ACCURACY OF RECORD SAMPLING IN THE 1977 NATIONAL HOSPITAL DISCHARGE SURVEY

Steven L. Botman and Iris M. Shimizu, National Center for Health Statistics

INTRODUCTION AND SUMMARY

Statistics from the National Hospital Discharge Survey (NHDS) are based on a stratified two-stage sampling design. The first-stage sampling frame consists of a stratified list of non-Federal short-stay hospitals. In the second stage, sample hospital medical records (i.e., discharges) are usually selected for the NHDS through a procedure that uses the last two digits of the hospital-assigned medical record number. All medical records whose identification number ends in certain NHDS-assigned two-digit combinations are selected for the NHDS sample. This work focuses both on the development of a method to measure the accuracy of record sampling and on the resulting assessment.

The method for assessing the accuracy of record sampling uses information from the NHDS data base that contains a magnetic tape record for each responding inscope record. Each record is identified with a hospital-assigned record number. Also used were the NHDS Sample Listing Sheets that identified all records sampled for the NHDS.

This method could not be applied to data from all NHDS sample hospitals; the procedure could only be applied to data from hospitals which at inpatient admission serially assign a unique medical record identification number to each inpatient episode of care. Excluded from such examination were the NHDS sample hospitals whose medical record numbers were assigned to people (i.e., not assigned to an individual's particular episode of care).

The data submitted for approximately one quarter of the hospitals in the NHDS sample were amenable to the sample record number analysis, which is later described. The results of the evaluation of record sampling among these hospitals extends to the remaining hospitals in the NHDS sample if one accepts the assumption that the accuracy of record sampling is independent of the scheme used by the hospital to assign record numbers. There is, however, no way to determine the validity of this assumption.

Overall, the accuracy of record sampling as assessed through an examination of the sample medical record identification numbers was quite good. Some 97.7 percent of the record numbers on the average that were eligible for inclusion in the NHDS sample were indeed sampled for the NHDS, with 96.3 percent of the eligible record numbers appearing on the NHDS data base. The proportion of the eligible sample record numbers that were sampled for the NHDS varied from 84.6 percent to 100.0 percent among the hospitals whose data were examined. Some deviation from the 100.0 percent figure is to be expected; on the other hand, a large deviation is not.

PRELIMINARIES AND PROCEDURES

An examination of the identification numbers for the sample records is very much dependent on the record-numbering scheme used by the sampled NHDS hospitals. As noted, the sample NHDS medical records (i.e., discharges) are usually selected by a procedure that uses the last two digits of the hospital-assigned medical record number.

Most NHDS sampled hospitals elect one of two schemes for assigning medical record numbers. In those hospitals with a unit-numbering scheme, each individual is assigned a number that is used to identify the medical records corresponding to the individual's inpatient care at all episodes of care, whether one or many.

In hospitals using a unit-numbering scheme the numerical pattern of record identification numbers has two characteristics: First, considerable overlap exists in the monthly ranges of record identification numbers for discharges (i.e., the interval defined by all numbers falling between the smallest and the largest sampled medical record number in the hospital's monthly data transmission). This overlap is due to the fact that identification numbers are assigned at the individual's first admission to the hospital--not the current admission. The record numbers assigned at an earlier admission date are much smaller than the record identification numbers assigned on the current admission date. Secondly, the same number may be used to identify two or more distinct sample records--each of which corresponds to a separate episode of inpatient care within the sampled hospital by a single individual.

These two factors preclude an examination of the record numbers among unit hospitals. Some 249 sampled hospitals were initially classified as using a unit-numbering scheme to assign medical record numbers in 1977. These hospitals are excluded from subsequent examination.

In NHDS sample hospitals using a serial-numbering scheme, each individual inpatient admission is serially assigned a medical record identification number. An individual who had several admissions at such a hospital would be assigned a medical record identification number at admission for each episode of care. In the 1977 NHDS, 191 hospitals were initially classified as assigning medical record identification numbers using a serial-numbering scheme. When the identification numbers for the sampled records were examined, duplicate record numbers were found to exist in some hospitals.

There are two possible reasons why duplicate record numbers were found. The first is that some hospitals initially classified as using a serial-numbering scheme during 1977 were in fact using a unit-numbering scheme. This was assumed to be the case in those hospitals for which the monthly ranges of sampled record numbers had considerable overlap. The second is that some hospitals using a serial-numbering scheme appear to be assigning the identification number of the medical record for a mother to her newborn. This was assumed to happen in hospitals where two records, each with the same identification number, were sampled and showed considerable

overlap in the time periods for the two episodes of care (viz., the mother giving birth during the episode of care and the baby being born during the episode of care, with the time spans for both of these overlapping). The important factor is that there are duplicate record numbers among the records sampled, not the particular reason that the duplicate record numbers occurred. The hospitals in which duplicate numbers occurred were eliminated from further scrutiny.

Whenever for a hospital the difference between the number of records actually sampled for the interval of eligible sample records and the number of sample records theoretically expected was greater than 20 percent, the differences were classified as representing an artifact of the facility's record management, rather than resulting from inaccuracies in record sampling. It appeared that the record numbers were not being uniquely assigned in a serial fashion and that the use of the medical records numbers to measure the accuracy of record sampling would yield erroneous results. Examples of the situations encountered in the sampled hospitals excluded from record number analysis include: The presence on a periodic or aperiodic basis of record numbers alien to the other record numbers--some individual sampled hospitals had some medical record numbers consisting of four, five, and six nonzero digits. In some hospitals it appeared that numbers were being assigned out of several serial sequences. In other hospitals, numerous gaps were found in the pattern of sampled record numbers. More than one of these problems appeared in the identification numbers carried by the sample records from some hospitals.

The assessment of the record sampling accuracy in the 1977 NHDS is based on the examination of the identification numbers for the sample records in 111 serial hospitals, with the identification numbers from 80 allegedly serial hospitals not used. About a third of the 80 hospitals were using a unit-numbering scheme; about a third had duplicate numbers present in the sampled records; and the remainder were excluded for various reasons, such as those mentioned in the prior discussion.

In addition, one other pattern was noticed among hospitals using a serial-numbering scheme that was an important factor in the subsequent analysis. The majority of the hospitals using a serial-numbering scheme continued assigning record identification numbers between calendar years; however, some did not. This fact had to be taken into account while determining the gaps in sample identification numbers. For those which did not continue the serial numbering pattern, the most common procedure was to reinitialize the numbering scheme at 1 on January 1 of each year, with the initial digit(s) of the record identification number indicating the year in which the patient was admitted.

Before describing in detail the procedure used in examining the identification numbers of the sample records, two procedures commonly used to to establish the second-stage sampling frame must be understood. In the vast majority of examined hospitals (101 for the entire survey year and 1 for half of the survey year), the sampling frame is established by listing discharges. Among these hospitals, it is somewhat difficult to determine a pattern among the identification numbers for the sampled records. The basic complication in this case is that we are systematically sampling by using the record number's terminal digits from a list of discharges that are not ordered in the same fashion in which these records have their identification numbers assigned. For example, two consecutive sampled record identification numbers which might have been assigned on a single day of admission may correspond to episodes of care with widely-different lengths of stay; accordingly, these two record numbers would appear on the discharge lists on two widely-separated dates. It is assumed for our purposes that gaps in sampled identification numbers at the very beginning or the very end of the survey year are attributable to the occurrence of a few episodes of care with either relatively long or relatively short lengths of stay terminating about the end of the survey year.

In the remaining examined hospitals, the second-stage sampling frame is established by listing admissions to the hospital. The sample records are selected from this list, generally in ascending numerical order, and recorded on the NHDS Sample Listing Sheet. The pattern of sample record identification numbers from such hospitals is readily apparent, with the nonsampled (i.e., missing) record numbers easily determined.

Any examination of the identification numbers of the sample records requires a systematic procedure to establish an interval of eligible sample numbers. This term should be interpreted as a numerical interval in which identification numbers of medical records with the specified NHDS-assigned terminal two-digit combinations without exception were expected to be sampled for the NHDS. Implicit within this procedure is a method to determine the outliers--the accuracy of record sampling about the outliers could not be evaluated using this type of procedure. Prior to the adoption of the selected approach, several others, including eyeballing, were taken; these other approaches did not allow the development of uniform and comparable results.

The initial endpoint of the interval of eligible sample numbers for a facility was determined by identifying a sampled number having the property that each subseqent eligible number could be expected to be sampled for the NHDS. The procedure basically consisted of determining the smallest set of three consecutive eligible numbers that were actually sampled; the first number of the set would then be the initial endpoint. The terminal endpoint was determined in an analogous fashion. Record numbers outside of the interval were considered outliers for our purpose and dropped from the analysis to determine the missing record identification numbers.

Once the interval of eligible sample numbers for each hospital was determined, the information needed to evaluate the accuracy of record sampling for the NHDS was computed for each hospital. First, the number of eligible sample record numbers in the interval was determined; then the records in the data base that fell within the interval; then the number of records

that were sampled but were not reflected in the data base because the particular records were not available or were incomplete or the particular record was out of the survey's scope or failed the survey's edits.

All medical record numbers with the appropriate terminal digits falling in each hospital's interval of eligible sample numbers were expected to have been sampled for the NHDS. Any record number eligible for sampling within a hospital's interval of eligible sample numbers that was not actually sampled for the NHDS was classified as a "nonsampled" number.

The number and proportion of nonsampled numbers in the interval of eligible sample numbers was taken as a measure of the accuracy of record sampling. If many eligible numbers or a relatively large proportion of the total count of them in the interval of eligible sample numbers were nonsampled, then we concluded there were inaccuracies in the sampling of records--or there were difficulties in the hospital's record management. If all of the eligible numbers in the interval of eligible sample numbers were sampled, then we concluded that the sampling of records was done accurately.

Another measure of the accuracy of record sampling in the NHDS may be obtained by comparing various candidates for the within hospital discharge total during the survey year. One candidate is the figure that is provided by the facility as the discharge total. Another candidate is the usual estimate of within hospital discharge total (i.e., the number of inscope sample discharges times the reciprocal of the probability of sample selection). The third candidate is a revised within hospital discharge total that is essentially the usual estimator coupled with adjustment for "nonsampled" records. These were computed for each examined sampled hospital.

FINDINGS

Among the 111 hospitals whose sample record identification numbers were examined, 60,379 identification numbers fell into the intervals of eligible sample numbers. That is, if every record number eligible for sampling and falling within in each hospital's interval of eligible sample numbers was sampled, then 60,379 identification numbers (or records) would have been sampled for the 1977 NHDS. Of course, all these identification numbers were not sampled.

Some 58,100 (96.2 percent) of these identification numbers were found on the NHDS data base. An additional 842 record numbers were identified as having been sampled for the NHDS but did not make it onto the data base. The remaining 2.4 percent of the record identification numbers were not sampled. The number and proportion of such nonsampled records are being taken as a measure of the accuracy of record sampling in the NHDS.

The average hospital whose data were examined for this research sampled some 531 record numbers in the hospital's interval of eligible sample numbers, with an average of 13 additional record identification numbers being not sampled within the interval. Thus, overall, the number and proportion of nonsampled record identification numbers are small. Moreover, some 47 out of the 111 hospitals had under 5 eligible record identification numbers nonsampled in their respective intervals of eligible sample numbers. Some of the hospitals did accordingly appear to have inaccuracies in record sampling, evidenced by the persistent occurrence of nonsampled record identification numbers.

Although the average length of stay for these hospitals was not estimated from the sample data (for a hospital to be in scope for the survey it had to have an average length of stay of 30 days or less), 1977 NHDS published estimates reported[1] that the average length of stay was 7 days. Accordingly, the error in our measure of record sampling accuracy that may be due to nonsampled records which corresponds to episodes of care occurring at the beginning and end of the year should be small.

The average number of nonsampled records in the interval of eligible sample numbers did not strictly increase with increases in the number of sampled records (see table 1). This was unexpected since with more records being sampled it was speculated that it would be more likely for a record to be not sampled. An average of 21 eligible record identification numbers were not sampled in their respective intervals of eligible sample numbers by hospitals having 600-799 sampled medical records in the 1977 NHDS data base--this was almost twice as large as the figure for hospitals in any other sample size category. This large number is partially due to the presence within this category of a few hospitals with relatively large numbers of nonsampled record numbers (see table 1). This again provides evidence that the overall accuracy of record sampling is good. In the cases where there was a relatively large number or proportion of nonsampled records, questions are raised on the accuracy of record sampling--or the facility's record management.

Another measure of the accuracy of record sampling in the NHDS can be obtained by using the hospital reported total number of hospital discharges during the survey year. We compared this figure with the usual estimate of within hospital discharge total.

This estimate of within hospital discharge total was computed for each of the 111 hospitals whose sampled record identification numbers were examined. We wanted to determine if this estimator was a good predictor of the reported total number of discharges. It was; the correlation coefficient between the two sets of figures was computed to be 0.998. Of course, there were cases identified where the difference between the two sets of figures was proportionally or numerically large. These cases can result from inaccuracies in record sampling (viz., undersampling) and difficulties in record management.

DISCUSSION

Evidence from this research indicates that there there were some relatively minor inaccuracies in record sampling resulting from certain eligible sample record numbers being not sampled. Such errors can lead to underestimation (not overestimation) of a within hospital

discharge total if ratio-adjustment to the hospital reported discharge totals was not made.[2] The procedure used in the examination of the record identification numbers enables one to quantify the extent that eligible sample identification numbers were not sampled.

The appropriateness of the approach was tested. It was speculated that the revised estimator of discharge totals would be better than the one previously constructed. Also, if the revised estimator was a better estimator than the previously constructed one, then this would lend additional credence to the suitability of the described measure of the accuracy of record sampling. The correlation coefficient between the revised set of within hospital discharge totals and the corresponding hospital reported totals was 0.997--slightly lower than the figure computed earlier. Not much can be concluded. It appeared from examining the data that the revised estimator was overcompensating for nonsampling of eligible records.

The relationship among the usual estimator of within hospital discharge totals, the revised estimator of within hospital discharge totals, and the hospital reported total of discharges was examined. The most common relationship among the three figures for a specific hospital was that the usual estimate of within hospital discharge totals was less than the hospital reported discharge total that in turn was less than the revised estimate of within hospital discharge totals. This provides evidence that there was undersampling of records for the NHDS (the usual estimate was less than the reported total) and that at the same time a large proportion of the nonsampled records were out-of-scope (the hospital reported discharge total was less than the revised estimate of discharge total).

While the described method of assessing the accuracy of record sampling in the NHDS may have some weaknesses, these weaknesses result not from a deficiency in the method but rather peculiarities in facility record management. These peculiarities are of themselves worthy of being identified. In some cases, significant questions are raised about the facility's record management. Additional research is needed to determine if the same hospitals exhibit these peculiarities year after year and if the peculiarities correspond to errors in sampling.

NOTES AND REFERENCES

[1] National Center for Health Statistics: "Utilization of short-stay hospitals, annual summary for the United States, 1977". Vital and Health Statistics. Series 13 - No. 41. DHEW Pub. No. (PHS) 79-1792. Public Health Service. Hyattsville. U.S. Government Printing Office, March, 1979.

[2] There may be reservations about this kind of adjustment for nonresponse--even with relatively low levels of nonresponse--if the characteristics of the sampled eligible records vary greatly from the characteristics of the nonsample eligible records. Additional research is needed to make this determination.

TABLE 1

Number and distribution of sampled hospitals by number of sampled records in the hospital's interval of eligible record numbers and the number of eligible numbers that were not sampled, and average number of nonsampled record numbers by number of sampled records in interval

Number of nonsampled numbers in interval	Total	Number of sampled records in interval					
		Under 200	200-399	400-599	600-799	800-999	1,000 or more
		Number of hospitals					
Total	111	5	31	37	29	4	5
Under 5	47	3	15	15	11	2	1
5-9	22	1	7	8	5	0	1
10-14	14	1	4	4	3	1	1
15-19	8	0	2	4	1	0	1
20-24	5	0	1	2	2	0	0
25-29	4	0	1	1	1	1	0
30-34	1	0	0	0	1	0	0
35 and over	10	0	1	3	5	0	1
		Average number of nonsampled record numbers in interval					
	13	5	8	12	21	11	10