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

March 10, 2009

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 old age, survivors and disability insurance (OASDI) 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 or match 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 analyze these data files through the NCHS RDC, the NCHS secure data access portal. 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 1994-2005 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), the 1999-2004 National Health and Nutrition Examination Survey data and the 1985, 1995, 1997, and 2004 National Nursing Home Survey (NNHS).

The **NHIS** has been conducted annually since 1957. Each year data are collected from approximately 40,000 households, including about 90,000 to 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 non-institutionalized 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.

NHANES 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.

In 1999, the National Health and Nutrition Examination Survey (NHANES) became a continuous annual survey. For a variety of reasons, including disclosure issues, the Continuous NHANES survey data is released on public-use data files in two-year increments (**NHANES 1999-2000, NHANES 2001-2002, NHANES 2003-2004**). The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year. The survey combines interviews with physical examinations.

The **National Nursing Home Survey (NNHS)** is a continuing series of nationally representative surveys of United States nursing homes, their services, staff, and residents. The NNHS was first conducted in 1973-1974 and repeated six times over the past 35 years, with the most recent release in 2004. The 1985, 1995, 1997 and 2004 NNHS were linked to the SSA administrative data. The National Nursing Home Survey provides information on nursing homes from two perspectives- that of the provider of services and that of the recipient of care. Data for the surveys were obtained through personal interviews with facility administrators and designated staff who used administrative records to answer questions about the facilities, staff, services and programs, and medical records to answer questions about the residents. For more information on the NNHS contents and methods, refer to http://www.cdc.gov/nchs/products/elec_prods/subject/nnhs.htm.

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 2006, more than 49.1 million people were receiving benefits that totaled more than $546.2 billion annually.

The old age, survivors, and disability social insurance program (OASDI), 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 2006, over 7.2 million people were receiving federally-administered SSI payments.

Descriptions of the Social Security programs come from the Annual Statistical Supplement, 2007 and the SSA publication No. 13-11700, April 2008. More information on SSA programs can be found in these documents and at www.ssa.gov.

The NCHS linked Social Security Administration benefit history data were extracted from five administrative records files: the Master Beneficiary Record (MBR) for the

years 1962-2007, the Payment History Update System (PHUS) for the years 1984 to 2007, the Supplemental Security Record (SSR) for the years 1974 to 2007, the 831 Disability Master File (831) for the years 1989 to 2007, and a special extract of summarized quarters of coverage from the Master Earnings File for the years 1953-2007.

The **MBR** file is the major administrative database for the Social Security OASDI 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 old age, 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 denials. The MBR includes information regarding the OASDI 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.

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 OASDI 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 OASDI 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.

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

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.

The 831 Disability Master file (**831 DMF**) contains the initial disability decision rendered by the Disability Determination Services (DDSs) for individuals applying for disability benefits under Title II (OASDI) and/or Title XVI (SSI). An 831 DMF record is established as soon as the DDS completes its initial disability decision. The 831 Disability file are primarily used for research on initial disability or continuing disability diagnoses and are available back to 1989.

Due to Internal Revenue Service (IRS) regulations, NCHS was only able to extract a small set of summarized annual variables regarding quarters of coverage from the Master Earnings File. These summarized variables describe an individuals' "insured status" based on their earnings history and can be found in the **Quarters of Coverage (QOC)** file made specifically for this NCHS linkage. Insured status is the minimum number of credits or quarters of coverage a worker must earn to become eligible for his or her own Social Security benefit. The QOC file contains data regarding these credits back to 1951.

A list of data items for all the linked SSA administrative files 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 2008 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 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

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

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 thus 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 number of records with benefit history data is less than the number of records 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 17% of participants from the NHANES surveys being ineligible to as high as 20% to 45% 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, 2008).

Table 1 provides the total survey sample size, the total number ineligible and eligible, the number successfully matched to the Numident file, and the total eligible sample linkage rates for each linked survey. 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 on-line at: http://www.cdc.gov/nchs/r&d/nchs_datalinkage/data_linkage_ssa.htm.

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. Due to the nature of the linked SSA benefit data, NCHS is unable to release public-use micro data files. Researchers who wish to analyze the NCHS-SSA linked data files must submit a research proposal to the Research Data Center, the NCHS secure data access portal.

Data Limitations

Social Security OASDI 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	OASDI from MBR	SSI payments from SSR	831 FILE
NHIS 1994	116,179	21,790	94,349	87,102	75.0%	92.3%	35,506	9,940	8,665
< 65 years	87,381	16,411	70,970	64,529	73.8%	90.9%	13,648	6,599	6633
65+ years	28,798	5,379	23,419	22,573	78.4%	96.4%	21,858	3,341	2032
NHIS 1995	102,467	21,209	81,258	73,813	72.0%	90.8%	29,013	8,839	7664
< 65 years	79,440	16,626	62,814	56,111	70.6%	89.3%	11,859	5,900	5807
65+ years	23,027	4,583	18,444	17,702	76.9%	96.0%	17,154	2,939	1857
NHIS 1996	63,402	16,140	47,262	42,444	66.9%	89.8%	16,498	5,094	4494
< 65 years	50,065	12,995	37,070	32,714	65.3%	88.2%	7,064	3,530	3494
65+ years	13,337	3,145	10,192	9,730	73.0%	95.5%	9,434	1,564	1000
NHIS 1997	103,477	32,285	71,192	63,213	61.1%	88.8%	24,284	7,738	6818
< 65 years	82,439	26,234	56,205	48,908	59.3%	87.0%	10,372	5,316	5309
65+ years	21,038	6,051	14,987	14,305	68.0%	95.4%	13,912	2,422	1509
NHIS 1998	98,785	37,041	61,744	54,126	54.8%	87.7%	20,226	6,408	5729
< 65 years	79,463	30,009	49,454	42,526	53.5%	86.0%	8,915	4,490	4462
65+ years	19,322	7,032	12,290	11,600	60.0%	94.4%	11,311	1,918	1267
NHIS 1999	97,059	38,889	58,170	50,160	51.7%	86.2%	18,100	5,887	5,284
< 65 years	79,035	32,075	46,960	39,833	50.4%	84.8%	8,043	4,122	4108
65+ years	18,024	6,814	11,210	10,327	57.3%	92.1%	10,057	1,765	1176
NHIS 2000	100,618	41,827	58,791	49,732	49.4%	84.6%	17,000	5,806	5165
< 65 years	83,018	34,501	48,517	40,351	48.6%	83.2%	7,866	4,144	4105
65+ years	17,600	7,326	10,274	9,381	53.3%	91.3%	9,134	1,662	1060
NHIS 2001	100,760	44,841	55,919	48,030	47.7%	85.9%	16,163	5,582	5106
< 65 years	84,194	37,689	46,505	39,311	46.7%	84.5%	7,694	4,083	4057
65+ years	16,566	7,152	9,414	8,719	52.6%	92.6%	8,469	1,499	1049
NHIS 2002[4]	93,386	23,914	69,472	53,495	57.3%	77.0%	16,643	5,690	5208
< 65 years	78,376	19,203	59,173	44,819	57.2%	75.7%	8,178	4,350	4249
65+ years	15,010	4,711	10,299	8,676	57.8%	84.2%	8,465	1,340	959
NHIS 2003	92,148	26,404	65,744	49,526	53.7%	75.3%	14,795	5,129	4693
< 65 years	78,163	21,380	56,783	42,038	53.8%	74.0%	7,495	3,945	3826
65+ years	13,985	5,024	8,961	7,488	53.5%	83.6%	7,300	1,184	867
NHIS 2004	94,460	34,708	59,752	46,151	48.9%	77.2%	13,850	4,881	4490
< 65 years	80,751	29,205	51,546	39,151	48.5%	76.0%	7,003	3,776	3662
65+ years	13,709	5,503	8,206	7,000	51.1%	85.3%	6,847	1,105	828
NHIS 2005	98,649	39,025	59,624	45,328	45.9%	76.0%	13,112	4,673	4329
< 65 years	84,930	32,988	51,942	38,878	45.8%	74.8%	6,809	3,666	3594
65+ years	13,719	6,037	7,682	6,450	47.0%	84.0%	6,303	1,007	735

[1]Age is the participant's assumed age at the time of the linkage (July 1, 2008).

[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]Beginning in 2002, NHIS changed its procedures for collecting personal identification information for survey participants. The effect of this change increases the number of NHIS survey participants who are considered eligible for record linkage but decreases the amount of identification information collected, resulting in lower match rates for the eligible population.

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	OASDI from MBR	SSI payments from SSR	831 FILE
LSOA II[4]	9,447	1,812	7,635	7,499	79.4%	98.2%	7,304	1,042	26
NHEFS[5]	14,407	867	13,540	12,974	90.1%	95.8%	11,434	2,382	1,033
< 65 years	1,906	98	1,808	1,728	90.7%	95.6%	864	285	344
65+ years	12,501	769	11,732	11,246	90.0%	95.9%	10,570	2,097	689
NHANES III	33,994	1,003	32,991	31,612	93.0%	95.8%	13,506	5,210	3,606
< 65 years	24,944	804	24,140	22,868	91.7%	94.7%	5,055	2,948	2,802
65+ years	9,050	199	8,851	8,744	96.6%	98.8%	8,451	2,262	804
NHANES 1999-2000	9,965	1,645	8,320	7,906	79.3%	95.0%	13,506	5,210	3,606
< 65 years	7,974	1,355	6,619	6,236	78.2%	94.2%	5,055	2,948	2,802
65+ years	1,991	290	1,701	1,670	83.9%	98.2%	8,451	2,262	804
NHANES 2001-2002	11,039	1,120	9,919	9,315	84.4%	93.9%	2,939	1,233	1,037
< 65 years	9,084	903	8,181	7,601	83.7%	92.9%	1,262	880	834
65+ years	1,955	217	1,738	1,714	87.7%	98.6%	1,677	353	203
NHANES 2003-2004	10,122	980	9,142	8,675	85.7%	94.9%	2,906	1,192	1,021
< 65 years	8,267	803	7,464	7,013	84.8%	94.0%	1,286	831	780
65+ years	1,855	177	1,678	1,662	89.6%	99.0%	1,620	361	241
NNHS[6] 1985	11,170	625	10,545	9,959	89.2%	94.4%	9,544	3,053	49
< 65 years	282	11	271	251	89.0%	92.6%	208	220	34
65+ years	10,888	614	10,274	9,708	89.2%	94.5%	9,336	2,833	15
NNHS[6] 1995[7]	8,056	705	7,351	6,526	81.0%	88.8%	6,299	1,935	251
< 65 years	266	46	220	192	72.2%	87.3%	155	163	79
65+ years	7,790	659	7,131	6,334	81.3%	88.8%	6,144	1,772	172
NNHS[6] 1997[7]	14,814	1,132	13,682	12,567	84.8%	91.9%	12,118	3,590	778
< 65 years	696	97	599	551	79.2%	92.0%	440	448	283
65+ years	14,118	1,035	13,083	12,016	85.1%	91.8%	11,678	3,142	495
NNHS[n] 2004′	13,507	120	13,387	13,180	97.6%	98.5%	12,739	4,012	11,925
< 65 years	1,188	38	1,150	1,129	95.0%	98.2%	946	900	704
65+ year s	12,319	82	12,237	12,051	97.8%	98.5%	11,793	3,112	878

[1]Age is the participant's assumed age at the time of the linkage (July 1, 2008).

[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 = National Nursing Home Survey.

[7] Total survey population counts are less than reported by health survey. This is due to lack of age data for purposes of these match rate tables.

Missing age at time of linkage for NNHS 1995, 1997 & 2004 survey participants (1995: n=2; 1997:n=57; 2004: n=26)