Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Social Security Administration

Introduction

Under an interagency agreement including the National Center for Health Statistics (NCHS), the Centers for Medicare and Medicaid Services (CMS), the Social Security Administration (SSA), and the Office of the Assistant Secretary for Planning and Evaluation, DHHS (ASPE), several NCHS population-based surveys were linked to Social Security administrative records. The linkage was undertaken to support various research initiatives of the participating agencies. The NCHS-Social Security linked files combine health and socio-demographic information from the surveys with administrative data on the receipt of retirement, survivors and disability insurance (RSDI) benefits and Supplemental Security Income (SSI) benefits obtained from the SSA. The linked files provide unique population-based information that can be used for an array of epidemiological and health services research that evaluates the needs of the elderly and persons with disabilities.

This report is intended to serve as a brief overview and to provide guidelines for using the NCHS-SSA linked data. The report describes the NCHS surveys and the Social Security administrative data files followed by a discussion of the linkage processes, linkage rates, and the linked data files. The linked NCHS-SSA data files are large and complex. Researchers are advised to read the documentation and supporting tabular data in order to understand the complexity of the data files before submitting a proposal to the RDC. Please see http://www.cdc.gov/nchs/r&d/nchs_datalinkage/data_linkage_ssa.htm.

Data Sources

National Center for Health Statistics

The following NCHS surveys were linked to Social Security benefit history data: the National Health Interview Survey (NHIS), the Second Longitudinal Study of Aging II (LSOA II), the NHANES I Epidemiologic Follow-Up Study (NHEFS), the Third

National Health and Nutrition Examination Survey (NHANES III), and the 1985 National Nursing Home Survey (NNHS).

The **NHIS** data included in the SSA linkage cover the years 1994 to 1998. The NHIS is the principal source of information on the health of the civilian, non-institutionalized population of the United States and has been conducted annually since 1957. Each year data are collected from approximately 40,000 households, including about 100,000 persons. The NHIS collects data on basic social and demographic items, health conditions and health behaviors, as well as health insurance, access to health care and utilization. In addition, the 1994 and 1995 NHIS included a supplement on disability. For detailed information on the NHIS's contents and methods, refer to http://www.cdc.gov/nchs/nhis.htm and for the NHIS Disability Survey http://www.cdc.gov/nchs/about/major/nhis_dis/nhisddes.htm.

The **LSOA II** is a prospective study of a nationally representative sample of civilian noninstitutionalized persons 70 years of age and over at the time of their 1994 NHIS interview, which served as the baseline for the study. The LSOA II study design included two follow-up telephone interviews, conducted in 1997-98 and 1999-2000. The LSOA II provides information on changes in disability and functioning, individual health risks and behaviors in the elderly, and use of medical care and services employed for assisted community living. For detailed information on the LSOA II contents and methods, refer to www.cdc.gov/nchs/about/otheract/aging/lsoa2.htm.

NHEFS is a national longitudinal study that includes the 14,407 participants who were 25-74 years of age when first examined in NHANES I (1971-75), which served as the baseline for the longitudinal follow-up study. The NHEFS study design included four follow-up interviews, conducted in 1982-84, 1986, 1987 and 1992, to investigate the relationships between clinical, nutritional, and behavioral factors assessed at baseline, and subsequent morbidity, mortality, and institutionalization. For detailed information on the NHEFS contents and methods, refer to http://www.cdc.gov/nchs/about/major/nhefs/nhefsdes.htm.

N HANES III is a nationwide probability sample of 33,994 persons ages 2 months and older and was conducted from 1988 to 1994. It was designed to provide national estimates of health and nutritional status of the civilian non-institutionalized population of the United States aged 2 months and older. NHANES III contains examination and laboratory data on diabetes, cholesterol and hypertension, dietary food recall data as well as information on health care access and utilization. For detailed information on the NHANES III contents and methods, refer to www.cdc.gov/nchs/about/major/nhanes/nh3data.htm.

The 1985 **NNHS** is the third in a series of national sample surveys of nursing homes, their residents, and their staff. The 1985 NNHS collected a variety of information about long-term care facilities and their residents. Data were collected on a sample of 11,170 patients who either were current residents (N = 5,195) at the time of contact with the facility or resident discharges (N = 5,975) that occurred within 12 months prior to the facility contact. For more information on the NNHS contents and methods, refer to http://www.cdc.gov/nchs/products/elec_prods/subject/nnhs.htm and http://wonder.cdc.gov/wonder/sci_data/surveys/nnhs/type_txt/nnhs85.asp.

Social Security Data

On August 14, 1935, President Franklin Roosevelt signed the Social Security Act into law. The new Act created a social insurance program designed to pay retired workers age 65 or older a continuing income. Social Security benefits are essential to the economic well-being of millions of individuals. Social Security benefits are paid to 90% of those 65 years and older and Social Security is the major source of income for 65% of the beneficiaries. At the end of December 2004, more than 47 million people were receiving benefits that totaled more than $493 billion annually.

The retirement, survivors, and disability social insurance program (RSDI), also known as Title II, provides monthly benefits to qualified retired and disabled workers and their dependents and to survivors of insured workers. Eligibility and benefit amounts are determined by the worker's contributions to Social Security. To become eligible for his or her benefit as well as for benefits for family members or survivors, a worker must earn a minimum number of credits (described as quarters of coverage) based on work in

covered employment or self-employment. To qualify for disability benefits, workers (excluding workers who are legally blind) must have recent work activity as well as have earned enough work credits to be eligible for Social Security benefits. For the purpose of Social Security benefits, disability refers to the inability to engage in any substantial gainful activity because of any medically determinable physical or mental impairment that can be expected to result in death or that has lasted or can be expected to last for a continuous period of not less than 12 months.

The SSA also administers the Supplemental Security Income (SSI) program, known as Title XVI, which is a needs-based program providing income support to persons aged 65 or older, blind or disabled adults, and blind or disabled children. As of 2005, over 7.1 million people were receiving federally-administered SSI payments.

Descriptions of the Social Security programs come from the Annual Statistical Supplement, 2005 and the SSA publication No. 21-059, August 2005. More information on SSA programs can be found in these documents and at www.ssa.gov.

The Social Security Administration benefit history data were extracted from three administrative records files: the Master Beneficiary Record (MBR) for the years 1962-2003, the Payment History Update System (PHUS) for the years 1984 to 2003, and the Supplemental Security Record (SSR) for the years 1974 to 2003.

The **MBR** file is the major administrative database for the Social Security RSDI program. The file includes data used to determine program eligibility as well as information for the calculation of benefit amounts and the maintenance of information about beneficiaries. The MBR contains information about each person who has applied for retirement, survivors, or disability benefits starting in 1962[1]. A MBR record is created whenever an individual applies for benefits; however, not everyone who applies will receive benefits and the MBR record will reflect the final decision about the initial claim, including

1 Although the Social Security Title II Act began in the 1930's, electronic record keeping did not begin until 1962.

denials. The MBR includes information regarding the RSDI benefit amount, payment status, dual entitlement (i.e. whether the person is entitled to benefits based on more than one person's work history), and, if applicable, information about disability entitlement, estimates and reports of earnings, and student entitlement. A list of data items can be found at http://www.cdc.gov/nchs/data/datalinkage/nchs_ssa_data_codebook.pdf.

Beginning in 1984, a portion of Social Security benefits became subject to federal income taxes. In order to provide beneficiaries with an IRS Form 1099 for income tax reporting, the total amount of benefit payments actually received per month was recorded in the **PHUS** file. The PHUS file maintains information on RSDI benefit payment amounts, including withholding information for Medicare Part B premiums. The file includes historical data on the monthly amount of Social Security benefit actually paid in a given month and only includes RSDI benefit payments.

The **SSR** file maintains information on all persons who have ever applied for SSI. Payments under SSI began in January 1974, replacing the former federal-state adult assistance program in the 50 states and the District of Columbia. Under SSI each eligible person is provided a monthly cash payment based upon statutory federal benefit rates. For those who have applied for SSI benefits, the file contains data about SSI eligibility and for those eligible for SSI, it contains basic demographic information, benefit information, actual payment amounts, as well as sources and amounts of other income information. A list of data items can be found at http://www.cdc.gov/nchs/data/datalinkage/nchs_ssa_data_codebook.pdf.

Researchers should refer to the data documentation and usage for more information on each file. In addition, researchers are encouraged to refer to the information on the use of Social Security Administration data for research purposes, please see http://www.ssa.gov/policy/docs/ssb/v65n2/v65n2p95.html.

Data Linkage

The linkage of NCHS survey respondents to their Social Security benefit history records was performed by SSA. The linkage was conducted in July 2001 and had approval from NCHS's Research Ethics Review Board[2]. The process of linking NCHS survey data with Social Security data began by matching individual survey respondents with Social Security's Numident file. The Numident file is a numerically ordered master file for each Social Security Number (SSN) ever issued and contains records for approximately 400 million SSNs, including personal identifying information.

To link NCHS survey respondents with their Social Security benefit histories, NCHS provided SSA with as many of the following individual identifiers that were available on the survey record for all eligible survey respondents:

- SSN
- Last name
- First name
- Middle initial
- Date of birth (month, day, year)
- Sex
- Father's surname (women only)
- State of birth
- Zip code

NCHS survey participants were considered ineligible for matching to the Numident file if they refused to provide their SSN at the time of the interview. Additional ineligibility criteria included refused, missing, or incomplete information on last name and date of birth.

The match process consisted of two steps. First, SSA verified whether the SSN received from NCHS was correct using the Enumeration Verification System (EVS). In the cases where the SSN of a survey participant was missing or could not be verified, SSA utilized

2 The NCHS Research Ethics Review Board, also know as an IRB, is an administrative body of scientists and non-scientists that is established to protect the rights and welfare of human research subjects.

an enhanced EVS matching algorithm to try to determine the correct SSN. The enhanced EVS matching algorithm was developed by the Office of the Actuary, SSA, to utilize additional identifying data elements collected during the survey interview and contained in other administrative records held by the SSA. It features a scoring system with a threshold score used to determine which potential matches are acceptable and provided the opportunity to increase the number of successful matches. For the NCHS records determined to be matched to the Numident file, SSA extracted data, where available, from the benefit history files. Since not all survey participants matched to the Numident have Social Security benefit history data, the records with available benefit history data is less than the number matched to the Numident.

Linkage Rates

Linkage rates are based upon successful matches to the Numident file, not to the individual SSA administrative benefit history files. The proportion ineligible for matching varied dramatically by survey, with as few as 3% to 7% of participants from the NHANES surveys being ineligible to as high as 20% to 35% of participants from the NHIS surveys. Due to the significant variation in the proportion of eligible survey respondents across surveys two linkage rates are provided: a total survey sample linkage rate and an eligible sample linkage rate. Additionally, linkage rates for each survey were examined overall and by two age groups – less than 65 years and 65 years and older. Age was defined as the survey participant's assumed age at the time of the SSA extraction (July 1, 2001).

Table1 shows for each survey, the survey sample size, the ineligible sample size, the eligible sample size, the number matched to the Numident file, and the two linkage rates. For eligible NCHS respondents, linkage rates overall were very good, about 90% for most of the surveys. Linkage rates were slightly higher for persons 65 years and older, which is to be expected since missing SSN is more common among younger persons, making them less likely to be matched. Additional information on the frequency of participants who are ineligible, not linked and linked by selected socio-demographic characteristics for each survey can be found in Tabular Data.

We are evaluating the linkage process to identify differences between respondents who are linked and those who are not linked to determine potential biases in the linked sample. Findings from this analysis will be published in a NCHS Vital Statistics Series 2 Report (expected publication date December 2006). Since the NHIS surveys have the largest proportion of ineligible respondents, the evaluation is restricted to the NHIS surveys. To determine the extent to which the linked sample reflects the larger NHIS sample population, we are examining subgroup patterns in linkage rates for the following selected socio-demographic and health characteristics: age, gender, race/ethnicity, educational attainment, family income, marital status, employment status, immigration status, health insurance, self-reported health status, activity limitation, and region. The non-linked group in all analyses includes those ineligible for linkage, as well as those eligible but not successfully linked. The ineligible sample makes up approximately 75 to 85 % of the total non-linked sample.

Data Confidentiality

The NCHS must provide safeguards for the confidentiality of its survey respondents. To ensure confidentiality, all personal identifiers have been removed from the NCHS-SSA linked data files. However, there remains the small possibility of re-identification and for this reason, the linked NCHS-SSA data are not available as public-use files. Researchers who want to access the NCHS-SSA linked data must submit a research proposal to the Research Data Center.

Data Limitations

Social Security RSDI and SSI data are extracted from files designed for program administration, and not for research. They are inherently not "user friendly" and are very complex. Users are urged to review the documentation carefully and to consult basic program information, such as the section on "SSA Program Rules" available at www.ssa.gov. In addition, see the Rand Corporation's *SSA Program Data User's Manual* (Panis et al., 2000) and the *Social Security Bulletin's Annual Statistical Supplement*.

Table 1. Sample Size and Non-response Information for NCHS Surveys Linked to Social Security Data by survey and age[1]: Unweighted Data

							Number of Respondents with Benefit History Data[3]	
	Total Person Sample	Sample Ineligible for Linking[2]	Sample Eligible for Linking	Sample Linked to SSA Numident File	Link Rate for Total Sample	Link Rate for Eligible Sample	RSDI from MBR	SSI payments from SSR
NHIS 1994	116,179	21,645	94,534	86,718	74.6%	91.7%	30,486	8,272
< 65 years	95,009	17,853	77,156	69,975	73.7%	90.7%	14,154	5,956
65+ years	21,170	3,792	17,378	16,743	79.1%	96.3%	16,332	2,316
NHIS 1995	102,467	21,034	81,433	73,492	71.7%	90.2%	24,789	7,404
< 65 years	85,874	17,824	68,050	60,653	70.6%	89.1%	12,235	5,389
65+ years	16,593	3,210	13,383	12,839	77.4%	95.9%	12,554	2,015
NHIS 1996	63,402	16,035	47,367	42,226	66.6%	89.1%	13,966	4,168
< 65 years	53,949	13,937	40,012	35,194	65.2%	88.0%	7,106	3,134
65+ years	9,453	2,098	7,355	7,032	74.4%	95.6%	6,860	1,034
NHIS 1997	103,477	32,516	70,961	62,429	60.3%	88.0%	20,465	6,275
< 65 years	88,819	28,543	60,281	52,218	58.8%	86.6%	10,497	4,709
65+ years	14,658	3,973	10,680	10,211	69.7%	95.6%	9,968	1,566
NHIS 1998	98,785	37,205	61,580	53,397	54.1%	86.7%	16,832	5,208
< 65 years	85,281	32,360	52,921	45,227	53.0%	85.5%	8,821	3,944
65+ years	13,504	4,845	8,659	8,170	60.5%	94.4%	8,011	1,264
LSOA II[4]	9,447	1,809	7,638	7,452	78.9%	97.6%	7,275	1,033
NHEFS[5]	14,407	864	13,543	12,811	88.9%	94.6%	10,342	2,243
< 65 years	4,050	227	3,823	3,652	90.2%	95.5%	1,660	539
65+ years	10,357	637	9,720	9,159	88.4%	94.2%	8,682	1,704
NHANES III	33,994	970	33,024	31,466	92.6%	95.3%	12,017	4,488
< 65 years	26,456	826	25,630	24,173	91.4%	94.3%	4,910	2,651
65+ years	7,538	144	7,394	7,293	96.7%	98.6%	7,107	1,837
NNHS[6]	11,170	627	10,543	9,841	88.1%	93.3%	9,457	3,043
< 65 years	416	22	394	359	86.3%	91.1%	287	303
65+ years	10,754	605	10,149	9,482	88.2%	93.4%	9,170	2,740

[1]Age is the participant's assummed age at the time of the linkage (July 1, 2001).

[2]Survey respondents are ineligible for linking if they refused to provide their Social Security number at the time of interview or if they are missing key identification data.

[3]Not all persons linked to the Numident will have Social Security benefit information.

[4]All persons in LSOA II are older than 65 years.

[5]NHEFS = NHANES I Epidemiologic Follow Up Study.

[6]NNHS = 1985 National Nursing Home Survey