Sample designs for measuring the health of small racial/ethnic subgroups

Marc N. Elliott[1,*,†], Brian K. Finch[2,‡], David Klein[1], Sai Ma[3,§], D. Phuong Do[4], Megan K. Beckett[1], Nathan Orr[1] and Nicole Lurie[1]

[1]*RAND Corporation, Santa Monica, CA, U.S.A.*
[2]*San Diego State University, San Diego, CA, U.S.A.*
[3]*Department of Population, Family and Reproductive Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, U.S.A.*
[4]*University of Michigan, Ann Arbor, MI, U.S.A.*

SUMMARY

Most national health surveys do not permit precise measurement of the health of racial/ethnic subgroups that comprise < 1 per cent of the U.S. population. We identify three potentially promising sample design strategies for increasing the accuracy of national health estimates for a small target subgroup when used to supplement a small probability sample of that group and apply these strategies to American Indians/Alaska Natives (AI/AN) and Chinese using National Health Interview Survey data. These sample design strategies include (1) complete sampling of targets within households, (2) oversampling selected macrogeographic units, and (3) oversampling from an incomplete list frame.

Stage (1) is promising for Chinese and AI/AN; (2) works for both groups, but it would be more cost-effective for AI/AN because of their greater residential concentration; (3) is somewhat effective for groups like Chinese with viable surname lists, but not for AI/AN. Both (2) and (3) efficiently improve measurement precision when the supplement is the same size as the existing core sample, with diminishing additional returns as the supplement grows relative to the core sample, especially for (3). To avoid large design effects, the oversampled geographic areas or lists must have good coverage of the target population. To reduce costs, oversampled geographic tracts and lists must consist primarily of targets.

These techniques can be used simultaneously to substantially increase effective sample sizes (ESSs). For example, (1) and (2) in combination can be used to multiply the nominal sample size of AI/AN or Chinese by 8 and the ESS by 4. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: health disparities; multiple-frame techniques; federal surveys; National Health Interview Survey; Chinese; American Indians

---

*Correspondence to: Marc N. Elliott, RAND Corporation, Santa Monica, CA, U.S.A.
†E-mail: elliott@rand.org
‡Professor of Sociology.
§Assistant Professor.

OVERVIEW

Despite dramatic gains in life expectancy and declines in morbidity since 1900, significant dispari-ties in health and mortality persist.[‖] Eliminating disparities is a major federal policy goal; however, most national health surveys have limited to no ability to measure the health of racial/ethnic subgroups smaller than Mexican-Americans [1]. We reviewed available literature to assess the accuracy and cost-effectiveness of alternative strategies for substantially increasing the accuracy of national health estimates for small racial/ethnic groups (defined as representing less than one percent of the U.S. population). Our review identified five promising sample design techniques, and we discuss the three that were evaluable with available data. We apply these three tech-niques to two examples of very small groups: American Indians/Alaska Natives (AI/AN) and Chinese-Americans (Chinese).

IDENTIFYING POTENTIAL STRATEGIES TO IMPROVE ESTIMATES FOR SMALL RACIAL/ETHNIC GROUPS

We conducted a systematic, English-language review using search terms related to sample design. We searched standard databases, conducted a snowball survey of experts, and reviewed docu-ments from authors' personal libraries, identifying and screening 1866 titles with potential to identify sample designs that could increase the sample size and effective sample size (ESS) for national estimates of a subpopulation of interest in a cost-effective way and/or decrease the cost of identifying/contacting members of the subpopulation of interest [2].

We identified five promising methods for improving ESS: oversampling at the macrogeographic level, oversampling from an incomplete frame with substantial coverage of the population, complete sampling within households, micro-geographic/adaptive techniques, and network sampling. We discuss these approaches below; for the first three approaches, we assessed their feasibility through simulations using existing data on AI/AN and Chinese subpopulations. Available data did not support assessment of the two remaining techniques.

Each of the techniques we identified involve oversampling the target subpopulation through stratified random sampling—that is, increasing the probability that members of the target subpopu-lation are approached relative to other members of the general population. As such, all are examples of disproportionate stratified random sampling (DSRS), which involves dividing all members of the population into a set of mutually exclusive and exhaustive categories (*strata*). Individuals are randomly sampled from each stratum, so that the selection probabilities for the members of the target subpopulation are higher than for the general population. If race/ethnicity is known *a priori*, one can use *direct oversampling*, where the strata are simply the racial–ethnic subcategories. Selection probabilities vary by racial/ethnic group, but not within these groups. In contrast, if race/ethnicity is not known *a priori*, *indirect oversampling* can be used, in which the strata, which are rich in the target subpopulation, are used as proxies for the racial–ethnic subgroups.

When sampling probabilities vary, weighted analyses become necessary, which decreases ESS through a *design effect* (DEFF). With direct oversampling, this means there is a trade-off between

---

[‖]Health disparities are defined as differences in incidence, prevalence, and mortality rates between disadvantaged and advantaged groups.

increased ESS in the target subgroup (e.g. AI/AN) and a loss of ESS in the overall population. In the case of indirect oversampling, DEFFs are increased both within the subgroup and in the overall population, although to different extents. The DEFF within the target population consumes some of the apparent gain in ESS from oversampling. Still, it is usually possible with a fixed total sample size to increase ESS for a subgroup by a large proportion with only a small proportionate loss of ESS in the overall population.

Oversampling at the macrogeographic level

Oversampling geographic clusters with high proportions of the target subpopulation is a common and effective indirect oversampling technique [3]. 'Macro-level' clusters are geographic levels (e.g. Metropolitan Statistical Areas) large enough that they could serve as primary sampling units (PSUs). 'Micro' clusters include geographic entities, such as Census blocks, that are usually used to locate targets within macroclusters. The geographic concentrations of racial/ethnic groups in each of these area units determine the potential gains from geographic oversampling for each small racial/ethnic subgroup.

A feasibility study of oversampling AI/AN and Asians using the 1990 decennial Census by Westat and NCHS statisticians for the National Health Interview Survey (NHIS) 1995–2004 redesign concluded that these populations were too dispersed for oversampling to be very effective. While blocks of telephone area code/exchanges correspond to geographic areas that could be used as the units for oversampling on random digit dialing (RDD) phone surveys, it may be difficult to obtain racial–ethnic information linked to blocks of phone numbers, especially as cell phones with 'portable' numbers that lose geographic meaning replace landlines [4].

Oversampling from an incomplete frame with substantial coverage of the population

A priori overlapping frames typically combine an incomplete sample frame that inexpensively over-represents members of the desired racial/ethnic subgroup in combination with a complete (but less target-dense) frame. Typically the complete frame is a geographic sample and the incomplete sample is some list, such as the intersection of voter registration lists, telephone directories, or social security files with an indicator of race/ethnicity. For example, California Health Interview Survey (CHIS) supplemented its complete RDD frame with several incomplete frames for Asian subgroups by matching Asian subgroup surname lists with residential White Pages listings [5].

Multiple-frame techniques are a cost-effective means to improve ESS if the incomplete frame is dense in the target subpopulation and the incomplete frame is sufficiently large that it contains a sizeable proportion of the total target subpopulation. However, if the first condition does not hold, there is little potential for increasing the concentration of the target group in the sample. If the second condition does not hold, meaningful over-sampling of the incomplete frame will produce a large DEFF for the subpopulation.

Surname lists have been developed using large administrative files so that the surnames listed have high sensitivity and specificity for detecting those who self-identify as a given race/ethnicity. In the case of Hispanic ethnicity, these lists achieve about 80 per cent sensitivity and about 90 per cent positive predictive validity. In other words, about 80 per cent of all Hispanics have surnames that appear on the list, and about 90 per cent of those who appear on the list self-identify as Hispanic [6]. This performance may be approaching the limits imposed by imperfect test–retest reliability in self-identification as Hispanic [7]. Performance of surname lists is somewhat lower for married females than for other combinations of sex and marital status, as a function of intermarriage

and changing patterns in surname adoption. Similarly, researchers have developed Asian surname lists for each of several national origin subgroups, including a Chinese surname list [8]. The sensitivity of Asian subgroup lists ranging from 70 to 80 per cent for Chinese, Japanese, and Vietnamese; is somewhat lower for Koreans, and is much lower (30–40 per cent) for Asian Indians and Filipinos. Positive predictive values (PPVs) can be over 75 per cent for all groups (and over 90 per cent for Japanese). As with Hispanics, performance tends to be lower for women, especially married women. No viable surname lists for AI/AN currently exist; hence, we explore the use of the surname approach only for Chinese. Similar performance might be expected for Vietnamese, with better performance for Japanese, and lower performance for other Asian subgroups as noted above. One might consider compiled tribal rolls as an alternative listed sample approach, but it would take a large number of tribal rolls to achieve substantial coverage of the AI/AN population and the logistical ease of obtaining a number of such lists is unclear.

Complete sampling within households

Complete sampling within households involves interviewing multiple members (perhaps all) of a sampled household when the household contains members of a target racial/ethnic group, instead of the more common practice of interviewing one individual per household. The potential savings from this approach stem from the high cost of locating members of the target group compared with actual interview costs. The high homogeneity of health within households, and the resulting DEFFs are the main concerns with this approach. Household health homogeneity is measured by intra-class correlations (ICCs). Because households are fairly small clusters, ICCs that would produce large DEFFs in other contexts might be acceptable in this context.

Microgeographic/adaptive techniques

At the macrogeographic level, the density of the target population is typically known *a priori*, so cluster selection probabilities and within-cluster sample sizes can be specified according to targets. Micro-level geographic data regarding race/ethnicity is not typically available below the level of the Census block group (typically several thousand individuals); hence, any attempt to capture microclustering, to the extent that it exists, may need to rely on *adaptive* techniques. In adaptive sampling, an initial portion of the sample is used to create micro-level prevalence estimates for the targeted subgroup, which then guide the allocation of the remaining sample, capitalizing on hidden microclustering. For example, if Asian subgroup identifiers were available for Census block groups but not for blocks, one might initially sample two households per block, and only complete blocks on which one of those households contained a Chinese resident. As with all DSRS in multi-stage designs, adaptive procedures only help if the target population is concentrated at the geographic level in question. The microclustering of these subgroups is not well understood; hence, the utility of these techniques cannot yet be fully assessed without access to microgeographic data.

Network and snowball sampling

Snowball sampling has traditionally been used to sample elusive, interconnected target populations, e.g. intravenous drug users. It begins with a small sample of members of the target subpopulation, ideally obtained through random sampling, but sometimes through a convenience sample. Eligibles then successively refer other members of the target subpopulation. In a common non-probability version of this technique, eligibles are interviewed as they are accumulated and the process

terminates when a quota sample size is achieved. Such an approach over-represents those who are highly interconnected within the population and who are closely linked to the 'seed' eligibles and is not applicable to the goals of obtaining a nationally representative sample.

Network sampling can be thought of as a rigorous variation of snowball sampling [9]. It begins with a 'seed' probability sample of appreciable size ('nodes'). Each 'node' then provides 'links,' names of others with defined relationships to the node. These multiple overlapping frames are the network sample, which is typically very dense in the target subpopulation. The defined relationship must be one for which the number of network members linked to each member of the initial sample is quantifiable. One must also be able to identify when cases appear in more than one network or when a primary sample member appears in a network. Network size can be used to calculate conditional sampling probabilities (and ultimately weights) based on original sample members, which may be further adjusted for multiplicity when network cases appear more than once. Finally, unlike snowball sampling, network sampling does not involve multiple successive rounds; links do not become new nodes.

In racial/ethnic applications, one could define the network as adult siblings, parents, and children, including those who do not reside with the node respondent. The advantage of this approach is the potentially inexpensive acquisition of additional eligible sample from nodes. Disadvantages include loss of efficiency through within-network clustering (e.g. of relatives), increased DEFF from weighting, and a potential breach of anonymity given the identification of potential cases by other sample subjects. Improvements in cost per ESS are sensitive to the ICC of the outcomes used, the average network size, the accuracy and cooperativeness of network nominations, and the response rates of network members compared with those directly sampled. Both NHIS and CHIS considered using network sampling to oversample minorities but declined because of uncertainties inherent in its application and concerns related to protection of anonymity. A future field simulation would allow one to explore the utility of network sampling in national health surveys.

EMPIRICAL ANALYSIS

Using data from the NHIS, Census data, and commercial data regarding the sensitivity and specificity of listed samples, we assessed the merits of three design approaches for large national samples, such as NHIS: (1) macrogeographic sampling, (2) incomplete frames, and (3) complete sampling within households. We summarize our findings in this section. We focus on these three sampling design strategies because they have the potential to measurably increase the ESS of small racial/ethnic groups. We considered each of these techniques not as stand-alone methods, but only as supplements of Chinese or AI/AN cases to be combined with 'naturally occurring' Chinese of AI/AN cases from an existing national probability sample, such as (but not limited to) the NHIS. We use ESS as a metric that accounts for efficiency losses to compare candidate approaches with what would be obtained from a simple random sample (SRS). In particular, the ESS of a candidate approach is the sample size of an SRS that yields the same accuracy as the candidate approach. We summarize our findings below.

Macrogeographic sampling

We first considered selection of 'high-target-density' Census tracts, without oversampling. We assume weighting to produce unbiased population estimates and consider the DEFF from

weighting.$^{\parallel}$ We considered what DEFF would result from sampling Census tracts that are randomly sampled with weights determined by the density of the target group in each tract. We focus on sampling schemes with probability proportionate to the percentage of the target group and briefly consider schemes with probability proportionate to the percentage squared and percentage cubed of the target race in that tract.** In combination, we also examine restricting the supplement to Census tracts with at least a given proportion of the target group but otherwise sampling proportionately to target density; this approach is a departure from equiprobable sampling within the target race. We assumed that supplements were paired with an original sample of 200 observations from the target race, obtained as a SRS (no clustering or disproportionate probability of inclusion).††

We calculate the probability of someone from the target race being sampled from a tract with density below the 'floor' and similarly the probability from tracts above the floor. The inverse of the appropriate probability of selection formed the basis of the weights we created, which in turn are used to calculate the DEFF from disproportionate sampling within the target race. As the supplementary sample comes entirely from tracts above the floor, the probability of someone from the target race who lives in a tract below the density threshold being sampled, P_{LD}, can be calculated simply.

Let $n=200$ be the number of core sample members of the target race. Let N_{LD} be the total number of members of the target race living in low-density tracts. Let N_{HD} be the total number of members of the target race living in high-density tracts. Let $N=N_{\mathrm{LD}}+N_{\mathrm{HD}}$ be the total number of members of the target race in the population.

Then

$$P_{\mathrm{LD}}=n*N_{\mathrm{LD}}/N$$

A member of the target race who lives in a high-density tract (one above the threshold) can appear via the core sample or the supplementary sample (which consists only of high-density tracts). Thus, the probability of selection for an individual in a high-density tract is

$$P_{\mathrm{HD}}=(n*N_{\mathrm{HD}}/N+\mathrm{delta})/N_{\mathrm{HD}}$$

Again, the inverse of these probabilities became our weights from which we calculated the DEFF for various values for the floor and for delta:

$$W=1/P_{\mathrm{LD}}=N/(n*N_{\mathrm{LD}}) \text{ for observations from low-density tracts}$$

$^{\parallel}$These analyses do not consider the additional DEFF from clustering, which would be negligible if the supplement involved a large number of tracts (and relatively few per tract) but nontrivial if it involved relatively few tracts sampled.

**These analyses assumed that census tracts were of equal total population (they are actually roughly equal by census mandate), a simplifying assumption that is unlikely to greatly alter the pattern of results presented here.

††2001 standard errors for a series of dichotomous outcomes (uninsured under 65, self-rated health status of fair or poor, adult diabetes, adult flu shot, adult current smoker) for AI/AN and Chinese subgroups suggests that a typical NHIS DEFF for these groups might be about 1.5, with the exception of insurance, which varies strongly by PSU, and hence has much larger DEFFs (2–5). This in turn suggests annual effective sample sizes of about $200/1.5=133$ for these groups (based on analyses of the 2000–2002 National Health Interview person and sample adult level data (J. Lucas, unpublished data, 2005). Because this section works with ratios of supplemental sample sizes to core sample sizes, these patterns will hold, even without taking these DEFFs for the existing NHIS into effect.

and

$$W = 1/P_{HD=} = N_{HD}/(n * N_{HD}/N + \text{delta}) \text{ for observations from high-density tracts}$$

We examined the distribution of the Chinese and AI/AN population using 2000 Census tracts. Only 277 of 65 443 Census tracts are at least 25 per cent Chinese American and only 327 tracts are at least 25 per cent AI/AN. These tracts contain 18.7 per cent of the Chinese and 29.4 per cent of the AI/AN population, respectively. The eight Census tracts that are at least 75 per cent Chinese contain 1.5 per cent of the national Chinese population, whereas the 78 tracts that are at least 90 per cent AI/AN contain 12.6 per cent of the national AI/AN population.

We also compared the concentration of the target group in sampled tracts for each of AI/AN according to the sampling schemes described above, employing varying floors and sampling rates. A random sample of all tracts yields the population rate of the target group, <1 per cent in each case. This would be an undesirable scheme for a supplement, as it is not proportionate to the target group. Sampling all Census tracts proportionately to the percentage of the target group would yield a self-weighting sample of the target group, with no additional DEFF within that group, but would still require substantial screening or inclusion of non-targets, yielding average sampled tracts that would be only 13.4 per cent Chinese and 30.3 per cent AI/AN. Sampling proportionately to the square or cube of the tract's percentage of the target group oversamples members of the target group in high-density tracts, as does restricting to tracts with at least a specified floor percentage of eligibles. Sampling all tracts proportionately to the squared proportion of eligibles increases Chinese density to 35.6 per cent and AI/AN density to 74.9 per cent . Similar levels of concentration (42.8 per cent Chinese and 74.6 per cent AI/AN) can be obtained by sampling the subset of tracts of which the target group comprises at least 25 per cent proportionately to the concentration of that target group.

Table I provides detailed descriptions of the implications of proportional-to-target schemes for AI/AN and Chinese, summarizing nine designs: factorial combinations of 10, 25, and 50 per cent 'floors' (minimum target group proportion for inclusion in the supplement), with supplemental samples of 200, 400, or 800 target completes (yielding 100 per cent, 200 per cent, and 400 per cent the size of the core target sample, respectively). The table shows the number of tracts eligible for sampling at each floor threshold, the DEFF on the total sample, the ESS improvement (beyond the core 200) from the supplement accounting for DEFF, the proportion of the supplemental sample that are members of the target group, and ESS improvement assuming no screening within supplemental sample and all eligibles are interviewed.

Excluding tracts with densities less than 25 per cent adds relatively little DEFF to a proportionate-to-target sampling scheme in return for tremendous gains in screening efficiency. The resulting supplemental samples are 43 per cent Chinese and 75 per cent AI/AN, suggesting that proportionate-to-target group sampling among Census tracts with at least 25 per cent density may be a useful approach. DEFF increases with the size of the supplement: a supplement that doubles the nominal sample size for the target group multiplies the ESS by about 1.7; a supplement that quintuples its size multiplies the ESS by about 3.3. A supplement of 800 target completes is equivalent to about 460 target completes acquired without oversampling high-density tracts; this will be cost-effective compared with unrestricted proportionate-to-target sampling if the increase in density from 30.3 to 74.6 per cent for AI/AN and the increase from 13.4 to 42.8 per cent for Chinese results in a reduction in costs per target of 42.5 per cent or more.

Proportionate-to-target group sampling in 25+ per cent target Census tracts is most efficient when the target population is most concentrated, such as with AI/AN. Along with 200 AI/AN

Table I. Weights and design effects for sampling proportionally to target group density, excluding low-density tracts.

Supplemental sample target completes	Floor	AI/AN					Chinese				
		Number of tracts above the floor	Design effect	Gain in effective sample size*	Per cent of supplemental sample that are AIAN	Additional target ESS per supplement complete (per cent)	Number of tracts above floor	Design effect	Gain in effective sample size*	Per cent of supplemental sample that are AIAN	Additional target ESS per supplement complete
200	0.10	758	1.36	146.9	60.8	44.7	985	1.37	146.2	29.4	21.5
	0.25	327	1.39	143.8	74.6	53.6	277	1.42	140.5	42.8	30.0
	0.50	205	1.41	141.8	84.2	59.7	37	1.48	135.1	68.8	46.5
400	0.10	758	1.56	255.9	60.8	38.9	985	1.57	255.0	29.4	18.7
	0.25	327	1.59	251.9	74.6	47.0	277	1.61	247.9	42.8	26.5
	0.50	205	1.60	249.4	84.2	52.5	37	1.65	241.8	68.8	41.6
800	0.10	758	1.74	460.4	60.8	35.0	985	1.74	459.6	29.4	16.9
	0.25	327	1.75	456.1	74.6	42.5	277	1.77	452.1	42.8	24.2
	0.50	205	1.76	453.6	84.2	47.8	37	1.79	446.1	68.8	38.4

*Above the core 200.

completes, one would acquire 67 other (non-target) completes in the supplemental sample. About 265 other completes would accompany 200 Chinese completes from a supplement of this sort. The fact that 277 tracts are eligible for this oversampling for Chinese and 327 tracts for AI/AN means that this procedure could be implemented in a way that would not result in large DEFFs from clustering in the supplemental sample.

Incomplete frames as oversampled strata

Incomplete frames such as lists based on surnames and telephone listings, which provide an inexpensive means to reach a portion of the target population, could be used to supplement a core probability sample in a dual frame approach. If n_1 is the core probability sample size available, one could allocate cases for $n_2 = a * n_1$ completes to an incomplete listed and complete frames, with $a \geqslant 1$ of most interest. For Chinese in the NHIS, $n_1 = 200$ for a given year in recent years; hence, $n_2 = 800$ corresponds to $a = 2$, for example. The listed stratum covers a proportion q of the population, $0 < q < 1$. This means that of the n_1 observations from the complete frame (e.g. NHIS cases obtained through normal means), an average of qn_1 of them also can be expected to appear somewhere on the lists and an average of $(1-q)n_1$ observations will not appear on the list. The latter quantity is the total sample size obtained for what might be called the *unlisted stratum* (those not on this list), but the total sample size from the *listed stratum* (those who exist on the list, whether obtained from the list or via the complete frame) also includes the n_2 cases obtained directly from lists, for a total of $n_1(q+a)$ cases in the listed stratum. In this approach the unlisted are undersampled relative to the listed; they comprise the fraction $(1-q)/(1+a)$ of the sample and are sampled with a probability that is $q/(a+q)$ the probability of those who are listed. This implies design weights that do not require knowing anything about how the ways in which the listed sample is unrepresentative, but only requires that we be able to find whether individuals in our complete frame exist on the (incomplete) listed frame. Table II evaluates this approach at list coverages $q = (0.1, 0.3, 0.4, 0.6, 0.9)$ and ratios of core to list-based of $a = (1, 1.50, 2.0, 3.0, 4.0, 10.0)$, assuming a core sample of 200 observations from a complete frame.

As seen in Table II, the list coverage proportion q sets limits on the maximum gains in ESS from listed samples, a limit that increases with q. At low levels of q, additions to ESS taper quickly with increasing list-based sample size. In fact, it can be demonstrated that the maximum multiplication of ESS from the use of a listed sample with coverage q is $1/(1-q)$. Coverage $q = 0.3$, roughly corresponds to the 33 per cent reported coverage of Chinese for commercial Chinese lists. Adding 200 listed completes to 200 completes from the complete frame ($a = 1$) results in an overall DEFF of 1.54, which means that the ESS is 260 based on 400 Chinese interviews—the ESS is only 60 higher than would have been obtained from the core alone. In other words, despite doubling the nominal sample size, ESS increases by only 30 per cent. Adding 600 more listed cases to the 200 in the last example ($a = 4$, for 800 list-based completes, 1000 total) marginally increases ESS from 260 to 277 (adding only 17 in ESS for 600 more completes)—this increase is clearly not cost-effective. The maximum ESS multiplication possible at $q = 0.3$ is $1/(1-0.3) = 1.43$.

At $q = 0.4$ (40 per cent coverage), adding 200 listed ($a = 1$) produces an overall design effect of 1.42 and an ESS of 280 from 400 completes, adding 80 to the ESS and multiplying ESS by 1.40. Adding 600 more listed cases ($a = 4$, for 800 list based, 1000 total) only increases ESS from 280 to 314 (adding only 34 in ESS for 600 more completes—20 from the next 200 and 17 from the

Table II. DEFF and ESS from combining listed sample with a complete frame of $n_2 = 200$ completes.

a—ratio of list based to core (complete frame) sample sizes	q = target population coverage of listed sample	DEFF	ESS multiplication	Gain in ESS per listed case if $n_2 = 200$
1	0.2	1.67	1.20	0.20
1.5	0.2	2.06	1.21	0.14
2	0.2	2.45	1.22	0.11
3	0.2	3.25	1.23	0.08
4	0.2	4.05	1.24	0.06
10	0.2	8.84	1.24	0.02
1	0.3	1.54	1.30	0.30
1.5	0.3	1.88	1.33	0.22
2	0.3	2.22	1.35	0.18
3	0.3	2.91	1.38	0.12
4	0.3	3.60	1.39	0.10
10	0.3	7.80	1.41	0.04
1	0.4	1.43	1.40	0.40
1.5	0.4	1.71	1.46	0.31
2	0.4	2.00	1.50	0.25
3	0.4	2.59	1.55	0.18
4	0.4	3.18	1.57	0.14
10	0.4	6.77	1.63	0.06
1	0.6	1.25	1.60	0.60
1.5	0.6	1.43	1.75	0.50
2	0.6	1.62	1.86	0.43
3	0.6	2.00	2.00	0.33
4	0.6	2.39	2.09	0.27
10	0.6	4.77	2.30	0.13
1	0.9	1.05	1.90	0.90
1.5	0.9	1.09	2.29	0.86
2	0.9	1.14	2.64	0.82
3	0.9	1.23	3.25	0.75
4	0.9	1.33	3.77	0.69
10	0.9	1.92	5.74	0.47

400 after that), so is clearly still wasteful. The maximum ESS multiplication possible at $q = 0.4$ is 1.67.

At $q = 0.6$ (60 per cent coverage), if such coverage were to become achievable, adding 200 list-based completes ($a = 1$) produces an overall DEFF of only 1.25, and an ESS of 320, multiplying ESS by 1.6. Adding 600 more listed cases ($a = 4$, as before) further increases ESS from 320 to 418 (adding 98 in ESS for 600 more completes—51 from the next 200 and 47 from the 400 after that)—this increase to 800 listed cases is probably not cost effective but adding a total of 400 listed for ESS = 371 (tripling the nominal sample size and multiplying ESS by 1.855) may be reasonable.

Surname/telephone lists are not a viable approach for AI/AN because of the low rate of landline telephone coverage for this group and the absence of a viable surname list. At 30–40 per cent coverage, incomplete commercial lists based on Chinese surname lists could be used to double the nominal sample sizes for Chinese in a weighted approach that produced DEFFs of 1.42–1.54,

so that ESS was multiplied by a factor of 1.3–1.4. Greater use of the incomplete frames at this coverage level leads to rapidly increasing DEFFs and does not add much to ESS. At higher levels of coverage (e.g. 60 per cent), it might be cost-effective to double or triple the nominal sample size through the use of lists and thereby multiply ESS by 1.60–1.85. The potential of this method to increase ESS would increase greatly if there were at least 70–80 per cent coverage. At typically observed levels of coverage, this technique would only have substantial impact on the precision of estimates in combination with other techniques, and only for Chinese and other groups well identified by surnames.

Complete sampling within households

We used 1989–1994 NHIS data to calculate mean family sizes and within-family ICCs for a variety of health outcomes, both overall and by race–ethnicity. From 1989 to 1994 the NHIS collected core information for everyone in each family interviewed. However, a single individual (proxy respondent) within each family reported on the health of other members of that family.[‡‡] Therefore, some care must be taken in generalizing from these ICCs to those that would result from individual self-reports, especially for subjective items, such as self-rated health, where such ICCs may be overstated as a function of respondent effects. Nonetheless, the collection of health status of multiple household members allows us to roughly assess household homogeneity. If ICCs are inflated the DEFFs from this approach will be overestimated slightly, and the potential benefits will be slightly understated. Sample sizes for Chinese and AI/AN in the NHIS provide inadequate power to distinguish the ICC for a given outcome in those subgroups from overall ICCs; hence, we employ overall ICCs for health outcomes but racial/ethnic-specific family sizes (mean number of adults from the same family residing in the household) in calculating the DEFFs and multiplication of ESS that result from this approach for nine NHIS outcomes. Because ICCs are outcome specific, our results are also outcome specific.

Mean family sizes (limited to adults in the household) are 3.48 for Chinese and 3.40 for Asians, 23–25 per cent larger than the overall average of 2.92 for all NHIS participants. The relatively large household sizes for these groups result in larger DEFFs, but also increases the potential multiplication of ESS. Table III presents DEFFs estimated from overall ICCs and group-specific family sizes (number of adults per household). Overall ICCs are high (0.138–0.223), but relatively small cluster (household) sizes keep the DEFFs moderate for both Chinese and AI/AN (1.33–1.55) when surveying all adults in the household. The ratio of family size to DEFF (2.20–2.59) denotes the amount by which this strategy would multiply ESS. Thus, complete sampling within households for AI/AN and Chinese populations could more than double ESS for these groups, and would be cost-effective if the costs of surveying all adults in a household were at least 25–34 per cent less per complete than one per household.

Combining approaches

The techniques described can be used in combination with one another or with changes in analysis. Sample design techniques can be combined to cost-effectively increase the ESS of small racial/ethnic groups by a factor of 5 or more—a meaningful improvement in precision. For example, one could both add a geographic supplement and attempt to interview all target group

---

[‡‡]This practice ended for the NHIS in 1997.

Table III. Design effects and multiplication of ESS with full household sampling, assuming overall ICCs apply to subgroups (NHIS data 1989–1994).

		Chinese		AIAN	
	Overall ICC	DEFF	ESS multiplication	DEFF	ESS multiplication
Health outcomes					
BMI	0.207	1.512	2.301	1.496	2.273
Bed days in the last 2 weeks	0.168	1.417	2.456	1.403	2.422
Bed days in the last 12 months	0.161	1.398	2.489	1.385	2.454
Doctor visit in the last 2 weeks	0.223	1.552	2.242	1.534	2.216
Doctor visit in the last 12 months	0.198	1.490	2.335	1.474	2.305
Short-stay hospital days in last 12 months	0.150	1.372	2.537	1.360	2.500
Short-stay hospital days in last 12 months excluding delivery	0.153	1.381	2.521	1.368	2.484
Number of days in short-stay hospital in the last 12 months for discharges in past 6 months	0.138	1.342	2.594	1.330	2.554
Number of days in short-stay hospital in the last 12 months for discharges in past 6 months excluding delivery	0.141	1.350	2.577	1.339	2.539

Note: ESS Multiplication was calculated as family size divided by DEFF.

adults within households. Similarly, one could both supplement with a listed sample and attempt to interview all within target households. In these combinations, one would interview all eligible target members wherever they appeared—in the core sample or in the geographic or listed supplement. Because the DEFF of the geographic and listed supplement is largely a function of the ratio of the supplement sample size to the core sample size, the gains would be roughly multiplicative, with ESS multiplications of about four possible with 8 times the nominal sample size and multiplications of about eight possible at 15 times the nominal sample size. Combination of listed and macrogeographic approaches would be more complex and not necessarily additive. In addition, these proposed design alternatives can be paired with analytic techniques to increase ESS, such as combining data over multiple years.

CONCLUSIONS

Our literature review identified five potentially useful sample design techniques for increasing the ESS of small racial/ethnic groups in national probability samples: (1) oversampling at the macrogeographic level, (2) oversampling from an incomplete frame with substantial coverage of the population, (3) complete sampling within households, (4) micro-geographic/adaptive techniques, and (5) network sampling. We directly assessed the effectiveness of the first three of these techniques through various statistical approaches.

All three techniques efficiently improve measurement precision when used as moderately sized supplements, but experience diminishing cost-effectiveness as the size of the supplement grows relative to the base sample; this point of diminishing returns comes fastest for listed samples, which are also limited in the groups to which they apply. Complete sampling within households has potential for both Chinese and AI/AN, and is likely to work well for other Asian subgroups. Macrogeographic sampling has potential for both groups, although it would be less expensive to implement for AI/AN because of their greater concentration. Using an incomplete frame in combination with a complete frame is a somewhat effective technique for groups like Chinese with viable surname lists, but is not helpful for AI/AN. Combined approaches can meaningfully improve the precision of estimates for these groups.

In many ways the surname approach parallels the macrogeographic approach. The macrogeographic approach oversamples an incomplete geographic frame (Census tracts at least 25 per cent target) and the surname approach oversamples an incomplete nongeographic frame (commercial ethnicity lists based on surnames and phone books). Both use weights to combine the incomplete frame with a complete geographic frame such as the NHIS core or a RDD frame to create a larger probability sample in which the oversampled incomplete frame is downweighted.

Just as oversampling *tracts* require good coverage of the target population in the 25%+ tracts, commercial lists need good coverage (sensitivity) to avoid large DEffs. Just as oversampling *tracts* have to generate a supplement that is predominantly targets, commercial lists need good specificity/PPV to save screening costs.

In short, there are various sampling approaches that can be used to increase the ESS of small racial/ethnic groups in health surveys, many of which can be cost-effectively applied to both current and revised data collection projects.

REFERENCES

1. Waksberg J, Levine D, Marker D. *Assessment of Major Federal Data Sets for Analyses of Hispanic and Asian or Pacific Islander Subgroups and Native Americans*: *Inventory of Selected Existing Federal Databases*. US Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation: Washington, DC, 2000.
2. Finch BK, Morton S, Elliott MN *et al. Evaluation of Statistical Methods for Data Collection and Analysis on Racial and Ethnic Minorities and Other Hard-to-reach Populations*. Literature review prepared for the Office of Minority Health, April 2005.
3. Kish L. *Survey Sampling*. Wiley: New York, 1965.
4. Blumberg SJ. Telephone coverage and health survey estimates: evaluating the need for concern about wireless substitutions. *American Journal of Public Health* 2006; **96**:5.
5. California Health Interview Survey (CHIS), 2003 Methodology Series. *Weighting and Variance Estimation*, *Report 5*, August 2006. Available from: http://www.chis.ucla.edu/pdf/CHIS2003_method5.pdf (accessed September 2007).
6. Perkins RC. Evaluating the passel-word Spanish surname list: 1990 Decennial Census Post Enumeration Survey results. *Population Division Working Paper No. 4*, U.S. Department of Commerce (Bureau of the Census, Population Division): Washington, DC, 1993.

7. Word DL, Perkins RC. Building a Spanish surname list for the 1990's—a new approach to an old problem. *Technical Working Paper No. 13*, U.S. Department of Commerce (Bureau of the Census): Washington, DC, 1996.
8. Choi BC, Hanley AJ, Holowaty EJ, Dale D. Use of surnames to identify individuals of Chinese ancestry. *American Journal of Epidemiology* 1996; **138**(9):723–734.
9. Kalton G. Sampling considerations in research on HIV risk and illness. In *Methodological Issues in AIDS Behavioral Research*, Ostrow DG, Kessler RC (eds). Plenum Press: New York, 1993; 53–74.