Where Have All the Smokers Gone?
A Simple Strategy for Targeting Rare Subgroups in RDD Samples

Mansour Fahimi (Marketing Systems Group)
Doug Currivan, Jeniffer Iriondo-Perez, and Matthew Farrelly (RTI International)

Abstract:

RDD-based studies are facing a growing number of challenges including eroding rates of coverage and response. Moreover, when the study population includes rare subgroups, surveys can require significant resources for screening. It is estimated that about 18 percent of New York adults are current smokers, defined by CDC as someone who has smoked at least 100 cigarettes and currently smokes. In order to increase the rate of identifying smokers in New York, a nonparametric segmentation technique was applied to prior survey data to identify high smoking areas. Subsequently, telephone exchanges associated with such areas were targeted for oversampling to increase the number of interviews with smokers. The employed disproportionate stratified design was implemented while managing the anticipated variance inflation due to unequal selection probabilities. This paper provides an overview of this design along with a summary of key findings.

Key words: RDD, CART, Oversampling, and Design Effect.

Contact Person:
Mansour Fahimi, Ph.D.
Vice President, Statistical Research Services
Marketing Systems Group
mfahimi@m-s-g.com
P: 240-477-8277
F: 215-653-7115

Background:

The New York Adult Tobacco Survey (NYATS) is a surveillance tool designed to monitor progress toward the New York tobacco control program goals by measuring tobacco use behaviors, attitudes, and related influences on tobacco use. Information from this survey has enabled researchers to follow relevant trends on a quarterly basis since the third quarter of 2003. RTI has overseen the administration of the quarterly NYATS on a calendar-year schedule, with approximately 2,000 telephone interviews completed each quarter since 2005.

The target population for the NYATS consists of all individuals aged 18 and older living in residential households in New York. All NYATS samples are selected based on random digit dialing (RDD) methodology using a dual-frame design that includes all 100-series telephone banks with at least one listed household and all directory-listed households in New York. These sampling procedures have been in place from the inception of the survey in 2003 through the third quarter of 2007. The enhanced design discussed here was implemented for the first time during the fourth quarter of 2007.

The NYATS follows a consistent set of best survey practices developed for the Behavioral Risk Factors Surveillance System (BRFSS) program and other reputable adult tobacco surveys. For each quarterly administration, all sample telephone numbers with matched valid addresses are sent a letter about one week before the start of data collection. The letter describes the purpose of the survey, provides an overview of the survey procedures including an incentive offer of $20 for completing the interview, and provides contact numbers in case survey participants want to obtain more information about NYATS.

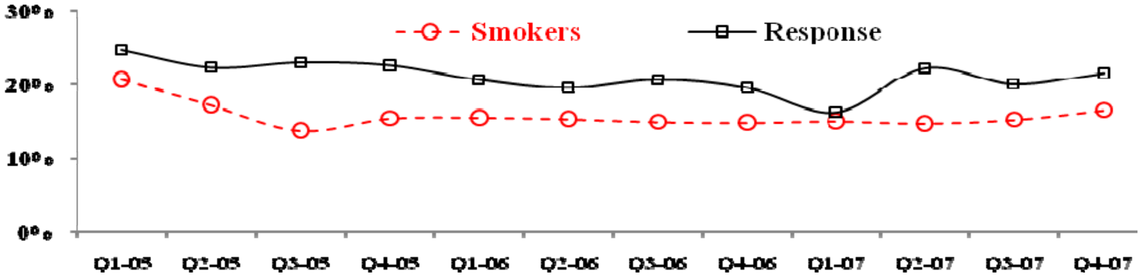
Interviews are conducted in either English or Spanish based on respondent needs and preferences. The interviews currently average about 22 minutes and are carried out in two call centers that follow identical protocols. The average interview length is about 5 minutes longer for current smokers who are asked additional questions about their smoking and quitting behaviors. About 80 percent of calls are made on weekday evenings and weekends and the other 20 percent are made on weekday days. Up to 15 calls are made to each sample number that has not yet been finalized.

Nature of the Problem:

The following figure shows the unweighted proportions of smokers and overall response rates for NYATS from 2005 to 2007. Accordingly, the proportion of smokers participating in the survey each quarter has been on a decline – with 20.7 percent in the first quarter of 2005 and nearly a 5 percentage point drop in later quarters. Despite this decline in the proportion of respondents who are smokers, overall response rates have remained relatively similar across quarters, with perhaps a slight downward trend. It has been concluded, therefore, that the significant decline in the proportion of smokers in the samples over the past three years can be attributed to:

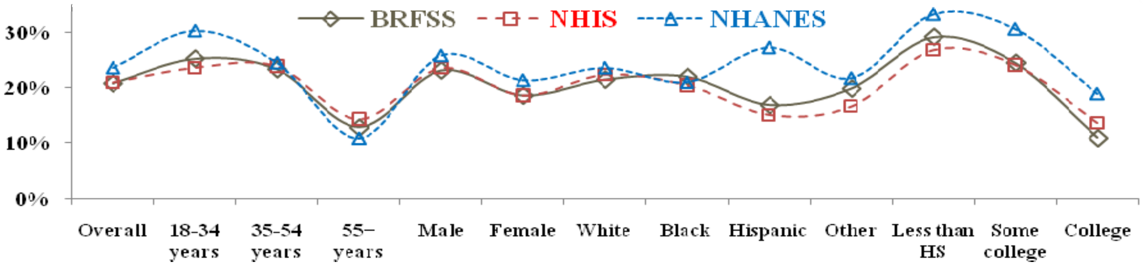
- A dramatic decline in the prevalence of smoking in New York;
- Differential decline in survey participation by smokers;
- Fewer smokers willing to identify themselves as such; or
- Decline in coverage rates for smokers.

Figure 1. Proportions of smokers in the sample and overall response rates from 2005 to 2007



According to the BRFSS survey estimates (CDC 2006) the rate of smoking in New York was 20.5 percent at the beginning of 2005 and 18.2 percent by the end of 2006. While the NYATS estimates over this same period indicate a sharper decline in smoking prevalence, it is recognized that there can always be differences in estimates obtained from different surveys. Among other factors, such discrepancies could be due to differences in sampling design, weighting procedures, mode of administration, offer of incentives, etc. For instance, the following chart shows significant differences in the national estimates of smoking prevalence among adults obtained from the 2004 BRFSS, NHIS, and NHANES (Fahimi 2008a). Nevertheless, the observed decline in smoking rates in New York cannot entirely explain the sizeable drop in survey participation rates among smokers.

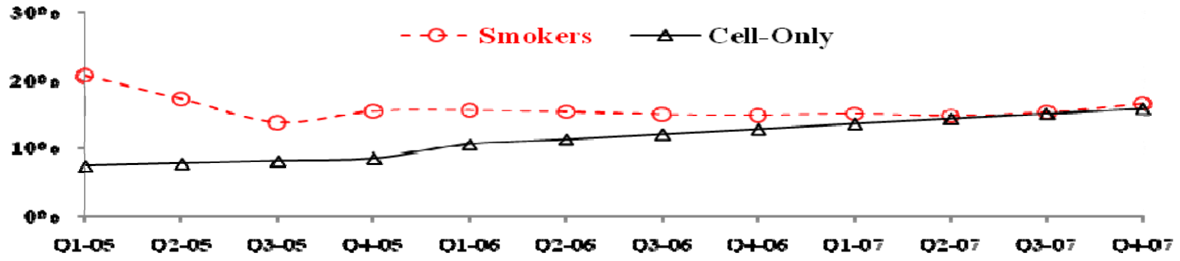
Figure 2. Estimates of smoking prevalence among adults obtained from 2004 BRFSS, NHIS, and NHANES overall and by demographic subgroups



On the other hand, it could be argued that for social desirability reasons fewer smokers are identifying themselves as such. However, without costly and intrusive studies it can be difficult to ascertain in a measurable fashion if, indeed, such emerging attitudes are partly the reason why fewer smokers are identified in the NYATS samples. After all, anti-tobacco campaigns have been in effect in New York and other states for many years now. It seems speculative that the impacts of such programs are now manifesting themselves via falsified responses to a rigorously implemented survey such as NYATS.

A more plausible explanation for the observed decline in the number of smokers in the NYATS samples could be due to the fact that the prevalence of smoking is higher among the "young-and-restless" adults. This is exactly the same flock of individuals who are steadily emerging in cell-only households – a growing coverage issue that is more pronounced in metropolitan areas such as New York. Since the NYATS samples do not include cell phone numbers, one can argue this is a major source of "leak" for smokers. The following chart shows the proportion of smokers in the sample and projected estimates for cell-only adults during the same period. Clearly, the rate of decline in the former seems to rhyme with the rate of incline in the latter. This phenomenon has been observed in other survey environments as well, such as recent political polls, causing survey estimates missing their targets by wide margins. The underlying coverage bias in landline RDD samples is no longer ignorable, not to mention the 20 percent of the households that are now not covered in traditional list-assisted RDD samples (Fahimi 2008b).

Figure 3. Proportions of smokers in the sample and projected estimates for cell-only adults[1]



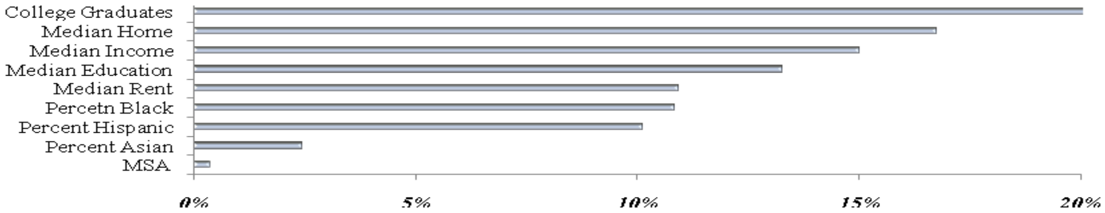
Whatever the reasons might be, the number of respondents identified as current smokers in the NYATS samples has decreased significantly during the past three years. Because smokers constitute a relatively rare subgroup, quarterly samples of NYATS have typically produced only 300 to 400 interviews with smokers. The decline in survey participation for smokers during the past three years has pushed this number down closer to 300 in most quarters. As a result, analytical possibilities for smokers or comparisons of smokers to non-smokers have been often limited by the small sample sizes of the smoker groups. For this reason, a strategy was needed to identify and oversample smokers in New York. This paper provides an overview of the methodology used to achieve this goal while keeping an eye on variance inflation due to unequal selection probabilities.

Methodology:

In order to increase the hit rate for smokers in New York, the nonparametric segmentation technique of Classification and Regression Tree (CART: http://www.salford-systems.com/cart.php) was applied to prior survey data to identify high smoking areas. The CART procedure is based on a CHAID-like segmentation methodology that automatically sifts a database to isolate patterns and relationships that could remain hidden to typical multivariate techniques due to existence of convoluted interactions or multicollinearity among variables. Starting with the many data items available for each respondent, only those that could be traced back to the RDD frame – telephone exchange characteristics – were used for this investigation. As such, beginning with a list of 35 potential predictors, 12 variables emerged relevant in distinguishing between smokers and nonsmokers - winnowing out the remaining 23 variables as having no predictive capacity. The relative importance of these potential predictors are ranked in the following figure, with the percentage of college graduates in the area being the most important and MSA status the least critical predictor.

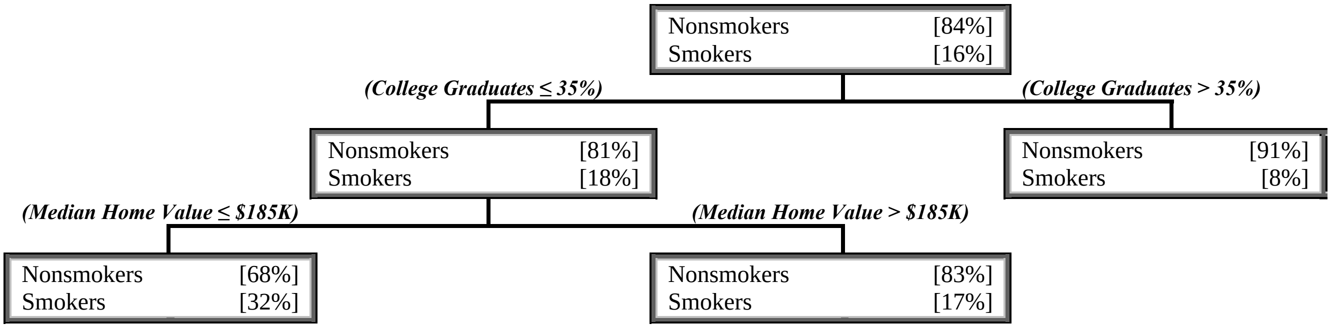
[1] These quarterly projections are developed based on the semi-annual estimates provided by the NCHS (Blumberg and Luke 2007): http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless200805.htm.

Figure 4. Relative importance of exchange characteristics for identifying smokers



In addition to calculating the relative importance of all predictors, CART conducts an exhaustive search to determine the cutoff point for each no-binary predictor that produces the *purest* separation between the two levels of the outcome variable (nonsmokers and smokers). The following diagram shows part of a decision tree that was produced when applying the CART procedure to the 2006 NYATS survey data that consisted of 1,497 nonsmokers and 278 smokers. Accordingly, the hit rate for smokers can increase from 16 percent to 32 percent if telephone exchanges serving areas where the percentage of college graduates among adults is less than 35 and the median home value is less than $85,000 are targeted. Conversely, this rate reduces to only 8 percent in areas where the percentage of college graduates among adults exceeds 35 percent.

Figure 5. Classification tree for the 2006 NYATS distinguishing between smokers and nonsmokers



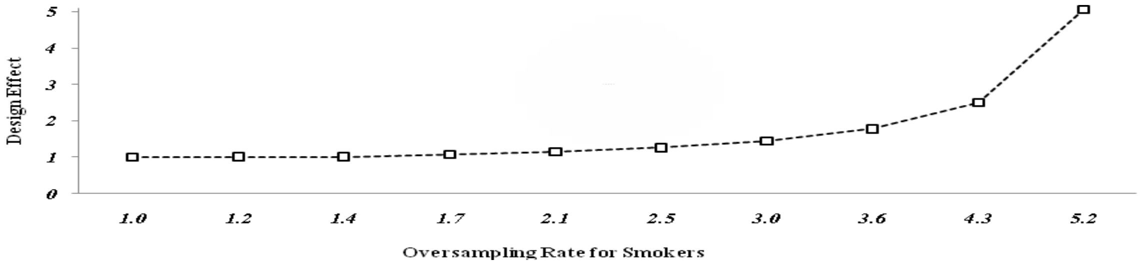
Implementation:

In order to increase the number of smokers in the NYATS sample for the last quarter of 2007, a disproportionate stratified design was employed that used higher sampling rates for telephone exchanges deemed to reach households with a higher likelihood of including smokers. Given that proportional allocation is the optimal sampling plan when overall estimates are of interest under an equal cost model, the loss in precision due to unequal sampling rates can be approximated by the following index:

$$\delta = n \times \frac{\sum n_h w_h^2}{\left(\sum n_h w_h\right)^2}$$

The above quantity, δ, provides a rough approximation for the factor by which the variance of the sample mean is increased due to unequal selection probabilities and, hence, more variability in the survey weights (Kish 1965). The following chart shows estimates of this factor, often refered to as design effect, as a function of the oversampling rate for smokers. Accordingly, it seemed advisable to keep the oversampling rate for smokers at a maximum of about 3.

Figure 6. Approximated design effect as a function of oversampling rate for smokers



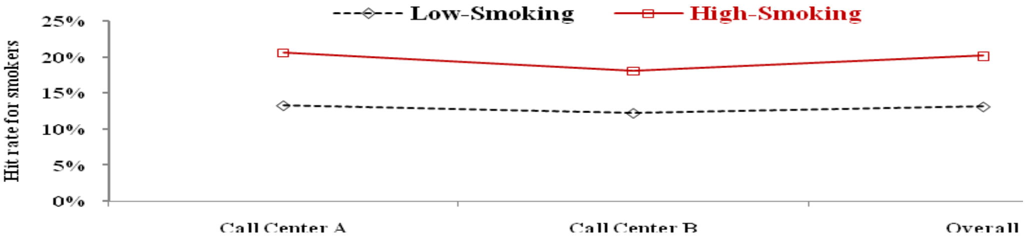
Results:

Relying on the guidelines developed from the above segmentation exercise, the sampling plan for the last quarter of the NYATS in 2007 consisted of random samples of 4,500 telephone numbers in each of the following four strata in New York[2]:

- 100-series banks within high-smoking exchanges with at least one listed household;
- 100-series banks within low-smoking exchanges with at least one listed household;
- Listed households within high-smoking exchanges; and
- Listed households within low-smoking exchanges.

As a result of the above allocation, high-smoking exchanges were oversampled at an overall rate of about 2.9 as compared to low-smoking exchanges. Consequently, upon administration of the survey it was possible to increase the hit rate for smokers to over 20 percent in the high-smoking stratum in comparison to a hit rate of only 13 percent in low-smoking stratum. As depicted in the following figure, similar results were obtained across the two call centers used for data collection.

Figure 7. Hit rates for smokers in high- and low-smoking strata by call center



It is worth noting that the above gain of over 7 percentage points in hit rate for smokers has been realized at the modest expense of an overall design effect of about 1.4. When dealing with a rare group this level of gain in productivity can result in significant cost and time savings. What is more, it would be possible to further increase the hit rates should a more aggressive oversampling strategy be employed. However, design effect projections depicted in Figure 6 suggest that points of diminishing returns approach very quickly when the oversampling rates for smokers exceed a value of about 4.

[2] All sample telephone numbers were first screened to exclude nonworking and business numbers using Marketing Systems Group's *GENESYSs-CSS* attended screening process.

Conclusions:

As mentioned at the onset, RDD studies are facing a number of serious challenges – so much so that a few researchers have questioned the future utility of RDD-based surveys in certain settings (Link 2006). What makes the situation more complicated is when the study population includes rare subgroups. The New York Adult Tobacco Surveys have been encountering declining hit rates for smokers - starting with almost 21 percent in 2005 down to less than 15 percent in 2007. This work shows that, using a simple segmentation technique, it is possible to identify telephone exchanges that are associated with higher rates of smokers. The emerging exchanges can then be targeted for oversampling purposes to increase the rates of identifying smokers for the survey.

Implemented judiciously, the above gain in productivity can be achieved without significantly reducing the precision of the resulting survey estimates; that is, the results of the employed segmentation technique can provide sampling statisticians with the information needed to develop a disproportionate stratification plan that will significantly reduce the screening requirements while managing the anticipated variance inflation due to unequal selection probabilities. Obviously, the strategy introduced in this paper is not limited to oversampling of smokers and can be applied to many surveys that focus on rare subsets of a population.

The above said, it is anticipated that one of the major reasons for the observed decline in the hit rates for smokers in NYATS samples has to do with the emergence of cell-only households (adults). As mentioned earlier, during the past three years this growing rate has increased from 5 percent to about 16 percent. Moreover, an equally sizable and growing number of households are becoming cell-mostly, resulting in 3 out of every 10 adults in the U.S. receiving all or nearly all of their calls on cell phones (Blumberg 2007). In that sense, supplementation of landline RDD samples with cell phone numbers is no longer a cosmetic enhancement but rather a virtual necessity. Given that smoking prevalence is higher among such adults, it is highly advisable for the NYATS studies to supplement their samples with cell phone numbers to reduce the existing coverage bias. This needed enhancement is not only expected to improve the overall quality of NYATS estimates, but also to help restore some of the decline in hit rates for smokers.

References:

Blumberg, S. J. and Luke, V. J. (2007). “Wireless Substitution: Early Release of Estimates from the National Health Interview Survey.” http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless200805.htm.

Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System 2006: http://apps.nccd.cdc.gov/brfss/page.asp?cat=XX&yr=2006&state=NY#XX

Fahimi, M., M. W. Link, D. Schwartz, P. Levy & A. Mokdad (2008a). “Tracking Chronic Disease and Risk Behavior Prevalence as Survey Participation Declines: Statistics from the Behavioral Risk Factor Surveillance System and Other National Surveys.” *Preventing Chronic Disease* (*PCD*), Volume 5: No. 3. (http://www.cdc.gov/pcd/issues/2008/jul/07_0097.htm)

Fahimi, M., Kulp, D. & Brick, J. M. (2008b). Bias in RDD Sampling: A 21st Century Digital World Reassessment. Presented at the *American Association for Public Opinion Research Annual Conference*, New Orleans, LA.

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

Link, M.W. and Kresnow, M. (2006). “The Future of Random-Digit-Dial Surveys for Injury Prevention and Violence Research” *American Journal of Preventive Medicine*, 31(5):444–450.