MAX and NCHS Survey Linkage, 1999–2009

Final Report

December 12, 2012

Julie Sykes
Kerianne Hourihan

MATHEMATICA
Policy Research

This page has been left blank for double-sided copying.

MAX and NCHS Survey Linkage, 1999–2009

Final Report

December 12, 2012

Julie Sykes
Kerianne Hourihan

Contract Number:
HHSM-500-2005-00025I

Task Number:
HHSM-500-T0002

Mathematica Reference Number:
06759.570

Submitted to:
Centers for Medicare & Medicaid Services
7111 Security Blvd, B2-27-00
Baltimore, MD 21244-1850
Project Officer: Cara Petroski

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Julie Sykes

MATHEMATICA
Policy Research

This page has been left blank for double-sided copying.

CONTENTS

ACRONYMS ix

I INTRODUCTION 1

II DATA SOURCES 5

A. Medicaid Data Source 5

B. Survey Data Sources 5

1. NHIS 6
2. NHANES 6
3. LSOA II 6
4. NNHS 7

C. Data-Transfer Process 7

D. Rounds of Linkage 8

III LINKAGE ALGORITHM 11

A. MAXNCHS1 11

B. MAXNCHS2 13

C. MAXNCHS3 14

D. MAXNCHS4 14

E. MAXNCHS5 14

F. File Contents 15

IV LINKAGE RESULTS 17

A. Preparation of the NCHS File 17

B. Preparation of the MAX PS File 17

C. Linking NCHS to MAX PS 17

D. Other Quality Measures 18

E. Final Record Counts 19

V CONCLUSION 27

REFERENCES 29

This page has been left blank for double-sided copying.

TABLES

IV.1 Preparation of NCHS Extract 20

IV.2 Preparation of MAX PS Files 21

IV.3 NCHS Records That Linked to a MAX PS Record 22

IV.4 Other Characteristics About NCHS Records That Linked to a MAX PS Record 23

IV.5 NCHS-MAX Record Counts 25

This page has been left blank for double-sided copying.

FIGURES

II.1 NCHS Data-Transfer Path 8

III.1 Linkage Process 12

This page has been left blank for double-sided copying.

ACRONYMS

ASPE	Assistant Secretary for Planning and Evaluation
CER	Comparative effectiveness research
CHIP	Children's Health Insurance Program
CMS	Centers for Medicare & Medicaid Services
DHHS	Department of Health and Human Services
DUA	Data use agreement
IAA	Inter-agency agreement
ID	Identification number
IP	MAX inpatient claims file
LSOA II	Second Longitudinal Study of Aging
LT	MAX institutional long-term care claims file
MAX	Medicaid Analytic eXtract
MSIS	Medicaid Statistical Information System
MSIS ID	Medicaid ID
NCHS	National Center for Health Statistics
NHANES	National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
NNHS	National Nursing Home Survey
OT	MAX other service claims file
PS	MAX person summary file
RDC	Research data center
RX	MAX prescription drug claims file
SSA	Social Security Administration

SSN	Social Security number

I. INTRODUCTION

The Centers for Medicare & Medicaid Services (CMS) have joined with the National Center for Health Statistics (NCHS), the Social Security Administration (SSA), and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) of the Department of Health and Human Services (DHHS) to support new comparative effectiveness research (CER) initiatives. One goal of this partnership is to link existing CMS administrative data files with national health-survey data collected by NCHS.

There are many advantages to linking health care administrative data with health-survey data. Medicaid administrative files, for example, contain a wealth of demographic, service-use, and payment information for Medicaid enrollees, but are not a good source for health-outcome variables or for overall descriptions of a person's health. Administrative claims files often contain diagnosis codes but only when they are associated with and recorded with the receipt of a particular medical service. Information about the progress of the disease or injury after receipt of service is usually not available from administrative data. In addition, administrative data typically does not contain information about health conditions or health-related behaviors for which an individual does not receive services.

In contrast, survey data, such as those collected by NCHS, are a rich source of information about an individual's health status, behaviors, and outcomes. Some surveys, such as the National Health and Nutrition Examination Survey (NHANES), include data from physical examinations conducted by physicians as well as blood test results and dietary-intake records. Many surveys provide information about health-risk factors such as tobacco and alcohol use, quantities and types of physical activity, sexual practices, and other behaviors that contribute to a person's health. However, survey data, by their nature, contain only limited, self-reported information about an individual's use of health care services prior to the measurement of the outcomes. They are not a good source of information regarding procedures or tests a person has received or the cost or payment source for those services. By linking administrative data with survey data, CMS, NCHS, and their partners will provide researchers with a new data source better suited to CER than either administrative or survey data alone.

During the past few years, the linked survey–administrative data sets have become available to the research community and have opened the door to a wide array of new research. For example, by comparing survey responses about Medicaid coverage to Medicaid administrative enrollment records, researchers examined the extent of and reasons for the underreporting of coverage in surveys (Davern et al. 2009; Call et al. 2012). By combining the sociodemographic and health-status information in the cross-sectional survey data with the longitudinal administrative data, researchers examined the relationship between insurance coverage before age 65 and the use of Medicare-covered services beginning at age 65 (Decker et al. 2012), and the effects of chronic obesity in adults on Medicare costs and mortality (Cai et al. 2010).

NCHS carefully protects the privacy of its survey respondents. Researchers who wish to use the linked survey–administrative data sets must apply to the NCHS for access. The researcher must specify the data sets, the years of data, and the data elements within each data set that are required for the research, and will be granted access to only that information. The work must be performed on site at one of the NCHS research data centers (RDC). Documentation going into and out of the RDC is closely scrutinized.

NCHS strongly recommends that researchers evaluate before applying whether the linked data contain a large enough sample size to generate statistically valid results. To facilitate this process, NCHS created feasibility study files. Even when the feasibility study files suggest the overall sample size is large enough, the research may not be able to be performed successfully.

Small sample sizes within the study's subpopulations may cause the predictive models to be unstable or cause certain results to be suppressed (Dodd 2012).

In this report, we focus on the linking of NCHS survey data to a set of research-oriented Medicaid administrative files, known as the Medicaid Analytic eXtract (MAX) files. In Chapter II, we describe the data sources. In Chapter III, we present the linkage algorithm. In Chapter IV, we examine the linkage results, and, finally, in Chapter V, we summarize the report and offer advice to researchers interested in using the linked NCHS-MAX files. The files were linked during the past few years and given to NCHS as they became available. NCHS, in turn, has made them available to the research community. In this report, we summarize the various rounds of linkage in one document.

This page has been left blank for double-sided copying.

II. DATA SOURCES

In this chapter, we describe the Medicaid data and NCHS surveys used in the linkage process. We describe how the data were transferred from NCHS to SSA to CMS and then back to NCHS, and we also explain why the data were linked in rounds (or batches).

A. Medicaid Data Source

The MAX files are research-oriented data files derived annually from Medicaid administrative data since 1999. There are five available MAX files for each state and calendar year. The person summary (PS) file contains one record for each person enrolled in either Medicaid or the Children's Health Insurance Program (CHIP) in the MAX calendar year. It contains eligibility and demographic information as well as summary information regarding expenditures and service use. The inpatient hospital (IP) file contains claims for inpatient hospital services. The long-term care (LT) file contains claims for long-term care received in institutions such as nursing facilities, intermediate care facilities for the mentally retarded, and psychiatric hospitals. The other services (OT) file contains claims for services provided in the community, hospitals, and long-term care facilities, as well as per-person capitation claims for services provided by managed care organizations. The prescription drug (RX) file contains claims for prescription drugs and durable medical equipment dispensed by a pharmacy.

B. Survey Data Sources

NCHS conducts a wide variety of surveys, including individual-, household-, and provider-level surveys, to measure national health and health care. For this linkage effort, NCHS extracted information about survey respondents from the following four NCHS surveys: (1) National Health Interview Survey (NHIS) 1994–2005, (2) NHANES 1999–2004, (3) Second Longitudinal Study of Aging (LSOA II) 1994–2000 (baseline and follow-up); and (4) National Nursing Home Survey (NNHS) for 2004.

1. NHIS

First conducted in 1957, NHIS is a cross-sectional household survey of a statistically representative sample of the civilian, noninstitutionalized U.S. population. It is a large sample consisting of more than 35,000 households and 75,000 individuals. Since 1982, the survey has been divided into two types of questions—core questions, which remain roughly the same from year to year, and supplemental questions, which cover a range of health topics of recent national interest. A revision of the survey in 1997 re-focused the core questions to cover demographic information as well as health status and limitations, health insurance coverage, health care utilization, and some health care behaviors. Recent supplements have included detailed questions about asthma, heart disease, immunizations, mental health, and many other topics of interest to those conducting CER (NCHS n.d.[b]).

2. NHANES

Since 1999, NHANES has been conducted as a continuous, nationally representative survey of approximately 5,000 individuals. Interviewees are asked about health status and risk factors such as alcohol and tobacco use, physical activity, sexual activity, and dietary behaviors. NHANES goes beyond traditional survey questionnaires to include body measurements, physical examinations, dental screenings, and laboratory test results. The files are released in two-year data sets (for example, 2001–2002, 2003–2004), but they contain a variety of weights designed to allow researchers to combine files across years (NCHS n.d.[a]).

3. LSOA II

The LSOA II provides a nationally representative sample consisting of nearly 10,000 civilian noninstitutionalized individuals 70 years of age and older. It consists of one baseline and two follow-up surveys. The baseline survey was conducted in conjunction with the 1994 NHIS, and the follow-up surveys were conducted in 1997–1998 and 1999–2000. Information was gathered on survivors and decedents from the 1994 baseline survey. Survey questions cover demographics, chronic health conditions, cognitive and physical impairments, health insurance coverage, and health care utilization (NCHS n.d.[g]).

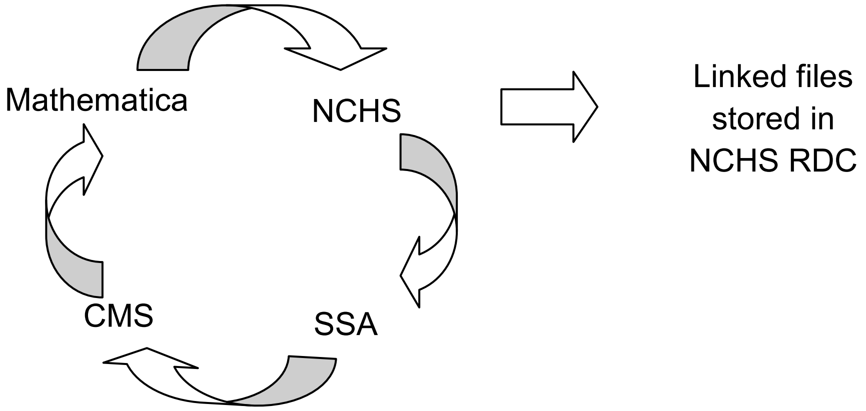
4. NNHS

The most recent NNHS, conducted in 2004, surveys both providers and recipients of care in nursing homes. Information regarding individuals residing in nursing homes is provided by facility administrators, who gather the information from the residents' medical records. The survey covers the health status of individuals, their prescribed medications, services received, and sources of payment. It also includes information about facilities, such as the size of the facility, services offered, and the facility's Medicare/Medicaid certification status (NCHS n.d.[c]).

C. Data-Transfer Process

The data-sharing agreement between NCHS, CMS and SSA to facilitate the linking of NCHS data to MAX files was defined by inter-agency agreement (IAA) 01-37 and the data use agreement (DUA) between CMS and Mathematica. In Figure II.1, we show how the data traveled between agencies, resulting in the final linked files stored in the NCHS RDC. To maintain the confidentiality of the survey data, NCHS did not send the full survey data sets to CMS. Instead, NCHS created an extract of the identification variables from each survey. To be eligible to be represented in the extract, the respondents must have provided sufficient personal identification information, including their Social Security number (SSN). The initial extract contained the NCHS linkage identification number (ID), the respondent’s name, SSN, date of birth, gender, state of birth, and zip code. For some respondents, NCHS created alternate records with slight variations in the respondent’s name, date of birth, or SSN. The extract did not contain the NCHS survey public-use ID, nor did it contain any information that SSA or Mathematica could use to determine the original survey source.

Figure II.1. NCHS Data-Transfer Path



The data from all four NCHS surveys were combined into one file and sent to SSA for verification purposes. SSA used its electronic verification system to confirm which record contained the correct SSN for each respondent. In some cases, SSA corrected the SSN or date of birth on the record. To preserve confidentiality, SSA removed the respondent's name, state of birth, and zip code before sending the revised extract to CMS. CMS loaded the extract onto its mainframe computer system. Mathematica linked the extract to the MAX files and kept only the records that successfully linked. Mathematica encrypted the linked files, put them on secure DVDs, and sent them to NCHS. From this pool of linked records, NCHS determined which ones were true matches and placed the final files in the NCHS RDC.

D. Rounds of Linkage

To balance NCHS's desire to get the linked survey–administrative data as soon as possible with its desire to have as much data as possible, Mathematica agreed to link and send the data to NCHS in four rounds (or batches). The first round included MAX data for 1999–2004 (delivered in March 2011). The second round included MAX data for 2005–2007 (delivered in April 2011). The third round included MAX data for 2008 (delivered in October 2011). The final round included MAX data for 2009 (delivered in December 2012). Unfortunately, the 2009 files provided data for only 44 jurisdictions (43 states and the District of Columbia). Seven states were not included because their Medicaid Statistical Information System (MSIS) files, which are the source files for MAX files, were unavailable or contained significant data problems. The seven excluded states are:

1. Hawaii
2. Idaho
3. Missouri
4. New Hampshire
5. Oklahoma
6. Utah
7. Wisconsin

This page has been left blank for double-sided copying.

III. LINKAGE ALGORITHM

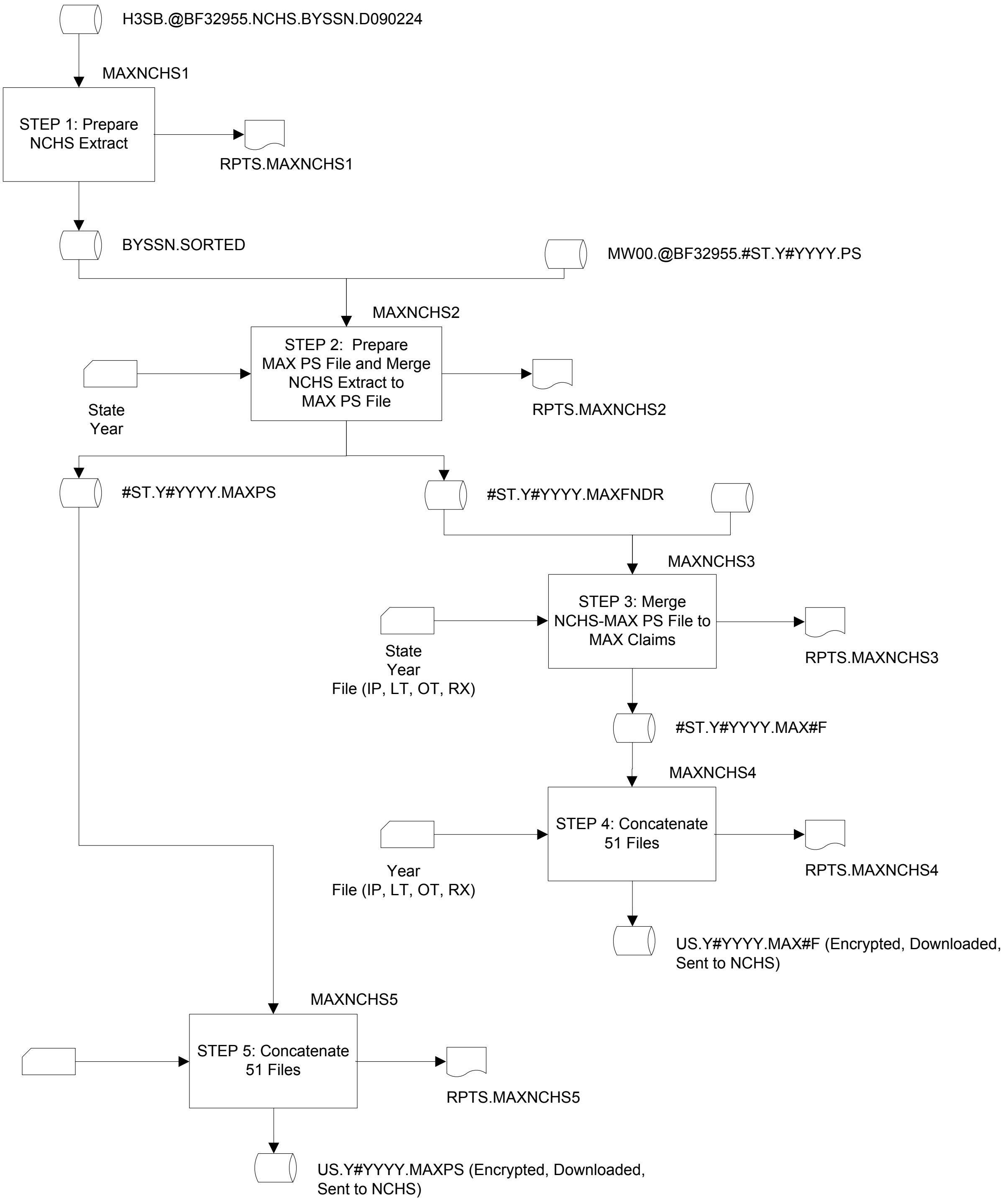
In this chapter, we describe how Mathematica created the linked NCHS-MAX files. The design is based in part on work previously performed by CMS and described in the design report (Hourihan 2010). Due to the content and structure of the NCHS extract, as revised by SSA, some changes were made to the initial linkage design. This chapter describes the final version of the design.

We divided the linkage process into five steps (Figure III.1). The first step (MAXNCHS1) prepares the NCHS extract for the linkage process. The second step (MAXNCHS2) prepares the MAX PS file for the linkage process and then merges the NCHS extract to the MAX PS file—separately for each state and year. Each record that matches to the NCHS extract will have a unique Medicaid ID (MSIS ID) within each state. The third step (MAXNCHS3) merges the MSIS IDs that matched to the NCHS extract to each of the four claims files (IP, LT, OT, and RX), separately for each state and year. The fourth step (MAXNCHS4) concatenates the 50 state-specific files and the District of Columbia into one file per claim file type and year—one file for each of the IP, LT, OT, and RX files for each year. The fifth step (MAXNCHS5) concatenates the 51 PS files for each year. We also generated quality control statistics throughout the process; the statistics will be presented in the next chapter.

A. MAXNCHS1

In the first step, we prepared the NCHS extract for the linkage process. This included a number of disparate things. We sorted the file by SSN. If the SSN was filled with 0, 8, 9, or spaces, which signify missing, we removed the record from the file. Because SSA fixed the SSN on some of the records, there were exact duplicates in the file. Because exact duplicates provide no value to the analysis and also prolong the program-execution time, we removed them from the

Figure III.1. Linkage Process


H3SB.@BF32955.NCHS.BYSSN.D090224
MAXNCHS1
STEP 1: Prepare NCHS Extract
RPTS.MAXNCHS1
BYSSN.SORTED
MW00.@BF32955.#ST.Y#YYYY.PS
MAXNCHS2
STEP 2: Prepare MAX PS File and Merge NCHS Extract to MAX PS File
State
Year
RPTS.MAXNCHS2
#ST.Y#YYYY.MAXPS
#ST.Y#YYYY.MAXFNDR
MAXNCHS3
STEP 3: Merge NCHS-MAX PS File to MAX Claims
State
Year
File (IP, LT, OT, RX)
RPTS.MAXNCHS3
#ST.Y#YYYY.MAX#F
MAXNCHS4
STEP 4: Concatenate 51 Files
Year
File (IP, LT, OT, RX)
RPTS.MAXNCHS4
US.Y#YYYY.MAX#F (Encrypted, Downloaded, Sent to NCHS)
MAXNCHS5
STEP 5: Concatenate 51 Files
RPTS.MAXNCHS5
US.Y#YYYY.MAXPS (Encrypted, Downloaded, Sent to NCHS)

Notes: Data files begin with H3SB.@BF32955.NCHS1, unless otherwise stated
#ST = state abbreviation
#YYYY = year
#F = file (IP, LT, OT, RX)

file. Because NCHS created alternate records with slight variations in the respondent's name, date of birth, or SSN, there can be more than one record in the file with the same NCHS ID. To help NCHS keep track of these duplicates, we created a flag that identifies them (DUPLICATE_FLAG = 1 (which means duplicate due to NCHS extract)). Because a survey respondent can participate in more than one survey (for example, NHIS and NNHS), that respondent can have more than one NCHS ID assigned. For the linkage process to work successfully, however, we had to ensure there was one and only one record in the file for each unique combination of SSN, year of birth, month of birth, and gender. When there was more than one record, we consolidated the records but preserved the NCHS IDs and the day of birth on the consolidated record.

B. MAXNCHS2

In the second step, we prepared the MAX PS file for the linkage process and then merged the file to the NCHS extract. During the preparation phase, we removed the PS record from the file if the SSN was filled with 0, 8, 9, or spaces, or if the date of birth was equal to zero. We took the remaining PS records and merged them to the NCHS extract by SSN, year of birth, month of birth, and gender. Among the records that matched, we kept the MSIS ID, the state code, and the preserved NCHS IDs. We also evaluated whether the day of birth matched. If it did, we set the exact-date-of-birth flag (EXACT_DOB_FLAG) equal to one. We added the flag to help NCHS identify records with an exact match on date of birth. We executed this second step separately for each state for each year.

C. MAXNCHS3

In the third step, we merged the linked NCHS-PS file created in the previous step to the MAX claims files by state code and MSIS ID.[1] We kept only the claims that matched the NCHS-PS file. We executed this third step separately for each file (IP, LT, OT, RX), state, and year.

D. MAXNCHS4

In the fourth step, we concatenated the 51 claims files into one national file. We generated one record for each NCHS ID on each claim. We recoded all of the personal identification information in MAX, including the SSN and the date of birth, to all 9s to preserve confidentiality. There was one exception. We kept the MSIS ID on the file to help NCHS identify when a Medicaid enrollee may have more than one Medicaid ID. Usually this happens when a person moves from one state to another, but there are instances in which a state assigned more than one Medicaid ID to the same person in the same year.[2] We executed this fourth step separately for each file (IP, LT, OT, RX) and year.

E. MAXNCHS5

In the final step, we concatenated the 51 PS files into one national file. We generated one record for each NCHS ID on each PS record. We kept the MSIS ID on the record and recoded the other personal identification variables. We also added two more values to the duplication flag created in the first step. We set DUPLICATE_FLAG equal to 2 if the duplicate was caused by the MAX file having more than one MSIS ID for the same person. As mentioned previously, this will occur primarily when the person is enrolled in more than one state in the same year. We

[1] MSIS IDs are unique within a state. Because we processed each state separately, the inclusion of the state code in the merge criteria was superfluous.

[2] The states have been instructed to assign one and only one Medicaid ID per person, but issues or limitations with the state's eligibility system can allow them to be generated more than once. This occurs infrequently.

set DUPLICATE_FLAG equal to 3 if the duplicate was caused by duplication in both the NCHS extract and by the MAX file having more than one record for the same person in the same year. This rarely occurred.

F. File Contents

The final set of linked NCHS-MAX files contains five files for each year. Each file contains the complete set of variables in the original MAX files, but the following personal identification information was overwritten with 9s:

1. SSN from MSIS (EL_SSN)
2. SSN from an external source (EXT_SSN)
3. Medicaid case number (EL_STATE_CASE_NUM)
4. Medicare health insurance claim number from MSIS (EL_HIC_NUM on the NCHS-MAX PS file and MDCD_HIC_NUM on the NCHS-MAX claims files)
5. Medicare health insurance claim number from Medicare enrollment database (EDB_HIC_NUM)
6. Date of birth (EL_DOB)

We added a few variables at the beginning of the files to aid NCHS in its processing of the file. To the NCHS-PS file, we added these four variables:

1. NCHS_LINKID
2. EXACT_DOB_FLAG
3. DUPLICATE_FLAG
4. ASTERISK

To the IP, LT, OT, and RX files, we added these two variables:

1. NCHS_LINKID
2. ASTERISK

The asterisk was added to denote the separation between the data elements created for NCHS and the data elements contained in MAX. The other variables were described previously in this chapter.

The data elements contained in each file vary by year because the MAX files changed periodically during 1999–2009. Specifically, MAX files for 1999–2004 have the same file layout. In 2005, many variables were added to the MAX files (all of the files) and, therefore, were added to the NCHS-MAX files. The 2006 files have the same file layouts as the 2005 files. In 2007, many variables on the RX file and a few variables on the PS file were changed; the overall record lengths did not change. And finally, the 2008 and 2009 files have the same file layouts as the 2007 files. For the complete set of MAX file layouts for 1999–2009, consult the MAX website (CMS 2012). The only difference between the file layouts on the MAX website and the actual NCHS-MAX files is the addition of the NCHS-related variables and the recoding of the personal identification variables described previously.

IV. LINKAGE RESULTS

In this chapter, we describe the linkage results. We primarily focus on the number of records that were removed, the number of duplicates, and the number written in the final files. We placed the tables at the end of the chapter to ease the readability of the text.

A. Preparation of the NCHS File

The NCHS extract, as produced by NCHS and then further modified by SSA, contained more than 800,000 records (Table IV.1). About 200 records (less than 0.1 percent) were removed from the file because they were missing an SSN or were an exact duplicate of another record in the file. Almost all of the records (99.6 percent) had one record per NCHS ID. The remaining 0.4 percent had more than one record per NCHS ID (these are the alternate records). To be linkable to the MAX files, however, we modified the NCHS extract to contain one record per unique combination of SSN, year of birth, month of birth, and gender. The final NCHS extract contained 791,553 records.

B. Preparation of the MAX PS File

The MAX PS files contained between 43.6 and 63.8 million enrollees (Table IV.2). Between 9.4 and 12.0 percent of the records were deleted because they were missing either an SSN or a date of birth. These deleted records are primarily children and persons who qualify only for emergency coverage. The final PS files contained between 39.5 and 57.0 million records. The numbers in 2009 are lower than in 2008 because seven states (Hawaii, Idaho, Missouri, New Hampshire, Oklahoma, Utah, and Wisconsin) were not included in 2009 because their MSIS files were unavailable or contained significant data problems.

C. Linking NCHS to MAX PS

The percentage of NCHS records that linked to a MAX PS record varied between 13.5 and 16.9 percent (Table IV.3). We did not expect all of the records to match because most of the survey respondents were not enrolled in Medicaid. We kept only the records that matched. Among those records, between 1.6 and 2.3 percent have more than one MAX PS record for a given NCHS record because the Medicaid enrollee is in more than one state in that year. These are people who moved and continued to be enrolled. In addition, between 0.2 and 0.5 percent have more than one MAX PS record for a given NCHS record because the state Medicaid agency assigned more than one Medicaid ID to the enrollee in that year. A typical example is a child enrolled in CHIP who then becomes enrolled in Medicaid. We did not reconcile or collapse multiple MAX PS records into one record per NCHS record, but we added the duplication flag to easily identify them.

D. Other Quality Measures

As explained in the previous chapter, before we linked the NCHS and MAX PS files, we consolidated the NCHS records to be one record per unique combination of SSN, year of birth, month of birth, and gender. We preserved the NCHS IDs on the consolidated record. Before sending the linked files to NCHS, however, we generated one record for each NCHS ID. We then tabulated various characteristics, which we deemed important for NCHS, about the linked files (Table IV.4).

First, the percentage of linked records that matched on the exact date of birth, including the day of birth, was consistently 96 percent. Only 3 to 4 percent matched using the year and month of birth but not the day (for many of these records, it was because the day was set to missing on the NCHS extract).

Second, the percentage of linked records with no claims files is between 8.0 and 11.2. These records are people who are enrolled in Medicaid but did not use any Medicaid services.

Third, the overwhelming majority of the records in the linked file (between 94.0 and 96.3 percent) had a one-to-one link between the NCHS file and the MAX PS file within each year.

The remaining records had more than one record per NCHS ID. Most of the duplicates (between 3.6 and 5.5 percent of the linked records) were caused by the MAX PS files. As mentioned earlier, this is primarily because a Medicaid enrollee is in more than one state in the same year, or, for a small number of cases, because the state Medicaid agency assigned more than one Medicaid ID to the enrollee in that year. Some of the duplicates (between 0.1 and 1.1 percent of the linked records) were caused by the NCHS ID having more than one record in the NCHS extract. And a very small number of records (between 2 and 12 of the linked records) were caused by duplicates in both the NCHS extract and the MAX PS files.

E. Final Record Counts

The final record counts for the linked NCHS-MAX records for each year are in Table IV.5. It is important to note two things. First, MAX 2009 does not include seven states (Hawaii, Idaho, Missouri, New Hampshire, Oklahoma, Utah, and Wisconsin) because their MSIS files were unavailable or contained significant data problems. Second, Maine was unable to accurately report its inpatient, long-term care, and other services in MAX 2005–2009, as the state did not have a fully functional Medicaid Management Information System. Consequently, Maine's MAX files for 2005–2009 contain only enrollment information and prescription drug claims.

Table IV.1. Preparation of NCHS Extract

Description	Number	Percent
Records in the initial extract file	802,868	100.0
Records deleted due to missing SSN	143	0.0
Records deleted due to exact duplicate of another record	66	0.0
Records with unique NCHS IDs	799,657	99.6
Records in the final extract file (unique SSN, year of birth, month of birth, gender)	791,553	98.6

Table IV.2. Preparation of MAX PS Files

Description	MAX 1999	MAX 2000	MAX 2001	MAX 2002	MAX 2003	MAX 2004	MAX 2005	MAX 2006	MAX 2007	MAX 2008	MAX 2009
Number of records in the initial person summary file	43,587,109	46,334,482	50,078,316	55,063,856	57,638,889	60,244,146	61,429,538	61,661,641	61,673,088	63,842,647	63,307,408
Number deleted due to missing SSN	4,088,349	5,032,270	5,917,272	6,557,086	6,480,799	6,724,150	6,537,005	6,796,452	6,979,841	6,822,820	6,475,650
Number deleted due to missing date of birth	22,939	26,919	20,649	25,883	19,253	30,363	27,620	16,420	4,394	5,704	15,884
Percent deleted	9.4	10.9	11.9	12.0	11.3	11.2	10.7	11.0	11.3	10.7	10.3
Number of records in the final person summary file	39,475,821	41,275,293	44,140,395	48,480,887	51,138,837	53,489,633	54,864,913	54,848,769	54,688,853	57,014,123	56,815,874

Note: MAX 2009 does not include the following 7 states because their MSIS files were unavailable or contained significant data problems: Hawaii, Idaho, Missouri, New Hampshire, Oklahoma, Utah, and Wisconsin.

Table IV.3. NCHS Records That Linked to a MAX PS Record

Description	MAX 1999	MAX 2000	MAX 2001	MAX 2002	MAX 2003	MAX 2004	MAX 2005	MAX 2006	MAX 2007	MAX 2008	MAX 2009
Total number of NCHS records	791,762	791,762	791,762	791,762	791,762	791,762	791,762	791,762	791,762	791,762	791,762
Number of NCHS records linked to a MAX PS record	115,947	117,409	121,416	128,920	132,033	133,794	130,831	123,523	116,635	114,468	107,207
Percent of NCHS records linked to a MAX PS record	14.6	14.8	15.3	16.3	16.7	16.9	16.5	15.6	14.7	14.5	13.5
Number linked to multiple MAX PS records, different state, same year	2,228	2,389	2,675	2,963	2,892	2,777	2,758	2,554	2,170	2,027	1,733
Percent linked to multiple MAX PS records, different state, same year	1.9	2.0	2.2	2.3	2.2	2.1	2.1	2.1	1.9	1.8	1.6
Number linked to multiple MAX PS records, same state, same year	268	457	501	632	364	404	340	249	223	187	192
Percent linked to multiple MAX PS records, same state, same year	0.2	0.4	0.4	0.5	0.3	0.3	0.3	0.2	0.2	0.2	0.2

Note: MAX 2009 does not include the following 7 states because their MSIS files were unavailable or contained significant data problems: Hawaii, Idaho, Missouri, New Hampshire, Oklahoma, Utah, and Wisconsin.

Table IV.4. Other Characteristics About NCHS Records That Linked to a MAX PS Record

Description	MAX 1999	MAX 2000	MAX 2001	MAX 2002	MAX 2003	MAX 2004	MAX 2005	MAX 2006	MAX 2007	MAX 2008	MAX 2009
Number of NCHS records linked to a MAX PS record	118,521	120,360	124,687	132,637	135,399	137,103	134,029	126,402	119,102	116,766	109,182
Number linked using exact date of birth	114,021	115,882	120,129	128,021	130,747	132,454	129,617	122,364	115,325	113,167	105,840
Percent linked using exact date of birth	96.2	96.3	96.3	96.5	96.6	96.6	96.7	96.8	96.8	96.9	96.9
Number with no MAX claims files	9,449	10,244	10,703	11,937	11,684	12,280	12,407	13,306	12,608	13,112	11,869
Percent with no MAX claims files	8.0	8.5	8.6	9.0	8.6	9.0	9.3	10.5	10.6	11.2	10.9
Number with one-to-one link between MAX and NCHS	112,165	113,515	117,391	124,652	128,243	130,210	127,428	120,539	114,140	112,151	105,175
Percent with one-to-one link between MAX and NCHS	94.6	94.3	94.1	94.0	94.7	95.0	95.1	95.4	95.8	96.0	96.3
Number with duplicated NCHS ID due to duplication in the MAX PS Files	5,051	5,764	6,406	7,270	6,592	6,468	6,276	5,661	4,843	4,495	3,886
Percent with duplicated NCHS ID due to duplication in the MAX PS Files	4.3	4.8	5.1	5.5	4.9	4.7	4.7	4.5	4.1	3.8	3.6
Number with duplicated NCHS ID due to duplication in the NCHS Finder File	1,295	1,071	878	703	552	417	319	196	117	116	117
Percent with duplicated NCHS ID due to duplication in the NCHS Finder File	1.1	0.9	0.7	0.5	0.4	0.3	0.2	0.2	0.1	0.1	0.1
Number with duplicated NCHS ID due to duplication in both NCHS and MAX Files	10	10	12	12	12	8	6	6	2	4	4

Table IV.4. (continued)

Notes: MAX 2009 does not include the following 7 states because their MSIS files were unavailable or contained significant data problems: Hawaii, Idaho, Missouri, New Hampshire, Oklahoma, Utah, and Wisconsin.

We consolidated the NCHS records to be one record per unique combination of SSN, year of birth, month of birth, and gender.

We preserved the NCHS IDs on the consolidated record. This table reflects the counts after we generated one record for each NCHS ID, which is why the total count in this table is higher than in Table IV.3.

Table IV.5. NCHS-MAX Record Counts

Description	MAX 1999	MAX 2000	MAX 2001	MAX 2002	MAX 2003	MAX 2004	MAX 2005	MAX 2006	MAX 2007	MAX 2008	MAX 2009
Number of linked NCHS-MAX PS records	118,521	120,360	124,687	132,637	135,399	137,103	134,029	126,402	119,102	116,766	109,182
Number of linked NCHS-MAX IP records	18,633	20,460	21,225	20,963	22,675	21,104	19,134	18,252	17,348	17,951	16,596
Number of linked NCHS-MAX LT records	165,068	170,805	173,449	185,907	209,626	263,590	301,605	236,821	197,461	176,732	144,193
Number of linked NCHS-MAX OT records	3,191,144	3,375,283	3,656,379	3,981,336	4,231,353	4,211,532	4,248,965	4,181,431	4,177,127	4,376,051	4,421,685
Number of linked NCHS-MAX RX records	1,796,094	1,879,446	2,017,224	2,194,037	2,464,761	2,752,030	2,729,182	1,075,119	1,017,169	1,038,509	1,009,002

Notes: MAX 2009 does not include the following 7 states because their MSIS files were unavailable or contained significant data problems: Hawaii, Idaho, Missouri, New Hampshire, Oklahoma, Utah, and Wisconsin.

Maine was unable to accurately report their inpatient, long-term care, and other services in MAX 2005-2009, as they did not have a fully functional Medicaid Management Information System. Consequently, Maine's MAX files for 2005-2009 only contain enrollment information and prescription drug claims.

We consolidated the NCHS records to be one record per unique combination of SSN, year of birth, month of birth, and gender. We preserved the NCHS IDs on the consolidated record. This table reflects the counts after we generated one record for each NCHS ID, which is why the total count in this table is higher than in Table IV.3.

V. CONCLUSION

In this report, we described the Medicaid and NCHS data sources, how we linked those files, and the linkage results. The linkage algorithm and the quality of the linkage variables (SSN, date of birth, and gender) drove the results. There was no additional or corroborating information that could be used to validate whether the matches were true matches or not. NCHS, with its access to the person's name and other personal information, performed the validation and removed false matches from each file before making the files available in the NCHS RDC.

Researchers interested in using the linked files at the NCHS RDC should go to the NCHS linkage website (NCHS n.d.[d]). The analytical guide published by NCHS, which describes the NCHS-MAX linked files, can help researchers formulate their study designs (Simon et al. 2011). The feasibility study files can be useful for determining if there is adequate sample size for specific research objectives (NCHS n.d.[e]). And finally, the application process to obtain access to the linked files is easily accessible on the NCHS website (NCHS n.d.[f]).

This page has been left blank for double-sided copying.

REFERENCES

Cai, Liming, James Lubitz, Katherine M. Flegal, and Elsie R. Pamuk. “The Predicted Effects of Chronic Obesity in Middle Age on Medicare Costs and Mortality.” Medical Care, vol. 48, no. 6, June 2010, pp. 510–517.

Call, Kathleen T., Michael E. Davern, Jacob A. Klerman, and Victoria Lynch. “Comparing Errors in Medicaid Reporting Across Surveys: Evidence to Date.” Health Services Research, forthcoming, published online ahead of print, July 20, 2012.

Centers for Medicare & Medicaid Services. “Medicaid Analytic Extract (MAX) General Information.” Available at [http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html]. Accessed November 2012.

Davern, Michael, Jacob A. Klerman, David K. Baugh, and George D. Greenberg. “An Examination of the Medicaid Undercount in the Current Population Survey: Preliminary Results from Record Linking.” Health Services Research, vol. 44, no. 3, June 2009, pp. 965–987.

Decker, Sandra L., Jalpa A. Doshi, Amy E. Knaup, and Daniel Polsky. “Health Service Use Among the Previously Uninsured: Is Subsidized Health Insurance Enough?” Health Economics, vol. 21, no. 10, October 2012, pp. 1155–1168.

Dodd, Allison Hedley. “Using the MAX-NHANES Merged Data to Evaluate the Association of Obesity and Medicaid Costs.” Washington, DC: Mathematica Policy Research, forthcoming 2012.

Hourihan, Kerianne. “MAX and NCHS Survey Linkage Design Report.” Washington, DC: Mathematica Policy Research, June 2010.

NCHS, Centers for Disease Control and Prevention. “About the National Health and Nutrition Examination Survey.’’ n.d. (a). Available at [http://www.cdc.gov/nchs/nhanes/about_nhanes.htm]. Accessed November 2012.

NCHS, Centers for Disease Control and Prevention. “About the National Health Interview Survey.” n.d.(b). Available at [http://www.cdc.gov/nchs/nhis/about_nhis.htm]. Accessed November 2012.

NCHS, Centers for Disease Control and Prevention. “About the National Nursing Home Survey.” n.d.(c). Available at [http://www.cdc.gov/nchs/nnhs/about_nnhs.htm]. Accessed November 2012.

NCHS, Centers for Disease Control and Prevention. “Data Access: NCHS Data Linked to CMS Medicaid Enrollment and Claims Files.” n.d.(d). Available at [http://www.cdc.gov/nchs/data_access/data_linkage/cms_medicaid.htm]. Accessed November 2012.

NCHS, Centers for Disease Control and Prevention. "Data Access: NCHS-CMS Medicaid Feasibility Study Files." n.d.(e). Available at [http://www.cdc.gov/NCHS/data_access/data_linkage/cms/cms_medicaid_feasibility.htm]. Accessed November 2012.

NCHS, Centers for Disease Control and Prevention. "Research Data Center." n.d.(f). Available at [http://www.cdc.gov/rdc/]. Accessed November 2012.

NCHS, Centers for Disease Control and Prevention. "The Second Longitudinal Study of Aging (LSOA II)." n.d.(g). Available at [http://www.cdc.gov/nchs/lsoa/lsoa2.htm]. Accessed November 2012.

Simon, A. E., A. K. Driscoll, C. Golden, R. Tandon, C. R. Duran, E. A. Miller, K. C. Schoendorf, and J. D. Parker. "Documentation and Analytical Guidelines for NCHS Surveys Linked to Medicaid Analytic eXtract (MAX) Files." Hyattsville, MD: National Center for Health Statistics, November 2011. Available at [http://www.cdc.gov/nchs/data/datalinkage/documentation_and_analytic_guidelines_nchs_survey_max_linked_data.pdf]. Accessed November 2012.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC