SAMPLE DESIGN FOR THE 1995 SURVEY OF COMMUNITY HEALTH CENTERS

Christopher L. Moriarity, NCHS, David W. Chapman, Klemm Analysis Group
Christopher L. Moriarity, National Center for Health Statistics,
6525 Belcrest Road, Room 915, Hyattsville, MD 20782

KEY WORDS: Simulation, VPLX, Substitution

Introduction

The Bureau of Primary Health Care (BPHC), located within the Health Resources and Services Administration (HRSA), an agency of the U.S. Department of Health and Human Services (DHHS), administers several Federal grant programs that provide support for primary health services to populations that live in medically underserved areas. A directory of the health care centers that receive aid from the programs, which includes a brief summary of the programs, is given in the reference section. One such program, the Community Health Center Program, provides grant funds to health centers under Section 330 of the Public Health Service Act. Under this program, community health centers (CHCs) provide primary and preventive services to underserved populations in both urban and rural areas.

There are approximately 500 CHCs receiving Section 330 funding. The CHCs provide data on clients in the annual application for funding; these data are provided on a voluntary basis, and may include counts of users by race/ethnicity, poverty level, and insurance status. The CHCs also submit annual reports to BPHC that contain information such as the number of medical users and summary counts of users by age and sex. Data from both the application and the annual report are entered into BPHC's database, BHCDANET. The application and the annual reports provide summary-level information and summary counts of users but do not include more detailed information such as health status and services utilized. In order to collect more detailed information about the CHC clientele, both for program evaluation and for comparison to national health survey estimates, HRSA decided to conduct a sample survey of the CHCs receiving Section 330 funding.

The National Center for Health Statistics (NCHS) is providing technical support and consultation for the CHC survey, including the sample design, under an interagency agreement.

CHC Survey Overview

The CHC sample survey contains two parts: A personal interview survey of medical users, and data abstraction of a sample of medical visit records (a "visit survey"). The questionnaire for the user survey is based primarily on the National Health Interview Survey (NHIS) core questionnaire and supplements; additional questions were taken from other national surveys. The data abstraction form for the visit survey is a modification of the form used in the National Hospital Ambulatory Medical Care Survey (NHAMCS). (The NHIS and NHAMCS are surveys sponsored by NCHS.) This design feature should help to provide for a valid comparison of estimates from the CHC survey to national estimates from the NHIS, NHAMCS, and other national surveys.

The overall survey universe is the 501 CHCs in the 48 contiguous states who are currently receiving funding under Section 330 of the Public Health Service Act and were funded and fully operational during fiscal year 1992 (i.e., submitted an annual report for calendar year 1992). CHCs in Alaska and Hawaii were excluded because of the possibility of high survey cost if selected into sample. Users in the eligible CHCs who made one or more medical visits to the CHC during calendar year 1994 constitute the user survey universe. Medical visit records in the eligible CHCs during calendar year 1994 constitute the visit survey universe.

Preliminary Sample Design Decisions

Limited resources were available to HRSA and us for the CHC sample design. As the current survey is the first of its kind, no cost data or other data from previous similar surveys were available. Such data would have simplified the planning for this survey.

We and HRSA made several preliminary sample size decisions on the basis of the projected budget and data requirements for the CHC survey. We and HRSA decided that a user survey sample of approximately 2000 persons and a visit survey sample of approximately 6000 visits were likely to satisfy the primary analytic objectives of the two surveys, and that the projected budget could support these sample sizes.

We and HRSA decided that the sample design for both surveys would involve multiple selection stages, where the CHC was the first-stage sampling unit. Also, both surveys would be carried out in the same sampled CHCs. As the CHC's cooperation was required to carry out both surveys, selection of the CHC as the first-stage sampling unit made sense.

The grant applications and annual reports filed by the CHCs provided summary-level information on user

characteristics but did not provide any information about characteristics of visits. For this reason, sample design research focussed primarily on the user survey, using data from the BPHC database BHCDANET.

Given a user survey sample size of approximately 2000 and that the CHC would be the first-stage sampling unit, we and HRSA decided that the best approach would be to select as many first-stage sampling units (CHCs) as possible, while selecting enough sample cases in each first-stage sampling unit to justify the expense of including the CHC in sample. In the absence of detailed cost data on the relative costs of recruiting CHCs versus users, and intraclass correlation data that would have allowed us to estimate the optimum number of users to select from each CHC, we and HRSA decided that the minimum user sample size in a sampled CHC should be 40, thereby determining a first-stage sample size of 50 CHCs. This decision was based in part on information from NCHS facility surveys that the recruitment of facilities was likely to be extremely costly relative to the recruitment of users.

Preliminary Sample Design Research

The BHCDANET database provided a count of total medical users, users by sex by several age groups, women in prenatal care, and users by the following race/ethnicity classifications: White Nonhispanic, Black Nonhispanic, Asian Nonhispanic, American Indian Nonhispanic, Other Nonhispanic, and Hispanic.

BPHC provided us with a file containing this information and geographic information for sample design research. For several reasons, subpopulation counts by race/ethnicity were missing for some CHCs and were inconsistent with the annual report counts of total users for others. BPHC provided additional information to us that resolved some of the missing/inconsistent data issues. However, some inconsistencies remained due to data not being reported systematically, and because of differences in reference periods for data reporting. These inconsistencies placed limitations on the utility of the data in the file for our purposes. The data file is called the "sample design research file" henceforth.

The geographic data in the sample design research file were employed to explore various geographic stratification possibilities. Note that geographic stratification would help to assure a well-dispersed sample, although sorting of the sample frame by geography followed by systematic selection can accomplish much the same effect. HRSA indicated an interest in urban/rural stratification, using the urban/rural indicator included in the sample design research file. One suggested stratification explored by us was defined by crossing urban/rural with the 10 DHHS regions for a total of 20 strata; however, this resulted in several potential strata with very few CHCs. We defined a coarser set of strata by crossing urban/rural with the 4 Census Regions (Northeast, South, Midwest, West) for a total of 8 strata; this was deemed suitable in the sense that no stratum contained too few CHCs. The rural stratum in the South was large enough that we decided to split it into two pieces; one piece was defined by the South Atlantic Census Division, the other piece was defined by the remainder of the South Census Region. This gave a total of 9 geographic sampling strata as a starting point for stratification. Several subsequent changes were made before the final sampling strata were defined, as described below.

An integrated sample design approach for a survey such as this one would employ modelling of both survey cost and precision of estimates. A design would be sought that provided maximum precision for several key statistics, given the fixed cost. However, several necessary ingredients were lacking for this survey: Sufficient cost data, other quality reference data, and a prespecified set of key statistics of paramount interest.

To answer the question of expected precision, we opted for an approach involving the selection of simulated samples, followed by the production of estimates and variance estimates in a fashion that we would follow in an actual survey. We felt that this strategy was likely to produce a somewhat realistic idea of what kind of precision could be expected from the user survey, although precise modelling may have led to a better design.

Sample Simulation

We developed a software system for selecting simulated samples and computing design-based variance estimates from those samples using SAS and VPLX, a replication variance estimation software system under development by Robert Fay of the U.S. Bureau of the Census (Fay, 1990). SAS software created the simulated samples and output them to a data file for input into VPLX. VPLX was then called from within SAS to compute variance estimates. Finally, a SAS program summarized the results of the simulations. More details follow below.

HRSA expressed interest in an estimate of the reliability of estimates related to Black Nonhispanics and Hispanics. In order to draw simulated samples of these two groups, the inconsistent/missing data in the sample design research file were edited/imputed. Where inconsistencies occurred such as the sum across race/ethnicity groups not being equal to the count of total medical users (which was assumed to be correct, as per a recommendation from HRSA), the

race/ethnicity counts were ratio-adjusted so the sum would agree with the count of total medical users. Missing race/ethnicity counts were imputed using the mean value of the proportion across cases in the same grouping defined by DHHS region by urban/rural if all race/ethnicity counts were missing. If some race/ethnicity counts were present and other counts were missing, the missing counts were assumed to be zero (as per a recommendation from HRSA).

The total sample of CHCs was then allocated to the strata. Initially, sample was allocated proportional to the number of CHCs in each stratum. We also explored allocation of sample proportional to size as measured by the number of medical users in each stratum. Results from sample simulations that implemented each allocation strategy suggested to us that the latter approach provided higher quality estimates. Hence, allocation proportional to size was the method employed. This method of allocation, followed by selection of 40 users within each sample CHC, led to an "epsem" or equal probability design under the assumption that the measures of size were correct. Under this design, within each stratum, the sampling weights used for estimation would be the same if the measures of size were correct, all sample cases were eligible and responded, and no weight adjustments such as ratio adjustment to control totals were done.

We rounded the allocations to integers that summed to the total sample size as the last step in the allocation process. The algorithm we employed in the last step of the allocation process checked to see if, after rounding, the allocated number of sample CHCs was exactly equal to the target allocation of 50. In the event that the allocation was not equal to 50, sample cases were added or subtracted one by one, as appropriate (depending on whether the allocation was too low or too high), using the sampling weight of stratum members as the selection criterion. For example, if the allocation was too high, the stratum with the lowest sampling weight was identified and a sample unit was removed from that stratum, thus increasing the sampling weight of sample cases for that stratum. The sampling weight was recomputed and, if necessary, the process was repeated until the target allocation was achieved.

In practice, the initial allocation was 1 too high; a sample unit was removed from the stratum with the lowest weight (Northeast Census Region, Rural), and the weight was recomputed. In hindsight, we would have used a different algorithm, one that would have explored the consequences of removing a sample unit from each stratum and selected the stratum with the smallest increase in the weight. If we had applied this criterion, we would have removed a sample unit from the Northeast Census Region, Urban stratum.

We selected simulated samples using systematic selection with probability proportional to size. In order to get a mix of larger and smaller CHCs in each sampling stratum, the CHCs were sorted by size within stratum. This selection method helped to respond to a well-founded concern raised by HRSA early on in the sample design process, that only large CHCs would be selected if probability proportional to size sampling of CHCs was used. HRSA initially favored selection of CHCs with equal probability in order to get some smaller CHCs into the sample, although most statistics of interest from the survey are related to users rather than to CHCs.

Selection of CHCs with probability proportional to size (the number of medical users) could create certainty selections (i.e., where the sampling interval was smaller than the measure of size). Within each stratum, after computing the sampling interval, we checked to see if any CHC had a measure of size that exceeded the sampling interval. (Alternatively, an overall check for certainty cases could have been done prior to sample allocation, and this approach would be preferable in some circumstances.) We found that one CHC was large enough that it should be a certainty selection. We placed this CHC in a separate certainty stratum and reduced by one the sample allocation to the stratum that had contained the CHC that was made a certainty selection. We then recomputed the sampling interval and checked to see if any additional CHCs should be certainty selections; this turned out not to be the case. Note that an alternative approach to designating certainty selections would have been to allow multiple "hits" of very large CHCs, but this could have led to large user survey interview workloads in the CHCs with multiple hits. Also, the number of CHCs could have been reduced if the multiple hit strategy had been used, with a corresponding reduction in the precision of variance estimates.

Once a CHC was selected into a simulated sample, a simulated sample of users needed to be drawn. This raised several issues. First, some CHCs contained multiple clinic sites. Pretest results suggested that CHCs usually (but not always) had a central listing of users that could be used for sample selection. Our simulation work assumed such a selection strategy was feasible although this approach has not always been used in the conduct of the survey. Note that it would have been impossible to realistically simulate the selection of clinic sites, followed by a sample of users, because we had an incomplete picture of the number of sites within each CHC and we had no information about the number and characteristics of the clientele of each site.

A second issue was related to the notion of

proportional representation in the sample of various groups. Proportional representation for at least some groups is achievable if the CHC's users are sorted or stratified by these groups prior to sample selection; otherwise, proportional representation will occur on the average but there is no guarantee of proportional representation in any particular sample. We assumed in our simulation work that groups such as Black Nonhispanics and Hispanics were represented in a CHC's sample in essentially the same proportion as was indicated by the sample design research file for the CHC.

Once the simulated sample of users within the simulated sample of CHCs had been selected, the input file for input to VPLX was prepared and VPLX was used to get estimates and sampling error estimates. We describe this in the next section, which also includes an overview of VPLX.

Use of VPLX to Compute Sampling Errors

VPLX is a software system designed to produce design-based standard errors, covariances, etc. using replication methods. If the input file contains stratum and cluster (primary sampling unit) identifiers, the default method used by VPLX is the stratified jackknife method (see Equation 4.5.6 in Wolter (1985), page 179). VPLX forms each replicate by leaving out one cluster and reweighting the other clusters in the stratum with the omitted cluster. The input file to VPLX must be sorted by cluster within stratum for the stratified jackknife method. No replicate weights are required on the input file; VPLX will automatically compute replicate weights if none are supplied by the user.

VPLX currently accepts "flat" data files as input. (More advanced features of VPLX permit more complicated input data file structures.) A FORTRAN format statement that is an integral part of the VPLX program describes the record layout and variable names to VPLX.

VPLX is set up in several modules, each designed for a specific purpose. Data are read into a special VPLX file in a "CREATE" step. During this step, VPLX sets up sums of squares and crossproducts in the special VPLX file in preparation for later calculations. In the simplest use of VPLX, a "DISPLAY" step follows the CREATE step, where user-specified statistics such as totals, means, proportions, etc. and their standard errors are produced from the special VPLX file. Another module called the "TRANSFORM" step allows the user to create statistics (of arbitrary complexity) that are functions of arithmetic operations (add, subtract, multiply, divide) applied to totals, means, proportions, etc. (A ratio statistic is a simple example of a statistic that can be defined in a TRANSFORM step.) If such statistics are desired by the user, the TRANSFORM step would be applied one or more times to the special VPLX file after a CREATE step and prior to a DISPLAY step.

Another useful feature of VPLX is the INCLUDE statement. This statement can appear anywhere in a VPLX program and allows the incorporation of VPLX statements and/or data contained in the external file pointed to by the INCLUDE statement. This feature allowed us to use one basic VPLX program for our simulations, making necessary changes by creating external files using the SAS PUT statement and pointing to these files using the VPLX INCLUDE statement.

The X command in SAS allowed us to call VPLX from within SAS, thereby enabling us to control the entire simulation procedure using SAS macro variables (a SAS macro do-loop).

In our implementation, each data file record was a summary at the CHC level, with a weight, total sample size, and various counts. Note that we would obtain an identical result if the input file contained stratum and cluster identifiers and one record for each sample person; in the CREATE step, VPLX creates cluster totals. We used the SAS PUT statement to create the data file, along with a file containing stratum finite population correction factors and other files that were pointed to by VPLX INCLUDE statements.

We defined the weight as the product of the reciprocals of selection probabilities at the first and second stage. Since the first stage of selection was probability proportional to size and the second stage always selected the same number of sample persons, the weights were equal within a given stratum, as mentioned above. As noted above, minor weight differences occurred across strata because of the requirement to sample an integer number of CHCs from each stratum.

Results of Simulations

The simulations indicated that estimates of totals and proportions for groups such as Black Nonhispanics, Hispanics, and women in prenatal care would have acceptable accuracy; i.e., oversampling was not necessary. (A coefficient of variation (CV) of less than 30 percent was deemed to be "acceptable", although a lower CV would be preferred, of course.) One other group of major interest, managed care users, was more of a problem.

The sample design research file did not contain information about managed care users. BPHC was able to provide some information about the number of managed care users in CHCs, with caveats about data accuracy and the possibility that the number of CHCs with managed care users may have increased since the

counts of managed care users were last compiled.

We ran simulations with the managed care user data and became concerned about the quality of survey estimates for that group. The simulation estimates varied widely, depending on whether certain CHCs with large proportions of managed care users were included in sample. After some data examination and some simulations involving various criteria, we proposed to HRSA that a separate "managed care" stratum of CHCs be formed consisting of those CHCs with more than 35% managed care users. Sampling simulations suggested that the formation of this stratum helped to stabilize estimates related to managed care users while not having much of an adverse impact on estimates related to Black Nonhispanics, Hispanics, and women in prenatal care.

Summary of Sample Design

Eleven sampling strata:

1. "Managed care"
2. Northeast Census Region, Rural
3. Northeast Census Region, Urban
4. South Atlantic Census Division, Rural
5. Remainder of South Census Region, Rural
6. South Census Region, Urban
7. Midwest Census Region, Rural
8. Midwest Census Region, Urban
9. West Census Region, Rural
10. West Census Region, Urban
11. Certainty Case

Some summary statistics and the sample allocation for the strata follow in Table 1, in the same order as is given above. "_FREQ_" is the number of CHCs in the sampling stratum. "NEWUSER" is the total count of medical users in the sampling stratum, after subtraction of a small number of users considered out of scope for the user survey. "MEAN" and "STD" indicate the average size and variation, respectively, of users per CHC in the stratum. Note that the average number of users in the urban strata is higher than the average number of users in the rural strata, and this resulted in a higher sampling rate (of CHCs) in urban strata. "NUMSAMP" is the sample allocation to the stratum. (Each noncertainty stratum was required to have at least 2 sample CHCs selected to allow for variance estimation; we would have collapsed strata as necessary to meet this requirement.) "SAMPINT" is the sampling interval used in the systematic sampling process. Recall that the variation in the sampling interval from stratum to stratum, and hence the stratum sampling weight ("WEIGHT"), was due to the requirement that an integer number of sample CHCs be selected from each sampling stratum. "WEIGHT" was computed under the assumptions that the CHC measures of size were correct and 40 users would be selected from each sample CHC.

Table 1:

FREQ	NEWUSER	MEAN	STD	NUMSAMP	SAMPINT	WEIGHT
19	273044	14371	9210	3	91014.7	2275.4
32	272972	8530	6293	2	136486.0	3412.2
63	875720	13900	13302	9	97302.2	2432.6
91	619654	6809	5085	6	103275.7	2581.9
70	528646	7552	4871	5	105729.2	2643.2
53	717775	13543	10055	7	102539.3	2563.5
42	308528	7346	6383	3	102842.7	2571.1
44	512855	11656	11197	5	102571.0	2564.3
57	463062	8124	6857	5	92612.4	2315.3
29	382037	16454	19872	4	95509.3	2387.7
1	111569	111569	0	1	111569.0	2789.2
501	5065862			50		

Sample Selection - Allowing for Substitution

HRSA requested that we select the sample in a way that would allow for substitution of a CHC who refused to participate in the survey. We agreed to this request with two conditions:

1. Substitution would be allowed only after extensive refusal conversion attempts had failed (i.e., that attempts to convert refusals would be as exhaustive as they would if substitutions were not allowed).

2. If substitution occurred, key survey estimates must be computed both with and without substitute CHCs. In the event that important differences are found, a decision must be made at the beginning of data analysis regarding whether to exclude the substitute data from all subsequent analyses.

Pros and cons of the use of substitutes are described in Chapman (1983).

We followed a relatively simple method of defining substitutes. Within a stratum, starting at the largest CHC, we paired it with the next largest CHC, and

defined a combined measure of size as the sum across the paired CHCs. In the event that the pairing would result in a certainty selection, the pairing was not allowed and the largest CHC had no substitute created. (Similarly, the certainty selection described above had no substitute.) The pairing proceeded in this fashion to the smallest CHCs in the stratum, where if necessary a combination of 3 CHCs was allowed so the smallest CHC was not "alone". During the sample selection process, groups were selected with probability proportional to group size, followed by random selection for the survey of one CHC within the group with probability proportional to CHC size. The CHC not selected in each pair would serve as the substitute for the CHC that was selected, if needed. If the substitute CHC refused, no other substitutes would be allowed.

Outcome of Sampling Process - Refusals

Two CHC refusals occurred (one in the Northeast Census Region, Urban stratum, the other in the South Atlantic Census Division, Rural stratum). In both cases, the substitute also refused. In both cases, we will make a nonresponse adjustment in the weights of the other cases in the sampling strata where the refusals occurred. Hence, the issue of what to do with data from substitute CHCs will not be an issue. As mentioned above, if data were collected from one or more substitute CHCs, we would have begun the estimation phase by examining key statistics resulting from the inclusion of the substitute versus not including the substitute coupled with a nonresponse adjustment. In the event that important differences would have been found, we would have had to decide whether to exclude the substitute data from all subsequent analyses.

Summary

We chose simulation as an aid in the sample design process for several reasons. We felt that since simulation mimics the actual process of producing estimates and corresponding variance estimates, we felt that we could make a realistic assessment of the sampling errors we could expect to get from the CHC user survey. Additionally, the prior information available to us did not appear to be of sufficient reliability and detail to construct a good model for cost and/or precision.

The flexibility of VPLX, coupled with several features of SAS, made it possible for us to create an integrated package of simulation software that would run from start to finish after specification of a few parameters such as the number of simulated samples to select. We usually set this parameter to 10 - a Sun SPARC 10 workstation at NCHS running under Sun OS 4.1.3 usually could complete 10 simulations in less than an hour.

Our simulation approach included the assumption of proportional representation of groups in the sample. The simulation methodology can be improved by relaxing this assumption, unless the sampling process guarantees proportional representation.

An additional limitation in our simulation work is the assumption of 100% response and 100% complete and consistent data, that is, no nonrespondents and no imputation for missing/inconsistent responses. Random mechanisms can be built into the simulation process to simulate both unit and item nonresponse; this was a refinement that we did not make, but would be advisable for surveys where problems with response are expected.

Our procedure for forming substitutes was simple but naive. Grouping CHCs by similar size is better than an arbitrary pairing but not as good as pairing on additional factors such as geography (e.g., same state) and clientele characteristics.

Acknowledgement

The authors would like to thank Patricia N. Royston of the Health Resources and Services Administration for her assistance in the preparation of this paper.

References

Chapman, D.W. (1983). "The Impact of Substitution on Survey Estimates", Incomplete Data in Sample Surveys, Volume 2, Madow, Olkin, Rubin (eds.), 45-61.

Fay, R. (1990). "VPLX: Variance Estimates for Complex Samples", Proceedings of the Section on Survey Research Methods, American Statistical Association, 266-271.

U.S. Department of Health and Human Services (1993). BPHC-Supported Primary Care Centers Directory, April 1993, Health Resources and Services Administration, National Clearinghouse for Primary Care Information.

Wolter, K. (1985). Introduction to Variance Estimation, Springer-Verlag.