ALTERNATIVE SAMPLE DESIGNS FOR THE 1993 NATIONAL MORTALITY FOLLOWBACK SURVEY

Joe Fred Gonzalez, Jr., and Katherine B. Offutt
Joe Fred Gonzalez, Jr., National Center for Health Statistics,
6525 Belcrest Road, Hyattsville, Maryland 20782

Sampling, Precision, Design Effects

INTRODUCTION

The 1993 National Mortality Followback Survey (NMFS) will be conducted by the National Center for Health Statistics (NCHS) in collaboration with funding co-sponsors agencies. The NMFS will be a nationally representative sample of adults and adolescents age 15 and over who died in 1993. Certain subgroups will be oversampled. The 1993 NMFS sampling design will primarily comprise death certificates selected from the Current Mortality Sample (CMS) of 1993. The CMS is a 10-percent sample of the States' death certificates received by NCHS about three months after they occur; this may be a 100-percent sample by 1993. Information will be gathered on approximately 24,000 deaths.

In addition to the limited information provided on the death certificate, the NMFS will collect data from the informant (usually the decedent's spouse, child, or other close relative) listed on the death certificate. Such data will include: information on health services used by the decedent prior to death; life style characteristics of the decedent; and, socio-economic characteristics of the decedent's household. Information will also be collected from the health care facilities, identified on either the Informant Questionnaire or the death certificate, where the decedent stayed in overnight in the year preceding his/her death. This data will include: medical information about diagnoses, procedures, surgeries, and other health conditions of the decedent.

Because of the need to collect better and more detailed data on minority groups in the United States, two sample designs were proposed to oversample Blacks and Hispanics. Design I is a Black and Nonblack sample design. Design II is a Hispanic, Nonhispanic Black, and Nonhispanic Nonblack Survey. The purpose of this paper is to compare the relative efficiency of these two alternative sample designs. Expected precision will be presented for Design I for a range of estimated proportions for assumed sample sizes for 10 selected causes of death by race (Black and Nonblack) and sex. Also, expected precision will be presented for Design II which includes Hispanics as a separate sampling stratum. Design effects were calculated to approximate the effect of differential sampling rates on variance estimation for both designs.

BACKGROUND

Mortality statistics are collected from the death certificates filed in the State vital statistics office for all deaths occurring the United States every year. Presently, the States provide the data to NCHS on magnetic data tape, microfilm or copies of the certificates through cooperative agreements. Plans are underway for all States to eventually provide their data on magnetic data tapes. Because the vital registration system under which death certificates are collected is a decentralized system, the death certificates are not uniform across all States but NCHS recommends a standard certificate of death which most States follow closely.

The United States Standard Certificate of Death collects only a limited amount of statistical data about the decedent. The two major categories of data collected for statistical purposes are demographic data and medical data. The demographic data include information about the decedent's age, race, sex, geographic residency, marital status, educational attainment, occupation and industry, place of birth, and date of death. The medical data include whether an autopsy was performed, hospital status, and cause-of-death data. Both the underlying cause of death and other multiple listed causes (up to twenty) are coded for each decedent. The underlying cause of death is the disease or accident that started the chain of events leading to the death. The multiple causes of death include other causes of death listed on the death certificate but not selected as the underlying cause. These multiple causes can be directly related to the chain of events leading to death or other significant conditions complicating or contributing to the death. The causes of death are coded according to the ninth revision of the International Classification of Diseases published by the World Health Organization.

Two separate files are produced from the death certificates. The first file is a provisional file based on a systematic sample of 10% of the deaths filed in each vital statistics office each month. These data are published by NCHS three months after the month of occurrence. The second file is the final mortality file based on all deaths in United States for an entire year. This file is usually available about two years after the end of the data year.

In order to supplement these two files, periodic surveys are conducted to supplement the data

available on the death certificate, and to evaluate the validity of the data collected on the death certificate. The first mortality survey was conducted in 1961[1]. Three additional surveys[2,3,4] were conducted in the sixties. The fifth survey was conducted in 1986[5] and the 1993 survey will be the sixth National Mortality Followback Survey.

The followback surveys conducted in the sixties had limited objectives and small sample sizes. The 1961[1] survey was based on 5,154 decedents and focused on the use of hospital and other institutional care in the last year of life. The 1962-63[2] survey was based on 10,822 decedents and focused on socioeconomic characteristics and hospital and institutional care in the last year of life. The 1964-65[3] survey was based on 10,408 decedents and data was collected on health insurance coverage, socioeconomic characteristics, and hospital and institutional care in the last year of life. The 1966-68[4] survey was based on approximately 19,526 decedents ages 35-84 and focused on smoking history and socioeconomic characteristics.

The 1986 NMFS[5] was based on 18,673 decedents about 1 out of 100 deaths in the United States. The survey had four major objectives: evaluation of the quality of demographic information on the death certificate; use and access to health institution and services during the last year of life; socioeconomic status of the household; and, health and behavioral risk factors. The sample design included oversamples of blacks, Native Americans, young decedents (25 to 54 years), and specified causes (Ischemic heart disease, Asthma, and rare cancers).

PLANNING FOR THE 1993 NMFS

The objectives of the 1993 NMFS are similar to those of the 1986 NMFS. The 1993 NMFS objectives include: evaluating demographic data from the death certificate; assessing health care needs in the last year of life; determining socioeconomic differentials in mortality; studying health risk behaviors in relations to mortality; and measuring disability in the last year of life. Although assessing changes since 1986 is an important aspect of the survey, the emphasis of the survey will contrast the 1986 survey by concentrating on both external causes (homicide, suicide, and accidents) and various diseases (AIDS, chronic obstructive pulmonary disease, heart disease, and cancer). In addition, a major goal of the 1993 NMFS is to improve data availability and reliability on causes affecting young persons (homicide, suicide, AIDS, and accidents) and to allow for comparisons among these causes by both race and sex.

To maximize comparisons by cause of death by race and sex, the desire to sample an equal number of cases for each race/sex/cause group was established. In addition to making reliable estimates for the Black population, the desire to make reliable estimates for the hispanic population was also established. Maurer, et al[6] described the mortality of the Hispanic-origin population for a 15-State reporting area during the period of 1979-81, by age, sex, and cause of death. This was the first such report from the national vital statistics system, based on information reported on the death certificate. Because this 15-State reporting area accounted for only 45% of the Hispanic population in the United States, the mortality data presented in that report may not be representative of the U.S. Hispanic-origin population.

A goal of the 1993 NMFS is to expand this knowledge of Hispanic mortality by collecting more detailed information beyond that provided on the death certificate.

For the 1986 NMFS, Hispanics could not be oversampled nationally since a Hispanic-origin identifier was not included in the death certificates for all States in the United States. Beginning with the 1989 final data, Hispanic origin should be available on the death certificates for all States. Presently, Hispanic origin is not coded for provisional data. For final 1988 data, which is the most recent data year available, Hispanic origin is only available for 26 States and the District of Columbia. Because the number of deaths for Hispanics is not available for the entire U.S., these numbers were estimated by inflating the available number of deaths by cause and sex by a factor of 1.22, the ratio of the number Hispanics in the U.S. in 1980 to the number of Hispanics in the States providing Hispanic data. Although in 1980 these States accounted for 82% of the Hispanic population, the sub-Hispanic groups were not equally represented in the States providing Hispanic data. These States accounted for 91% of the Mexican population, 79% of the Puerto Rican population, 35% of the Cuban population, and 72% of the "Other Hispanics." Because these Hispanic subgroups are extremely diverse, the number of estimated deaths for some causes may not be reliable. The average annual age-adjusted death rate for Mexicans (67.1) is almost twice as high as the similar rate for Cubans (36.9) in 1979-1981 for 15 reporting States.

METHODOLOGY FOR COMPARING RELATIVE EFFICIENCIES OF DESIGNS

As mentioned earlier, the main objective of this paper is to compare the relative efficiencies of the two proposed sample designs for the 1993 NMFS.

The following is a summary of the two sample designs:

Design I

A. Total Sample Size = 24,000 death certificates

B. Population Groups: (Blacks and NonBlacks) x Both Sexes x 10 Causes of Death = 40 design cells

C. Desired sample size per cell = 24,000/40 = 600

Design II

A. Total Sample Size = 24,000 death certificates

B. Population Groups: (Hispanics, Blacks, and NonHispanic-NonBlack) x Both Sexes x 10 Causes of Death = 60 design cells

C. Desired sample size per cell = 24,000/60 = 400

The efficiencies of both proposed designs were evaluated by first assuming that the selection of sample death certificates would be from the final 1993 100% CMS data tape. Secondly, the sample size would be equally allocated to each population group by race, sex, and by 10 causes of death in order to maximize reliable comparisons. However, depending on the timeliness of the availability of the 100% CMS data, it may be necessary to select an initial sample from the 10% CMS and then select a supplemental sample from the 100% CMS, especially for those causes of death where the 10% CMS provided an insufficient sample size for the NMFS.

Estimates for the NMFS would then be produced by applying a dual frame estimation procedure which would form a weighted average of corresponding estimates derived from the 10% CMS and the 100% CMS.

For this investigation, the precision of a range (3%, 10%, 20%, and 50%) of estimated proportions was evaluated for each design. The usual unbiased estimator for the variance of a sample proportion, $\hat{p}$, derived from a finite population assuming simple random sampling (SRS), was used:

$$v(\hat{p}) = (1 - \frac{n}{N}) \frac{pq}{n}$$

Where n = sample size
N = population size
$\hat{p}$ = sample proportion
$\hat{q} = 1 - \hat{p}$
$1 - \frac{n}{N}$ = finite population correction factor (fpc)

Since the quantity pq is a maximum when p=q=.5, the following formula was used to produce conservative estimates of variances assuming srs:

$$v(\hat{p}) = (1 - \frac{n}{N}) \frac{.25}{n}$$

Due to the oversampling of Blacks and Hispanics in order to achieve equal sample sizes per design cell, there will be wide variability in the sampling weights. This differential weighting increases the variances of estimates for domains collapsed across design strata. To approximate the increase in variance for estimated proportions due to differential sampling rates, a method outlined in Kish [6] and in Westat memo written by Joseph Waksberg [7] was used to calculate design effects (DEF) resulting from disproportionate sampling.

As part of the precision analysis for both designs I and II, the following were calculated for the range of estimated proportions for each population group by sex and by cause of death: standard errors, relative standard errors, and 95% confidence intervals.
Due to space limitation on this paper, the corresponding tables for this aspect of the precision analysis are not shown, but are available from the lead author upon request.

Table 1 shows the sampling intervals for Design I that had to be applied to the race x sex x cause groups in order to achieve the desired cell sample size. Table 3 shows the required sampling intervals for Design II. As can be seen in Tables 1 and 3 there is a wide range in the sampling intervals. For example, in Table 1 for white males the sampling intervals range from 2.69 to 575.97. In table 3, for All other Non-Hispanic females the range of the sampling intervals is from 1 to 812.52. This large variability in the sampling intervals will increase the variance of estimates based on domains collapsed across design strata. Table 2 presents estimated design effects due to differential sampling in Design II. For example, in Table 2 for all races both sexes, the design effect across all causes of death is 4.2. Table 3 shows that the corresponding estimated design effect for all races both sexes is 6.1 for Design II. So even though Design II allows us to make reliable estimates for Hispanics, there is an overall

increase in the variance which is shown in the estimated design effect.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Jeffrey Maurer, Harry Rosenberg, and Kenneth Kochanek for their technical assistance in the preparation of this paper. Also, the authors are grateful to Karen Whitaker and Janice Melvin for their help in preparing this manuscript.

References

1. Sutton, Gordon F., "Hospitalization in the Last Year of Life, United States, 1961" National Center for Health Statistics. Vital and Health Statistics Series 22 (1). 1965.

2. Mathis, Evelyn S., "Socioeconomic Characteristics of Deceased Persons, United States, 1962-1963" National Center for Health Statistics. Vital and Health Statistics Series 22 (9). 1969.

3. Timmer, Elaine J., "Health Insurance Coverage of Adults Who Died in 1964 or 1965, United States" National Center for Health Statistics. Vital and Health Statistics Series 22 (10). 1969.

4. Enstrom, J.E., Godley, F.H., Cancer Mortality Among a Representative Sample of Nonsmokers in the United States during 1966-68. Journal of the National Cancer Institute, Vol. 65, No. 5, November 1980, p. 1175-83.

5. Seeman, Isadore, Gail S. Poe, and Eve Powell-Griner, " Development, Methods and Response Characteristics: 1986 National Mortality Followback Survey" In review.

6. Maurer Jeffrey D., Harry M. Rosenberg, and Joan B. Keemer, "Deaths of Hispanic Origin, 15 Reporting States." National Center for Health Statistics. Vital and Health Statistics Series 20 (18). 1990.

7. Kish, L., "Survey Sampling." John Wiley and Sons, New York, 1965.

8. Memorandum by Joseph Waksberg, Westat, December 15, 1977.

Table 1: Sampling Intervals for 1993 NMFS Based on 1987 Mortality Data, Design I

	Ages 15 and Over						
Underlying	All Races	Whites and other			Blacks		
Cause	both sexes	Total	Males	Females	Total	Males	Females
Homicide	8.33	9.11	13.24	4.97	7.55	12.13	2.97
HIV Infection	5.52	7.78	14.52	1.00	3.27	5.41	1.12
Drugs	1.70	2.04	2.69	1.00	1.34	1.53	1.00
Alcohol-induced	7.42	11.46	17.37	5.55	3.37	5.04	1.71
Suicide	14.37	23.83	37.42	10.24	2.10	2.69	1.00
Accidents	35.46	62.17	84.10	40.23	8.76	12.73	4.80
Heart Disease	316.10	567.80	575.97	559.63	64.41	64.47	64.36
All Cancer	197.96	351.94	373.03	330.85	43.98	49.66	38.30
COPD	32.56	60.99	72.73	49.25	4.14	5.43	2.84
All Other	244.30	427.05	383.94	470.16	61.54	59.11	63.96
Total	89.03	155.41	157.50	153.23	20.90	21.82	19.90

Table 2: Design Effects for the 1993 NMFS Based on 1987 Mortality Data, Design I.

	Ages 15 and Over						
Underlying	All Races				Whites and other	Blacks	
Cause	both sexes	Total	Males	Females	Total	Males	Females
Homicide	1.3	1.2	1.0	1.0	1.4	1.0	1.0
HIV Infection	2.0	1.8	1.0	1.0	1.4	1.0	1.0
Drugs	1.2	1.2	1.0	1.0	1.0	1.0	1.0
Alcohol-induced	1.6	1.3	1.0	1.0	1.2	1.0	1.0
Suicide	2.1	1.3	1.0	1.0	1.1	1.0	1.0
Accidents	1.8	1.1	1.0	1.0	1.2	1.0	1.0
Heart Disease	1.6	1.0	1.0	1.0	1.0	1.0	1.0
All Cancer	1.6	1.0	1.0	1.0	1.0	1.0	1.0
COPD	1.8	1.0	1.0	1.0	1.1	1.0	1.0
All Other	1.6	1.0	1.0	1.0	1.0	1.0	1.0
Total	4.2	2.7	2.6	2.9	2.4	2.2	2.7

Table 3: Sampling Intervals and Overall Design Effect for Design II.

	Ages 15 and Over								
Underlying	All Other Non-Hispanic			Blacks Non-Hispanic			Hispanic		
Cause	Total	Male	Female	Total	Male	Female	Total	Male	Female
Homicide	8.94	11.70	6.18	10.79	17.28	4.31	3.67	6.06	1.00
HIV Infection	9.41	17.46	1.00	4.62	7.62	1.62	3.04	4.00	1.00
Drugs	*	*	*	*	*	*	*	*	*
Alcohol-induced	*	*	*	*	*	*	*	*	*
Suicide	33.44	52.10	14.78	2.50	3.63	1.00	2.11	2.54	1.00
Accidents	82.69	110.54	54.84	12.13	17.45	6.82	6.38	10.27	2.49
Heart Disease	821.20	829.87	812.52	93.48	93.22	93.74	19.54	21.54	17.54
All Cancer	507.62	537.81	477.44	63.89	72.25	55.52	12.95	13.72	12.17
COPD	89.03	106.13	71.93	5.96	7.84	4.08	1.59	1.89	1.28
All Other	*	*	*	*	*	*	*	*	*

* Not able to estimate
Design Effect for all races both sexes = 6.1