

## **Supporting Statement: National Sample Survey of Registered Nurses**

### **B. Statistical Methods**

#### **1. Respondent Universe and Sampling Methods Response Rates**

The primary target populations of interest for the 2008 NSSRN are the currently-licensed registered nurses (RNs) in the United States and the subpopulations living in each of the 50 States and the District of Columbia. The previous NSSRN conducted in 2004 estimated that there were roughly 2.9 million currently-licensed RNs at the time of the survey. Separate sample frames will be constructed for each of the 50 States and D.C. from listings of the currently-licensed RNs provided by the corresponding State Boards, eliminating any records deemed to be ineligible for the study prior to sampling.

The overall sample size for the United States as a whole is to be approximately 54,000 RNs, about the same number fielded for the 2004 NSSRN. An important sample allocation criterion used for the 2004 NSSRN was to obtain an effective sample size of RNs working in each State of at least 400 (the choice of 400 was made in order to achieve a half-width of a confidence interval of 5 percent for an estimate of 50 percent—that is, the resulting confidence interval would be from 45 percent to 55 percent). This objective was not achieved for 9 of the 51 governmental entities sampled from (the 50 States plus D.C.) while another 13 had an effective number of those employed as an RN below 500. A reduction in the overall sample size would mean allocating even more of the total sample to the smaller States, which are already oversampled at a national level, adding to the design effect for national estimates and reducing the precision for the larger States from what would have been obtained with a sample allocation closer to proportionate. Thus, it is prudent to sample at least as many RNs in 2008 as in 2004.

State-specific allocation strategies will be developed to help achieve desired precision levels for important state level estimates. Note that a State's RN workforce comes from both the RNs living in the State and from those living in neighboring States who commute to the State in question for purposes of employment, traveling nurses and

those who provide Telehealth consulting in the State in question. The 2008 design readily permits allowance to be made for the contributions made from the independent samples of RNs selected from neighboring and other States to a particular State's RN workforce. The sample design, therefore, involves taking account of both the increase in actual sample sizes and the impact on design effects arising from using different sampling rates in the different States. Another consideration is that some of the larger States were undersampled, for national estimates, in 2004. In the allocation process for the 2008 sample, steps will be taken to limit the effect on the precision of estimates for such States (due to reduced sample sizes in these States that will be necessary to help meet the effective sample size target of 400 working RNs per State) and national estimates (with increased design effects arising from the continuing "undersampling" of larger States).

The new sample design has no clustering component, so some gains in precision will arise compared to the 2004 survey. In the 2004 NSSRN, alphabetic clustering of names was used for sampling purposes. As a result, for some estimates this clustering made a non-trivial contribution to the sample variance. This is particularly true for estimates associated with the race/ethnicity of the RN.

The 2004 NSSRN obtained a response rate of 70.47 percent. For 2008, the Health Resources and Services Administration (HRSA) is setting a target of 80 percent. To the extent that HRSA can achieve this rate, it will serve to increase effective sample sizes at both the State and national levels.

## **2a. Sampling and Estimation Methodologies**

Systematic random samples of RNs will be selected from each of the 51 sample frames (the 50 States plus D.C.). Some RNs are licensed in multiple States and thus will have multiple chances of selection for the NSSRN. Hence, To identify RNs licensed in multiple states, probabilistic matching of records will be employed using software designed to indicate which records found in different files are highly likely to represent the same RN will be employed. Such matching will be employed both for records that reflect these multiple licenses will be cross-matched for both sampling and estimation

purposes. A detailed description of the matching strategies is provided below in the context of its use for estimation purposes.

For sampling, HRSA will employ cross-matching of match records found on different Sstate sample frames to the extent feasible; focusing on those States where a non-negligible contribution to one State's RN workforce is made by RNs residing in another State and licensed in both States, as discussed above. Some State frames will be partitioned into explicit strata reflecting those licensed in neighboring States as well, so that sampling rates for such RNs with multiple licenses can be determined to help control design effects arising from differential sampling rates. Examples of groups of such Sstates are New York, New Jersey, and Pennsylvania, as well as D.C., Maryland, and Virginia. Explicit strata will also be formed on each Sstate frame for those RNs whose Sstate of residence is Alaska, Hawaii, and New Mexico. Residents of these three Sstates are to be sampled at the same rate as RNs licensed in those Sstates to help reduce the variability of the weights that was experienced in the 2004 NSSRN. RNs who work in federal facilities in a Sstate need not be licensed in that Sstate, and this appears to have been a contributing factor to the high variability of the 2004 NSSRN weights for these three Sstates.

For estimation purposes, in order to determine the overall probability of an RN licensed in the United States of being sampled for participation in the 2008 NSSRN, HRSA will also employ probabilistic matching software. Each State sample will be matched to the sample frame of RNs established for each of the other States. Name and address will be available for matching for all States. Most States will also provide other variables such as date of birth and gender. And some may provide the Social Security number (SSN). (When provided, the SSN will be used only for matching purposes and never released as part of a sample data file.) Clearly, the SSN is the preferred matching variable since it is unique across State lines and provides an opportunity for accurate exact matching. When the SSN is not available, HRSA will implement alternate strategies for obtaining as high a matching rate as possible using available variables. Date of birth, when available, also provides an opportunity to have an accurate exact match. Those matching on date of birth can then be further matched using either full or

partial fields related to name and address to get highly accurate matches that identify most all of the RNs actually associated with the two States in question.

Whenever a sampled RN is found on the sample frame for a State, a flag will be set for that RN's record corresponding to that State. Thus, the probability of selection of an RN can be computed to reflect the RN's multiple opportunities for selection taking into account all the State frames on which a sampled RN appears. For example, in the simplest case, if registered nurse  $i$  ( $RN_i$ ) were sampled from State A and was also found on State B's list but not on the list of any other State, then the probability of selection  $p_i$  of  $RN_i$  would be computed as

$$p_i = 1 - (1 - p_{Ai})(1 - p_{Bi})$$

where

$p_{Ai}$  = the probability of selection of  $RN_i$  from State A's list, and

$p_{Bi}$  = the probability of selection of  $RN_i$  from State B's list

Respondents will be asked to verify in which States they are licensed. Their responses will serve as a quality control check on the matching effort. If the matching effort happened to miss a State, the probability of selection will be updated with the new information. If a State characterized as having a matching record is not listed, HRSA will investigate the match. Again, the probability of selection can be updated if needed. However, it may be that the responding RN has omitted a State where he or she appears on the sample frame. In that case, the probability of selection will be maintained.

HRSA will monitor the proportion of records requiring each type of update plus the percentage of records where a questionnaire response did not provide complete information on the number of sample frames on which RNs were found. This will help assess the degree to which the sample weights of the nonrespondents (for whom a quality control check is not possible) may not fully reflect the overall chance of selection of a nonrespondent. If HRSA notices any patterns in the types of errors made, then the matching algorithm will be updated to better identify the States in which the

nonrespondents are found and re-implement the matching algorithm prior to finalizing the sample weights.

~~A systematic sample from a frame sorted on key variable(s) of interest helps to achieve a proportionate or approximately proportionate sample distribution across the sort variables, in this way helping to increase the precision of the survey estimates (by reducing the variation associated with these variables). Thus, prior to sample selection, to achieve an implicit stratification within explicit sample strata, sorting on one or more of the variables age group, ZIP code, and race, depending on their availability from a State's licensure file, will be undertaken~~ will be undertaken to achieve an implicit stratification within explicit sample strata. When all three variables are available the sort order will be: age group; within age group by ZIP code; and within ZIP code by race. Some States provided only name and address, so ZIP code was the only sort variable used for such States. Few States provided race and, even when provided, there was often a substantial amount of missing data on race, so the effectiveness of the implicit stratification by race will generally be limited. ~~the sample frame for a given State will be sorted on variables such as age group, type of nurse (nurse practitioner, etc.), ZIP Code, or other variables, depending on the variables available on a State's frame. The choice of variables will depend on the analytic interest and needs. A systematic sample from a frame sorted on key variable(s) of interest will help achieve a proportionate or approximately proportionate sample distribution across the sort variables, in this way helping to increase the precision of the survey estimates (by reducing the variation associated with these variables).~~

HRSA is also interested in comparing the 2008 NSSRN estimates to those from the previous NSSRNs to help assess whether there have been any changes over time. The 2008 and previous sample designs permit appropriate inference to be made to the populations of licensed RNs at both the State and national levels. Estimated standard errors of study estimates will appropriately reflect the sample design for the 2008 NSSRN. The 2008 NSSRN sample will be independent of the samples for all previous NSSRNs, making for straightforward comparisons between estimates from the 2008 NSSRN and earlier administration cycles of the NSSRN. For example, HRSA can

estimate the standard error of the difference between two estimated averages as the square root of the sum of the variances of each estimated average.

There are a number of ways to develop sample weights for use with the 2008 NSSRN under this new design. HRSA will evaluate several approaches, and will make a decision as to which to use.

One approach would be to develop the weights as follows. Assign the base weights reflecting multiple chances of selection as described above. RNs would then be assigned to nonresponse adjustment cells constructed to represent those licensed in a single State only and then cells representing each combination of States where RNs are licensed in multiple States. (Those cells representing RNs licensed in a single State can be further partitioned by such variables as age groups, Metropolitan Statistical Area (MSA) status, etc. The opportunities for doing such partitioning in initial cells containing RNs with multiple licenses may be limited by the number of RNs found in such an initial cell.) Nonresponse adjustments could then be computed and final weights assigned. The resulting nonresponse adjusted weights, when limited to a particular State, would add up to the total number of RNs on the sample frame for the State while the sum across eligible respondents in all States will represent an estimate of the total number of currently-licensed RNs in the United States. However, adjustment cell sizes for some “multiple State licenses cells” could be small, and some adjustment factors in such cells could be undesirably large with such an approach. In such circumstances, the survey staff would examine (1) the extent to which collapsing across “multiple State licenses cells” would be necessary and (2) the implications of such collapsing.

Another approach would be to assign the base sample weights to all sampled RNs in the United States for whom eligibility status (eligible or ineligible) was ascertained. Then raking could be used to achieve a poststratification to control totals representing State sample frame totals, State frame totals by age, or perhaps some other set of dimensions. If there are nonrespondents for whom eligibility was ascertained, then, after dropping those RNs characterized as ineligible, a nonresponse adjustment will be assigned to the eligible respondents, accounting for the eligible nonrespondents. The issue for the raking approach would be whether convergence to the control totals can be

achieved, and, if so, the nature of the dimensions that can be used. HRSA will consider these and other approaches before making a final choice.

**Reliability of Estimates.** For the 2008 NSSRN, HRSA plans to use the Jackknife Replication methodology to establish replicate weights for variance estimation purposes, as has been done in past surveys. As in previous surveys, estimated sampling errors will be obtained for important items of interest, as will average design effects across multiple estimates of interest. In addition, a generalized variance estimation procedure will be provided for items for which direct estimates of sampling error are not computed.

**Nonresponse Bias Analysis.** HRSA plans to undertake analyses to help evaluate the extent to which there is the potential for bias arising in study estimates attributable to nonresponse. The information available for undertaking such analyses varies by Sstate. As mentioned above, some Sstates provided only name and address while others provided a number of other variables. These variables included sex, race, date of birth, date of last renewal, and advanced degree information. Even when a variable was provided, there can be an extensive amount of missing data on that variable. To undertake such analyses, it may be necessary to focus on RNs sampled from a small set of Sstates where there is a relatively large amount of data on nonrespondents. To the extent that data on RNs can be obtained from external sources, these could be used to supplement the data available from the sample frames for analytic purposes.

## 2b. Information Collection Procedure

The multi-mode data collection procedures are as follows:

- **First direct respondent contact sent by mail.** HRSA will make the first direct respondent contact a packet sent via first class U.S. mail to all sampled members with valid addresses. The packet will invite the sampled RN to participate in the study via the Internet or by using the enclosed paper questionnaire. The questionnaire will include the Web survey URL,

username, and password for the respondent to use to access the Web questionnaire.

- **Second mailing to sample members who have not yet responded.** A second packet containing a cover letter, the paper version of the questionnaire, and Web survey login information will be mailed approximately 2 weeks after the first mailing. This letter will emphasize that HRSA hasn't yet received the completed questionnaire and the need to respond soon.
- **Call nonrespondents.** Using a CATI instrument, the survey contractor's telephone interviewers will administer the full questionnaire to nonrespondents and attempt refusal conversion. Where possible, the interviewer will encourage the respondent to complete the Web survey.
- **Use FedEx to send out the nonresponse letter and a third copy of the paper questionnaire to all sample members who were not reached by telephone.** The final survey package is meant for those sampled RNs for whom a working telephone number could not be located. The package will contain the Web survey login information. HRSA will send this package 3 weeks after the second mailing to those with no working telephone number. The packet will be sent by FedEx since previous research experience shows that material sent by this method is more likely to be opened and read than materials sent using other types of mailing.

### **3. Methods To Maximize Response Rates**

The response rate for the 2004 National Survey of RNs was 70.5 percent. HRSA has set a target of 80 percent for the 2008 NSSRN. The improved response rate is anticipated due to the improvements in the questionnaire, and more resources set aside for nonresponse follow-up.

To ensure an 80 percent response rate to the survey, the data collection methodology will include the following:

- All mailings by first class U.S. mail, except for special subgroups of chronic nonrespondents with no telephone numbers;

- Follow-up telephone calls to nonrespondents with published phone numbers;
- Use of FedEx for chronic nonrespondents with no telephone numbers;
- Tracing efforts using commercial locating databases and directory assistance in an effort to obtain either new addresses or updated phone numbers for instruments returned as undeliverable, or to locate nonrespondents; and
- Use of a multi-mode design, meant to cater to the needs to respondents' preferences.

#### **4. Tests of Procedures**

The same type of data items were used in prior RN sample surveys and have yielded effective results, thus eliminating the need for a pretest. In addition, many of the recommendations implemented in the draft 2008 questionnaire for simplification, changes in instructions, and question wording were made by nursing workforce experts.

#### **5. Statistical Consultants\_**

This study methodology and overall sampling design were developed through contracts with consulting firms having expertise in designing complex sampling designs for large-scale surveys. A number of statisticians from HRSA, the National Center for Health Statistics (NCHS), and other Federal agencies have reviewed the methodology.

HRSA staff who participated in the design included the following:

- Marshall Fritz, Statistician and Project Officer 2008 NSSRN, Bureau of Health Professions, 301-443-6317
- Robert Oshel, Operations Research Analyst, Associate Chief for Resolution and Disputes, Practitioner Data Banks Branch, 301-443-6535
- Sarah Richards, Acting Branch Chief, Evaluation and Analysis Branch, Bureau of Health Professions, 301-443-5452
- William Spencer, Statistician and Evaluations Officer for Bureau of Health Professions, 301-443-6316

Outside consultation was provided by the following:

- Ralph DiGaetano, Senior Statistician, Westat, 301-294-2062
- Sherm Edwards, Vice President, Westat, 301-294-3993
- Stephanie Fry, Senior Study Director, Westat, 240-401-4643
- Jim Green, Senior Statistician, Westat, 301-251-4295
- Graham Kalton, Senior Vice President and Senior Statistician, Westat, 301-251-8253
- Vasudha Narayanan, Senior Study Director, Westat, 415-264-7064
- Joanne Spetz, Associate Director, Center for California Health Workforce Studies at University of California San Francisco (CCHWS at UCSF), 415-502-4443