Some Research on Sampling Design Alternatives for a Redesigned 2010 National Hospital Discharge Survey

Iris Shimizu[1] and Rong Cai [2]

[1]National Center For Health Statistics, 3311 Toledo Road, Room 3123, Hyattsville, MD 20782

[2]Substance Abuse and Mental Health Services Administration, 1 Choke Cherry Road, Rockville, MD 20850

Abstract

A redesign of the National Hospital Discharge Survey (NHDS) is planned for 2010. The new design will use a two stage sample of hospital discharges in which a stratified list sample of hospitals will be selected at the first stage. In this paper sample allocations to strata for samples of different sizes are optimized using either a Neyman or a nonlinear programming method. The different sample allocations are evaluated by estimating and comparing expected relative standard errors (RSEs, aka coefficient of variation or CVs) from the allocations for a set of discharge level and hospital level variables. This research uses sample discharge data available from the 2005 NHDS and hospital data available from a 2006 list of hospitals eligible for the 2010 NHDS universe.

Key Words: Cluster sampling, optimum sample allocation, nonlinear programming, relative standard errors

1. Introduction

The National Hospital Discharge Survey (NHDS) is conducted by the National Center for Health Statistics to produce nationally representative estimates of the characteristics of discharges, lengths of stay, diagnoses, surgical and non-surgical procedures, and patterns of use of care in hospitals in various regions of the country (DeFrances, Cullen, and Kozak, 2007). The universe for the survey consists of the discharges occurring at non-institutional, non-Federal hospitals, acute care and short stay hospitals having six or more beds staffed for inpatient care in the 50 states and the District of Columbia. Short stay hospitals are those for which the average length of patient stay is 30 days or less. The original sample for the current NHDS was implemented in 1988 with an area cluster sample of about 500 hospitals. Since then, the sample has been updated every third year to reflect developments in the hospital universe. Because most of the data currently collected in NHDS is in what is known as the Uniform Bill, the NHDS has been able collect computerized data in a uniform format and, thus, reduce response burden for hospitals willing to give computerized data while also reducing the cost of survey operations.

There are plans to implement a redesigned NHDS in 2010. Several changes are planned for that redesign. First there are plans to expand the data collection to include variables not included in the Uniform Bill, making data collection more expensive because the expanded data items are not uniformly available in a computerized format across hospitals. Second, a smaller hospital sample size is planned due to the expected increase in data collection costs. Thirdly, a list sample of hospitals (instead of a cluster sample) is planned. Finally, the hospital universe will be expanded to include hospitals regardless of their average lengths of stay. This expansion is designed to stabilize the universe so hospitals will not go in and out of the NHDS universe when the length of their patient's stays change.

A number of sampling designs have been considered for the redesigned survey. This paper discusses the research conducted to derive and evaluate the initial potential sampling designs considered for the redesign. For the research, a sampling frame of hospitals satisfying criteria for inclusion in the new universe was constructed from commercial hospital data base files available for 2006 from Verispan L.L.C. This frame is referred to as the "proxy frame," in the following sections. Sample discharge data from the 2005 NHDS were also used.

Section 2 outlines objectives which guided the research discussed in this paper. Section 3 discusses the methods used to derive sampling designs (define strata and sample allocation) and the precision expected from the derived designs. Section 4 discusses the results and comparisons between sampling designs considered while Section 5 presents a summary and conclusions from the research covered in this paper.

2. Design Objectives

To guide development of sampling designs discussed in this paper, it was assumed that discharge level statistics (e.g. percent of patients with Medicare) had priority over hospital level statistics (e.g. percent of hospitals in non-Metro areas). This means decisions were made in favor of patient level statistics when objectives for both patient level and hospital level statistics could not be met, simultaneously. For example, because the optimum sample allocation for one statistic may not be optimum for another statistic, only patient level statistics were used to optimize sample allocations discussed later in this paper.

Estimation domains defined in terms of hospital characteristics were among objectives expressed for the redesign. Estimates are desired for the following hospital domains listed in order of priority expressed for the initial research:

1. Type of service (General medical and surgical, psychiatric, children's other than psychiatric or long stay, and "other" which primarily includes long stay hospitals other than psychiatric hospitals).
2. Area type/size (MSA with one million or more population, MSA with less than one million population, non-MSA areas where MSAs are metropolitan statistical areas defined by the Office of Management and Budget.)
3. Bed size (6-49 beds, 50-99 beds, 100-299 beds, 300-499 beds, and 500 or more beds)
4. Ownership (government /non-government)
5. Region (Northeast, Mid-West, South, West).

Another design objective was to obtain expected relative standard errors (RSEs) (also known as coefficients of variation or CVs) of 30 percent or less for selected discharge and hospital variables. The study variables were selected to include a range of estimate magnitudes and, for discharges, a variety of demographic characteristics. The discharge variables included both percent and aggregate variables and an average (for days of stay in short- stay hospitals). The selected discharge variables and their estimates from the 2005 NHDS are shown in Table 1, where the binomial variables are listed in order of their estimated percent of discharges. The selected hospital variables included only percent estimates ranging from 2.5 to 33 percent of hospitals and are listed in Table 2. The study variables were limited to those for which data are available from the current NHDS or the proxy frame. For example, there are no data available for variables such as "average days of stay in long-stay hospitals" because the current NHDS universe excludes long-stay hospitals.

3. Sampling Designs and Methods Used to Derive Them

Strata for the initial sampling designs were defined using some of the estimation domains from the list above. Separate strata were first defined for each of the four service types. Because over 75 percent of the hospitals in the frame are general hospitals, the general hospitals were stratified further by area type and then by bed size within area type. Bed size classes containing small numbers of hospitals and/or discharges relative to other bed size classes in the same area type were collapsed with adjacent bed size classes to reduce variation in strata sizes. A total of 13 strata were defined (ten strata for general hospitals and one for each of the three remaining service types). The numbers of hospitals in the proxy frame and in the 2005 sample are shown in Tables 3 and 4 for the four service type domains and strata defined for general hospitals, respectively.

Sample allocations to the 13 defined strata were optimized. In this optimization, three hospital sample sizes (120, 240, and 480) and four discharge sample sizes (100, 200, 300, and 400) per hospital were considered.

In the following discussion, M and m denote the hospital counts in the population and sample, respectively, while N and n denote discharge counts in the population and sample, respectively; h denotes a stratum, i denotes a hospital, and j denotes a discharge. Let X denote the characteristic of interest for the population and x denote the sample estimate of X.

Two methods were investigated for optimizing hospital sample allocations to sampling strata. The first method applied is attributed to Neyman (Cochran, 1977). That method minimizes the variance of the estimated mean $\overline{x}$ for a fixed sample size m if the number of hospitals selected from stratum h is:

$$m_h = m\left[M_h S_h\left(X\right)\right] \Big/ \sum\nolimits_h M_h S_h\left(X\right) \tag{1}$$

where $S_h(X) = \sqrt{\sum\nolimits_{i=1}^{M_h}\left(X_{hi} - \bar{X}_h\right)^2 /(M_h - 1)}$ and h = 1, 2, …13. (2)

Under the assumption that variances of most discharge aggregate statistics would likely be minimized when the variance for total discharges was minimized, Neyman's method was used with X_{hi} being the total discharges recorded for hospital i in stratum h of the proxy sampling frame. To prevent understatement of stratum variances, only hospitals with total discharge counts recorded in the proxy frame were included in calculations of $S_h(X)$ in (2).

The second optimization method is a nonlinear programming (NLP) procedure discussed in Nocedal and Wright (1999). For this method, a function F is defined as

$$F = Var\left(\bar{x}\right) - \lambda\left(\sum\nolimits_h m_h - m\right) \tag{3}$$

with the constraint $\sum\nolimits_h m_h = m$ and is minimized for a fixed sample size m. For stratified cluster samples, it can be shown that

$$Var\left(\bar{x}\right) \doteq \frac{1}{N^2}\sum\nolimits_h \frac{a_h\left(x\right)}{m_h} \ , \tag{4}$$

where $a_h(x) = N_h^2\left\{\frac{1}{\bar{n}_h} s_h^2(x)[1+\left(\bar{n}_h - 1\right)\delta_h\left(x\right)] + \bar{x}_h^2 B_{Nh}^2\right\}$, $s_h^2(x) = \dfrac{\sum_i \sum_j \left(x_{hij} - \bar{x}_h\right)^2}{(n_h - 1)}$, (5)

δ is the intraclass correlation, and $B_{Nh}^2 = \sum\nolimits_i\left(N_{hi} - \bar{N}_h\right)^2 \Big/ \left[\left(M_h - 1\right)\bar{N}_h^2\right]$. When divided by the number of sample hospitals, the last term in a_h is the contribution to the stratum variance due to variation in hospital discharge volumes (Hansen, Hurwitz, and Madow, 1953). The intraclass correlation may be expressed as $\delta(X) = \sigma_B^2(X)/[\sigma_B^2(X) + \sigma_W^2(X)]$ where σ_B^2 and σ_W^2 are the between and within hospital variances, respectively (Foy 2004). Substituting (4) in (3), the function F and the variance of $\bar{x}$ are minimized when the number of hospitals selected from stratum h is:

$$m_h = m\sqrt{a_h(x)} \Big/ \sum\nolimits_h \sqrt{a_h(x)} \ . \tag{6}$$

The NLP procedure was used to optimize eleven sample allocations, one for each of the study binomial discharge variables other than the variable "% Non-metro hospital." For the *$a_h(x)$* in (5), discharge totals recorded in the proxy frame were used to compute B_{Nh}^2 and only hospitals for which those totals were recorded were included in the calculations to prevent understating B_{Nh}^2. Also in (5), 2005 NHDS sample discharge data were used to compute the $s_h^2\left(x\right)$ and the $\delta_h\left(x\right)$. For binomial variables, the $s_h^2\left(x\right)$ variances in (5) were calculated using the expression $\bar{x}_h\left(1 - \bar{x}_h\right)$. For all discharge variables, the intraclass correlations were computed using the following formula from Foy (2004):

$$\hat{\delta} = \left(MS_B - MS_W\right) \Big/ \left[MS_B + \left(n' - 1\right)MS_W\right] \tag{7}$$

where

$$MS_B = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\,(\bar{x}_{i.} - \bar{x}_{..})^2 \Big/ (m-1)\,,\ MS_W = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2 \Big/ \sum_{i=1}^{m}\left(n_i - 1\right) \text{ and} \tag{8}$$

$$n' = \left(\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} n_i^2 \Big/ \sum_{i=1}^{m} n_i\right) \Big/ (m-1)\,.$$

The w_{ij} in (8) denotes the sample weight included in the data file for discharge ij. Continuing the tradition of excluding newborns from NHDS estimates, newborns were excluded from the computations except when the variable was "newborn." Unweighted averages were used for sample means $\bar{x}$ within each hospital and stratum under the assumption the unweighted means will not adversely affect the results for the study variables.

Neyman hospital sample allocations were produced for each of three total sample sizes (120, 240, and 480 hospitals) and four discharges sample sizes (100, 200, 300, and 400) per hospital. The NLP allocations were optimized assuming a total sample of 120 hospitals and 100 discharges per sample hospital. Several constraints were imposed on the optimization results. First, the optimization procedures were applied to only 80 percent of assumed total sample sizes to allow for expected sample losses due to ineligible and refusal hospitals. In other words, only the allocation of the expected respondent sample, not the total sample, was optimized and those optimum sizes were then inflated to produce the total sample allocated to each stratum. Second, two or more respondent hospitals were required from each hospital stratum to assure representation from each stratum and the ability to approximate variances. If the optimization assigned fewer than two hospitals to a stratum, the sample size for that stratum was increased to two at the expense of strata assigned the largest sample(s). Third, integer values were required for each stratum sample size. The individual optimized stratum sample sizes were rounded up or down to integer values while retaining the fixed total sample size.

To evaluate the different sampling designs, the expected relative standard errors (RSEs) $\left(= \sqrt{\text{var}(\bar{x})} \big/ \bar{x}\right)$ from those designs were computed and compared for each study variable. For discharge variables, the variances were approximated using (4) and (5) together with the 2005 NHDS discharge data and the discharge volumes from the proxy frame. The variance for each estimated percent $\hat{p}$ of hospitals was approximated using hospital counts from the proxy frame in the formula $Var(\hat{p}) = \left(1/M^2\right)\sum_h M_h^2 p_h\left(1 - p_h\right)/m_h$.

4. Results

Table 1 presents RSEs for discharge variables when the Neyman method was used to optimize sample allocations. Columns C, D, E, and F show RSEs expected for a total sample of 120 hospitals when the number of discharges selected per hospital are 100, 200, 300 and 400 discharges, respectively. Column G in that table shows the difference between the maximum and the minimum of the four RSEs for each discharge variable. Except for the smallest magnitude estimates (≤ 4 percent of discharges in Column A), the RSEs differed by 0.5 percent or less. Hence, it was decided to use 100 discharges per hospital in the redesigned sample and also in the remaining sampling design research.

Also in Table 1, Columns C, H, and I show the RSEs expected for discharge statistics when 100 discharges were assumed selected per hospital from total samples of 120, 240, and 480 hospitals, respectively. Columns J and K present the decrease in RSEs when the hospital sample size of 120 is doubled and quadrupled. These differences are 30 and 50 percent, respectively, relative to the RSEs from the 120 hospital sample. This observation confirms that adding hospitals to the sample benefits precision of discharge statistics more than does adding discharges per hospital. This can be expected due to the variation in total discharge volumes among hospitals within strata.

Table 2 presents the RSEs for hospital statistics for the three samples sizes when the Neyman was used to optimize the sample allocations. For the sampling strata defined for the research in this paper, it can be seen samples of 120 hospitals are not sufficient to meet the targeted precision standard (RSE$\leq$30 percent) for the smallest hospital percent estimates. It is believed that precision standard was not met for hospitals having 500 or more beds because those hospitals shared some sampling strata with other bed size domains. (See strata defined in Table 4 for general hospitals). Had the 500-plus-bed hospitals been in a stratum (or strata) by themselves, the RSE for estimated percent of hospitals with 500 or more beds would have been zero, as was the case for the "other service type" hospitals, which were in a sampling stratum by themselves.

Table 5 presents RSEs from sample allocations optimized using the NLP method. Due to space limitations, it is not practical to present the RSEs from all 11 allocations. Instead, Columns A and B of Table 5 show the minimum and maximum RSEs, respectively, among those 11 allocations for each of the study discharge variables. The differences between the maximum and minimum RSEs are less than 2 percent except for the variables "Patient aged <15 years" (difference is 3.3 percent) and "Depression & Bipolar" (difference is 2.6 percent). From data not shown in this paper, it was observed that the allocation optimized for the variable "Patients aged < 15 years" yielded RSEs "near" the minimum RSEs for the greatest number of study discharge variables. The RSEs from that allocation are shown in Column C of Table 5 while Column D shows the differences calculated as "Column C RSE minus the minimum RSE (in Column A)." The differences in Column D are less than 1 percent for all but the variable "Depression & Bipolar," implying the "patients aged < 15 years" variable is a reasonable one to use when using the NLP method to optimize sample allocations.

To compare the Neyman and NLP allocations, the RSEs from the Neyman allocation of 120 sample hospitals are shown in Column E of Table 5 while Column F subtracts the minimum NLP RSEs (in Column A) from the Neyman RSEs. It can be seen the differences ranged from 0.0 to 2.0 percent suggesting the NLP method may be better for optimizing sample allocations for discharge statistics but the differences in RSEs are small, especially for the larger magnitude estimates.

5. Summary

This paper discusses some initial design research for a stratified hospital sample to be used in the 2010 redesigned National Hospital Discharge Survey (NHDS). The discussed research assumed discharge level statistics have a higher priority than hospital level statistics. For that research thirteen hospital sampling strata were defined by hospital service type, area type/size, and hospital bed size. Two methods, a method attributed to Neyman and a nonlinear programming method, were used to optimize allocations in hospital samples of different sizes and different numbers of discharges selected per hospital. The allocations were optimized with constraints on the minimum numbers of hospitals selected from each stratum. The relative standard errors (RSEs) expected from each sample allocation was computed and compared across allocations for each of selected discharge- and hospital-level variables. While the hospital universe for the redesigned NHDS includes hospitals without regard to length of stay, research involving discharge variables was limited to variables applicable at general and short stay hospitals included in the current NHDS universe.

The RSEs estimated in the research demonstrated:

- Increasing numbers of sample hospitals causes greater reduction of RSEs than does increasing numbers of sample discharges per hospital.
- For discharge-level variables, a nonlinear programming method for optimizing hospital sample allocations yielded lower RSEs than did the Neyman method, but the differences in RSEs were small, especially for larger magnitude statistics.
- When estimates are required for small hospital domains, separate sampling strata are advised for those small domains.

More research is planned on samples with alternate definitions for the hospital sampling strata. Improvements in precision of hospital level statistics for key small hospital domains will be targeted in those sampling designs.

References

Cochran WG (1977). *Sampling Techniques, Third Edition.* John Wiley and Sons.

DeFrances CJ, Cullen KA, Kozak LJ (2007). National Hospital Discharge Survey: 2005 annual summary with detailed diagnosis and procedure data. National Center for Health Statistics. Vital Health Stat 13(165)

Foy P (2004). "Intraclass correlation and variance components as population attributes and measures of sampling efficiency in PIRLS 2001." The 1st IEA International Research Conference.

Hansen, Hurwitz & Madow (1953). *Sample Survey Methods and Theory.* John Wiley and Sons. (Chapter 8.4).
Nocedal J and Wright SJ (1999). *Numerical Optimization.* Springer Science and Business Media Inc.

Table 1. Study discharge variables, their 2005 NHDS estimates and expected RSEs with alternative sample sizes: hospital sample allocation optimized by Neyman method

	Estimates from 2005 NHDS		Total number of sample hospitals							Decrease in RSE from RSE for 120 hospitals &100 discharges/hosp.	
			120					240	480		
		Targeted	Discharges per hospital				MAX-MIN*	Discharges Per hospital =100			
	Percent	Estimate	100	200	300	400					
Discharge Variable	A	B	C	D	E	F	G	H	I	J=C-H	K=C-I
			RSE %				DIFF.	RSE %		DIFFERENCE*	
%ED admission	41.4%	41.4%	5.7	5.6	5.6	5.6	0.1	4.0	2.8	1.7	2.8
%Medicare	39.8%	39.8%	3.5	3.4	3.4	3.3	0.2	2.5	1.8	1.0	1.8
Female aged 15-44 years	20.6%	7,794,615	4.6	4.3	4.2	4.2	0.4	3.2	2.3	1.3	2.3
%Non-metro hospital	17.1%	17.1%	1.6	1.5	1.5	1.5	0.2	1.2	0.8	0.5	0.8
Black	10.9%	4,108,786	15.6	15.5	15.4	15.4	0.2	11.0	7.8	4.6	7.8
Patient aged <15 years	6.4%	2,430,926	11.4	11.1	11.0	11.0	0.4	8.0	5.7	3.3	5.7
%Delivery with C-section	3.6%	3.6%	9.0	8.1	7.8	7.7	1.3	6.4	4.5	2.6	4.5
Depression & Bipolar	2.5%	944,599	15.4	14.8	14.6	14.5	0.9	10.9	7.7	4.5	7.7
Acute myo-cardial infarction	1.8%	682,699	11.2	9.5	8.9	8.5	2.6	7.9	5.6	3.3	5.6
Asthma	1.3%	488,594	13.5	11.5	10.8	10.4	3.0	9.5	6.7	3.9	6.7
%C-difficile	0.2%	0.2%	25.3	19.9	17.7	16.5	8.9	17.9	12.7	7.4	12.7
Newborn	10.6%	3,999,112	8.1	7.8	7.7	7.6	0.5	5.7	4.1	2.4	4.1
Average days per stay in short stay hospitals		4.8	3.1	2.9	2.8	2.8	0.3	2.2	1.5	0.9	1.5

* Differences in Column G, J, and K were calculated with unrounded RSE values and may not equal the differences calculated with the rounded RSE values shown in Columns C, D, E, F, H and I.

Table 2: Study hospital variables, their 2005 NHDS estimated percent and expected RSEs with alternative hospital sample sizes: Hospital sample allocation optimized by Neyman method

		Total sample size		
Hospital variables	Estimated percent	120	240	480
		RSE %		
Total	100.00%			
In Non-metro areas	33.70%	7.6	5.4	3.8
Owned by State/local govt.	23.50%	25.7	18.2	12.9
In West region	18.20%	30.2	21.4	15.1
Have 500 or more beds	2.50%	41.2	29.2	20.6
Are “Other” Service type	10.70%	0	0	0

Table 3: Hospital service type domains/strata with hospital counts from the proxy frame for the new universe and the 2005 sample

Service type/sampling strata	Proxy frame	2005 sample
	Hospital counts	
All	6517	444
General	5061	426
Psychiatric	662	6
Children’s (not psychiatric or long stay)	99	11
Other	695	1

Table 4. Strata for general hospitals and their hospital counts from proxy frame for the new universe and the 2005 sample

MSA 1+million population			MSA <1 million population			Non-MSA		
Bed-size	Proxy frame	2005 sample	Bed-size	Proxy frame	2005 sample	Bed-size	Proxy frame	2005 sample
All	1662	264		1339	107		2060	55
6-99 beds	492	24	6-99 beds	574	15	6-49 beds	1361	24
100-299 beds	840	125	100-299 beds	566	50	50 -99 beds	426	23
300-499 beds	236	71	300+ beds	199	42	100+ beds	273	8
500+ beds	94	44						

Table 5. Selected RSEs in percent for discharge estimates from optimized allocations: 120 sample hospitals with 100 discharges per hospital

	Optimizing method					
	NLP*		NLP for <15 years age		Neyman's	
	MIN RSE %	MAX RSE %	RSE %	Excess RSE %	RSE %	Excess RSE %
Discharge variable	A	B	C	D=C-A	E	F=E-A
%ED admission	5.5	6.1	5.6	0.1	5.7	0.1
%Medicare	3.4	3.7	3.5	0.1	3.5	0.1
Female aged 15-44 years	4.4	4.8	4.4	0.0	4.6	0.2
%Non-metro hospital	1.1	1.4	1.4	0.4	1.6	0.6
Black	14.5	16.2	14.9	0.4	15.6	1.1
Patient aged <15 years	9.4	12.6	9.6	0.3	11.4	2.0
%Delivery with C-section	8.5	9.6	8.6	0.1	9.0	0.5
Depression & Bipolar	13.5	16.1	14.9	1.4	15.4	1.9
Acute myocardial infarction	10.3	11.9	10.9	0.6	11.2	0.8
Asthma	11.9	13.7	12.1	0.2	13.5	1.5
%C-difficile	23.8	25.4	24.4	0.7	25.3	1.6
Newborn	7.9	8.7	7.9	0.1	8.1	0.3
Average days per stay in short stay hospitals	3.1	3.6	3.1	0.0	3.1	0.0

* The NLP method was used to optimize eleven sample allocations, one for each of the binomial discharge variables other than the variable "% Non-metro hospital."