The Centers for Medicare & Medicaid Services' Office of Research, Development, and Information (ORDI) strives to make information available to all. Nevertheless, portions of our files including charts, tables, and graphics may be difficult to read using assistive technology.

Persons with disabilities experiencing problems accessing portions of any file should contact ORDI through e-mail at ORDI_508_Compliance@cms.hhs.gov.

# MAX and NCHS Survey Linkage Design Report

Final Report

June 30, 2010

Kerianne Hourihan

MATHEMATICA
Policy Research, Inc.

This page is intentionally left blank.

# MAX and NCHS Survey Linkage Design Report

Final Report

June 30, 2010

Kerianne Hourihan

Contract Number:
CMS Contract HHSM-500-2005-00025I,
Task Order HHSM-500-T0002

Mathematica Reference Number:
06759.550

Submitted to:
Centers for Medicare & Medicaid Services
ORDI, Mail Stop C3-23-05
DRSPSP
7500 Security Blvd
Baltimore, MD 21244-1850
Project Officer: Susan Radke

Submitted by:
Mathematica Policy Research
600 Maryland Avenue, SW
Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Julie Sykes

MATHEMATICA
Policy Research, Inc.

This page is intentionally left blank.

# CONTENTS

# TABLES

This page is intentionally left blank.

# FIGURES

This page is intentionally left blank.

# ACRONYMS

| | |
|---|---|
| ASPE | Assistant Secretary for Planning and Evaluation |
| BIC | Beneficiary Identification Code |
| CDC | Centers for Disease Control and Prevention |
| CER | Comparative effectiveness research |
| CHIP | Children’s Health Insurance Program |
| CLIST | Command list |
| COBOL | Common Business-Oriented Language |
| CMS | Centers for Medicare & Medicaid Services |
| DASD | Direct access storage device |
| DHHS | Department of Health and Human Services |
| DOB | Date of birth |
| DUA | Data use agreement |
| EDB | Medicare Enrollment Database |
| HIC | Medicare Health Insurance Claim number |
| IAA | Inter-Agency Agreement |
| ID | Identification number |
| IP | MAX inpatient claims file |
| JCL | Job control language |
| LSOA II | Second Longitudinal Study of Aging |
| LT | MAX institutional long-term care claims file |
| MAX | Medicaid Analytic eXtract |
| MAX-PDQ | MAX Production, Enhancement, and Data Quality |
| MSIS | Medicaid Statistical Information System |
| 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

## A. Background

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) in support of 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. This report focuses on one particular effort within that goal—linking NCHS survey data to a set of research-oriented Medicaid files known as Medicaid Analytic eXtract (MAX) files. CMS has contracted with Mathematica Policy Research to undertake this effort as part of the MAX-Production, Enhancement, and Data Quality (PDQ) contract. The resulting files will be added to a growing list of linked survey-administrative datasets housed in the NCHS Research Data Center (RDC). Researchers who wish to use the data may apply to NCHS for access.

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 of health outcome variables or 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 regarding 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 that a person has received, nor the cost or payment source for those services. By linking administrative data with survey data, CMS, NCHS, and their partners will provide researchers with an exciting new data source better suited to CER than either administrative or survey data alone.

## B. 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)[1] 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

[1] CHIP records are available only if the state includes CHIP enrollment in their quarterly MSIS files.

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 prescribed by a pharmacist.

## C. Survey Data Sources

NCHS conducts a wide variety of surveys, including individual-, household-, and provider-level surveys, to measure national health and health care. The first round of links between MAX files and NCHS are already underway, using data primarily from four NCHS surveys: (1) National Health Interview Survey (NHIS) 1994-2005, (2) NHANES 1999-2004, (3) Second Longitudinal Study of Aging (LSOA II) with baseline data collected in 1994; and (4) National Nursing Home Survey (NNHS) for 2004.[2] NCHS provided data from all four of these surveys in one combined file. If additional rounds of linking are conducted, NCHS may choose to provide a new file with data from a different set of surveys.

### 1. NHIS

First conducted in 1957, NHIS is a cross-sectional household survey of a statistically representative sample of the civilian, non-institutionalized U.S. population. 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 that of 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

[2] The available NCHS survey data also include some respondents from much older NCHS surveys. Since MAX data are only available for 1999 and later, NCHS may choose to limit the data available through the RDC to the survey responses after 1998.

other topics of interest to those conducting CER. Information is collected about the household, the family, one sample adult in the household, and one sample child in the household (NCHS. About NHIS). The 2005 sample consisted of 38,509 households (CDC/NCHS, 2006).

### 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 datasets (for example, 2001–2002, 2003–2004), but they contain a variety of weights designed to allow researchers to combine files across years (NCHS. About NHANES).

### 3. LSOA II

The LSOA II is a nationally representative sample consisting of nearly 10,000 civilian non-institutionalized individuals aged 70 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 both 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. LSOA II).

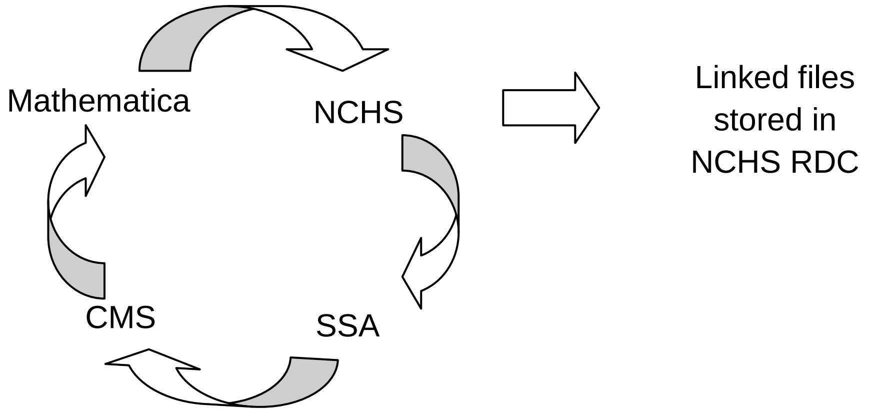
### 4. NNHS

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

## D. Data Transfer Process

The inter-agency data sharing agreement to facilitate the linking of NCHS data to MAX files is defined by Inter-Agency Agreement (IAA) 01-37 and the MAX-PDQ data use agreement (DUA). Figure I.1 shows how the data travels between agencies, resulting in the final linked files stored in the NCHS RDC. To maintain the confidentiality of the survey data, NCHS does not send the full survey datasets to CMS. Instead, NCHS creates a finder file[3] from each survey. The finder file contains only the NCHS linkage identification number (ID) and other identifying variables, such as Social Security number (SSN) and date of birth, which are necessary for linking.

**Figure I.1. NCHS Data Transfer Path**



The finder file does not contain the NCHS survey public-use ID, nor does it contain any information that Mathematica could use to determine the original survey source. The data from

[3] A finder file is a file containing only the identification variables for a particular population or dataset. Finder files are frequently used to find and extract data for a subpopulation from a larger dataset.

all four NCHS surveys are combined into one file, which NCHS sends to SSA for verification or assignment of SSNs using the SSA Numerical Identification System (Numident) file. After SSA completes the verification process, the NCHS finder file is sent to CMS and loaded onto their mainframe computer system. The records in the NCHS finder file are then linked to the MAX files and a subset of each MAX file is created, containing only the records that have been successfully linked to the NCHS finder file. The files are then encrypted, written on secure DVDs, and sent to NCHS. NCHS will complete the final link between the MAX subsets and the unique survey data sources. The resulting linked files will be housed in the NCHS RDC, where researchers can analyze the data either onsite or through a remote connection after obtaining permission from NCHS.

## E. Schedule and Current Progress

CMS has contracted with Mathematica to conduct three rounds of linking NCHS data with MAX. Table I.1 provides a detailed schedule for each round of MAX-NCHS links. The first round will combine the pre-selected NCHS survey data, described in Section C, with every MAX file between 1999 and 2009. The finder file for the first round of NCHS data was prepared and transferred to CMS in 2008. The current agreement, IAA 01-37, specifies that there must be three years between data transfers. Therefore, Mathematica anticipates receiving a finder file for the second round of NCHS data in 2011. If CMS, NCHS, and their partners decide to modify the IAA or develop a new IAA to allow for more frequent data transfer, Mathematica may also conduct a third round of linking under the MAX-PDQ contract.

Prior to the start of the MAX-PDQ contract, CMS completed part of the first round by linking the NCHS finder file with the MAX files for 1999 to 2004. Those links were made using an SSN-based NCHS finder file containing 802,868 records. NCHS also provided a secondary finder file (995 records) that used the Medicare Health Insurance Claim (HIC) number rather than the SSN as the primary identifier. However, due to the small number of records and to limited resources at CMS, the secondary file has not yet been linked to MAX 1999-2004.

**Table I.1. NCHS-MAX Linkage Schedule and Deliverables**

| Deliverables | Schedule |
|---|---|
| **First Round of Links** | |
| Link NCHS data to MAX 1999-2009 | September 2010–December 2011 |
| Assessment Report | February 2012 |
| **Second Round of Links** | |
| Link NCHS data to MAX 1999-2010 | January 2012–November 2012 |
| Assessment Report | January 2013 |
| **Third Round of Links (if IAA permits)** | |
| Link NCHS data to MAX 1999-2011 | January 2013–June 2013 |
| Assessment Report | August 2013 |
| **Summary of Linkage Activities** | |
| Final Assessment Report | August 2013 |

Mathematica plans to modify the software provided by CMS according to the recommendations in Chapter III of this report. Mathematica will then re-create the links between the SSN-based NCHS finder file and the 1999-2004 MAX files. Mathematica will also link the HIC-based NCHS finder file to MAX 1999-2004. Mathematica will then continue the series, linking both of the first round NCHS finder files to MAX 2005 through MAX 2009. Since MAX data through 2007 are already available, Mathematica will complete the links between the first round of NCHS data and MAX 1999-2007 and provide CMS with a memorandum presenting an interim assessment of the linkage results. The links between the first-round NCHS data and MAX 2008 will be completed after the MAX 2008 data are available. Finally, the links between the first-round NCHS data and MAX 2009 will be completed after MAX 2009 data are available. A report summarizing the first-round linkage activities and assessing the results will be delivered to CMS in February 2012.

The second round of NCHS links to MAX files, based on the new NCHS data expected to be available in 2011, is scheduled to begin in January 2012. That data will be linked to MAX 1999 through 2010. Since MAX 2010 production will be underway at the same time, Mathematica will begin by linking the new NCHS data to the earliest MAX year (1999) and work forward. The second round links, between MAX (1999-2010) and the second NCHS finder file(s), are expected to be completed in November 2012. If a third round of links is possible, either through a modified IAA or through a new IAA, the third round would begin in December 2012 and would cover MAX 1999 through the latest MAX year available in June 2013.

## F. Organization of the Design Report

The second chapter of this report describes the linking algorithm and software that CMS employed in matching NCHS records to MAX records. The third chapter describes Mathematica's plan for modifying the software and enhancing software tracking and documentation. It also describes the data quality checks that Mathematica will perform and the analytic files that NCHS will create to support researchers wanting to use the linked data.

# II. CMS LINKAGE DESIGN AND SOFTWARE

At the start of the MAX-PDQ contract, CMS provided Mathematica with the software used in linking the first SSN-based NCHS finder file to MAX 1999 through 2004. This chapter describes the linking algorithm used by CMS and provides complete documentation of each program in the software system provided by CMS.

## A. Linking Algorithm

Due to the large volume of records in each MAX file, CMS linked the SSN-based NCHS finder file separately for each state by file type and calendar year. CMS then created a subset of each MAX file containing only the records that successfully linked to the NCHS data, and combined the subsets into one national file for each file type and calendar year.

The first step in the CMS linking algorithm was to remove records with missing identifiers. Most NCHS records in the first-round (SSN-based) finder file contained an SSN that was verified or assigned by SSA, although a small subset of records (0.1%) had a missing SSN. In contrast, MAX records with missing SSNs are more common because many states are not able to provide SSNs for some enrollees. For example, in 2006, approximately 10 percent of MAX PS records nationwide had a missing SSN. The majority of those without an SSN are children, persons who qualify only for family planning benefits, and aliens who qualify only for emergency coverage. CMS removed records without SSNs from each MAX PS file before attempting to link it to the NCHS finder file.

The second step was to link the NCHS finder file to each MAX PS file. In the software provided by CMS, the files were merged and only records for which there was an exact match on SSN, date of birth, and sex were kept. The output file contained one record for every NCHS record that was successfully linked to the MAX PS file, and included all of the NCHS finder file variables as well as the full set of MAX PS variables. From that output file, CMS created a file containing only the Medicaid Statistical Information System (MSIS) ID and state abbreviation for the successfully linked records. CMS then used the list of MSIS IDs to extract the associated claims from each of the four MAX claims files from the respective state. Files of the same MAX file type and calendar year were then combined across all states, resulting in a single national file for each file type and year.

## B. Software

There were 14 programs in the MAX-NCHS linking system provided by CMS.[4] The programs were designed to run in the CMS mainframe environment and were written primarily in Job Control Language (JCL), Common Business-Oriented Language (COBOL), and SAS. The software also relied heavily on mainframe utility programs, including SYNCSORT and CSSELDUP, for sorting datasets. Many of the programs were controlled by command lists (CLIST), which allowed the same code to be executed efficiently across multiple states and calendar years. Table II.1 lists the CMS programs and their input and output files, while Figure II.1 presents a flow chart of the entire system.

### 1. SORTSSNF

SORTSSNF was used to sort the NCHS finder file by SSN, full date of birth, and sex. The input file was the original file provided by NCHS, including zero-filled SSNs, and the output file was a sorted copy of the NCHS finder file.

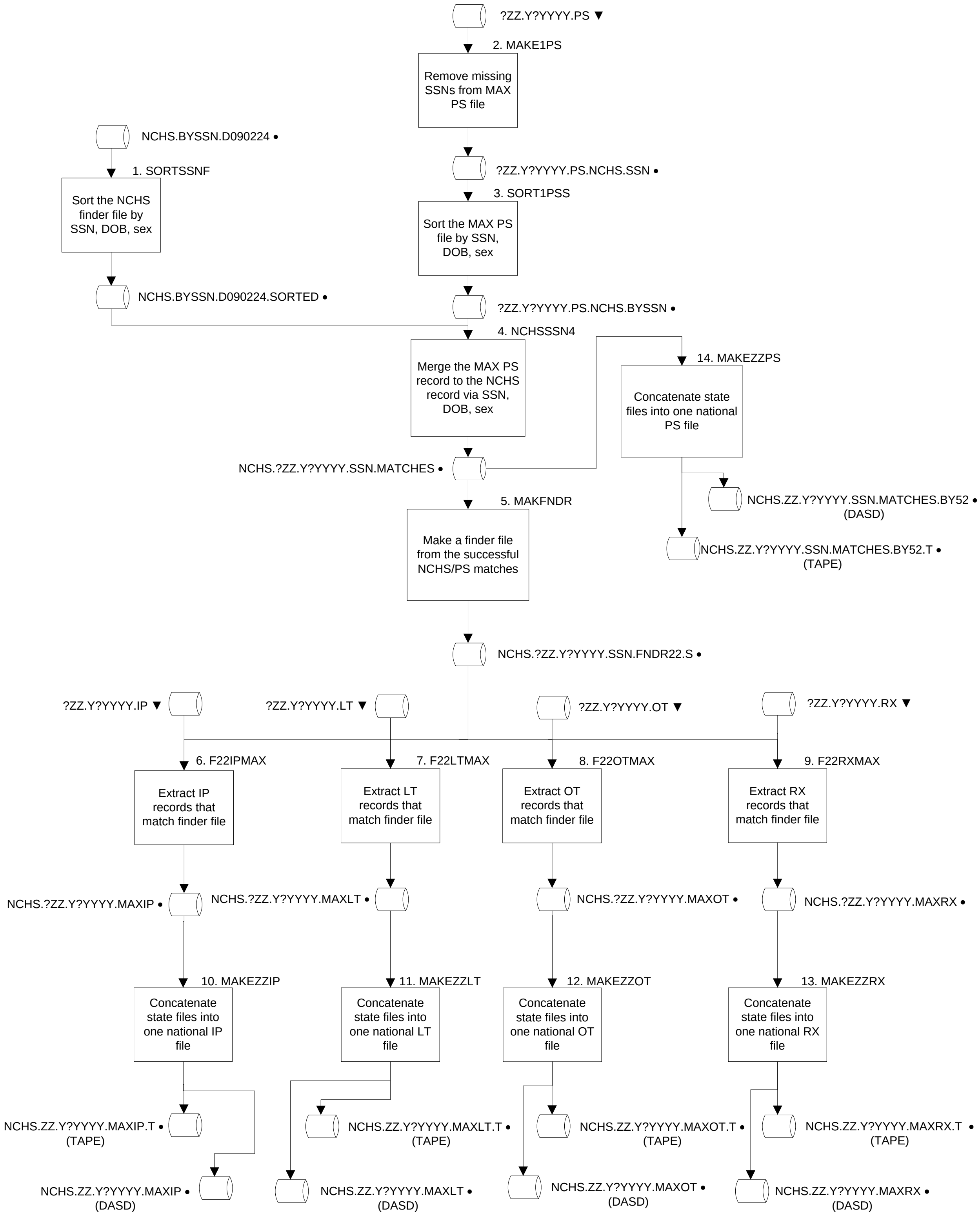
[4] A few additional programs written by CMS were used for handling special cases such as non-numeric non-alphabetic characters in MSIS IDs. Those programs are not covered in detail in this report.

**Table II.1. List of Programs in CMS MAX-NCHS Linking Software**

| # | Program | Language/Utility | Purpose | Input(s) | Output(s) |
|---|---|---|---|---|---|
| 1 | SORTSSNF | SYNCSORT | Sort NCHS finder file by SSN, DOB, sex | XXXX.NCHS.BYSSN.D090224 | XXXX.NCHS.BYSSN.D090224.SORTED |
| 2 | MAKE1PS | CSSELDUP | Remove missing SSNs from MAX PS files | MW00.@BF32955.?ZZ.Y?YYYY.PS | XXXX.?ZZ.Y?YYYY.PS.NCHS.SSN |
| 3 | SORT1PSS | SYNCSORT | Sort MAX PS files by SSN, DOB, sex | XXXX.?ZZ.Y?YYYY.PS.NCHS.SSN | XXXX.?ZZ.Y?YYYY.PS.NCHS.BYSSN |
| 4 | NCHSSSN4 | COBOL | Merge MAX PS record to NCHS records via SSN, DOB, sex | XXXX.?ZZ.Y?YYYY.PS.NCHS.BYSSN<br>XXXX.NCHS.BYSSN.D090224.SORTED | XXXX.NCHS.?ZZ.Y?YYYY.SSN.MATCHES |
| 5 | MAKFNDR | SAS | Make a list of MSIS IDs corresponding to the successful NCHS/PS matches | XXXX.NCHS.?ZZ.Y?YYYY.SSN.MATCHES | XXXX.NCHS.?ZZ.Y?YYYY.SSN.FNDR22.S |
| 6 | F22IPMAX | SAS | Extract IP records that match finder file | XXXX.NCHS.?ZZ.Y?YYYY.SSN.FNDR22.S<br>MW00.@BF32955.?ZZ.Y?YYYY.IP | XXXX.NCHS.?ZZ.Y?YYYY.MAXIP |
| 7 | F22LTMAX | SAS | Extract LT records that match finder file | XXXX.NCHS.?ZZ.Y?YYYY.SSN.FNDR22.S<br>MW00.@BF32955.?ZZ.Y?YYYY.LT | XXXX.NCHS.?ZZ.Y?YYYY.MAXLT |
| 8 | F22OTMAX | SAS | Extract OT records that match finder file | XXXX.NCHS.?ZZ.Y?YYYY.SSN.FNDR22.S<br>MW00.@BF32955.?ZZ.Y?YYYY.OT | XXXX.NCHS.?ZZ.Y?YYYY.MAXOT |
| 9 | F22RXMAX | SAS | Extract RX records that match finder file | XXXX.NCHS.?ZZ.Y?YYYY.SSN.FNDR22.S<br>MW00.@BF32955.?ZZ.Y?YYYY.RX | XXXX.NCHS.?ZZ.Y?YYYY.MAXRX |
| 10 | MAKEZZIP | CSSELDUP<br>SYNCSORT | Concatenate state files into one national IP file | XXXX.NCHS.?ZZ.Y?YYYY.MAXIP | XXXX.NCHS.ZZ.Y?YYYY.MAXIP<br>XXXX.NCHS.ZZ.Y?YYYY.MAXIP.T |
| 11 | MAKEZZLT | CSSELDUP<br>SYNCSORT | Concatenate state files into one national LT file | XXXX.NCHS.?ZZ.Y?YYYY.MAXLT | XXXX.NCHS.ZZ.Y?YYYY.MAXLT<br>XXXX.NCHS.ZZ.Y?YYYY.MAXLT.T |
| 12 | MAKEZZOT | CSSELDUP<br>SYNCSORT | Concatenate state files into one national OT file | XXXX.NCHS.?ZZ.Y?YYYY.MAXOT | XXXX.NCHS.ZZ.Y?YYYY.MAXOT<br>XXXX.NCHS.ZZ.Y?YYYY.MAXOT.T |
| 13 | MAKEZZRX | CSSELDUP<br>SYNCSORT | Concatenate state files into one national RX file | XXXX.NCHS.?ZZ.Y?YYYY.MAXRX | XXXX.NCHS.ZZ.Y?YYYY.MAXRX<br>XXXX.NCHS.ZZ.Y?YYYY.MAXRX.T |
| 14 | MAKEZZPS | CSSELDUP<br>SYNCSORT | Concatenate state files into one national PS file | XXXX.NCHS.?ZZ.Y?YYYY.SSN.MATCHES | XXXX.NCHS.ZZ.Y?YYYY.SSN.MATCHES.BY52<br>XXXX.NCHS.ZZ.Y?YYYY.SSN.MATCHES.BY52.T |

?ZZ = State abbreviation
?YYYY = MAX calendar year
Files with names beginning with XXXX are stored under a personal CMS user ID.

## Figure II.1.  Flow of CMS NCHS-MAX Linking Software


?ZZ.Y?YYYY.PS ▼
2. MAKE1PS
Remove missing SSNs from MAX PS file
NCHS.BYSSN.D090224 •
1. SORTSSNF
Sort the NCHS finder file by SSN, DOB, sex
?ZZ.Y?YYYY.PS.NCHS.SSN •
3. SORT1PSS
Sort the MAX PS file by SSN, DOB, sex
NCHS.BYSSN.D090224.SORTED •
?ZZ.Y?YYYY.PS.NCHS.BYSSN •
4. NCHSSSN4
Merge the MAX PS record to the NCHS record via SSN, DOB, sex
14. MAKEZZPS
Concatenate state files into one national PS file
NCHS.?ZZ.Y?YYYY.SSN.MATCHES •
NCHS.ZZ.Y?YYYY.SSN.MATCHES.BY52 • (DASD)
5. MAKFNDR
Make a finder file from the successful NCHS/PS matches
NCHS.ZZ.Y?YYYY.SSN.MATCHES.BY52.T • (TAPE)
NCHS.?ZZ.Y?YYYY.SSN.FNDR22.S •
?ZZ.Y?YYYY.IP ▼
?ZZ.Y?YYYY.LT ▼
?ZZ.Y?YYYY.OT ▼
?ZZ.Y?YYYY.RX ▼
6. F22IPMAX
7. F22LTMAX
8. F22OTMAX
9. F22RXMAX
Extract IP records that match finder file
Extract LT records that match finder file
Extract OT records that match finder file
Extract RX records that match finder file
NCHS.?ZZ.Y?YYYY.MAXIP •
NCHS.?ZZ.Y?YYYY.MAXLT •
NCHS.?ZZ.Y?YYYY.MAXOT •
NCHS.?ZZ.Y?YYYY.MAXRX •
10. MAKEZZIP
11. MAKEZZLT
12. MAKEZZOT
13. MAKEZZRX
Concatenate state files into one national IP file
Concatenate state files into one national LT file
Concatenate state files into one national OT file
Concatenate state files into one national RX file
NCHS.ZZ.Y?YYYY.MAXIP.T • (TAPE)
NCHS.ZZ.Y?YYYY.MAXLT.T • (TAPE)
NCHS.ZZ.Y?YYYY.MAXOT.T • (TAPE)
NCHS.ZZ.Y?YYYY.MAXRX.T • (TAPE)
NCHS.ZZ.Y?YYYY.MAXIP • (DASD)
NCHS.ZZ.Y?YYYY.MAXLT • (DASD)
NCHS.ZZ.Y?YYYY.MAXOT • (DASD)
NCHS.ZZ.Y?YYYY.MAXRX • (DASD)

?ZZ = state abbreviation
?YYYY = MAX calendar year

▼Filename begins with MW00.@BF32955
• Filename begins with CMS User ID

### 2. MAKE1PS

MAKE1PS was used to remove records with missing SSNs from a MAX PS file. The input file was a single MAX PS file for one state and one calendar year. The output was a temporary file containing the MAX PS records that had a non-missing SSN. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 3. SORT1PSS

SORT1PSS was used to sort the MAX PS file by SSN, full date of birth, and sex. The input file was the temporary MAX PS file, produced by MAKE1PS (program 2), with the missing SSNs removed. The output file was a sorted copy of the MAX PS file. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 4. NCHSSSN4

NCHSSSN4 was the primary linking program. There were two input files: the sorted NCHS finder file, created by SORTSSNF (program 1), and the sorted MAX PS file for a single state and year, created by SORT1PSS (program 3). The program searched through the two files and compared the SSNs until it found records with matching SSNs. It next compared the full date of birth and sex on each pair of records. If the program found a record with matching SSN, date of birth, and sex, it wrote a record to the output file containing the full NCHS record, followed by an asterisk, followed by the full MAX PS record. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 5. MAKFNDR

MAKFNDR was used to make a new MAX-based finder file containing only the MSIS ID and state for each record in the MAX PS file that matched a record in the NCHS finder file. It sorted the new finder file and removed duplicate MSIS IDs. The input file was the NCHS/PS file for a single state and year, as created by NCHSSSN4 (program 4). The output file contained only the MSIS ID and state abbreviation for each record. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 6. F22IPMAX

F22IPMAX was used to extract the MAX IP claims records that corresponded to the finder file created by program MAKFNDR (program 5). There were two input files: the state- and year-specific finder file created by MAKFNDR and the MAX IP claims file for a single state and year. The output file was a subset of the claims file, containing every MAX IP record belonging to MSIS IDs listed in the finder file. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 7. F22LTMAX

F22LTMAX was used to extract the MAX LT claims records that corresponded to the finder file created by MAKFNDR (program 5). There were two input files: the state- and year-specific finder file created by MAKFNDR and the MAX LT claims file for a single state and year. The output file was a subset of the claims file, containing every MAX LT record belonging to MSIS IDs listed in the finder file. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 8. F22OTMAX

F22OTMAX was used to extract the MAX OT claims records that corresponded to the finder file created by MAKFNDR (program 5). There were two input files: the state- and year-specific finder file created by MAKFNDR and the MAX OT claims file for a single state and year. The output file was a subset of the claims file, containing every MAX OT record belonging to MSIS IDs listed in the finder file. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 9. F22RXMAX

F22RXMAX was used to extract the MAX RX claims records that corresponded to the finder file created by MAKFNDR (program 5). There were two input files: the state- and year-specific finder file created by MAKFNDR and the MAX RX claims file for a single state and year. The output file was a subset of the claims file, containing every MAX RX record belonging to MSIS IDs listed in the finder file. This program was repeated, using a CLIST, for every state and calendar year in MAX 1999-2004.

### 10. MAKEZZIP

MAKEZZIP was used to concatenate the 51 state-specific IP claims extracts into one national IP file for one MAX calendar year.[5] The input files were state-specific IP claims extracts created by F22MAXIP (program 6). The national IP file for the calendar year, with multiple records per person, was written out in two formats (tape and direct access storage device (DASD)).

### 11. MAKEZZLT

MAKEZZLT was used to concatenate the 51 state-specific LT claims extracts into one national LT file for one MAX calendar year. The input files were the state-specific LT claims extracts created by F22MAXLT (program 7). The national LT file for the calendar year, with multiple records per person, was written out in two formats (tape and DASD).

### 12. MAKEZZOT

MAKEZZOT was used to concatenate the 51 state-specific OT claims extracts into one national OT file for one MAX calendar year. The input files were the state-specific OT claims

[5] Each of the MAKEZZ claims programs also created a national file with duplicate MSIS IDs removed, resulting in a file with only one claim per person. This file was used for a different project and is not needed for the NCHS linking task.

extracts created by F22MAXOT (program 8). The national OT file for the calendar year, with multiple records per person, was written out in two formats (tape and DASD).

### 13. MAKEZZRX

MAKEZZRX was used to concatenate the 51 state-specific RX claims extracts into one national RX file for one MAX calendar year. The input files were the state-specific RX claims extracts created by F22MAXRX (program 9). The national RX file for the calendar year, with multiple records per person, was written out in two formats (tape and DASD).

### 14. MAKEZZPS

MAKEZZPS was used to concatenate the PS records into one national PS file for one MAX calendar year. The 51 input files were state-specific PS files created by NCHSSSN4 (program 4). The national PS file for the calendar year was written out in two formats (tape and DASD).

# III. MATHEMATICA'S LINKAGE DESIGN AND SOFTWARE

This chapter describes Mathematica's recommended modifications to the linking algorithm and software provided by CMS. There are three primary reasons for making revisions at this time. First, some modifications to both the linking algorithm and the software are required due to new information provided by NCHS. Second, modifications to the software are necessary because of substantial changes to the MAX record layout in 2005. Third, Mathematica believes that the software will be easier to document and maintain if written in a different programming language. After making modifications to the software, Mathematica will re-link the first SSN-based NCHS finder file with MAX 1999-2004. Due to changes in the linking algorithm, the revised software may link more records for MAX 1999-2004 than the original software provided by CMS. In order to evaluate the new algorithm, Mathematica will incorporate a variable into the record layout indicating whether each link is due to the original linking algorithm (provided by CMS) or the new algorithm (written by Mathematica).

## A. Linking Algorithm

During the first round, Mathematica recommends making three changes to the linking algorithm. The first change is in response to new information from NCHS regarding the date of birth variable. On some records with a missing day of birth, NCHS may report an imputed value of 15. With this in mind, Mathematica believes the most appropriate linking algorithm should use year and month of birth only, rather than exact date of birth. That is, when matching the SSN-based finder file to MAX, Mathematica will link records for which there is a match on SSN, month and year of birth, and sex.

Second, Mathematica recommends linking the first-round HIC-based finder file (in addition to the SSN-based finder file) to MAX 1999-2009. After equating the Beneficiary Identification Codes (BIC) in each file, Mathematica would link the HIC provided by NCHS (through SSA) to the Medicare Enrollment Database (EDB) HIC that is recorded in MAX data. Records linked via the HIC would also be required to have the same month and year of birth and same sex.

Third, NCHS has requested that the NCHS linkage ID be appended to each of the MAX records, and that the MSIS ID be removed from each of the linked records. Mathematica will modify the linkage algorithm accordingly.

After the first round of NCHS-MAX links, Mathematica will assess the linking algorithm and determine whether to recommend any further changes. For example, if there are records in the NCHS finder files that link to more than one MSIS ID in the same state and year, Mathematica may recommend combining the PS records for that person based on an algorithm already used in other MAX-related tasks. It is also possible that an individual record in the NCHS finder file will match Medicaid enrollment records in more than one state in a given year. In general, when the same person has more than one PS record, the information from the most recent enrollment date is considered to be the more accurate data. Mathematica will monitor the number of NCHS records that match multiple PS records during the first round of linkage and provide a count of such records in the assessment report at the end of the first round.

## B. Software

Programs written with COBOL, as well as those using utility programs for sorting, typically require less processing power and have shorter run times than programs written in SAS. However, SAS provides more control over options when sorting and merging files, which are the primary functions in this software system. Likewise, SAS provides a better environment for writing data quality tabulations to check the validity of linked files. SAS programs are also easier to document and better understood by a much larger population of programmers. For these reasons, Mathematica believes that the best course of action is to rewrite the programs in SAS. In doing so, Mathematica will be able to reduce the total number of programs from fourteen to four and create a system that is easier to maintain and datasets that are easier to validate.

In addition to converting all of the existing programs to SAS, Mathematica recommends incorporating data quality checks and combining programs wherever possible. First, Mathematica will modify SORTSSNF (program 1) to work with the HIC-based finder file in addition to the SSN-based finder file. The two finder files will be handled separately, and data from the two files will not be combined during this program.

Additionally, we will modify SORTSSNF (program 1) to verify that there are no records in the NCHS finder file with identical SSN (or HIC), date of birth, and sex. CMS reported that they found no duplicate records in the first-round of NCHS data. However, NCHS confirmed that records with identical SSNs (or HICs) but different NCHS linkage IDs could be included if a person participated in more than one of the surveys that contribute to the finder files. Currently, the check for duplicate records is in NCHSSSN4 (program 4), but Mathematica recommends moving it to SORTSSNF (program 1). If duplicates are found in the NCHS data, they will be handled appropriately so that the linked MAX PS files also contain duplicate records with different NCHS linkage IDs.

Next, Mathematica recommends combining MAKE1PS, SORT1PSS, NCHSSSN4, and MAKFNDR (programs 2 through 5) into a single SAS program. This program will be written to incorporate data from both the HIC-based finder file and the SSN-based finder file. The program, which will still be state- and year-specific, will sort the MAX PS file, remove records with missing identifiers, and merge records with non-missing identifiers to the NCHS finder file using the new linkage criteria (SSN or HIC, month and year of birth, and sex). This program will combine the links that result from the SSN-based finder file and the HIC-based finder file, and will assign a flag to indicate the primary identifier used in linking. The program will have two output files: one will contain the combined NCHS and PS record for each person in the linked file, and the other will contain only a list of the MSIS IDs and state abbreviations for the same records. We will add quality assurance tabulations to validate the linked records, such as counting the number of NCHS records linked to multiple MSIS IDS. We will also monitor the links made using the old linkage criteria, so that we may compare the first-round MAX 1999-2004 links with those constructed by CMS.

Mathematica also recommends combining F22IPMAX, F22LTMAX, F22OTMAX, and F22RXMAX (programs 6 through 9) into a single SAS program. This program will be repeated separately for each MAX claims file for each state and calendar year of MAX. We will use a CLIST parameter to repeat the program for each of the four MAX claims file types. We will also use this program to record the number of linked NCHS-PS records for which no claims are available.

Finally, Mathematica recommends combining MAKEZZIP, MAKEZZLT, MAKEZZOT, MAKEZZRX, and MAKEZZPS (programs 10 through 14) into a single SAS program. This program will be repeated for each of the five MAX file types. The output will be national files that combine all of the state files for that file type. All output files will be written as text files. Those files will be compressed, encrypted, and sent to NCHS on DVDs.

In addition to the coding changes, Mathematica will generally enhance the current system of software documentation. We plan to streamline the CLISTs, using them to modify a single copy of each program rather than writing a new copy for each state and year. We will keep a detailed run log, including record counts for each input and output file and run dates for each program. We will also preserve the job logs and summary output, and monitor those files to ensure that the software system is functioning as it should. These will only contain information necessary to check that the software is functioning correctly, and will never contain information that could be used to identify individual records.

## C. Data Quality and Reporting

At the end of each round of linking, Mathematica will assess all NCHS-MAX linkage activities and report their findings to CMS. Ideally, when conducting any sort of file linking activity, we would also assess the quality of the links by examining the linked records. However, to protect the confidentiality of the NCHS survey respondents, the linked data are available only through the NCHS RDC. Any analysis of the linkage results must be conducted within the RDC and the output must be reviewed and approved by NCHS before publication.

In place of detailed linkage analysis, Mathematica will rely on record counts in each of the input and output files. Table III.1 shows the data quality measures that Mathematica will use to monitor the linkage activity. Mathematica will record the number of records in the PS file and in each NCHS finder file. We will also record the number of links made using the old and new linkage criteria, and the number of NCHS records that match to more than one MSIS ID in the in the same year. The information that Mathematica gathers about the linked records will be used only to assess or improve the linking software.

Before applying for access to the linked files in the NCHS RDC, researchers may wish to know more about the demographic characteristics or geographic distribution of the linked records. To assist them in assessing the population represented in the linkage results, NCHS will release a set of public-use feasibility study files, which will contain the NCHS survey public-use ID, as well as a small number of variables from MAX. Mathematica will assist NCHS in identifying the MAX variables most likely to be of interest to researchers using Medicaid data.

If CMS determines that additional analyses are necessary to assess the quality of the linkage or to assist researchers who wish to use the linked files, it may be possible for Mathematica to analyze the linked files through the NCHS RDC. Before conducting such additional analyses, CMS, NCHS, and Mathematica will meet to discuss any privacy or confidentiality concerns. If the analyses are conducted, NCHS will review the results to determine whether they maintain an acceptable level of confidentiality for survey respondents.

**Table III.1. Recommended Data Quality Measures**

| Measures | MAX 1999 | MAX 2000 | MAX 2001 | etc. | MAX 2009 |
|---|---|---|---|---|---|
| Number of Person Summary Records | | | | | |
| Number of NCHS Records in Finder Files | | | | | |
| NCHS SSN finder file | | | | | |
| NCHS HIC finder file | | | | | |
| Number of Person Summary Records Linked to NCHS File | | | | | |
| Linked via NCHS SSN finder file | | | | | |
| Using date of birth and gender | | | | | |
| Using year and month of birth and gender | | | | | |
| Linked via NCHS HIC finder file | | | | | |
| Using date of birth and gender | | | | | |
| Using year and month of birth and gender | | | | | |
| Number of NCHS Records Linked to Person Summary File | | | | | |
| Linked to Person Summary file but not claims files | | | | | |
| Linked to more than one Person Summary record - same state, same year | | | | | |
| Linked to more than one Person Summary record - different state, same year | | | | | |

# REFERENCES

CDC/NCHS. "NHIS Survey Description, National Health Interview Survey, 2005." Hyattsville, Maryland: National Center for Health Statistics, Centers for Disease Control and Prevention, 2006.

NCHS, Centers for Disease Control and Prevention. "About the National Health and Nutrition Examination Survey". Available at [http://www.cdc.gov/nchs/nhanes/about_nhanes.htm]. Accessed May 2010.

NCHS, Centers for Disease Control and Prevention. "About the National Health Interview Survey." Available at [http://www.cdc.gov/nchs/nhis/about_nhis.htm]. Accessed May 2010.

NCHS, Centers for Disease Control and Prevention. "About the National Nursing Home Survey." Available at [http://www.cdc.gov/nchs/nnhs/about_nnhs.htm]. Accessed May 2010.

NCHS, Centers for Disease Control and Prevention. "The Second Longitudinal Study of Aging (LSOA II)." Available at [http://www.cdc.gov/nchs/lsoa/lsoa2.htm]. Accessed May 2010.

MATHEMATICA
Policy Research, Inc.

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