Linking Medicaid Administrative Records over Time and Space: Methods and Findings, 2005 to 2007

Shinu Verghese* and John L. Czajka**

*Mathematica Policy Research
600 Alexander Park
P.O. Box 2393
Princeton, NJ 08543-2393
sverghese@mathematica-mpr.com

**Mathematica Policy Research
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221
jczajka@mathematica-mpr.com

Introduction

To provide health policy researchers with access to Medicaid administrative data in a form that is suitable for research, the Centers for Medicare & Medicaid Services (CMS) has funded and overseen the development of an annual Medicaid Analytic Extract (MAX). MAX includes enrollment and claims information for each person enrolled in Medicaid and a subset of those enrolled in the Children's Health Insurance Program (CHIP) through a separate child health program (S-CHIP). While the MAX data have supported extensive research on state Medicaid programs and enabled detailed cross-state comparisons, their application to national-level and longitudinal research has been limited by the fact that the files do not identify records belonging to the same individual over time or across states.

To address this limitation of the MAX data, CMS contracted with Mathematica Policy Research to design and construct a set of research files that would identify unique Medicaid enrollees both within and across states and provide a reliable means of linking the records of unique enrollees over time. Mathematica produced such files for 2005, 2006, and 2007. This paper documents the results of that work.

Unduplication Methodology

The goal of unduplication was to determine which records in an annual MAX file and in MAX files from multiple years represent the same person. Once we made this determination, by evaluating pairs of records to decide if they should be linked or not, we assigned these records a common identifier—a MAX Enrollee Master file or MAXEM ID (hereafter just "ID")—that was unique to the person these records represented.

Linkage Algorithm. The Medicaid records that states submit to CMS lack names and addresses, so the linkage variables consisted of numeric identifiers and demographic variables. Two records were linked and assigned a common ID if they satisfied pre-specified, deterministic linkage criteria. Through these pair-wise links, all records that appeared to represent the same individual were ultimately assigned the same ID.

Several "linkage" variables were used to evaluate whether two records represented the same person. These were the Medicaid Statistical Information System (MSIS) ID, the MAX Social Security number (SSN), the Medicare Enrollment Database (EDB) SSN, the Health Insurance Claim number (HIC), and two demographic variables: date of birth (DOB) and sex. The linkage algorithm linked a pair of records if they agreed on any one of the following:

- MSIS ID
- MAX SSN + sex + at least two of (DOB year, DOB month, DOB day)

- EDB-SSN + sex + at least two of (DOB year, DOB month, DOB day)
- EDB-HIC + sex + at least two of (DOB year, DOB month, DOB day)

The four criteria were applied sequentially. That is, all record-pairs were evaluated to determine which pairs satisfy the first criterion. Following that, the same record-pairs were evaluated to determine which pairs met the second criterion, and so on. For record-pairs within the same state, 99 percent of the linkages were determined by the MSIS ID while for record-pairs from different states, 99 percent of the linkages were determined by the MAX SSN and the demographic variables, as the MSIS ID is state-specific and cannot be used to link records across states.

2005 and 2006. Before the first linkage pass, each record was assigned a preliminary ID. For the initial implementation of our unduplication methodology, this ID consisted of three components in the following order: (1) a scrambled state code, (2) a year code equal to the year less 2000, and (3) a record number from MAX. The sequence of components mattered because each record was ultimately assigned the lowest of its preliminary ID and all of the IDs on the records to which it linked. Records were first linked within states, both within the same year and across years. At the conclusion of the within-state linkages the IDs were reassigned, using the logic just described, and all records with the same ID were edited to replace missing or inconsistent values for five of the linkage variables (all but the MSIS ID) and a race/ethnicity code. Records with the same ID within a state and year were then consolidated (combined into a single record) so that any given ID occurred no more than once within a state and year. After this, the unduplicated records were linked across states, both within and across years, but they were not edited and they were not consolidated. To produce unduplicated counts of Medicaid enrollees nationally, the number of states in which each ID appeared was enumerated, and the inverse of this state count was used as a weight. When this weight was summed across all records, an ID that appeared in two states would be counted twice with a weight of one-half each time, yielding a total count of one for that ID.

2007. In extending our unduplication efforts to MAX 2007, we made several adjustments to the procedures that were used to produce the 2005 and 2006 research files in addition to adding a third year of data. These included:

- Redefining the format of the ID to place the year component first
- Retaining records with no enrollment data through the unduplication process
- Assigning common IDs at the conclusion of each major linkage step
- Applying a first round of editing to replace missing and inconsistent values immediately after the completion of linkages based on the MSIS ID rather than editing only at the conclusion of within-state linkages
- Editing the near-final IDs to eliminate within-state duplicates created during cross-state linking

The last three revisions were designed to (1) reinforce the linkages that were based on the MSIS ID, which prior research had shown to be the most reliable of the four types of linkages allowed by our linkage algorithm, and (2) reduce the influence of incomplete or inconsistent linkage variables in determining what records are linked or not.

In extending unduplication to 2007, then, we recreated the unduplicated research files for 2005 and 2006. The full unduplication methodology involved the following steps:

- Assignment of an initial ID to all records
- Performance of within-state links
- Application of edits
- Consolidation of records representing the same person in the same state and year
- Performance of cross-state links
- Assignment of “final” IDs

- Review and editing of within-state duplicates created by the cross-state linking

For each step we designed diagnostic tabulations to document key outcomes and provide guidance in selecting subsets of records for review.

Implementation

The implementation of unduplication for 2007 can be divided into unduplication within states and unduplication across states. These production activities were followed by an evaluation of the linkages before the final unduplicated research files were created.

Unduplication within States. Eight states use SSNs as MSIS IDs. When new enrollees lack SSNs, temporary MSIS IDs are assigned, which the states later correct once the recipients have been issued SSNs. In many cases the corrected MSIS IDs will not be identified until after the state has submitted the original records to CMS, so the state will compile and send corrections. For 2007, we received corrections to 49,000 records from 2005 and 112,000 records from 2006. We applied these corrections and then ran our linkage program using just the MSIS ID within each of the states that submitted corrections. For the two years we identified 46,500 linked pairs.

Following these preliminary steps we performed the full cross-year linkages for all states, again using just the MSIS ID. We identified 50.3 million linked record-pairs between 2005 and 2006, another 50.3 million between 2006 and 2007, and 42.2 million between 2005 and 2007. In a change from our initial procedures for 2005 and 2006, we reassigned the IDs at this point and ran our edit routine in order to reduce inconsistencies among the remaining linkage variables. We then repeated the cross-year links using the three additional linkage criteria—that is, other than the MSIS ID—listed earlier, which involved the MAX SSN, EDB-SSN, or EDB-HIC in combination with sex and DOB. This produced a total of 1.0 million new cross-year linkages. Reassigning the IDs at the end of this step produced a number of within-year linkages—about 421,000—between records that linked to common records in other years. We performed within-year linkages with these same three linkage criteria, obtaining 69,000 additional linkages. Overall, linkages by MSIS accounted for 99.3 percent of the cross-year linkages but only 8.7 percent of the within-year linkages.

To complete the within-state unduplication we reassigned the IDs again, performed a second edit step, and then consolidated records within states. Consolidation removed 199,000 records from the 61.4 million MAX records in 2005; 179,999 records from 61.7 million MAX records in 2006; and 151,000 records from 61.7 million records in 2007.

Unduplication across States. Records were linked across states, both within and then across years using the SSN and HIC criteria. Within year we identified a total of 4.1 million linked pairs over the three years, with 99.9 percent of them being based on the MAX SSN (with sex and DOB). Across years we identified more than twice that number for a total of 8.9 million over the three pairs of years. Here, too, 99.9 percent of the linked pairs were identified with the MAX SSN. The IDs were reassigned at the end of this process.

The assignment of a common ID to all records that represent the same person makes it possible to generate unduplicated counts of unique individuals enrolled in Medicaid in any year or across multiple years. Taking into account the duplicates across states, we identified 59.8 million unique individuals in the 2005 MAX data, 60.1 million in 2006, and 60.3 million in 2007. These unduplicated counts of unique Medicaid enrollees represent between 97.35 and 97.76 percent of the total records in MAX in each year. In other words, multiple records for the same individuals were about 2.5 percent of the total MAX records, on average, over the three years.

Quality of Linkage within States. We evaluated the quality of the linkages within states by comparing records linked by MSIS ID with respect to the MAX SSN, DOB, race/ethnicity and basis of eligibility (BOE). Only 58 percent of the records that were linked within state as a result of the state-supplied corrections had full agreement on DOB and sex while 90 percent had the same nonmissing MAX SSN. For record-pairs linked across years (for all states) by MSIS ID, 92 percent had the same, nonmissing MAX SSN, and 98.7 percent had the same DOB and sex. Agreement on race and BOE was lower. Depending on the pair of years, between 86 percent and 93 percent of

record-pairs had the same race/ethnicity and BOE while another 7 to 8 percent agreed on one of the two variables with the other being missing.

Record-pairs that were linked by the other three linkage criteria could not be evaluated with respect to DOB and sex, as these were used as linkage variables. Agreement on race/ethnicity and BOE was 56.3 percent overall for cross-year linkages and 66 percent for within–year linkages. Including cases that agreed on either race/ethnicity or BOE, with the other missing, raised the level of agreement to between 76 percent and 78 percent for cross-year links and 82 percent for within year links.

Quality of Linkage across States. Linkages across states could not use the MSIS ID; they were based entirely on the other three linkage criteria. To assess the quality of these linkages, we examined agreement on race/ethnicity and BOE. The record-pairs linked across states had somewhat higher levels of agreement on race/ethnicity and BOE than the record-pairs linked within states with the same variables (SSNs or HICs in combination with sex and DOB). For all within-year links, 73 percent agreed fully on race/ethnicity and BOE. For cross-year links, 72 percent agreed fully on these two variables. Adding cases that agreed on one of the two while the other was missing raised the levels of agreement to 84 percent and 85 percent, respectively.

About 10 percent of the records in 2005 and 2006 and 11 percent in 2007 were missing SSNs and therefore could not be linked to records in other states. Nearly two-thirds of the missing SSNs were from California, due in large part to a restricted benefits program in which immigrants could participate without providing SSNs. We estimated that 70,000 of the 6.1 million records with missing SSNs would have linked to records in other states if valid SSNs had been present.

Medicaid Enrollment: Analyses with Unduplicated Data

Several analyses illustrate what can be learned about Medicaid enrollment patterns by reducing MAX data to unique enrollees.

Enrollees by Eligibility Group. Within the population of Medicaid enrollees there are differences in the impact of unduplication by eligibility group—that is, among the aged, disabled, child, and adult enrollees. In all three years, duplicate records among disabled and child enrollees occur at about twice the frequency as among aged enrollees and about 50 percent more often than among adult enrollees. In 2007, for example, duplicates within and across states accounted for 1.15 percent of aged enrollee records, 2.73 percent of disabled enrollee records, 2.61 percent of child enrollee records, and 1.56 percent of adult enrollee records.

Geographic Movement. The cross-state linkages performed as the final stage of the unduplication of Medicaid enrollment records provide detailed information on the movement of Medicaid enrollees between states. One way that such movement is reflected is in persons enrolled in more than one state in the same year. In 2005, 2.31 percent of the 59.8 million unique individuals in the MAX PS file had records in more than one state; in 2006, 2.23 percent of the 60.1 million unique enrollees had records in more than one state; and in 2007 1.98 percent of the 60.3 million unique enrollees had records in more than one state. Nevada and Wyoming led all states with six to seven percent enrolled in other states during the year. California was lowest with less than one percent enrolled in other states during any year while New York was next lowest with between 1.20 and 1.37 percent enrolled in other states. Four additional states—Hawaii, Massachusetts, Michigan, and Pennsylvania—had less than two percent of their enrollees enrolled in other states.

When we use the common ID to pair up records that have been linked across states over time, we are able to separate the flows of enrollees from state A to state B from the flows of enrollees from state B to state A. Over the three years the largest movement was that of 66,000 individuals from Louisiana to Texas between 2005 and 2006, reflecting the impact of Hurricane Katrina. The second largest flow in that year was 45,000 persons in the reverse direction. Three other pairs of states had flows in excess of 30,000 in at least one pair of years. About 35,000 enrollees moved from California to Arizona between 2005 and 2007. Over that same period, nearly 33,000 moved from New York to Florida, and more than 30,000 moved from Florida to Georgia. All three pairs of states had smaller but still substantial flows of enrollees in the reverse direction over the same years. Most of the remaining large flows involved a fairly small set of states, and most of these shared borders.

We calculated migration rates out of and into each state for each pair of years. Nevada and Wyoming stand out with double-digit rates in both directions in all three pairs of years. Louisiana had an out-migration rate of 12 percent and an in-migration rate of nearly 10 percent between 2005 and 2006 but smaller rates in later years. California had the lowest migration rates in both directions in all three pairs of years, with out-migration of about 2 percent and in-migration hovering around 1.5 percent. New York also had comparatively low migration rates in both directions, as did Pennsylvania, Maine, Massachusetts, Vermont, Michigan and Hawaii.

States' outflows and inflows were generally comparable to each other, regardless of magnitude. Only three states and the District of Columbia had net in-migration rates (in-migration minus out-migration) in excess of two percent in either direction for any pair of years, and none of the four did so in more than one pair of years. Household survey data from the American Community Survey (ACS), which asks respondents where they lived one year ago, shows similar patterns among persons who were enrolled in Medicaid at the time of the survey. While the ACS captures a lower volume of migration (in part because in asking about location one year earlier it misses seasonal migration), it too shows flows of similar magnitude in both directions in a substantial majority of states as well as generally low net migration.

Turnover in Medicaid Enrollment. With Medicaid records that have been unduplicated at the state and national levels, it becomes possible to examine turnover in Medicaid enrollment more rigorously than with other Medicaid administrative data.

Comparisons of annual-ever enrollment and average monthly enrollment in 2005 show differing amounts of turnover by age. For all ages combined the ratio of the former to the latter is 1.26, implying that annual-ever enrollment was 26 percent higher than average monthly enrollment. For children, annual-ever enrollment is 23 percent higher than average monthly enrollment. For nonelderly adults, annual-ever enrollment is 32 percent higher than average monthly enrollment; for elderly adults, annual-ever enrollment is 15 percent higher. Removing enrollees who received institutional care has no discernible effect except among the elderly; annual-ever enrollment among the non-institutionalized is 14 percent higher than average monthly enrollment.

Using records that were unduplicated within each state, we estimated the combinations of years that unique enrollees were enrolled within the same state. Of those who were enrolled in 2005, 82.2 percent were still enrolled one year later, and 66.8 percent were still enrolled two years later (that is, enrolled in all three years). A very small fraction, 2.0 percent, were not enrolled in 2006 but resumed enrollment in 2007 while 15.8 percent were not enrolled in either 2006 or 2007. Among those who were enrolled in both 2005 and 2006, 81.2 percent remained enrolled for an additional year. That this is barely lower than the proportion of 2005 enrollees who were still enrolled a year later suggests that the rate of disenrollment from Medicaid may not increase appreciably with the duration of enrollment.

The possibility that former Medicaid enrollees whose coverage has ended may remain on the rolls in some states has been suggested as a possible explanation for why survey estimates of Medicaid coverage do not compare more closely with program administrative estimates. To evaluate this idea, we examined patterns of service use in 2007 by enrollment duration and the number of states in which an individual was enrolled. With respect to duration of enrollment, we found that in all but three states, service use was higher among enrollees with three or more consecutive years of coverage than among those enrolled for just a single year. This runs counter to the prediction, but this evident tendency for service use to increase with years of enrollment would obscure any reduction in service use due to persons remaining on the rolls past the end of their eligibility. With respect to the number of states in which an individual was enrolled, however, we do find that service use declined between those enrolled in only one state and those enrolled in three or more states. This may indicate simply that people enrolled in multiple states distribute their annual service use across the states, but further investigation with more extensive measures of service use and with the expenditure data available in the MAX files could shed further light on this finding.