Title
Second Report of the California Hospital Outcomes Project (1996): Acute Myocardial Infarction Volume Two: Technical Appendix-Chapter004

Authors
Romano, Patrick S
Remy, Linda L
Luft, Harold S

Publication Date
1996-03-21

CHAPTER FOUR: LINKING HOSPITALIZATION RECORDS

The purpose of this chapter is to describe the linkage methods developed for the acute myocardial infarction study. The goal of the linkage process is to identify relevant hospital discharge records, order them temporally and logically, then create a linked single-record analysis file summarizing information from all related records. Additionally, the linkage must protect patient confidentiality.[1]

These linkages are important for several reasons. First, linkages with subsequent records help identify the outcome for each patient (e.g., death within 30 days). Otherwise, hospitals that transfer their sickest AMI patients might have unduly low outcome rates. Second, linkages make it possible to identify fresh AMIs as described in Chapter Three. Third, linkages provide important information about clinical risk factors. Diabetes and other chronic comorbidities are not always coded on discharge abstracts, so more complete information can be obtained when multiple records are available.

OVERVIEW OF THE LINKAGE STRATEGY

The main steps in record linkage were to: (1) identify records which meet initial selection criteria, (2) find all additional records with linkage potential, (3) delete duplicate records and resequence record sets, (4) order records in the period around the admission, and (5) create the linked single-record analysis file.

1. Identify Records Which Meet Initial Selection Criteria

The first step in record linkage was to create a condition file containing all records that (a) met preliminary inclusion criteria and (b) were within the time window used to select cases.[2] These preliminary inclusion criteria

---

[1]OSHPD has interpreted this to mean that patient identifiable data can be returned only to the hospital originally submitting the data. This means the linked single-record in the analysis file must permit discrimination between data derived from the index admission and other admissions.

[2]The master OSHPD database was used to create the condition file. Before starting the search, all records with valid SSNs were extracted from the master OSHPD hospital discharge database and divided into discrete files containing all records with valid SSNs for each month. The monthly files were sorted by SSN to simplify searching and to improve mainframe data

are described in Chapter Three. For the AMI study, the window period included an admission between August 26, 1990 and May 31, 1992 (inclusive).

At this point, records in the condition files were only *candidates* for study. Whether a specific record would be used for the study, whether more records existed for this patient, and where in the sequence of care the record(s) fit were still unknown. For example, patients could have had more than one AMI in the window, so more than one record with the same encrypted social security number (SSN) could have been in the condition file. Further, because of coding ambiguities or errors, a specific record could have been coded as a fresh AMI, when in fact it might have been a prior or subsequent admission.

2. **Find All Additional Records with Linkage Potential**

The goal of this step was to find any additional records within the study frame that might link with the AMI records identified in the previous step. The frame is a specific time period before and after the hospitalization record in the condition file. For AMI, the frame extended from eight weeks before the admission date to one day after the discharge date.

To start the search, the condition file was divided into two subfiles. One subfile contained records with an SSN, and the other contained records missing the SSN.[3]

Two AMI lookup files were constructed using the Step 1 AMI condition file as a base. These lookup files were used to search for candidate records within the study frame which might be related to those already pulled. Lookup file 1 contained one entry for each unique SSN and all associated admission dates and birth dates. The maximum number of birth dates found for any given SSN was two. This lookup file had 65,176 entries.

Lookup file 2 contained one entry for each unique combination of birth date, sex, and 5-digit ZIP code; it had 64,963 records.[4] This lookup file was used to search for additional records related to candidate records without an SSN. If records in the searched file matched records in the

---

management of the extremely large OSHPD master dataset. The master file and monthly files were used at different times in the search process.

[3]Issues with respect to both present and missing SSN are discussed in more detail later in this chapter in the section "Reliability of the Dataset for Linkage".

[4]The 1993 AMI study used 3-digit ZIP code. This year it was discovered that the 3-digit ZIP code in combination with the other variables pulled matches with multiple SSN. By using the 5-digit ZIP code, the number of records with the same combination but different SSN was reduced.

lookup file, the record was pulled as a candidate and the associated SSN (if available) was assigned to the condition-file record lacking the SSN.

The lookup files were used to locate all potential records for the study. This process involved four steps:

2.1. Using lookup file 1, all records with an exact match on SSN were extracted if they matched on 2 of the 3 birth date elements (i.e., month, day, year). If an SSN was associated with two birth dates, the second date also was checked to see if it matched on 2 of the 3 birth date elements. This relaxed criterion was used to pull records which otherwise indicated a match but where a data entry error may have occurred on one or more records. This step found 159,059 records with SSNs.

2.2. Lookup file 2 was matched against the AMI condition file that had no SSNs (5,271 records). An exact match was required on each unique combination of birth date, sex, and 5-digit ZIP code. Records in the condition file lacking SSNs that matched entries in the lookup file were assigned the SSNs associated with those entries. As a result, SSNs were assigned to 302 records. This left 4,867 records with valid birth dates, gender, and ZIP codes, but no SSNs. No further searching was possible for 102 records missing one of those essential data elements.

2.3. The 4,867 records above were matched against the monthly files based on birth date, sex, and 5-digit ZIP code. All exact matches were checked for AMI diagnosis codes, for same day or 1-day transfers, or same day readmissions. If criteria were met, the record was pulled and an SSN was assigned. This step assigned SSNs to 415 records.

2.4. The remaining 4,554 records which appeared not to have been involved in multiple admissions were assigned a simulated SSN with a first digit of #. These records then were combined with all the records found in Steps 1 through 3.

A simulated SSN field was created to keep track of assigned SSNs. If the record had an SSN, the value of the SSN and the simulated SSN were identical. If a missing SSN was found by the second lookup file, the SSN field was blank and the simulated SSN field contained the found SSN. Lastly, where no SSN was found, the SSN field was blank and a simulated SSN was created with a first digit of "#". The simulated SSN was used to group related records into independent periadmission periods. For the rest of this chapter, the term SSN refers to the value in the simulated SSN field.

3. **Delete Duplicate Records and Resequence Record Sets**

The files created in Step 2 above were joined and sorted by SSN, admission date, discharge date, date of birth, sex, and OSHPD facility number. The purpose of sorting by these variables was to identify any duplicate records with identical SSNs, admission and discharge dates, birth dates, genders, and hospital-identification numbers. Fourteen such pairs were reviewed manually.[5] The record from each pair with a longer or more precise list of diagnoses or procedures was retained. If both records had equally numerous and precise diagnoses and procedures, the record with higher total charges or a more heavily weighted DRG was retained. (OSHPD editing procedures have since changed to correct this problem).

A manual review also was performed on 19 record sets with the same SSN, birth date, gender, and admission (and/or discharge) dates, but at different hospitals. These patients were apparently admitted to one acute care hospital, transferred to another, and then discharged, all on the same day. Each set was manually sequenced based on the discharge disposition and admission source. For 13 pairs, one of the records had a disposition of "death", so it was sequenced last. For another 5 pairs, one record had a disposition of "general acute care hospital", so it was sequenced first. The order could not be determined for one pair, so it was not resequenced. In addition, one SSN had three records with the same admission date but different discharge dates. This made it appear to be two different periadmission periods when it was really one. This was not discovered until after linking was completed. After reviewing all the records in the set, the records were manually resequenced into the proper order.

After dropping duplicate records and resequencing sets, the file was divided into a subfile containing SSNs with only one record (which did not require linkage) and another subfile containing SSNs with multiple records.

4. **Order Records in the Period Around the Admission**

All records for a given SSN were extracted in Step 2, including some admissions that were irrelevant to the AMI study. For example, a person treated for AMI could have been admitted several months later for

---

[5]All numbers cited in this chapter come from analyses performed before certain hospitals were excluded for unresolved transfers and extreme coding practices. These numbers may therefore differ from numbers that would be obtained from analysis of the final data set.

appendicitis. The goals of this step were to identify the periadmission period, which consists of the "true" index admission and the records around it, and to delete irrelevant records. Defining the periadmission period was done in four steps: (1) the index admission was identified, (2) the outcome record was identified, (3) prior admissions were identified, and (4) the periadmission number was assigned.

The first step in establishing a periadmission period was to identify records which included the condition of interest as described in Chapter Three. The first record for an SSN in the ordered multiple record file that met selection criteria was marked as the index admission. At this point, some admissions and their subsequent transfers or readmissions were marked for exclusion, as described in Chapter Three.

The next step was to identify the outcome record. This process began by classifying all records that followed an index admission as transfers. Some patients experienced several transfers during the periadmission period; the last transfer represented the outcome record (as long as it occurred within 30 days of the AMI).

Very specific criteria were established to classify subsequent hospitalizations as transfers. These criteria were necessary because most subsequent hospitalizations after AMI relate to evaluation or surgical therapy of coronary artery disease and do not belong to the periadmission period. Subsequent SNF/ICF admissions also do not belong to the periadmission period. The specific criteria used to evaluate potential linkages with subsequent hospitalizations varied as follows:

4.1. **Candidate records with a "report type" of skilled nursing and intermediate care (3), psychiatric care (4), alcohol/drug care (5), or rehabilitation care (6) were not evaluated**.

Lookup file 1 pulled many records that were not from general acute care hospitals. These were used to identify prior admissions, but were not used to identify transfers.

4.2. **Candidate records with a "report type" of general acute care (1) were categorized according to the discharge disposition of the immediately prior index hospitalization and included or excluded, as follows**:

a. *Intermediate care facility (03) or skilled nursing facility (04)*. No records subsequent to the index record were linked.

b. *Other facility (05)*. OSHPD's 1988 reabstraction study showed that some cases with this discharge disposition were incorrectly labeled. They actually were transfers to acute care hospitals and should have been assigned a disposition of 02. Therefore, subsequent records were linked when: (1) the admission date was the same as the index discharge date, and (2) the hospital identification number was different from that on the index record (suggesting that the patient may have remained at the same level of care), and (3) the principal diagnosis on the candidate transfer record was neither rehabilitation (V57.xx) nor psychiatric (290.x-319).

c. *Acute care hospital (02).* Some cases with this discharge disposition appear to have been transferred to lower levels of care. Therefore, subsequent records were linked only when: (1) the hospital identification number was different from that on the index record (suggesting that the patient may have remained at the same level of care), and (2) the admission date was the same as or one day later than the index discharge date (allowing for late night transfers), and (3) the principal diagnosis on the candidate transfer record was neither rehabilitation (V57.xx) nor psychiatric (290.x-319). Although a patient may have been readmitted to an acute care hospital more than one day after a prior discharge, the second hospitalization was regarded as a separate episode of care and not a transfer.

d. *Routine (01), against medical advice (06), or home health service (07)*. Some patients were discharged to home or left against medical advice and returned to a hospital later the same day. These patients were still in the acute phase of care when they were readmitted, so their hospitalizations needed to be linked. Subsequent records were linked only when: (1) the admission date was the same as the index discharge date, and (2) the principal diagnosis on the candidate transfer record was neither rehabilitation (V57.xx) nor psychiatric (290.x-319).

At this point, all valid transfer and index hospitalizations had been identified. These records were grouped to define an episode of care (i.e., index admission and transfer(s), if any).

Next, all records that preceded an index record but fell within the study frame were classified as prior admissions. A lookup file was created to determine if a record was an admission 0 to 56 days before the index admission. If so, it was flagged as a prior admission. The prior, index, and transfer admissions were grouped into a periadmission period. Records not flagged as a prior, index, or post (i.e., transfer for AMI) were discarded.

After the multiple record file was ordered, it was recombined with the single-admission file from Step 3 to create the periadmission file. A new variable was created to group sets of records (prior, index, post) into distinct periadmission periods. This grouping variable was needed for patients with multiple periadmission periods within the study frame.[6] The periadmission file contained one-to-n periadmissions composed of one-to-n records for each SSN.

5. **Create the Linked Single-Record Analysis File**

The purpose of this step was to transform the periadmission file into a linked analysis file containing one record per periadmission. The transformation began by running programs which used all clinical information from all records in the periadmission file to summarize the frequency of all diagnoses and procedures, and their relationship to the study outcomes. All hospitals with evidence of unusual coding or high proportions of missing out-transfers, as described in Chapter Six, were excluded at this stage. After deleting these hospitals, 68,102 periadmission periods for 65,994 people remained. Of these periadmission periods, 19.5% had one or more transfers, and 8% had one or more prior admissions.

Next, the periadmission file was used as input for a complex program summarizing the diagnoses and procedures into clinical risk factors, which may be obtained from prior, index, or post records.[7] Ethnicity and date of birth can be recorded differently from one record to another, and source of payment can change from one hospital to another. Therefore, index-record values for these variables were retained. The linked single-record analysis file was the product of the program creating the clinical risk factors. After eliminating hospitals with unusual coding (Chapter Six) and creating random subsets of the file (Chapter Eight), the linked analysis file was ready for statistical modeling.

---

[6]A flag was created at this point for AMI cases to identify transfers that were not found. A subsequent program identified hospitals in which 20% or more of transferred cases had no further record. All records associated with a periadmission period were excluded at this point if the index record came from one of these hospitals. This is described in Chapter Seven: Selection and Exclusion of Hospitals.

[7]Only variables from the index admission can be returned to the index hospital. The risk factor program flags cases which will require special handling. Two variables are created in the linked analysis file to count records obtained from prior and later admissions. If either of these counters are greater than zero, clinical risk factor variables that could have been obtained from those admissions are set to missing in the file returned to hospitals.

RELIABILITY OF THE DATASET FOR LINKAGE

Using linked records, it is possible to summarize some reliability issues associated with using the OSHPD database to identify related hospitalizations. This section summarizes the reasons why certain demographic variables were used in the linkage process and others were not.

Reliability of Variables Considered but Not Used to Match

Expected principal source of payment was considered but not used because a patient's insurance status may change from one hospitalization to the next.

Race was not used because the definitions may be subjectively applied, and the overall error rate was reported as 6% with 56% underreporting of Asian ancestry (according to OSHPD's 1988 reabstraction study).

In models based on linked datasets, the decision was made to use the race reported on the index record for modeling, since that is the record returned to the admitting hospital. Of 13,857 AMI periadmissions with more than one record (20% of all AMI periadmissions), race differed from the index record for 1,561 (11.3%). Of these, 610 index records indicated White race but had a different value on another record. Black race was indicated on 64 index records with a different value on another record. Similar discrepancies were found for 269 Hispanic, 18 Native American, and 159 Asian index records. An additional 273 AMI cases were missing race on the index record but had a value on another.

The problem of using race as a linking or modeling variable is highlighted this year in the AMI Priors Model B analysis. In a regression, cases missing information on any variable are dropped from the model. The fact that race was missing or discrepant on one or more records within the periadmission period suggests that the coefficients reported for the Priors Model B would be different if race was reported reliably.

Reliability of Variables Used to Match

In the OSHPD database, candidate variables that can be used for linkage are limited to SSN, date of birth, sex, and ZIP code. Problems were identified with each of these variables.

The AMI condition file contained 75,895 records. Of those, 5,271 (6.9%) were missing SSNs. After completing linkage, there were 68,012 AMI periadmission records, of which 13,587 (20%) were constructed from multiple records. An SSN was assigned to one or more records for 691 multi-record periadmission periods (5.0%) that initially had been lacking one.

A match on two of three birth date elements (i.e., month, day, year) was used to confirm linkage of records based on SSNs. Date of birth discrepancies occurred in 756 multi-record periadmission periods (5.5%). Overall, 20.2% of multi-record AMI periadmissions periods had discrepancies on SSN, race, or date of birth. Of these periadmission periods, 2,579 were discrepant on one variable, 213 were discrepant on two variables, and one was discrepant on all three.

Several hundred records were found with the same SSN as index records, but with different values for various demographic variables. For example, a 21 year old Black female and a 75 year old Hispanic male, reportedly with the same SSN, were admitted to the same hospital. The former patient had a normal delivery; the latter patient had an AMI. Possible explanations for this problem include: (1) these SSNs correspond to invalid social security numbers that were not identified by OSHPD staff before encryption; (2) hospital employees entered social security numbers incorrectly; (3) multiple people used the same social security number; and (4) patients reported incorrect social security numbers.

A list of valid SSNs was obtained from the Social Security Administration, and a series of diagnostic programs were written to test the reliability of the SSNs. Analysis of the OSHPD master file found, for example, one California facility that assigned the same SSN to every emergency room admission over a 4-month period. A program was written to flag and set to missing records with three types of invalid SSN: (1) a constant false SSN assigned by a facility for all cases (presumably) missing SSNs (i.e., 111-22-3333), (2) SSNs associated with multiple dates of birth derived from multiple records, and (3) SSNs outside the valid values provided by the Social Security Administration.

Patient social security numbers were added to the OSHPD database beginning July 1, 1990. A series of analyses showed that Hispanic patients were more likely to be missing SSNs than white patients. Patients in southern California and those admitted to large public hospitals were most likely to be missing SSNs. Reporting practices have not changed substantively over time, except at certain Kaiser facilities in northern California that experienced difficulty implementing the SSN reporting requirement. These findings indicate that patients without SSNs differ systematically from patients with reported SSNs. Although an algorithm for linking records without SSNs was developed, this algorithm is probably less effective than that based on SSNs. As a result, systematic underestimation of transfer rates among patients without SSNs may have occurred.