Two Phase Sampling Design for Selecting Representative Samples from Large Administrative Databases

Skip Camp†, Srinath, K.P†., Arday, S‡. and Elizabeth Axelrod†.
†Abt Associates Inc., Washington D.C.
‡Centers for Medicare and Medicaid Services, Baltimore, Maryland

Key words: EDB, Sample, CATI, Influenza, Medicare, PRO

Introduction:
Sampling from large administrative databases, like those maintained by federal agencies, requires the use of efficient sampling methods that result in representative samples while minimizing resources required for selection. In this paper, large administrative databases are defined as databases containing millions of records with at least 80 independent variables. The Centers for Medicare and Medicaid Services (CMS), formally known as the Health Care Financing Administration (HCFA), maintains and updates the Enrollment Database (EDB) files. The EDB is home to over eighty million records, both active and dormant, with over 120 independent variables[1]. Efficient use of resources is required due to the large number of records and variables as well as the run time, storage space, transfer limitation, and program time to select the sample.

Round one of The Implementation of the Peer Review Organizations (PRO) sixth Scope of Work Pneumococcal Pneumonia and Influenza Immunization Remeasurment Survey (Remeasurment Survey) was conducted by Abt Associates Inc. (Abt) on behalf of CMS. This national survey required Abt to select from the EDB a nationally representative sample of all living Medicare beneficiaries currently enrolled in Medicare in 50 states and two territories at the time of selection. In this study there are two phases of data collection. In 2001 (phase I) data was collected from 36 states and will be collected from 16 states in 2002 (phase II). To achieve these goals, a two-phase sampling method was designed that first sampled a representative 5% sample from all eligible records (the first phase) on the EDB and then selected the survey sample from the 5% sample (the second phase).

Study Background:
The Pneumoccocal Pneumonia and Influenza Immunization Remeasurement Survey is administered by telephone to a sample of Medicare beneficiaries randomly selected from each of the 50 states and two territories. The purpose of this survey is to estimate the vaccination rates of influenza and pneumococcal pneumonia vaccines among Medicare beneficiaries. These rates will be used to evaluate the vaccine promotion work performed by Peer Review Organizations under the Medicare sixth Scope of Work (SOW). CMS administers the Peer Review Organization program, which is designed to monitor and improve utilization and quality of care for Medicare beneficiaries. The program consists of a national network of fifty-three (53) PROs (also known as Quality Improvement Organizations) responsible for each U.S. state, territory, and the District of Columbia. Each PRO maintains a staff of highly qualified, multi-disciplinary experts in medicine, quality improvement, health information management, statistical analysis, computer programming and operations, communications, public relations, and clerical/administrative support. Their mission is to ensure the quality, effectiveness, efficiency, and economy of health care services provided to Medicare beneficiaries.

Under the Peer Review Organization sixth Scope of Work (SOW) Health Care Quality Improvement Program (HCQIP), CMS has charged the PROs to decrease morbidity and mortality in six national clinical priority areas: acute myocardial infarction (AMI), heart failure, breast cancer, diabetes, stroke,

[1] This number was derived from the CMS EDB data dictionary data January 1, 1999 located at www.hcfa.gov.

The authors acknowledge the contribution of Linda Piccinino for her support and editing.

and pneumonia and influenza. Baseline rates have been supplied for all six clinical areas, and PROs are responsible for crafting and implementing interventions and demonstrating that they have achieved statistically significant improvement over the baseline rates within the three-year contract cycle.

Baseline influenza and pneumococcal pneumonia immunization rates for outpatient elderly Medicare beneficiaries' in the 50 states and 2 territories (the District of Columbia and Puerto Rico) have been obtained from the Centers for Disease Control and Prevention's (CDC) 1997 and 1999 administration of the Behavioral Risk Factor Surveillance System (BRFSS). However, the BRFSS cannot be used by the PROs for remeasurement due to mismatches between the PROs' timetable for evaluation under their contract and that of subsequent BRFSS survey administrations and data release from those administrations.

Unfortunately, the BRFSS survey did not collect immunization information in the year 2000 and this information is not currently scheduled to be collected in the year 2002. Furthermore, while the annual BRFSS survey data will be collected for influenza and pneumococcal pneumonia immunizations in the year 2001, CDC and the states' departments of health will not make this data available to any outside organization until the October 2002 at the earliest, which is too late for use by the PROs in their evaluation process. For these reasons, CMS has proposed conducting its own Pneumoccocal Pneumonia and Influenza Immunization Remeasurement Survey.

In order to obtain remeasurement survey data comparable to the 1999 BRFSS data used for the baseline, CMS and Abt replicated the 1999 BRFSS methodology as closely as possible. CMS supplied Abt with a CMS-modified version of the CDC's 1999 BRFSS questionnaire; in essence, a subset of questions from the 1999 BRFSS instruments (English and Spanish) that includes the influenza and pneumococcal pneumonia immunization questions as well as critical demographic information needed for analysis purposes were included in the survey questionnaire. To this subset of questions, Abt added questions on consent, screening and questions that identified reliable sources of information to complete the survey instrument.

Replication of the mode of administration is not so straightforward. The BRFSS is a telephone-only survey. It runs continuously on an annual cross-sectional basis, for twelve contiguous calendar months. The BRFSS telephone interviewing cycle starts every January 1. The BRFSS employs random digit dialing (via the Mitofsky-Waksberg or Disproportionate Stratified Sampling methods) to contact the general community-dwelling adult population (age >= 18 years) with residential telephones within each state or territory. Only one person per household is interviewed for the BRFSS in a given calendar year. Proxies are not allowed. Therefore, the data for the over-65 Medicare population in the annual BRFSS is a subset of the data collected from the entire sample.

In contrast, CMS, through its Medicare Enrollment Database (EDB), which contains the names and addresses of the beneficiaries and some basic insurance and demographic data, can generate a list of all elderly Medicare beneficiaries for each state and territory. Using systematic random sampling, it is then possible to select a sample of community-dwelling beneficiaries to be contacted in each state (or territory) for participation in a telephone survey. Using a two-phase sampling design (described below) Abt selected a representative sample of all living enrolled Medicare beneficiaries age 65 or older in each state.

The goals of this survey were to sample 2000 beneficiaries and complete at least 500 interviews in each state while achieving an overall 70% CASRO response rate. Each selected beneficiary was first notified of the study by United States first class mail in a pre-notification letter on the Department of Health and Human Services letterhead signed by the Chief Clinical Officer. Since telephone numbers are not maintained on the EDB, telephone number generations procedures were implemented prior to the start of data collection. Complementing these procedures a self-addressed and stamped return postcard was included with each prenotification letter that allowed potential respondents to supply a telephone number through which they can be contacted on. In addition, a study Web site (www.flustudy.org) containing the pre-notification letter, informative links, and answers to frequently asked questions was launched prior to contacting the respondents. The address to reach this Web site was included in the advance letter. Each consenting survey respondent was screened for eligibility and administered a subset of replicated BRFSS questions using Computer Administered Telephone Interviewing (CATI) to determine if subjects

The authors acknowledge the contribution of Linda Piccinino for her support and many hours of editing each draft.

received an influenza or pneumococcal pneumonia shot in the past year.

Sampling Procedures:
The target population for the survey is all enrolled living Medicare beneficiaries age 65 or older listed on the July file of the EDB maintained by CMS. The sampling frame for the first phase selection was the Medicare Enrollment Data Base. The EDB has information on age, sex, and race/ethnicity of each beneficiary in addition to their names and addresses.

For the first phase selection, a 5% systematic sample was selected from the universe. All persons in this 5% sample who are identified as out of scope were excluded from the list prior to selection. Certain other beneficiaries were also excluded from the list for second phase selection because of participation in other surveys. But the excluded persons will still be part of the first phase sample and therefore second phase weight adjustments will be made to take this into account.

The sampling frame for the selection of the second phase sample was the list constructed using the 5% sample. This list was first stratified by age. There were three age strata - which are 65-74, 75-84 and 85-115 and approximately 2000 beneficiaries were selected from each state and territory.

The second phase sample was first allocated to each stratum in proportion to the number of persons in each stratum in the first phase sample. Within each stratum, the list of beneficiaries was sorted by gender, date of birth, race/ethnicity and a systematic sample of 2000 observations was selected and allocated to stratum.

The sample was divided into five replicates of unequal size. The first replicate in each state was of size 750, the second replicate of 500, and the last three replicates were of size 250 each. To divide the overall sample into replicates, we used systematic sampling to ensure that the replicates are similar with respect to the distribution by various characteristics that were used for sample selection.

Selecting the Sample:
The first-phase read 81,533,333 EDB records and selected a 5% sample of eligible Medicare beneficiaries. This was achieved by selecting beneficiaries meeting the above requirements and having a Health Insurance Claim (HIC) number containing one of the following numerical combinations in the 8th and 9th digit of the 11 digit number: 05, 20, 45, 70 or 95. To select a 10% sample only one digit is required since numerical values of 0 to 9 naturally divide the database into ten 10% segments. Selecting a second digit further divides the database into 1% segments and therefore five pairs are required to achieve a 5% sample. It is important to note that only the 6th, 7th, 8th and 9th digits of the 11-character HIC number can be used to achieve a random selection. Numbers found in the 1st, 2nd, 3rd, 4th and 5th locations are not randomly assigned[2] and if used will not yield an unbiased sample. After identifying records with the required numerical combinations in the 8th and 9th digit, additional records were excluded if it was determined that the person was under 65 years of age, dead, or not enrolled in Medicare. These procedures resulted in selection of a first-phase sample of 1,734,892 observations. This sample was further allocated into survey rounds by state with 1,240,081, 479,570 and 15,241 observation allocated into rounds 1,2 and other[3] respectively.

In the second phase, a sample of 72,280 beneficiaries was selected from 1,204,081 in the random sample selected for the 36 round one states.

The Enrollment Database:
The EDB is CMS's repository for person-level Medicare beneficiary enrollment and entitlement records for all Medicare recipients either currently or historically enrolled in Medicare Hospital Insurance (Part A (HI)) or Supplementary Medical Insurance (Part B (SMI). The EDB file is a subset of the Social Security Administration's (SSA) Master Beneficiary Record (MBR) and is updated daily. The EDB was created in 1991 and contains benefit, enrollment, termination, claims numbers, address, demographic and other information maintained in the SSA MBR for all Medicare beneficiaries since the establishment of Medicare in

[2] HIC numbers use social security numbers as their foundation. Therefore, the first five-digit field of the HIC number is the same as the first five digits of the person's social security number. The first three numbers indicate the state in which the application was made and the next two indicates which numerical bank the last four digits (6th, 7th, 8th and 9th) that are randomly generated are stored.

[3] Round other contains observations selected in the 5% sample but that are not located in one of the 50 states or 2 territories that make up round 1 & 2.

The authors acknowledge the contribution of Linda Piccinino for her support and many hours of editing each draft.

1965[4], including Social Security Retirement and Disability Insurance, End Stage Renal Disease (ESRD), and Railroad Retirement Board (RRB) beneficiaries. Updates to the EDB are transferred from the SSA, RRB, and Common Work File (CWF) host sites as well as by CMS staff.

To ensure compliance with Privacy Act requirements request to use EDB files[5] containing individual identifiers requires execution of a "Data Use Agreement". Access to the EDB is through the EDB Workbench (EDBW) by authorized users. The Unloaded Enrollment Database (UEDB) is a flat file mirror version of the EDB updated monthly in the fourth week.

The EDB is maintained as an M204 database at the CMS Data Center (CDC) and contains two types of records: "full" and "Skeleton". A full record contains eligibility and entitlement information for all current and some historical records. Skeleton records are for inactive beneficiaries, and have much missing data.

Quarterly UEDB files are retained for 30 months with the last two quarters (6 files) maintained as monthly files.

The EDB contains over 81 million records[6] with over 129 data elements and approximately 150,000 new beneficiaries are added each month. Data can be selected using unique identifiers like the Social Security Number (SSN) or Health Insurance Claim (HIC) number. Record retrieval using SSN returns data for all family members associated with the SSN. Record retrieval using the HIC number returns records for individuals matching the HIC number.

When the status of a beneficiary changes[7] it is possible that the HIC number may change. The standard SSA format for HIC numbers is nine digits followed by one or two alpha or alphanumeric suffix combinations. All HIC numbers originating under the Railroad Retirement Benefits (RRB) Act have been updated into standard SSA format

The EDB supports various CMS and external research efforts. These combined efforts require regular access and sample selection from the EDB. CMS regularly maintains an extract drawn from the UEDB quarterly called HSKEW. HSKEW represents an extract of data items from the complete UEDB file.

This file is routinely used by CMS or made available to others for research purposes to analysis and generate statistics on Medicare enrollment data. The UEDB is stored in ten files. Each file contains HIC numbers identified using the eight digit. For example, file "A" contains HIC numbers with an eight digit of 0; file "B" contains files with an eight digit of 1 and so on. To read the entire EDB all ten files must be included.

Table 1 below, called 100% EDB and 5% UEDB July File Counts From December 31, 2000, represents the 100% and 5% counts by file of all the records contained in the EDB and the 5% sample selected from the July 2000 UEDB file. To use resources efficiently, Abt submitted a COBOL program that first read the entire UEDB and then executed a number of logic commands[8] to select all eligible records for inclusion in the 5% sample.

Table 2 below, called Comparison of 100%, 5% and Study Sample Statistics compares demographic

[4] Beneficiaries are enrolled in Medicare under Title XVIII and subsequent amendments of the Social Security Act of 1965.

[5] Other enrollment files under the EDB umbrella include the Third Party File (TPEARTH) and General Health Plan (GHP) master files, the Health Insurance Skeleton Eligibility Write-off File (HISKEW), Denominator (combines information from the EDB and TPEARTH files), and Name and Address files.

[6] Abt submitted COBOL programs on December 31, 2000 via EDBW that read all available records (active and inactive) stored on the July UEDB file.

[7] Changes in an individual's entitlement to Social Security or Railroad Retirement benefits may result in a new HIC number being assigned.

[8] The logic commands Abt used in sequential order to select eligible records for inclusion in the 5% sample were as follows: read and count all records, exclude records that do not contain one of the selected 8th and 9th digit combinations, records with a date of death (DOD), records with a DOB on or before 1935, records not enrolled in part A and/or B, not in round 1 or 2 states and keep remaining records as the 5% sample.

The authors acknowledge the contribution of Linda Piccinino for her support and many hours of editing each draft.

variables for race, age and gender across the three different list.

Table One: 100% EDB and 5% UEDB July File Counts From December 31, 2000

Number of Records with an 8th digit of	0	1	2	3	4	5	6	7	8	9	Totals
Records Read	8,157,575	8,157,132	8,155,145	8,157,550	8,150,710	8,151,386	8,147,967	8,154,742	8,153,843	8,147,283	81,533,333
Not Sampled	7,908,877	8,157,132	7,906,901	8,157,550	7,903,121	8,151,386	8,147,967	7,907,081	8,153,843	7,899,394	80,293,252
UEDB 5%	347,707		346,602		346,829			346,476		347,278	1,734,892
Round One 5%	**248,698**		**248,244**		**247,589**			**247,661**		**247,889**	**1,240,081**
Round Two 5%	95,853		95,390		96,278			95,753		96,296	479,570
Round Other	3,156		2,968		2,962			3,062		3,093	15,241
8&9th digit pair used to select the 5% sample	05		25		45			70		95	Totals

Records Read Records read from the UEDB on 12/31/2001 without exclusions

Not Sampled Records excluded for any and all reasons that include: not in the 5% sample, dead, DOB on or before 1935, not currently enrolled in part A and/or B and not in 1 of the 36 Round 1 states

UEDB 5% Records kept for sampling that make up the 5% sample

Round One 5% Records in one of the 36 round one states meeting all selection criteria and allocated to round one 5% sample

Round Two 5% Records in two of the 16 round two states meeting all selection criteria and allocated to round one 5% sample

Round Other Records in states or territories not in round one or two or invalid or unknown states

NOTE: UEBD 5% plus Excluded does not equal Records Read since Round 2 and Other are counted twice, once in Excluded and as separate categories.

Table 2: Comparison of 100%, 5% and Study Sample Statistics

Category	Percent From 100% EDB	Percent From 5% Sample	Percent from Study Sample	Category	Percent From 100% EDB	Percent From 5% Sample	Percent from Study Sample	Category	Percent From 100% EDB	Percent From 5% Sample	Percent from Study Sample	Category	Percent From 100% EDB	Percent From 5% Sample	Percent from Study Sample
Age Both Sexes/All Races--Hispanic/Non-Hispanic				Both Sexes/White--Hispanic/Non-Hispanic				Both Sexes/Black--Hispanic/Non-Hispanic				Both Sexes/Other--Hispanic/Non-Hispanic			
All Ages	100%	100%	100%	All Ages	100%	100%	100%	All Ages	100%	100%	100%	All Ages	100%	100%	100%
Ages 65-74	48%	50%	50%	Ages 65-74	47%	50%	50%	Ages 65-74	52%	49%	49%	Ages 65-74	55%	71%	61%
Ages 75-84	38%	35%	35%	Ages 75-84	38%	36%	36%	Ages 75-84	34%	33%	33%	Ages 75-84	33%	16%	22%
Ages 85-115	15%	14%	14%	Ages 85-115	15%	13%	13%	Ages 85-115	14%	17%	17%	Ages 85-115	11%	14%	17%
Male/All Races--Hispanic/Non-Hispanic				Male/White--Hispanic/Non-Hispanic				Male/Black--Hispanic/Non-Hispanic				Male/Other--Hispanic/Non-Hispanic			
All Ages	100%	100%	100%	All Ages	100%	100%	100%	All Ages	100%	100%	100%	All Ages	100%	100%	100%
Ages 65-74	53%	56%	56%	Ages 65-74	52%	56%	56%	Ages 65-74	57%	55%	54%	Ages 65-74	59%	76%	63%
Ages 75-84	37%	34%	34%	Ages 75-84	37%	35%	35%	Ages 75-84	33%	33%	33%	Ages 75-84	32%	16%	25%
Ages 85-115	10%	10%	10%	Ages 85-115	11%	9%	9%	Ages 85-115	10%	12%	13%	Ages 85-115	9%	8%	13%
Female/All Races--Hispanic/Non-Hispanic				Female/White--Hispanic/Non-Hispanic				Female/Black--Hispanic/Non-Hispanic				Female/Other--Hispanic/Non-Hispanic			
All Ages	100%	100%	100%	All Ages	100%	100%	100%	All Ages	100%	100%	100%	All Ages	100%	100%	100%
Ages 65-74	44%	47%	47%	Ages 65-74	43%	47%	47%	Ages 65-74	48%	46%	46%	Ages 65-74	52%	67%	60%
Ages 75-84	38%	36%	36%	Ages 75-84	39%	37%	37%	Ages 75-84	35%	34%	34%	Ages 75-84	34%	16%	20%
Ages 85-115	17%	17%	17%	Ages 85-115	18%	16%	16%	Ages 85-115	16%	20%	20%	Ages 85-115	13%	17%	20%

The authors acknowledge the contribution of Linda Piccinino for her support and many hours of editing each draft.

Conclusion:

We found that using a two-phase sampling design to first select a 5% sample from the UEDB and then using this list to select a study sample results in a sample that is representative of the universe in terms of age, race, and sex even without controlling for any of these factors in selection. Table two indicates that some differences exist between the 100% counts and the first phase 5% sample. However, these differences are small and can be accounted for once adjustments are made for logic commands executed to exclude records not meeting all the conditions of the desired population. For example, Table 2 indicates that for the category Age Both Sexes/All Races—Hispanic/Non-Hispanic in the 65-74 and 75-84 strata 48% and 38% of the universe reside in these strata. However, for the same categories and strata in the 5% sample 50% and 35% of the sample reside in these strata. Since counts taken from the 100% EDB include inactive and out of frame records and counts taken from the 5% sample do not, we find the variance in these statistics to be reasonable and easily explained. Comparing counts from the 5% and study sample reveal almost identical statistics with the exception of the other categories. A review of the EDB revealed that between the time the 100% counts were taken and the time the 5% sample was selected, a reclassification of records held in the other categories was completed by CMS to redistribute records based on correct race classifications.

Sources Consulted:

Arday, S. L., D. S. Arday, S. Monroe, and J. Zhang. (2000). HCFA's Racial and Ethnic Data: Current Accuracy and Recent Improvements. *Health Care Financing Review*, 21(4): 107-109, Summer 2000.

R. Hicks. Enrollment DataBase (EDB). http://www.os.dhhs.gov/progorg/aspe/minority/minhcfa1.htm.

"The Social Security Number" (Pub. No. 05-10633)

The authors acknowledge the contribution of Linda Piccinino for her support and many hours of editing each draft.