Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Centers for Medicare and Medicaid Services

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 Medicare administrative records. The linkage was undertaken to support various research initiatives of the participating agencies. The NCHS-Medicare linked files combine health and socio-demographic information from the surveys with claims information from the Medicare program, resulting in unique population-based information that can be used for an array of epidemiological and health services research.

This report provides researchers with an overview of the NCHS-Medicare linked data. The report describes the NCHS surveys and the Medicare administrative data files followed by a discussion of the linkage processes, linkage rates, and the linked data files. The linked NCHS-Medicare 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/rdc.htm.

Data Sources

National Center for Health Statistics
The following NCHS surveys were linked to Medicare 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 Second National Health and Nutrition Examination Survey (NHANES II), and the Third National Health and Nutrition Examination Survey (NHANES III).

The **NHIS** data included in the Medicare 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 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 http://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.

The **NHANES II** is a nationwide probability sample of persons from 6 months-74 years of age. The Medicare linkage includes information on 9,252 participants 30-75 years of age who completed a medical examination during the NHANES II survey period between

1976 and 1980. The NHANES II included a standardized physical examination, laboratory tests, and questionnaires that covered various health-related topics.

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.

Medicare Data

Medicare is the primary health insurance program for people age 65 or older, people under age 65 with certain disabilities, and people of all ages with End-Stage Renal Disease (permanent kidney failure requiring dialysis or a kidney transplant). Nearly all Medicare beneficiaries receive Part A hospital insurance benefits, which helps cover inpatient hospital care, skilled nursing facility stays (not custodial or long-term care), home health and hospice care. Most beneficiaries also subscribe to Part B medical insurance benefits, which help to cover physician services, outpatient care, durable medical equipment and some home health care[1].

Medicare claims information was extracted for the years 1991-2000. **Denominator** files are annual files containing demographic and enrollment information about beneficiaries enrolled in Medicare during each calendar year. It does not contain information on all beneficiaries *ever* entitled to Medicare, only those entitled during the calendar year. The Denominator file is used to determine beneficiary demographic characteristics, entitlement, and beneficiary participation in Medicare Managed Care Organizations.

---

1 http://www.cms.hhs.gov/MedicareGenInfo/

There are two Medicare files, known as MedPAR files, that contain final action claims for (1) inpatient hospitalizations stays and (2) skilled nursing facility stays (SNF). The MedPAR Hospial Stay file contains all Part A short and long stay hospitalization claims for each calendar year and the MedPAR SNF file contains claims for each skilled nursing facility stay. MedPAR Hospital Stay records are based on year of discharge, whereas MedPAR SNF records are based on year of admission. Records on both MedPAR files include ICD-9 clinical modification (CM) diagnoses and procedures codes associated with the health care facility stay.

The Outpatient files contain Part B claims data submitted by institutional outpatient providers, such as hospital outpatient departments, rural health clinics, comprehensive outpatient rehabilitation facilities, community mental health centers, and ambulatory surgical centers for each calendar year. **Home Health Care** files contain claims data submitted by Home Health Care Agency providers and include information on the number of visits, type of visit (skilled-nursing care, home health aides, physical therapy, speech therapy, occupational therapy, and medical social services), the dates of visits, reimbursement amount, HHA provider number. The **Hospice** files contain claims data submitted by Hospice providers and includes information on the level of hospice care received (e.g., routine home care, inpatient respite care), the dates of service, reimbursement amount, Hospice provider number, and beneficiary demographic information. Each Outpatient, Home Health Care and Hospice record is at the individual claim (or bill) level.

Carrier files (formerly known as the Physician/Supplier Part B file) contain claims data submitted by non-institutional providers, such as physicians, nurse practitioners, independent clinical laboratories, and stand-alone ambulatory surgical centers as well as durable medical equipment (DME) claims processed by carriers who also process physician claims. Separate **Durable Medical Equipment** (DME) files containing claims processed by authorized DME Regional Carriers (DMERC's) are also provided. DME claim records can contain claims for medical equipment such as oxygen, walkers, and wheelchairs. Information contained in the DMERC file includes diagnosis codes,

description of equipment, dates of service, and reimbursement amount. DME claims on the carrier and DMERC files are for separate services and are not duplicates. For more information on DME claims in the Carrier and DMERC files, please refer to the documentation.

Researchers should refer to the data documentation and usage for more information on each file. In addition, researchers are encouraged to refer to the Research Data Assistance Center (ResDAC), a CMS contractor providing free assistance to researchers interested in using Medicare data for their research.

Data Linkage

The linkage of NCHS survey respondents to their Medicare claims records was performed by NCHS and CMS and is not the responsibility of researchers using the data. The linkage was conducted in July 2001 and had approval from NCHS's Research Ethics Review Board[2]. The process of linking each NCHS survey with Medicare data began by matching individual survey respondents with Medicare's Enrollment Database. Medicare's Enrollment Database (EDB) is a master enrollment file of all people ever entitled to Medicare. The EDB's records are comprehensive and updated daily.

To link NCHS survey respondents with their Medicare data, NCHS provided CMS as many as of the following individual identifiers that were available on the survey record for all eligible survey respondents:

- SSN
- Health Insurance Claim (HIC) number
- Last name
- First name
- Middle initial
- Date of birth (month, day, year)

---

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.

- Sex
- Father's surname (women only)
- State of birth
- Zip code

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

CMS identified potential matches between NCHS survey participants and records in the EDB. CMS based potential matches on whether NCHS records matched EDB records on (1) HIC number, (2) SSN, or (3) name and date of birth. For these potential matches, NCHS employed a deterministic matching algorithm to determine which matches were correct. All potential matches were assigned a score based upon whether the identifying information provided matched between the NCHS and CMS records. For example, if NCHS provided a participant's SSN and it matched the SSN in a potential EDB match record, then a predetermined point value was added. If a SSN was submitted and the potential EDB match record for that participant did not match the SSN provided, no points were assigned for that identifying data element. This process was done for each of the identifying data elements submitted. Based upon this process, a total match score was established for all potential EDB matches.

Next, the scored matches were classified according to which identifying data elements matched, reflecting the fact that concordance between some identifying data elements, e.g. SSN, is more important than others. Within each class category, matches above a threshold score were considered acceptable, while matches below a threshold were considered non-matches. Matches falling between these thresholds were manually reviewed to determine the match status. For the NCHS records determined to be matched to the Medicare EDB, CMS extracted data, where available, from each of the Medicare

claims files for those records for the years 1991-2000. Since not all survey participants matched to the EDB have claims information for the years 1991-2000, the number of records with available Medicare claims data is less than the number matched to the EDB.

Linkage Rates

Linkage rates are based upon successful matches to Medicare EDB, not to the individual Medicare claims 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 40% of participants from the NHIS survey. 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 CMS extraction (July 1, 2001), because some survey participants became age eligible for Medicare after the time of their initial survey interview. For example, a respondent in the 1994 NHIS who was 59 years of age and not eligible for Medicare would be 66 to 67 years at the time of the linkage and eligible to be enrolled in Medicare.

Table 1 shows for each survey, the total survey sample size, the ineligible sample size, the eligible sample size, the number of eligible survey respondents linked to the EDB, and the total sample and eligible sample linkage rates. For NCHS respondents who were eligible for the Medicare linkage and 65 years of age or older as of July 2001, the linkage rates to the Medicare EDB were very good, ranging from a low of about 80% in the NHANES II survey to over 90% for most of the other surveys. Linkage rates for most of the surveys for those less than 65 years were about 5%. This is to be expected since Medicare eligibility for the under age 65 group is limited to persons meeting the Social Security Administration criteria for disability benefit entitlement or persons diagnosed with end-stage renal disease. 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 and Health 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 and to respondents 65 years or older at the time of the linkage. 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 65 years and older.

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-Medicare linked data files. However, there remains the small possibility of reidentification and for this reason, the linked NCHS-Medicare data are not available as public-use files. Researchers who want to obtain the NCHS-Medicare linked data must submit a research proposal to the Research Data Center.

Data Limitations

Although Medicare provides coverage for a wide range of services, there are health care services not covered by Medicare as well as a number of cost sharing requirements for Medicare beneficiaries. Examples of services not covered include routine physical

exams, long-term care, and some cancer screening procedures. These gaps in coverage and required cost-sharing mean that there are no claims records for these services or for certain time periods. You may find more information on what is not covered by Medicare at www.cms.gov or www.aarp.org/health/medicare.

CMS generally does not receive claims data for Medicare beneficiaries who enroll in managed care plans (including private fee-for-service plans paid on a capitation basis). During the time covered by the linked database, enrollment in managed care increased from approximately 6% of beneficiaries in 1991 to 17% in 1999.

Monthly managed care enrollment can be identified on the Denominator file from the HMO indicator variables (HMO_INDICATOR01- HMO_INDICATOR12) one for each month.   During periods of managed care enrollment, beneficiaries do not generate claims when using Medicare-covered services, except for selected services. Enrollees in cost-based plans may also generate some claims for inpatient hospital services. Utilization of most Medicare-covered services is unobservable from Medicare claims data during periods of managed care enrollment. Therefore, in general, studies based on analysis of claims data should exclude managed care enrollees from their beneficiary samples. For more information on how to create an analytic sample that excludes Medicare beneficiaries enrolled in a managed care plan, refer to a technical brief written by ResDAC or contact ResDAC, which provides free consultation for researchers using Medicare files,www.resdac.umn.edu.

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

	Total Person Sample	Sample Ineligible for Linking2	Sample Eligible for Linking	Sample Linked to CMS Enrollment Database (EDB)	Link Rate for Total Sample (%)	Link Rate for Eligible Sample	Number of Respondents with Information on the Medicare Denominator File[3,4]
NHIS 1994	116,179	21,647	94,532	20,102	17.3%	21.3%	17,639
< 65 years	95,009	17,855	77,154	3,980	4.2%	5.2%	1,900
65+ years	21,170	3,792	17,378	16,122	76.2%	92.8%	15,739
NHIS 1995	102,467	21,039	81,428	15,758	15.4%	19.4%	13,788
< 65 years	85,874	17,829	68,045	3,342	3.9%	4.9%	1,631
65+ years	16,593	3,210	13,383	12,416	74.8%	92.8%	12,157
NHIS 1996	63,402	16,035	47,367	8,717	13.7%	18.4%	7,562
< 65 years	53,963	13,937	40,026	1,954	3.6%	4.9%	959
65+ years	9,439	2,098	7,341	6,763	71.6%	92.1%	6,603
NHIS 1997	103,477	33,125	70,352	12,380	12.0%	17.6%	10,658
< 65 years	88,819	28,620	60,199	2,871	3.2%	4.8%	1,396
65+ years	14,653	4,504	10,149	9,505	64.9%	93.7%	9,262
NHIS 1998	98,785	37,913	60,872	9,766	9.9%	16.0%	8,360
< 65 years	85,281	32,467	52,814	2,321	2.7%	4.4%	1,143
65+ years	13,504	5,446	8,058	7,445	55.1%	92.4%	7,217
LSOA II[5]	9,447	1,923	7,524	7,235	76.6%	96.2%	7,114
NHEFS[6]	14,407	959	13,448	9,028	62.7%	67.1%	5,577
< 65 years	4,050	227	3,823	858	21.2%	22.4%	265
65+ years	10,357	732	9,625	8,170	78.9%	84.9%	5,312
NHANES II[7]	9,252	2	9,250	6,004	64.9%	64.9%	4,309
< 65 years	2,123	0	2,123	229	10.8%	10.8%	141
65+ years	7,129	2	7,127	5,775	81.0%	81.0%	4,168
NHANES III	33,994	970	33,024	8,281	24.4%	25.1%	7,484
< 65 years	26,456	826	25,630	1,194	4.5%	4.7%	608
65+ years	7,538	144	7,394	7,087	94.0%	95.8%	6,876

[1]Age is the 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 or Medicare number at the time of interview or if they are missing key identification data.

[3]Not all persons linked to the EDB will have Medicare claims information.

[4]Linkages to the Denominator file could be in any or multiple years from 1991-2000.

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

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

[7]Only the NHANES II participants who were aged 30 years and older and who agreed to a physical examination are included.