Published in final edited form as:

Public Opin Q. 2009 January 1; 73(1): 56–73. doi:10.1093/poq/nfp009.

USE OF EXPERT RATINGS AS SAMPLING STRATA FOR A MORE COST-EFFECTIVE PROBABILITY SAMPLE OF A RARE POPULATION

Marc N. Elliott

RAND, 1776 Main Street, Santa Monica, California 90401, USA Phone: (310) 393-0411 Fax: (310) 451-7006 elliott@rand.org

Daniel McCaffrey

RAND, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213, USA Phone: (412) 683-2300 Fax: (412) 683-2800 danielm@rand.org

Judith Perlman

RAND, 1776 Main Street, Santa Monica, California 90401, USA Phone: (310) 393-0411 Fax: (310) 451-7006 judyp@rand.org

Grant N. Marshall

RAND, 1776 Main Street, Santa Monica, California 90401, USA Phone: (310) 393-0411 Fax: (310) 451-7006 grantm@rand.org

Katrin Hambarsoomians

RAND, 1776 Main Street, Santa Monica, California 90401, USA Phone: (310) 393-0411 Fax: (310) 451-7006 katrin@rand.org

Abstract

We consider situations in which externally observable characteristics allow experts to quickly categorize individual households as likely or unlikely to contain a member of a rare target population. This classification can form the basis of disproportionate stratified sampling such that households classified as “unlikely” are sampled at a lower rate than those classified as “likely,” thereby reducing screening costs. Design weights account for this approach and allow unbiased estimates for the target population.

We demonstrate that with sensitivity and specificity of expert classification at least 70%, and ideally at least 80%, our approach can economically increase effective sample size for a rare population. We develop heuristics for implementing this approach and demonstrate that sensitivity drives design effects and screening costs whereas specificity only drives the latter. We demonstrate that the potential gains from this approach increase as the target population becomes rarer. We further show that for most applications, unlikely strata should be sampled at 1/6 to 1/2 the rate of likely strata.

This approach was applied to a survey of Cambodian immigrants in which the 82% of households rated “unlikely” were sampled at 1/4 the rate as “likely” households, reducing screening from 9.4 to 4.0 approaches per complete. Sensitivity and specificity were 86% and 91% respectively. Weighted estimation had a design effect of 1.26 so screening costs per effective sample size were reduced 47%. We also note that in this instance, expert classification appeared to be uncorrelated with survey outcomes of interest among eligibles.

Keywords

Model-based estimation; Rare populations; Sensitivity; Specificity; Survey costs

Introduction

In the context of area-based probability sampling, screening a population for a relatively rare characteristic can be a major source of total survey costs. United States Decennial Census data is publicly available at the level of the Census Block Group (a collection of residential blocks containing about 1500 residents) but finer geographic information typically cannot be obtained. Where it can, techniques described by Waksberg et al. (Waksberg, Judkins, and Massey 1997) may increase the efficiency of the sampling approach.

Another strategy that can be used for rare or hidden populations is network sampling (Kalton 1993b; Sirken et al. 1975; Sudman and Kalton 1986; Sudman, Sirken, and Cowan 1988), a probability sampling method that begins with a traditional probability sample of “nodes” such as a systematic random sample of every tenth household on selected blocks. Eligible nodes would then be asked to supply a list of all eligible members with a defined network of known size; applications have used siblings, members of a military platoon, or other groups where the network size for a given node is unambiguous. All members of the network would then be approached for sampling. The success of this approach depends upon various factors including the cultural comfort of respondents with naming network members for a research study and the response rate effects of being approached on the basis of a referral. Failure to identify members of the network may result in biased estimation.

If distinct micro-geographic clustering exists by blocks within block groups, then adaptive techniques (Kalsbeek 2003; Kalton 1993a) in which sampling probabilities and allocations are, in part, determined recursively or interactively on the basis of eligibility information, may realize some gains in efficiency. For example, the Mitofsky-Waksberg method for sampling residential phone numbers (Waksberg 1978) uses initial sample allocations to adaptively classify blocks of phone numbers as likely or unlikely to contain eligible cases on the basis of the eligibility of sampled cases. Subsequent allocations are determined on the basis of these initial results. The efficiency gains of this implicit group classification over simple random sampling increase with the heterogeneity of eligibility within the units of aggregation, with no gains possible if the units of aggregation are homogeneous with respect to eligibility. Related adaptive approaches make inferences over geographic domains (Thompson 1991; Thompson 1996). These approaches similarly capitalize on spatial heterogeneity in eligibility.

We consider the situation in which externally observable characteristics of a household allow an expert to quickly categorize individual households as likely or unlikely to contain a member of the target population. This information supplements the available frame and can become a basis for disproportionate stratified sampling of individual households (Kalton and Anderson 1986; Kish 1965; Waksberg 1973). While excluding low probability households entirely would result in a non-probability sample of unknown and unmeasured bias under any condition other than perfect classification, sampling low probability households at a lower rate can generate a probability sample, allowing for unbiased inference through the use of design weights. We consider how the sensitivity and specificity of the classifications and the rate at which low probability households are undersampled affect the design effects and screening costs of the study. We apply this approach to a study of Cambodian immigrants to the United States (primarily refugees) who had lived in Cambodia during the Khmer Rouge era. This study (Marshall et al. 2005) required a probability sample with an effective sample size of 300–500

individuals residing in the community with the largest concentration of Cambodian refugees in the US.

This same approach might be successively applied in a number of other settings in which an expert could classify cases as likely or unlikely to be eligible with high sensitivity and specificity. The approach can be used in place of an adaptive approach when accurate expert classification is possible but heterogeneity in eligibility by geographic or other units of aggregation is low or uncertain. First, imagine that a researcher wanted a probability sample of victims of domestic abuse who had received medical treatment over the past year. Such a condition would be unlikely to be recorded very often in medical records but an expert such as a nurse trained in such issues could classify randomly sampled medical records as likely or unlikely to represent such cases. In a second example, researchers might want a multi-stage probability sample of people being treated for a new disease. After the selection of geographic regions in the first stage of sample, key informants for each area might classify providers as likely or unlikely to have such patients. In a third example, researchers might want a probability sample of secondary schools using a particular mathematics curriculum. Although curricula tend to differ by schools within districts, district office personnel would be able to classify individual schools as likely or unlikely to use the curriculum in question.

Approach

Definition of parameters

Let a represent the proportion of all households that are eligible (i.e. contain at least one member of the target group). An expert rapidly classifies each household as *likely* or *unlikely* to contain an eligible resident. Let t , the sensitivity of classification, be the proportion of eligibles who reside in likely households, which also equals the probability that an eligible household is classified as likely if the mean number of eligible residents does not differ between eligible households classified as likely or unlikely.¹ Let v , the specificity of classification, be the proportion of ineligibles who reside in unlikely households, which also equals the probability that an ineligible household is classified as unlikely under these same assumptions about residents per household. Excluding all *unlikely* households would result in a non-probability sample of unknown and unmeasured bias unless $t=1$. Sampling *unlikely* households at a lower rate than *likely* households can result in a probability sample that can produce unbiased estimates with the use of design weights in a manner that is potentially more cost-effective than a standard approach. Poor sensitivity would result in many eligibles being sampled from the unlikely households with large design weights, contributing to a large design effect for estimated parameters. Poor specificity would erode the potential savings in screening households because there would be many ineligible households classified as likely.

To avoid perverse situations that complicate analytic results but are not of interest for applications, we will only consider further cases such that households in the likely stratum are eligible at a higher rate than the overall prevalence rate. This requirement implies that $t \geq 0.5$. We will further restrict attention to cases such that $v \geq 0.5$ to exclude cases that would actually increase the amount of screening per complete.

We describe an approach in which an initial sample is selected, followed by classification of households and subsampling households on the basis of this classification. In what follows, we assume that within sampled blocks, one samples all likely households but only the proportion $1/s$ of unlikely households; in other words, any necessary subsampling would ideally take place prior to expert classification.² This stage produces design weights of 1 for

¹For simplicity of exposition, we make this assumption throughout.

eligibles within the likely stratum and $s \geq 1$ for eligibles in the unlikely stratum; initial subsampling and nonresponse will further modify these weights.

The benefits of the proposed sampling strategy, after accounting for the costs of classification, depend upon two competing factors:

1. the reduction in the number of households that need to be screened per eligible case included in the sample; versus
2. the increase in the variability of estimates due to unequal sampling probabilities for eligible cases that occurs when eligible cases reside in households in both the likely and unlikely strata, i.e., a greater design effect (DEFF) due to weighting for the resulting estimates made from the sample.

To achieve large screening cost savings, we will want to sample the unlikely stratum at a low rate (use a large s); however, a smaller sampling fraction will result in greater DEFF when eligible cases exist in both strata. The proportion of eligible households in the two strata depends on the sensitivity and specificity of classification into strata and the overall proportion of eligibles in the population. To evaluate the proposed design, we first derive a formula for the DEFF and then derive a formula for the reduction in screening costs relative to screening all sampled households ($s = 1$). We then use these formulas to consider the cost savings in the application of our approach to sampling Cambodian immigrants.

Design effects and screening rates

Assuming simple random sampling without replacement in both the likely and unlikely strata, the design effect (DEFF) for the mean from weighting equals the number of eligible responses times the sum of the squared design weights divided by the squared sum of these same weights (Kish 1965: 257–259). Thus, the DEFF can be calculated noting that: 1) we expect the number of eligible cases in the likely stratum to be proportional to at – the probability that a case is eligible and classified as likely and 2) we expect the number of cases in the unlikely stratum to be proportional to $a(1-t)/s$ —the probability that a unit is eligible and classified as unlikely. It follows that the expected number of eligible cases in the sample is proportional to $a(t + (1-t)/s)$; the sum of the squared weights is proportional to $at + a(1-t)s$; and the sum of the weights is proportional to a . From these quantities, one can derive the DEFF:

$$\text{DEFF} = 1 + \frac{t(1-t)(s-1)^2}{s}, \text{ for } 0.5 \leq t \leq 1.0, s \geq 1 \quad (\text{Equation 1})$$

Study of this equation yields several conclusions. First, the design effect, as expected, is one only when sensitivity is one ($t = 1$) or there is no undersampling of low probability units ($s = 1$). Second, near these conditions (i.e., near $s = t = 1$), variance inflation increases just less than linearly with $1-t$ and s . Variance inflation increases with both s and $1-t$ throughout their range but as s and $1-t$ increase, the variance inflation grows at a decreasing rate.

Reduction in cases screened

Let us assume that the response rate is uniformly equal to u independent of the expert classification. That implies that the expected number of approaches needed per complete response in the absence of disproportionate stratification would be $1/(au)$ and the proportionate reduction in number of screenings (PRS) needed per complete can be expressed as

²One could also sample a proportion f , $0 < f < 1$, of likely households and the fraction f/s of unlikely households and the formulas that follow would largely hold.

$$PRS = \frac{(1-a)(t+v-1)(s-1)}{1+t(s-1)} \text{ for } t, v \geq 0.5. \quad (\text{Equation 2})$$

The costs of screening with $s > 1$ relative to $s = 1$ are $1 - PRS$. The absolute reduction in cases screened would be $(1-a)(t+v-1)(s-1) / [au(1+t(s-1))]$.

Equation (2) has a number of implications. First, $t + v - 1 = t - (1 - v)$, which equals the difference in probability of an eligible household and an ineligible household being classified as likely. The proportionate reduction in screening grows linearly with specificity (v) and with the prevalence of the non-target group ($1 - a$). Savings increases less than linearly with sensitivity (t) and with the design weight for the unlikely group (s). In particular, savings are nearly linear with $1 - 1/s$. Savings are zero at $t + v = 1$ and reach their maximum possible value for a given prevalence and sampling fraction of $(1-a)(1-1/s)$ for $t = v = 1$ which corresponds to perfect classification. As $1/s$ approaches 0 and the unlikely sample approaches exclusion, the proportion of screenings saved approaches $(1-a)(1-(1-v)/t)$, its maximum value for a given sensitivity and specificity. That is, when no unlikely households are sampled, we obtain the greatest saving in screening since every likely household is expected to provide more eligible units than every unlikely household. As expected, the maximum cost savings for $v = 1$ approaches $1 - a$ as $1/s$ approaches 0, which corresponds to only sampling eligibles.

Choice of inverse sampling rate (s) and screenings per effective sample size

When choosing the sampling rate, one must balance between reducing the value of the completed cases by inflating the DEFF and reducing the cost per completed case by reducing the expected number of households that must be screened to obtain an eligible case. To make the values of these two competing factors comparable, we consider the effective sample size (ESS) of the final sample in terms of the precision of estimated means (ESS = sample size divided by the design effect) relative to the ESS of a sample with no subsampling in the unlikely stratum (a sample with $s = 1$) and screening costs relative to a design with $s = 1$. The relative ESS equals $1/DEFF$.

We can also consider the screening costs (in terms of the number of households screened) per ESS relative to the same ratio for a sample with $s = 1$ in order to produce a single summary metric of the cost per useful information in the resulting sample. If screening costs are high, with many households screened per completed case or the value of sample units is low (i.e., ESS is small because of a large DEFF), then the screening costs per ESS will be high. To interpret this ratio, we compare it for designs with $s > 1$ to a design with $s = 1$.

Screenings per ESS relative to a sample with $s = 1$ is thus equal to $(1 - PRS) \times DEFF$, or from (1) and (2):

$$SPS = \left(1 - \frac{(1-a)(t+v-1)(s-1)}{1+t(s-1)}\right) \times \left(1 + \frac{t(1-t)(s-1)^2}{s}\right) \text{ for } 0.5 \leq t \leq 1.0, s \geq 1 \quad (\text{Equation 3})$$

Application to Cambodian Example

Background

The application involved a study of the mental health of Cambodians aged 35–75 who had lived in Cambodia under the Khmer Rouge (1975–1979), immigrated to the United States prior to 1993, and lived in a defined region of the United States at the time of the interview. In-person interviews, carried out in Khmer, were conducted from October 2003 through February

2005 (see Marshall et al., 2005 for a more complete description of this study). We drew an area sample from Long Beach, California, the largest Cambodian refugee community in the United States. Within Long Beach, this area included five contiguous Census tracts containing 15,000 households. Precision requirements suggested that the study required a probability sample with an effective sample size (ESS) of 300–500 individual respondents. It was estimated that 12% of households in the defined area contained an eligible resident.

Sample design and response rate

In the first stage of a three-stage sampling process, we randomly sampled 37% of all blocks (80 of 217 Census blocks). A local community expert, at no cost to the project, then rapidly classified all 5555 individual residences contained on these blocks as *likely* (18%) or *unlikely* to contain an eligible resident. He did so using externally (publicly) observable cultural indicators that included footwear outside the front door, Buddhist altars on the front porch, and lemon grass, bamboo, or banana trees in the front yard. In a second stage, all *likely* households and one-fourth of *unlikely* households were sampled ($s = 4$). Ninety-seven percent of these households were successfully screened to determine whether they contained an eligible individual. Of these, 29% contained an eligible Cambodian, 14% contained only ineligible Cambodians, and the remainder did not contain Cambodians. In the third stage, a single eligible individual was selected at random from each household, 90% of whom agreed to participate in the survey, for a $90\% \times 97\% = 88\%$ overall response rate. Within a multivariate logistic regression, variables available for comparison (census tract, age, and gender) were not significantly associated with nonresponse ($p > .05$ in all cases); thus, no nonresponse weights were constructed. Final design weights combined inverse probability weights that accounted for the number of eligibles per household with the weights based on classification of households but their variance was primarily a function of the classification stratum.

Parameters obtained from the application

Only 12.1% of households in the defined area contained an eligible resident ($a = 0.121$). Values of sensitivity and specificity were $t = 0.864$ and $v = 0.914$, respectively. Plugging these values into (1) yields $DEFF = 1 + 0.117(s-1)^2/s$, $s \geq 1$, and with the current $s = 4$, $DEFF = 1.26$. Plugging into (2), screenings per completed interviews were reduced 57% (from 9.4 to 4.0 approaches per complete, given a response rate of $u = .88$). Thus the screening costs per ESS were reduced 47%.

Screening costs and efficiency as variations of these parameters

In this section, we illustrate the cost-effectiveness of this approach in this application as a function of (a) the sampling rate of the unlikely stratum, (b) sensitivity and specificity of classification, and (c) the prevalence of the target group. Because different designs achieve equal ESS with different mixes of screening and interviews, we later consider the relative *total* costs per ESS of the various designs and the role of (d), the ratio of screening costs (the cost of accessing household members to determine eligibility) to interview costs. We develop formulas that generalize these same relationships to other applications of this approach.

Effects of inverse sampling rate (s) at application values of other parameters

We start by examining the effect of s on the number of screenings per complete (the ESS) and their ratio (screenings per ESS), each relative to a sample with $s = 1$, fixing t , v , and a at the values observed in this application. Figure 1 presents the results for $s = 1$ to 30. As sampling fraction in the unlikely stratum falls from 1 (and s increases from 1), screening costs initially fall faster than ESS; screenings per ESS minimized for these parameters at roughly $s = 6$, but with little variation in the range of $s = 4$ to 8, before ESS begins to fall faster than screening

costs. Screening costs per ESS exceed those for $s = 1$ by $s = 30$ because the DEFF becomes very large at these extreme levels of undersampling in the unlikely stratum.³

Inverse sampling rate (s) and screenings per effective sample size, as a function of sensitivity/specificity

Table 1 examines how sensitivity and specificity affect the screenings per ESS given the value of a observed in the Cambodian application (12.1% target group prevalence). Here we vary s for 6 levels of sensitivity and specificity. Column 2 corresponds to perfect sensitivity and specificity, column 3 to near perfect values, column 4 to what was observed in the application, columns 5–7 to intermediate levels of performance, and column 8 to 50% specificity and sensitivity. We examine proportionate reductions in screenings per ESS at fixed values of s as well as noting the values of s that minimize this ratio and the values of $s > 1$ that are equivalent to $s = 1$ in this ratio.

As expected, within any given row, performance is better at higher levels of sensitivity and specificity. Relatively modest oversampling of the likely stratum does not take full advantage of the potential savings when sensitivity and specificity are very high but also does not risk inefficiency if they are low. Also as expected, the breakeven point (the value of s for which sampling at a rate $1/s$ in the unlikely stratum provides no saving in terms of screening per ESS compared to screening all households) increases with sensitivity and specificity, exceeding $s = 100$ at 95% sensitivity/specificity and falling to $s = 4$ by 69% sensitivity/specificity. Understanding the breakeven point provides some sense of what values of s might be harmful to precision if there is substantial uncertainty in the estimating sensitivity and specificity and thus in determining optimal s .

The maximum possible savings rise with sensitivity and specificity, from 2% at 58.5% sensitivity/specificity through 47% at the observed values, to 65% at 95% sensitivity/specificity. The optimal s at which these savings are achieved similarly rises from $s = 2$ through $s = 5.5$ to $s = 10$ respectively. Because gains slow near optimal s , most of the possible reduction, even at 95% sensitivity/specificity, is achieved by $s = 6$. Furthermore, even though Table 1 only considers $a = .121$, the finding that there is little gain from sampling fewer than one in six unlikely households should extend to other prevalence rates (other values of a) because DEFF is not a function of a as shown in (1) but savings are roughly linear in a as noted in (2).

Screenings per ESS as a function of target group prevalence

Table 2 examines the minimum screening cost per ESS at different levels of sensitivity and specificity as a function of the prevalence of the rare group a . In this table, optimal values of s are employed for given values of t , v , and a . We examine the observed prevalence as well as a higher level (20%) and a lower level (1%). As can be seen, rarer groups increase potential costs savings (and the optimal s will be higher). For example, with $t = 0.864$ and $v = 0.914$, the maximum savings per ESS would have been 24% at $a = 40\%$ (at $s = 3$), 40% at $a = 20\%$ (at $s = 4.5$), 47% at $a = 12.1\%$ (at $s = 5.5$), 55% at $a = 5\%$ (at $s = 7$), and 60% at $a = 1\%$ (at $s = 8$). The value of high sensitivity and specificity is greatest for the rarest target groups. At $t = v = 0.95$ and $a = 0.01$, an extreme situation in terms of the optimal value of s , savings are maximized (79%) at $s = 17.5$ but 73% savings are achieved at $s = 6$. At extremely low values a and very high sensitivity and specificity, there might be appreciable reductions in screening per ESS for values of s exceeding 6 but in most circumstances there will be little gain from increasing s above 6.

³In the limit, we sample only from the likely stratum, have an infinite design effect (and thus do not have a probability sample), and are limited by only the sensitivity in screening cost reductions.

Incorporating interview and classification costs

Thus far we have focused on variable screening costs per ESS rather than total costs per ESS. Total costs included the fixed cost of classifying the set of households for stratification and the variable cost of an interview itself (Groves 2004). Expert-guided sampling will result in fewer households being screened but will require more completed interviews than the traditional approach for any given ESS, because it increases DEFF. Thus, total costs can have different implications for design alternatives than screening costs alone. Consider the expression

$$\text{Total Cost/ESS} = \text{DEFF} \left(m c_{\text{screen}} + c_{\text{interview}} + c_{\text{classify}} / n \right) \quad (\text{Equation 4})$$

where m is the number screened per complete, n is the number of anticipated completes, $c_{\text{interview}}$ is the cost of conducting an interview with someone who has agreed to participate, c_{screen} is the cost of determining the eligibility of one sampled household through in-person screening, and c_{classify} is the cost of classifying all households as *likely* or *unlikely*. If c_{classify} is small, then for a reasonably large n , this last factor is very small or can be considered part of $c_{\text{interview}}$, so that only an estimate of $c_{\text{interview}} / c_{\text{screen}}$ is needed to optimize (4). This was the case in the present application.

The ratio of interview cost to screening costs is a function of the length of the interview, the difficulty of accessing households, and the costs of required personnel. If timelines are short or if there are a limited number of qualified interviewers (e.g., individuals with strong Khmer and English language skills), then the relatively expensive personnel used to conduct interviews must also be used for screening. Otherwise, less expensive personnel can screen and schedule appointments for more expensive interviewers. Typical values range from 2 to 4, with 4 being more typical when households are screened and scheduled, as in the present application.

We demonstrate the effect of interview costs on the choice of an optimal sampling rate for the unlikely stratum by considering total cost per ESS for a variety of values of s at $c_{\text{interview}} / c_{\text{screen}} = 1, 2, 4$, and 8, where a value of 1 is a logical minimum (the interview is the same cost as the screening) and a value of 8 would be rather extreme. We compare cost per ESS to the value for a sample of $s = 1$. Because costs depend on the number of completes and the number screened, the cost per ESS at $s > 1$ relative to the cost per ESS at $s = 1$ depends on the number of completes at $s = 1$ which requires an assumption about the response rate. We assume a constant 88 percent as was observed in the Cambodian immigrant application. Other values needed to assess ESS per cost are sensitivity, specificity, and prevalence; these also take on the values observed in the Cambodian immigrant application.

The first panel of Figure 2 plots total costs relative to $s=1$; the second panel plots total costs per ESS, also relative to $s=1$. As the cost of an interview relative to the cost of screening rises from a 1-to-1 ratio (shorter interviews/expensive screening) to an 8-to-1 ratio (longer interviews/inexpensive screening), the optimal s that minimizes total cost per ESS drops slowly from about 4 to about 2.5. The maximum cost savings relative to $s = 1$ drops from about 39% to about 15% at the same time because expert-guided sampling saves money primarily by reducing screening costs. Also note that as the interview-to-screening cost ratio rises, sensitivity to the correct value of s increases; a wide range of s values are beneficial at a 1-to-1 ratio but a much narrower range are advantageous at an 8-to-1 ratio. In the present application, at $s = 4$ and an estimated 4–1 cost ratio, ESS per cost was increased approximately 32%.

In other applications, expressions (3) and (4) could be combined to determine optimal s under various assumptions regarding sensitivity, specificity, prevalence, response rates, and cost ratios.

Evaluating the bias-variance trade-offs of the design weights

Even greater gains in ESS per fixed cost are possible if the *unlikely* eligible cases are not different from the *likely* eligible cases with respect to the parameters being estimated. In that case, one could argue for abandoning design-based inference so that weights and design effects would not apply and ESS per cost would be the inverse of relative costs. Four of the five results tables in Marshall et al. (2005) lend themselves to straightforward estimates of bias reduction through weighting. These tables estimate means and prevalences of demographic variables (Marshall et al. 2005, Table 1), prevalence of types of trauma (Table 2), prevalence of psychiatric disorders (Table 3), and the log odds of either of two major psychiatric disorders as a function of each of several specific trauma indicators (Table 5). Within each table, each parameter estimate was calculated with and without design weights. The squared difference of these estimates was used as a conservative overestimate of the squared difference in bias in unweighted estimates (inflated upwards by sampling variance). These estimates were then converted to an estimated absolute bias for each item by dividing by unweighted standard deviations. Using the observed sample size of 527 completes, each such standardized absolute bias was converted to the DEFF at which the weighted and unweighted estimates would have the same mean squared error (MSE). The standardized absolute bias and “MSE-neutral DEFF” were averaged across all parameters (rows) in a given table from Marshall et al. and are summarized in the present Table 3. The latter can be compared to the actual DEFF of 1.26 from the design weights.

The mean absolute bias ranged from <0.001 to 0.020 standard deviations across the four classes of parameters. In all cases, these estimated bias reductions were so modest that design effects of more than 1.01–1.21 would worsen the MSE of weighted estimates compared to unweighted estimates; for no set of parameters would the observed design effect of 1.26 result in weighted estimates with lower MSE than for unweighted estimates. This pattern, in turn, suggests that the expert's false negatives in classification were uncorrelated with any characteristics of interest (perhaps even completely at random) so that eligibles from the unlikely stratum may be exchangeable with likely eligibles. If design weights were not employed, then ESS per total costs would have been increased by (1/.60–1) or 67% in this application under assumptions noted above.

Discussion

In an application to a survey of Cambodian immigrants, a community expert rapidly classified 18% of households (with 86.4% sensitivity and 91.4% specificity) as likely to contain the target group which was 12.1% prevalent in the defined area of interest. The likely stratum was sampled at four times the rate as the remainder, reducing screening costs by 57% and total costs by 40%. The resulting design weights had a DEFF of 1.26 so that ESS per total cost was increased 32% under design-based inference and by 67% under what may be justifiable unweighted inference.

Classification occurred within a simple random sample of Census tracks. More complex sampling of tracks, such as a sample based on the methods of Waksberg (1973), might have yielded a greater density of Cambodian households in the sample and further reduced screening costs. However, given the very high levels of specificity and sensitivity of the expert classification, second stage stratification based on classification would still have been cost effective in this study.

Analyses suggest that in this instance, the classification strata may be uncorrelated with parameters of interest. Community experts tell us that migration within this area is very common and may well be completely at random. Classification errors may therefore not reflect distinctions between more- and less- acculturated or difficult-to-classify Cambodian

immigrants. Instead, errors may be a function of random migration that occurred between when a household was classified and when it was screened for eligibility if sampled. Given that our eligibility criteria also included age, some false positives may also be attributable to the difficulty in determining age eligibility of households containing Cambodian immigrants based on externally observable characteristics.

These patterns suggest that in this application, design weights need not (and perhaps should not) be used. In such an instance, the savings we realized in screening costs would come at no cost in design effect as long as we are willing to give up what would, under these circumstances, be a very conservative assurance of robustness to model specification through strict design-based inference. Because one cannot assume exchangeability *a priori*, one should plan for design-based inference, with the knowledge that one could have a “precision windfall” at analysis.

A limitation of our results is that we ignore the variability in estimated effects due to sampling error in the observed number of eligible cases. In addition, our proposed analyses ignore the sample design conditional on the observed sample sizes. Ignoring the sampling error in the sample size underestimates the standard error of resulting estimates; however, when estimating means or proportions the bias in the standard errors is likely to be trivially small and would not affect ESS calculations or inferences made from the sample.

Conclusions

We describe an approach that classifies individual households as likely or unlikely to be eligible and oversamples the former relative to the latter. With very good sensitivity and specificity (at least 70%, and ideally at least 80%), expert classification of households can be a cost-effective means of maximizing the ESS for a rare population. The accuracy of expert classification is critical to the success of this approach and to the selection of the rate of undersampling unlikely households. Thus, the sensitivity and specificity of expert classification should either be established via pilot testing or checked via a replicate approach in which expert classification is first used and validated on a random subset of the intended sample. Knowledge of classification accuracy would provide an opportunity to adjust parameters in later replicates.

This strategy can be used as an alternative to adaptive approaches when the heterogeneity of eligibility by units of aggregation that these techniques require is uncertain or holds only weakly. Alternatively, expert classification can be used in conjunction with adaptive techniques when the conditions for both approaches hold. For example, expert classification of an initial sample from a cluster could take the place of screening a few cases for eligibility to determine allocations and could also serve to validate the sensitivity and specificity of that classification in real time.

For expert classification, sensitivity drives design effects and screening costs whereas specificity only drives the latter. Experts should therefore be encouraged to err on the side of inclusion in their classifications. The inverse sampling parameter for the unlikely stratum (s) is a critical parameter as it determines potential gains and risks of large design effects. The approach can be inefficient relative to the reference of $s = 1$ for large values of s , especially when the sensitivity or specificity of classification is low. Moreover, even when sensitivity and specificity are very high, most of the benefit is obtained at moderate values of s . Thus, given that sensitivity and specificity are typically unknown, one should choose s conservatively unless one has great confidence in the classification from previous applications or pilot testing. For most applications, values of s from 2 to 6 would be best (sampling unlikely strata at 1/6 to 1/2 the rate of likely strata). Below $s = 2$, potential gains are small; above $s = 6$ the risk of inefficiency due to excessively large DEFF is great and the potential for additional gains from

reduced screening is small. Expert-guided sampling will be most-effective when interviews are relatively brief or inexpensive relative to screening.

Maximization of ESS per total cost should explore a range of possible values of sensitivity, specificity and s . A Bayesian approach might be particularly well-suited here with prior distributions for parameters other than s . In such an instance, one might choose s in order to maximize the 10th percentile of ESS per total cost as a means of providing a sample that can provide greater ESS given available resources while being robust to the uncertainty of the unknown parameters.

Acknowledgments

This study was supported by grants R01MH059555 from the National Institute of Mental Health and R01AA013818 from the National Institute on Alcohol Abuse and Alcoholism. Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056). The contents of the publication are solely the responsibility of the authors and do not necessarily reflect the official views of the sponsors. The authors would like to thank Kate Sommers-Dawes for assistance with the preparation of the manuscript.

References

- Groves, Robert M. Survey Errors and Survey Costs. John Wiley & Sons; New York: 2004.
- Kalsbeek, William D. Sampling Minority Groups in Health Surveys. *Statistics in Medicine* 2003;22:1527–1549. [PubMed: 12704614]
- Kalton, Graham. INT-92-P90-16E. United Nations Department for Economic and Social Information about Policy Analysis; New York: 1993a. National Household Survey Capability Programme: Sampling Rare and Elusive Populations.
- Kalton, Graham. Sampling Considerations in Research on HIV Risk and Illness. In: Ostrow, DG.; Kessler, RC., editors. *Methodological Issues in AIDS Behavioral Research*. Plenum Press; New York: 1993b.
- Kalton, Graham; Anderson, Dallas W. Sampling Rare Populations. *Journal of the Royal Statistical Society (A)* 1986;149:65–82.
- Kish, Leslie. Survey Sampling. John Wiley & Sons; New York: 1965.
- Marshall, Grant N.; Schell, Terry L.; Elliott, Marc N.; Berthold, Sarah M.; Chun, Chi-Ah. Mental Health of Cambodian Refugees 2 Decades after Resettlement in the United States. *Journal of the American Medical Association* 2005;294:571–579. [PubMed: 16077051]
- Sirken, Monroe G.; Inderfurth, Gail P.; Burnham, Clinton E.; Danchik, Kathleen M. Household Sample Surveys of Diabetes: Design Effects of Counting Rules. *American Statistical Association Proceedings of the Social Statistics Section* 1975:659–663.
- Sudman, Seymour; Kalton, Graham. New Developments in the Sampling of Special Populations. *Annual Review of Sociology* 1986;12:401–429.
- Sudman, Seymour; Sirken, Monroe G.; Cowan, Charles D. Sampling Rare and Elusive Populations. *Science* 1988;240(4855):991–996. [PubMed: 17731711]
- Thompson, Steven K. Adaptive Cluster Sampling: Designs with Primary and Secondary Units. *Biometrics* 1991;47(3):1103–1115.
- Thompson, Steven K. Adaptive Cluster Sampling Based on Order Statistics. *Environmetrics* 1996;7(2): 122–133.
- Waksberg, Joseph. Paper read at Social Statistics Section. American Statistical Association; Alexandria, VA: 1973. The Effect of Stratification with Differential Sampling Rates on Attributes of Subsets of the Population.
- Waksberg, Joseph. Sampling Methods for Random Digit Dialing. *Journal of the American Statistical Association* 1978;73(361):40–46.
- Waksberg, Joseph; Judkins, David; Massey, James T. Geographic Based Oversampling in Demographic Surveys of the United States. *Survey Methodology* 1997;23:61–71.

Fig. 1. Screenings (solid line), ESS (dashed line), and screenings per effective sample size (SPS, dotted line) for subsampling relative to sampling all household ($s=1$) as a function of the reciprocal of the sampling rate (s) for $a=0.121$, $t = 0.864$ and $v = 0.914$.

Fig. 2.

The total cost for subsampling relative to sampling all households ($s=1$) by the reciprocal of the subsampling rate (s) for different interview-to screening-cost ratios ($c_{\text{interview}}/c_{\text{screen}}$), Left Panel. The total costs per ESS for subsampling relative to sampling all households ($s=1$) by the reciprocal of the subsampling rate (s) for different interview-to-screening cost ratios ($c_{\text{interview}}/c_{\text{screen}}$), Right Panel. Both panels set $a=0.121$, $t = 0.864$, $v = 0.914$ and $u = 0.880$.

Table 1

Screenings per Effective Sample Size (ESS) for $a=.121$ Prevalence, Relative to $s=1$

	$t=v=1$	$t=0.95$	$v=0.95$	$t=0.864$	$v=0.914$	$t=v=.825$	$t=v=0.69$	$t=v=0.585$	$t=v=0.5$
$s=1.5$	0.71		0.74		0.78	0.82	0.91	0.98	1.04
$s=2$	0.56		0.61		0.67	0.74	0.89	1.02	1.13
$s=3$	0.41		0.48		0.58	0.68	0.92	1.14	1.33
$s=4$	0.34		0.42		0.54	0.67	1.00	1.29	1.56
$s=5$	0.30		0.39		0.53	0.68	1.09	1.18	1.80
$s=6$	0.27		0.37		0.53	0.71	1.18	1.63	2.04
$s=7$	0.25		0.36		0.54	0.74	1.28	1.80	2.29
At optimal s	0.12		0.35		0.53	0.67	0.89	0.98	1.00
Break-even s^1	none		101		29	14	4	1.75	none
Optimal s^2	Infinity		10		5.5	4	2	1.5	1

¹The value of $s>1$ with screenings per ESS equivalent to that for $s=1$

²The value of s that minimizes screenings per ESS

Table 2

Minimum Screenings per ESS, by Prevalence of Rare Group, by Sensitivity and Specificity, Relative to $s=1$

Proportion of population that is eligible (a)						
Sensitivity(t) and Specificity (v)	$a=0.400$	$a=0.200$	$a=0.121$	$a=0.050$	$a=0.010$	
$t=v=0.95$	0.63	0.44	0.35	0.27	0.21	
$t=0.864, v=0.914$	0.76	0.60	0.53	0.45	0.40	
$t=v=0.825$	0.83	0.73	0.67	0.62	0.59	
$t=v=0.690$	0.95	0.91	0.89	0.87	0.86	
$t=v=0.585$	0.99	0.98	0.98	0.97	0.97	

Table 3

Mean Bias and MSE-neutral DEFF at n=527 for the Cambodia Study (see Marshall et al., 2005)

Table from Marshall et al.	Parameters	Mean Absolute Bias (in standard deviations)	MSE-neutral DEFF
1	Demographics	0.005	1.02
2	Trauma Prevalence	<0.001	<1.01
3	Psychiatric Disorder Prevalence	0.020	1.21
5	Log Odds of Disorder from Trauma	0.012	1.07