Two Options for Oversampling in the National Health Interview Survey

Karen E. Davis, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Room 3107, Hyattsville, MD 20782

Key Words: Sample survey, oversampling

Introduction

The National Health Interview Survey (NHIS) is one of the major data collection programs of the National Center for Health Statistics (NCHS). The sample design for the NHIS traditionally has undergone a redesign every 10 years to address new and continuing data needs at both the subnational level and for minority and economic subdomains of the population. The ability to produce reliable annual estimates for the elderly population age 65-74 and 75+, by race and ethnicity, is a major design objective. Two options for oversampling sample adult (SA) elderly minority persons to meet the goal of improving the precision of estimates while retaining the same precision for non-elderly estimates and keeping the overall sample size constant will be reviewed. The first option involves selecting more than one sample adult per household in some households. The second option retains the current protocol of selecting one sample adult per household, but gives elderly minority persons an increased probability of selection. This paper describes the research that has been conducted to assess the effects of these two options, and to give a picture of the expected increase in sample yield for elderly minority persons.

Background

The National Health Interview Survey (NHIS) is a multi-purpose health survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), and is the principal source of information on the health of the civilian, noninstitutionalized, household population of the United States. The NHIS has been conducted continuously since its beginning in 1957. The data collected in the NHIS are obtained through a complex sample design involving stratification, clustering, and multistage sampling. Both the black and Hispanic populations are oversampled to allow for more precise estimation of health in these growing minority populations. Within each sample SSU all households containing black or Hispanic persons are selected for interview, while only a subsample of the other households are selected for interview. NCHS initiated a redesign of the NHIS questionnaire that was implemented in 1997. Some information is collected on all household members, and other information is obtained only for a sampled person. For example, the sample adult questionnaire requires that one sample adult (SA) per family be selected for interview.

The current design can produce annual estimates for non-Hispanic blacks and Hispanics with satisfactory precision. However, equivalent precision is not attained for subdomains defined by sex and age within a race-ethnic group (e.g.,Hispanic males and females 75+ years of age). The ability to produce reliable annual estimates for the elderly population by race and ethnicity, within the age groups 65-74 and 75+, is a major design objective (Ezzati-Rice, et al., 2001). As a result, an effort to research several methods for oversampling elderly minority persons to meet the goal of improving the precision of estimates while retaining the same precision for non-elderly estimates and keeping the overall sample size constant, was undertaken. Minority persons are defined as non-Hispanic Asians, non-Hispanic blacks, and Hispanics.

Methods

Two methods for oversampling sample adult (SA) elderly minority persons to meet the goal of improving the precision of estimates were assessed. First research was conducted to investigate the effect of the gain if we made changes to the sample adult protocol that resulted in more than one sample person per family to be selected in some cases. Data from the 1995 NHIS and the 1997 NHIS were used. The 1995 NHIS did not employ sample adult protocols, all questions were asked about all family members. The 1995 data were used to sample all eligibles in the household. Analytic variables representing 8 health characteristics (see Davis, et al., 2001) were selected from the sample adult section of the 1997 NHIS and compared to similar health conditions in the 1995 NHIS. Prior to 1997 the NHIS covered health conditions across six condition lists. All persons were not asked these questions in 1995, instead a subsample was assigned to each condition list. Although many of the 1997 condition questions are very similar to, if not identical to, those asked in the 1995 NHIS, questions are quite different for several conditions, and these changes must be considered when comparing condition prevalence estimates (National Center for Health Statistics, 2000).

Design effects for each of the 8 selected health variables were estimated using SUDAAN software for all sample adults by age, race, and ethnicity. The median design effects for the variables of interest were then calculated to provide an indication of the level of household clustering and to measure the change from the 1995 procedure to the current method. We assumed that if the sample adult protocol is expected to increase the sample adult size by a certain percentage, then the expected change in the design effect would be that percentage of the way from the median 1997 design effect to the median 1995 design effect. Nominal sample sizes were divided by design effects to obtain effective sample sizes (see Table A in appendix).

In contrast to selecting more than one adult per household, research was also conducted that retained the current protocol of selecting one sample adult per household, yet gave elderly minority persons an increased probability of selection. This method assesses the effect of increasing the selection probabilities for elderly minority persons to the sample adult protocol, to give a picture of the expected increase in sample yield for minority elderly persons. Data from the 1997 NHIS were used. We focused on the minority households that contained both elderly and non-elderly persons, since adjusting the selection probabilities in these households would have the only impact. In other words, households that did not contain both elderly and non-elderly persons could not provide the opportunity to vary the selection probabilities that would yield more elderly persons. We assumed each household was one family, and assigned each adult a probability of selection as a sample adult. Then the expected yield of elderly and non-elderly persons was calculated using SAS software to obtain the sample sizes for the current sampling protocol. Next, the elderly were given a higher chance of selection by doubling their sample count, and assigning selection probabilities based on these counts. The expected yields were calculated to obtain the sample sizes for this "doubling" sample protocol. Finally, the elderly were given an increased chance of selection by tripling their sample count, and assigning selection probabilities based on these counts. The expected yields were calculated to obtain the sample sizes for this "tripling" sample protocol. Further, for comparative purposes, we then assigned the elderly minorities a selection probability equal to one, to obtain the maximum sample yield for the elderly (see Table B).

The design effects were calculated to provide an indication of the increase in variance due to the unequal sampling rates (Kish, 1965). Table B provides the estimated design effects for the total minority sample, the expected yield of elderly persons, and the expected yield of non-elderly persons. Note that the expected yields do not include an adjustment for nonresponse, and in 1997 the response rate for sample adults was 80.4%.

Results

For the first method, where more than one sample adult is selected, the results in Table A indicate there is an increase in design effects between the 1995 protocol and the 1997 method. The net effect of household clustering and the subsampling for sample adults give an estimated 9.7% increase in design effect for All Persons. For all elderly persons, 65-74 years old, there is a 10.6% increase in design effect. However, for NH Blacks in this age category, there is an estimated 27.3% increase in design effect, and for Hispanics there is an estimated 35.5% increase. For Asians, 65 years or older, the increase in design effect is about 21.2%. Effective sample sizes are shown in Table A to show the gain as nominal sample sizes are increased.

For the second method, where elderly minority persons had an increased probability of selection, the sample yields in minority households that contained both elderly and non-elderly persons were assessed. Adjusting the selection probabilities in these households would have the only effect. The results in Table B indicate that the nominal sample sizes for the elderly minority persons increase as the selection probabilities increase, with a corresponding decrease in design effect for the elderly. Table C provides the effective sample size yields for the elderly and non-elderly minority persons. When the "doubling" sampling protocol is used, the total effective sample size decreases by an estimated 1.6% with a corresponding 13.4% increase in elderly yield. However, when the "tripling" sampling protocol is used, the total effective sample size decreases by about 7.1%, with an even greater decrease of 12.9% for NH-Asians and 11.4% for NH-Blacks. It is clear that there is a limit in the extent to which the selection probabilities can be increased for elderly minorities, due to the detrimental impact in the total effective sample yields.

Discussion

Selecting multiple sample adults per household can provide a greater increase in the elderly minority sample size, thereby improving precision of estimates. Further, this would reduce weight variability by less subsampling and provide a decrease in design effects. Research indicates an overall decrease from the current sampling protocol of about 10% in design effects. However, interviewing multiple sample adults per household would increase the interview length, thereby increasing field costs and possibly having a negative effect on household response rates. A sampling protocol with multiple sample adults could likely further reduce the sample adult response rate (currently 73.8% for the 2001 NHIS). It would necessitate a decrease in the overall sample of Non-Hispanic Others to maintain a cost neutral design. Further, new field training procedures would need to be developed for conducting multiple sample adult interviews, and new sample selection and data collection procedures would need to be developed, implemented, and tested for conducting multiple sample adult interviews.

Selecting one sample adult with an increased probability for elderly minority persons would not increase the interview length. The field interviewer would not expend additional time conducting multiple sample adult interviews, thereby maintaining current field cost levels. The current interviewing protocol would remain the same, with only minor changes in the CAPI sample selection program. Using the current protocol and moderately increasing the selection probability for the elderly will provide a substantial increase in the effective elderly sample size without too much of a penalty for the effective nonelderly sample size. However, if we take only one sample adult per household, the net effect of household clustering and subsampling would provide an increase in design effects relative to sampling more than one sample adult in some households.

Acknowledgments

Christopher Moriarity provided important comments on this research.

References

Davis, K. E., Ezzati-Rice, T., Gonzalez, J.F., Jones, C., Moriarity, C., Tompkins, L., NCHS Internal Report, *2005-2014 NHIS Sample Redesign Research Report: Current Precision of Selected Health Statistics*, May 15, 2001

Ezzati-Rice, T., Moriarity, C., Katzoff, M., Parsons, V., *Overview of Sample Design Research for the National Health Interview Survey*, 2001 Proceedings of the Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association.

Kish, L., (1965), *Survey Sampling*, New York: John Wiley and Sons, pg. 424-433

National Center for Health Statistics (2000). *Data File Documentation, National Health Interview Survey, 1997 (machine readable data file and documentation)*. National Center for Health Statistics, Hyattsville, Maryland.

Appendix

Table A. Effective Sample Size for NHIS Sample Adults at Varying Levels

	1995 Median Design Effect	1997 Median Design Effect	1997 Nominal Sample (SA)	1997 Effective Sample	Effective Sample (20% increase)	Effective Sample (30% increase)
All Persons						
All	1.39	1.54	36116	23452	28701	31405
65-74	1.18	1.32	3820	2894	3548	3886
75+	1.15	1.32	3152	2388	2941	3229
NH Blacks						
All	1.32	1.52	5087	3347	4125	4530
65-74	1.09	1.50	468	312	396	442
75+	1.01	1.35	320	237	300	333
Hispanics						
All	1.51	1.48	5685	3841	4591	4963
65-74	1.20	1.86	340	183	236	266
75+	1.13	1.20	190	158	192	209
NH Asians						
All	1.09	1.22	892	731	896	982
65+	0.93	1.18	80	68	85	94
NH Others						
All	1.31	1.43	24452	17099	20869	22803
65-74	1.14	1.25	2957	2366	2890	3159
75+	1.12	1.30	2617	2013	2484	2730

Table B. Selection of Sample Adults from Minority Households (1997 NHIS)

Current Probability of Selection

Race	Total Sample*	Elderly Yield	Non-Elderly Yield	Total Design Effect	Elderly Design Effect	Non-Elderly Design Effect
All	14169	1604	12565	2.029	2.035	2.026
Hispanic	7075	630	6445	1.646	1.664	1.643
NH-Asian	1169	111	1059	1.369	1.363	1.370
NH-Black	5925	864	5061	1.558	1.549	1.558

Double Probability of Selection for Elderly Minorities

Race	Total Sample*	Elderly Yield	Non-Elderly Yield	Total Design Effect	Elderly Design Effect	Non-Elderly Design Effect
All	14169	1752	12417	2.061	1.924	2.053
Hispanic	7075	696	6379	1.675	1.543	1.666
NH-Asian	1169	123	1046	1.420	1.233	1.404
NH-Black	5925	933	4992	1.602	1.456	1.590

Triple Probability of Selection for Elderly Minorities

Race	Total Sample*	Elderly Yield	Non-Elderly Yield	Total DesignEffect	Elderly Design Effect	Non-Elderly Design Effect
All	14169	1832	12337	2.172	1.897	2.086
Hispanic	7075	733	6342	1.773	1.514	1.694
NH-Asian	1169	130	1039	1.546	1.203	1.445
NH-Black	5925	969	4956	1.736	1.434	1.630

Maximum Probability of Selection for Elderly Minorities

Race	Total Sample	Elderly Yield	Non-Elderly Yield
All	2615	2592	23
Hispanic	1092	1080	12
NH-Asian	209	207	2
NH-Black	1314	1306	8

*Note: The total sample column shows the available nominal sample and does not reflect nonresponse.

Table C. Effective Sample Sizes of Sample Adults from Minority Households (1997 NHIS)

Current Probability of Selection

Race	Total Sample	Elderly Yield	Non-Elderly Yield
All	6984	788	6201
Hispanic	4299	378	3923
NH-Asian	854	81	773
NH-Black	3802	558	3248

Double Probability of Selection for Elderly Minorities

Race	Total Sample	Elderly Yield	Non-Elderly Yield	Total (%Change)	Elderly (%Change)	Non-Elderly (%Change)
All	6874	911	6048	-1.6	13.4	-2.5
Hispanic	4223	451	3829	-1.8	16.2	-2.5
NH-Asian	823	99	745	-3.7	18.5	-3.7
NH-Black	3697	641	3139	-2.8	13.0	-3.5

Triple Probability of Selection for Elderly Minorities

Race	Total Sample	Elderly Yield	Non-Elderly Yield	Total (%Change)	Elderly (%Change)	Non-Elderly (%Change)
All	6522	965	5914	-7.1	18.4	-4.8
Hispanic	3990	484	3743	-7.8	21.8	-4.8
NH-Asian	756	108	719	-12.9	24.7	-7.4
NH-Black	3413	676	3041	-11.4	17.5	-6.8