from a state Medicaid agency of addresses of child beneficiaries with asthma-related prescriptions, a second list coming from a market-research company that identifies households likely to have children, and a third low-density frame of all remaining addresses not on either of the two high-density lists. Other applications of this method could entail using voter registration lists as the high-density frame for an address-based sample of likely voters for a local election, and the low-density frame could be all the remaining addresses. Market-research companies and administrative data sources offer ample opportunities to take advantage of this kind of methodology, as many lists are available to stratify the sampling frames into high-density and low-density strata that presumably have differing eligibility rates for members of the rare domain *D*. Lists used for stratification could target information on age, race, ethnicity, people who purchased and registered specific products (e.g., insulin pumps or asthma prescriptions), disease registries, voter registration lists, and lists of households who redeem specific coupons.

The methods presented in the foregoing section for sampling and interviewing members of a rare domain therefore have application to at least two related problems:

1. A comprehensive sampling frame exists, which contains information that permits the population to be partitioned into two or more sampling strata that vary in their density of the rare domain, *D*.
2. There are initially two (or more) sampling frames: one containing a complete list of the overall population, and one (or more) containing only a subset of the first list that is rich in members of the rare domain, *D*. By matching the second list(s) to the first, a revised sampling frame can be obtained that identifies two or more sampling strata: cases on the second list (the high-density stratum) and cases not on the second list (the low-density stratum).

The lists used to stratify the sampling frame (e.g., an age-targeted list from a market-research company or Medicaid enrollment data on likely asthma patients) are subject to error, including the telephone numbers or addresses of households that do not actually have the rare attribute (false positives), and excluding the telephone numbers or addresses of households that do have the attribute (false negatives). Due to their origin in the market-research field, some lists may be skewed towards more affluent households that have landline telephone numbers, register automobiles, and buy things on credit. As long as the entire population of *D* is covered by at least one of the lists or sampling strata, there is no bias in estimators of population parameters of interest.

## 4. Application: The National Immunization Survey

As an illustration of the method of disproportional sampling, we apply the concept of age targeting to the design of the National Immunization Survey (NIS). The NIS uses two phases of data collection to obtain information for a large national probability sample of young children: a random-digit-dialing (RDD) telephone survey designed to identify households with children between 19 and 35 months, followed by a mail survey of the vaccination providers of the children identified in the household survey (called the Provider Record Check), which obtains provider-reported vaccination histories for the children. At the close of the telephone interview the interviewer asks the respondent, the child(ren)'s parent or guardian, for consent to contact providers and for their names and addresses, and the Provider Record Check is conducted only for children for whom oral consent is given. Data from the Provider Record Check yield each child's number of doses for each of eleven vaccines. These counts are compared to the recommended number of doses for each vaccine (CDC 2010) to determine whether the child is up to date (UTD).

The NIS is designed to produce direct, sample-based estimates of *vaccination coverage rates* (UTD children as a proportion of all age-eligible children) within each of 56 estimation areas, consisting of 46 whole states, six large cities, and four rest-of-state areas (CDC 2012b). The estimation areas are the primary sampling strata in the NIS sampling design. A dual-frame RDD sampling design is used within each estimation area. The landline RDD sample has been conducted since 1994, while the cell-phone RDD sample was introduced in the fourth quarter of 2010.

The NIS deploys a new and independent RDD sample every calendar quarter. Vaccination coverage rates, $R$, are estimated using the combined sample from an annual time period. The estimator within a given estimation area is a ratio of the form $\hat{R} = \hat{Y}/\hat{X}$, where $\hat{Y} = \sum_{i \in s_c} W_i Y_i$ is an estimator of the total number of children who are UTD with respect to a given vaccine, $\hat{X} = \sum_{i \in s_c} W_i X_i$ is an estimator of the total number of age-eligible children, $s_c$ is the set of children for whom the NIS interview (including PRC) is complete within the annual time period, $Y_i$ is an indicator variable signifying whether the $i$ th child is UTD, $X_i = 1$ for age-eligible children and $= 0$ for all other units in the population, and $W_i$ is the survey weight taking into account the probability of selection, adjustments for both household and provider nonresponse, and calibration to known population counts. See the NIS Data User's Guide (CDC 2012b) for a description of the methods of weighting.

The population domain studied in the NIS is considered to be rare. In 2011 only about 18 percent of the resolved telephone numbers in the landline sample were working residential numbers and two percent of the completed screening interviews resulted in finding eligible children age 19–35 months. Given the rarity of the domain, it is reasonable to examine whether it would be possible to gain cost efficiency by using a disproportionate sampling design within high- and low-density sampling strata within each estimation area.

In what follows, we work with age-targeted lists of landline telephone numbers compiled by Marketing Systems Group (MSG) from consumer databases maintained by the marketing-research companies InfoUSA, Experian, Acxiom, and Targus. MSG and other vendors have the capability to produce lists that target various age ranges. We have conducted research for the NIS using lists targeted at ages 0–5 and 0–17 and find that both

lists yield similar results. We report the results of our investigation of the list that targets households with someone age 0–17. Because the large NIS screening sample is also used for a companion survey of American adolescents aged 13–17 years, called the NIS-Teen, we report the results of our investigation of the list that targets households with someone age 0–17. This list should support the needs of both the NIS and the NIS-Teen. However, we continue this brief illustration only for the NIS sample. Because age-targeted lists are not available for cell phones, we work only with the landline sample in this illustration.

In some applications of consumer databases in sampling rare populations, it may be possible to classify the units in the overall population into three strata: (i) in the targeted domain, (ii) not in the targeted domain, and (iii) domain status indeterminate. For the current application, however, we were only able to classify telephone numbers into two categories: on or not on the age-targeted list.

Because the NIS is an important national healthcare survey that must represent the entire population of age-eligible children to the greatest extent feasible, we use the age-targeted list for stratification purposes rather than for purposes of restricting the sampling frame. The set of all telephone numbers on the landline sampling frame that are also on the list shall be deemed the high-density stratum ($h = 1$), and the set of all other numbers on the landline sampling frame that are not on the list shall be deemed the low-density stratum ($h = 2$), with $L = 2$. We observe that some market-research surveys that target consumers in a specific age range may choose to restrict the sampling frame by selecting the sample solely from an age-targeted list. This practice saves screening costs while incurring potentially large errors of undercoverage (failing to represent persons actually in the age range but not on the list). Our approach aims to achieve both complete representation of the population and some efficiency in data collection through the use of disproportionate sampling.

We illustrate the optimum allocation in terms of the annual sample size for a single, typical estimation area. A strategy of oversampling (undersampling) the high-density (low-density) stratum will tend to result in both (i) a higher observed eligibility rate in the sample and more productive data-collection operations, and (ii) a weighting effect (due to disproportionate sampling) in the estimation of population parameters of interest and, therefore, a larger sample size to maintain variance at a fixed level. A key question before us is to what extent total data collection cost can be reduced as the net effect of these two factors, one of which tends to decrease cost while the other tends to increase it.

We determine the optimum allocation under the following ideal assumptions: (a) that there is no nonresponse in the household or provider surveys, (b) that each household in the landline population of households is connected to one and only one landline, and (c) there is at most one child aged 19–35 months in the household. If the methods cited here were used in actual practice, the sample sizes would have to be adjusted for these various factors.

The model for data collection costs is (1), where $L = 2$ and $n_h$ is the sample size of households in stratum $h$. The per-unit cost components, $t_h$, reflect numerous features of the NIS design, including the cost per telephone number for obtaining the age-targeted flag, the cost per telephone number for sample preparation and sending advance letters; the cost per telephone number for the screening interview (including both resolution of residential telephone number status and age screening); the cost per incentive given; and the cost per age-eligible household for the main interview and the PRC. The per-unit cost components

must be loaded with both the costs directly expended on completed cases and a pro-rata share of the costs of all efforts expended on unproductive cases, for example, households and providers that break off or otherwise fail to complete the survey. We have analyzed recent NIS cost data and determined that the ratio of the per-unit cost components is $t_1/t_2 = 5.1$. Thus the per-unit cost of data collection in the high-density stratum is about 5 times the per-unit cost in the low-density stratum. This result is to be expected, because, as we will show, the overall eligibility rate is much higher in the high-density stratum, and therefore this stratum requires more interviewing effort than does the low-density stratum.

The vaccination coverage rates in the high- and low-density strata are quite similar, usually differing by only one or two percentage points. Thus, given the foregoing assumptions, the variance of the estimated vaccination coverage rate, $\hat{R}$, is given approximately by (2), where $r_h$ is the overall eligibility rate within stratum $h$ (encompassing both the age-eligibility rate and the rate of working residential numbers among the resolved telephone numbers in the selected sample), $r$ is the overall eligibility rate across both of the sampling strata within the estimation area, $W_h = N_h/N$ is the proportion of landlines on the area-specific sampling frame that are classified in stratum $h$, $S^2 = R(1 - R)$ is the variance component in the domain of age-eligible children.

With the cost and variance models in hand, the optimal allocation of the total sample size to the two sampling strata within an estimation area is given by (3) and (4) and the total sample size by (5).

We estimate the overall eligibility rates and population proportions using NIS data from the third and fourth quarters of 2010 (henceforth referred to as Q3–Q4 2010). Since we actually conducted the NIS in these two quarters, we know which of the selected landline telephone numbers were associated with a household with a resident child in the eligible age range, and we have since been able to determine retrospectively which of the selected landline numbers were on the age-targeted lists in those quarters. The overall eligibility rates and population proportions are given in Table 1.

While the overall eligibility rate is not high in absolute terms in either stratum, the rate in the high-density stratum is relatively much higher than the rate in the low-density stratum. The rate in the high-density stratum is almost 14 times greater than that in the low-density stratum, and about 58 percent $= r_1W_1/r_1W_1 + r_2W_2$ of the population of age-eligible children is classified in the high-density stratum. While the statistics presented in Table 1 are at the national level, we will take them to be appropriate for calculating the optimum allocation for a single, typical estimation area.

The Centers for Disease Control and Prevention have specified that the NIS sample size in an estimation area shall be large enough so that the coefficient of variation of the estimated vaccination coverage rate is 7.5 percent when the true rate is 50 percent. Thus, we can take $V^o = 0.001406$ as the value of the fixed variance. When the true vaccination coverage rate is 0.50 (or 50 percent), the variance component for eligible children is $S^2 = R(1 - R) = 0.25$.

Plugging the foregoing parameter values into (3), (4), and (5) gives the optimum allocation to the high-density stratum, $n_1^o = 3,824$, the low-density stratum, $n_2^o = 22,875$, and the total sample size $n^o = 26,699$, which are cited in Table 2. The optimum allocation is expected to result in 320 completed interviews in the estimation area, with 223 in the high-density stratum and 97 in the low-density stratum.

*Table 1. Overall eligibility rates and population proportions at the national level: NIS Q3–Q4 2010*

| Parameter | Low-density stratum, $h = 2$ | High-density stratum, $h = 1$ | Overall landline RDD sampling frame |
|---|---|---|---|
| Eligibility Rate, $r_h$ | 0.30% | 4.10% | 0.65% |
| Proportion of the Landline RDD Sampling Frame, $W_h$ | 0.9075 | 0.0925 | 1.0000 |

The ratios of the optimum sample sizes and the optimum sampling fractions are displayed in Table 3. Optimality calls for the high-density stratum to be sampled at a rate 1.64 times the rate of sampling in the low-density stratum. While the ratio of the population sizes is about 0.10, the ratio of the sample sizes is about 0.17.

By comparison, if we were to use the sampling design that actually was used for the NIS, which is essentially a proportional-allocation design, the corresponding total sample sizes given our assumptions would be those that appear in Table 4. The same sampling precision can be achieved in two different ways: (a) use of the current design, or (b) use of the optimum-allocation design. The latter design requires about 11,266 fewer telephone numbers in the released sample, because we have oversampled the high-density stratum that has the higher eligibility rates. However, the optimum-allocation design introduces a disproportionate allocation of the completed interviews and a corresponding weighting effect, and thus it requires about 15 more completed interviews to achieve the specified level of precision.

Our methods may be contrasted to those of Srinath et al. (2004), who previously tested the use of the Experian list for improving the efficiency of NIS sampling. They determined a method of sample allocation to minimize the variance of the estimated vaccination coverage rate subject to fixed sample size, and concluded that the estimator suffers from a loss of precision due to the weighting effect. From our work in Sections 2 and 3, it is clear that the optimum allocation, which involves disproportionate sampling, requires more interviews to maintain a constant level of precision. It is also clear that optimum allocation can maintain precision while reducing data-collection costs, at least for the age-targeted lists studied here.

Plugging the expected sample sizes in Table 4 into the cost model, we find that the ratio of data-collection costs, $T^o/T^p$, is about 0.87. The optimum allocation is expected to save about 13 percent in data-collection costs relative to the current NIS design for the landline

*Table 2. Optimum allocation and expected sample sizes in a typical estimation area to minimize total cost subject to the specified variance constraint (7.5 Percent coefficient of variation for the estimated vaccination coverage rate)*[a]

| Landline RDD sample components | Low-density stratum, $h = 2$ | High-density stratum, $h = 1$ | Total landline sample size |
|---|---|---|---|
| Sample size, $n_h^o$ | 22,875 | 3,824 | 26,699 |
| Eligible households with complete NIS interview | 97 | 223 | 320 |

[a] The sampling sizes are computed using the national rates in Q3–Q4 2010.

*Table 3. Ratios of population sizes, optimum aample sizes, and sampling fractions*

| | |
|---|---|
| $W_1/W_2$ = high-density population size/low-density population size | 0.1019 |
| $n_1^o/n_2^o$ = high-density sample size/low-density sample size | 0.1672 |
| $f_1^o/f_2^o$ = high-density sampling fraction/low-density sampling fraction[a] | 1.6406 |

[a] The sampling fraction is $f_h^o = n_h^o/N_h$.

RDD sample. This percentage translates into considerable potential cost savings across 56 estimation areas per year. Most telephone surveys do not have, and thus do not bear the costs of, a second phase of data collection like the PRC. To test our methods in this more common setting, we repeated all of the calculations in this section assuming no PRC costs, and found that the resulting cost savings relative to proportional allocation amount to about 15 percent.

## 5. Summary

In this study of the use of disproportionate stratification for sampling a rare domain $D$, we made a number of assumptions, including that (a) the sampling frame covers a general population that contains both members and nonmembers of the rare domain; (b) domain membership is not known at the time of sampling; (c) the sampling design involves simple random sampling within two or more strata that vary in the density of the rare domain; (d) the parameter of interest is the mean of the rare domain; (e) the estimator of the domain mean is the standard ratio estimator; (f) classification of sampling units in or out of the rare domain based on the screening interview is conducted without error; (g) the cost of data collection arises as in (1); and (h) the variance of the ratio estimator can be represented by (2). We focused on the optimum allocation of the sample size to the several strata when one's object is to minimize the cost of data collection subject to a constraint on the variance of the ratio estimator (we also briefly treated the optimization problem when the object is to fix cost or to fix the number of interviews achieved for members of the rare domain). We find the optimum allocation to a stratum is proportional to the size of the stratum and to the root of the eligibility rate in the stratum, and is inversely proportional to the per-unit cost of data collection in the stratum. Given our assumptions, the optimum-allocation design, which oversamples the high-density stratum, introduces no bias into the

*Table 4. Expected sample sizes within a typical estimation area to achieve the specified variance constraint (7.5 percent coefficient of variation for the estimated vaccination coverage rate) for two allocation regimes*

| Landline RDD sample components | Expected sample size given current NIS design | Expected sample size given optimum-allocation design | Difference in expected sample size (current design minus optimum-allocation design) |
|---|---|---|---|
| Sample size, $n_h^o$ | 37,965 | 26,699 | 11,266 |
| Eligible households with complete NIS interview | 305 | 320 | −15 |

ratio estimator of the domain mean. Because the optimum-allocation design, by definition, minimizes the cost of data collection, it must result in non-negative cost savings relative to a proportional-allocation design. The cost savings could be small unless (a) the eligibility rates in the high-density strata are much higher than those in the low-density strata; (b) a substantial portion of the rare domain is classified in the high-density strata; and (c) the per-unit cost of the main interview is high relative to the screening cost. While the optimum-allocation design potentially saves cost, it does so through disproportionate sampling of the strata, which creates a weighting effect. Thus it actually requires more completed interviews than does the less efficient proportional-allocation design.

We illustrated the optimum-allocation design using the NIS, in which the rare domain is children 19–35 months and the parameters of interest are vaccination coverage rates for this domain. Results for the NIS are limited to the age-targeted lists obtained from the MSG vendor for the period Q3–Q4 2010.

Other surveys operating in future time periods and targeting different domains of interest should test the lists available to them. The method of disproportional stratification is broadly applicable to lists available from market-research companies as well as those derived from administrative data sources. Examples include targeted lists of people or households defined by age, race, ethnicity, income, disease registry, health insurance claims data, and voter registration status.

In deciding whether to use the optimum-allocation design, the survey statistician should be mindful of any secondary objectives for the rare-population survey, other than those embodied in the optimized objective function. For estimating other population parameters of interest, such as means for crosscutting domains, the optimum-allocation design could result in a decrease in sample size and an increase in the standard error of the estimator. These issues should be tested before the decision to implement the optimum design is taken.

## 6. References

Centers for Disease Control and Prevention (CDC). 2010. "Recommended Immunization Schedules for Persons Aged 0–18 Years – United States, 2010." *Morbidity and Mortality Weekly Report* 58 (51 & 52); 1–4. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5851a6.htm (accessed 10/13/2015).

Centers for Disease Control and Prevention (CDC). 2012a. "National, State, and Local Area Vaccination Coverage Among Children Aged 19-35 Months – United States, 2011." *Morbidity and Mortality Weekly Report* 61: 689–696. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6135a1.htm (accessed 10/13/2015).

Centers for Disease Control and Prevention (CDC). 2012b. National Immunization Survey: A User's Guide for the 2011 Public-Use Data File. Available at: http://www.cdc.gov/nchs/nis/data_files.htm

Cochran, W.G. 1977. *Sampling Techniques*, (3rd ed.). New York: John Wiley & Sons.

Kalton, G. 2009. "Methods for Oversampling Rare Subpopulations in Social Surveys." *Survey Methodology Journal* 35: 125–141.

Kalton, G., and D.W. Anderson. 1986. "Sampling Rare Populations." *Journal of the Royal Statistical Society, Series A* 149: 65–82. Doi: http://dx.doi.org/10.2307/2981886.

Pasek, J., S.M. Jang, C.L. Cobb, J.M. Dennis, and C. DiSorga. 2014. "Can Marketing Data Aid Survey Research? Examining Accuracy and Completeness in Consumer File Data." *Public Opinion Quarterly* 78: 889–916. Doi: http://dx.doi.org/10.1093/poq/nfu043.

Srinath, K.P., M.P. Battaglia, and M. Khare. 2004. *A Dual Frame Sampling Design for an RDD Survey that Screens for a Rare Population*, In Proceedings of the Survey Research Methods Section, American Statistical Association, Toronto, August 8–12, 2004. (pp. 4424–4429). Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000462.pdf (accessed 10/13/2015).

Sudman, S. 1972. "On Sampling of Very Rare Human Populations." *Journal of the American Statistical Association* 67: 335–339. Doi: http://dx.doi.org/10.1080/01621459.1972.10482383.

Valliant, R., F. Hubbard, S. Lee, and C. Chang. 2014. "Efficient Use of Commercial Lists in U.S. Household Sampling." *Journal of Survey Statistics and Methodology* 2: 182–209. Doi: http://dx.doi.org/10.1093/jssam/smu006.

Waksberg, J. 1973. *The Effect of Stratification with Differential Sampling Rates on Attributes of Subsets of the Population*. In Proceedings of the Social Statistics Section, American Statistical Association, New York City, December 27–30, 1973. (pp. 429–434). Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/srms/Proceedings/ (accessed 10/13/2015).

Received May 2013
Revised February 2015
Accepted March 2015