Record matching between the National Hospital Care Survey and the National Death Index

Shaleah Levant[1], Monica Wolford[1]
[1]National Center for Health Statistics, CDC, 3311 Toledo Road, Hyattsville, MD 20782

Abstract

Linking the National Hospital Care Survey (NHCS) with the National Death Index (NDI) provides information on the outcomes of hospitalizations and allows for analysis of individual and provider characteristics associated with in-hospital and post-discharge mortality. We test the viability of confirming hospital mortality through the linkage of preliminary 2011 NHCS data for “known dead” inpatient discharges (i.e., patients that died during a hospitalization) with the NDI, assessing the true match rate and the quality of the match. We then expand the analysis to identify patients with a 30-, 60-, and 90-day post-discharge mortality. The true match rate for the “known dead” is 94 percent.

Key Words: data linkage; National Hospital Care Survey; National Death Index

1. Background

In May 2011, the National Center for Health Statistics (the Center) began recruitment for a new data collection effort—the National Hospital Care Survey (NHCS). The purpose of the NHCS is to collect data on patient care in hospital inpatient, emergency department (ED), and outpatient department (OPD) settings. Data in 2011 were collected electronically through Uniform Bill (UB)-04 administrative claims, including personally identifiable data (PII) such as name, address, and Social Security Number, when available. With the collection of PII in NHCS, the Center has been able to link individual discharge records across health care settings and with other national datasets. Linking the National Hospital Care Survey (NHCS) with the National Death Index (NDI) provides information on the outcomes of hospitalizations and allows for analysis of individual and provider characteristics associated with in-hospital and post-discharge mortality. This paper explores the preliminary linkage of NHCS patient-level data with the National Death Index (NDI), also maintained by the Center.

1.1 De-duplicating Claims Data and Creating Patient Identifiers

Using UB-04 claims data presents the challenge of de-duplication of claims, since one discharge can have multiple claims. The initial de-duplication is performed at the hospital level, developing unique processes to identify duplicate claims for the same discharge within a hospital. In order to develop the de-duplication method for a hospital, claims were grouped in each of these three ways: (1) by Patient Control Number (PCN); (2) by Date of Admission + Medical Record Number (MRN); and (3) by Date of Admission + Date of Birth + Patient Name. The purpose of this processing was to evaluate whether PCN could be used to accurately identify duplicate claims for the same discharge, controlling for matching data elements for date of admission, MRN, date of birth, and patient name. If the number of duplicate groups of claims identified by PCN was close to the other counts produced, PCN was used to de-duplicate the claims. However, if the values of these variables in a group of duplicates was not similar, then the hospital's claims were further assessed through a manual review of the duplicate groups and a final determination of de-duplication method for the hospital was made.

After de-duplication of claims, a probability-based record linkage method was used to identify patients. In the first round of patient identification, two records were compared by name (first, last, middle initial), date of birth, sex, hospital identifier, medical record number, Social Security Number (SSN), and ZIP code. If there was sufficient agreement of data elements between the two records, the records were retained as a pair (i.e. record pair) for further analysis. The second round compared the record pairs, controlling for agreement in the hospital identifier and Medical Record Number (Pass 1), SSN if reported (Pass 2), and for agreement in sex, year and month of birth, soundex of last name, and state abbreviation (Pass 3). The record pairs compared in the three passes had match weights assigned to 11 matching variables (see Table 1). The match weights are likelihood ratio scores based on the probability of agreement in the records retained as record pairs and the probability of agreement in the records that were not included in any record pairs. A match weight of 60.98 indicated perfect agreement in all the comparison fields. Pairs with a match weight above a threshold of 30 were retained as likely matches, based on selection thresholds suggested in Winglee, Valliant, and Scheuren (2005).

Table 1: Match Rates

Match Key	*Match Outcome*	*Agreement Weight*	*Disagreement Weight*
Last name	string score* between 0 and 1	9.18	-3.51
Standardized first name	string score between 0 and 1	9.16	-3
Middle initial	1 agree, 0 disagree, missing	3.21	-2.96
Sex	1 agree, 0 disagree, missing	0.64	-2.3
Year of birth	1 agree, 0 disagree, missing	4.55	-2.99
Month of birth	1 agree, 0 disagree, missing	2.43	-2.91
Day of birth	1 agree, 0 disagree, missing	3.38	-2.96
ZIP code	1 agree, 0 disagree, missing	6.48	-1.05
Hospital identifier	1 agree, 0 disagree, missing	3.77	-1.59
Medical Record Number	1 agree, 0 disagree, missing	8.99	-1.61
Social Security Number	1 agree, 0 disagree, missing	9.19	-3.91

*The greater the score the closer the match, 0 is no match and 1 is a perfect match.

Two additional reviews were conducted for records pairs of children aged less than ten years old at the time of discharge. The first review targeted newborn infants whose first names contained "BABY," "GIRL," "BOY," "FEMALE", and "MALE." The newborn pairs were then subject to one of three adjustments. (1) Pairs that contained records where names may be "BABYGIRL" in one record and a real name, e.g. "JANE," in another were accepted as a match when the hospital identifier and medical record number were the same. (2) Pairs with the same hospital identifier, service date, and patient address, but different medical record numbers, were identified as twin or multiple birth records and manually split. (3) Pairs with different last names and medical record numbers were manually reviewed and split if the pair was determined to be false.

At the end of the patient identification process, 314,878 likely individuals were identified in 321,474 NHCS records.

2. National Death Index Submission and Results Files

2.1 Submission File

The Center submitted the 321,474 NHCS records for matching to the NDI by NDI staff. The file consisted of 12 variables for matching: SSN, first name, middle initial, last name, sex, birth month, birth day, birth year, state of residence, state of birth, race, and marital status (Table 2).

2.1.1 Variable missingness

When the SSN was missing for adults over age 65, an effort was made to recover it using the Medicare subscriber number on the UB-04 claim. This increased the percent of records with an SSN from 20 to 40 percent. The NDI uses race in the matching process, and marital status can be used in quality control for the returned matches. However, neither race nor marital status are required on the UB-04 and therefore not collected in the NHCS.

Table 2: Variable missingness

Variable	*Percent missing*
SSN	60%
First name	0.01%
Middle initial	51%
Last name	0.01%
Sex	0.02%
Birth month	0%
Birth day	0%
Birth year	0%
State of residence	0%
Race	100%
Marital status	100%

2.1.2 Match outcomes

Submitted records had one of three outcomes: rejected, a potential match, and no match. A record was rejected by the initial NDI edit program if it did not contain at least one of the following combinations of variables:

1. First and last name and SSN;
2. First and last name and month and year of birth;
3. SSN and date of birth and sex.

All NHCS records that were not rejected by the edit program were included in the NDI search. To qualify an NDI record as a potential match, both the NHCS record and the NDI record had to satisfy at least one of the following criteria:

1. Match on SSN;
2. Exact month and +/- 1 year of birth, first and last name;
3. Exact month and +/- 1 year of birth, first and middle initials, last name;
4. Exact month and day of birth, first and last name;
5. Exact month and day of birth, first and middle initials, last name;

6. Exact month and year of birth, first name, father's surname;
7. If the subject is female: exact month and year of birth, first name, last name (on NHCS record), and father's surname (on NDI record).

Of the 321,474 records submitted, 55,831 records had a potential NDI match, 265,491 records were not matched, and 152 records were rejected.

2.1.3 Rejected Records

All of the rejected records were missing SSN and had a problem with at least one other variable. Date of birth was missing for 134 of the records, names were missing or incorrect (e.g. "TGIRL1" for an unnamed twin girl or "ER TRAUMA") for 35 records, and sex was missing for 35 records. One particular hospital accounted for over 80 percent of the rejected records.

2.2 NDI Results

NDI assigns both a class and a score to all potential matches. The five classifications groups developed by NDI are:

- Class 1: Exact match on at least eight of the nine digits of the SSN, first name, middle initial, last name, sex, state of birth, birth month, and birth year.
    - Note- There are no Class 1 matches in the NHCS-NDI match because state of birth is not collected as a discrete element in the NHCS.
- Class 2: SSN matches on at least seven digits, and one or more of the other items from Class 1 may not match.
- Class 3: SSN unknown but eight or more of first name, middle initial, last name, father's surname (for females), birth day, birth month, birth year, sex, race, marital status, or state of birth match.
- Class 4: Same as Class 3 but fewer than eight items match.
- Class 5: SSN is known but does not match.

The score assigned to each match is a probabilistic score: the sum of the weights assigned to each of the identifying data items used in the NDI record match, where the weights reflect the degree of agreement between the information on the NHCS record and the NDI death record.

$$\begin{aligned}\boldsymbol{Score} = {} & W_{SSN} + W_{firstname\ x\ sex\ x\ birthyear} + W_{middleinitial\ x\ sex} + W_{lastname} \\ & + W_{race} + W_{sex} + W_{maritalstatus\ x\ sex\ x\ age} + W_{birthday} + W_{birthmonth} \\ & + W_{birthyear} + W_{stateofbirth} + W_{stateofresidence}\end{aligned}$$

Weights are positive when the data items on the NHCS record and the NDI record agree; and negative when there is no agreement. If the data item is missing on either the NHCS or NDI record, then the weight is zero. More information about class and score can be found in *National Death Index User's Guide*.

3. Selecting Matches

3.1 Test Case of "Known Dead"

The true match rate between the standard submission file of NHCS "known dead"—individuals with a final discharge status of dead (DS_STATUS = '20') in the 2011 NHCS data—and the NDI death certificate records was evaluated in order to determine the quality of the linkage. Since each eligible NHCS discharge may have multiple submission records and each submission record may return one or more potential matches to a NDI record, NHCS staff employed a strategy to provide the single best NDI match record for inclusion on the linked mortality file, as discussed in the following paragraph.

First, NHCS-NDI potential match records with a score of less than or equal to zero were considered false matches and eliminated from the pool of potential matches—potential matches with negative scores had more weighted disagreement among data elements than agreement[1]. Next, potential NDI match records with a date of death greater than one day previous or one day post (to allow for small discrepancies on NHCS records and NDI records) the date of discharge on the NHCS record were considered false matches and eliminated. Among the remaining pool of potential matches, duplicate death certificate numbers (i.e. match records that referred to the same death certificate) were eliminated within PATIENT_ID, retaining the match with the highest score. Some discharge records, however, still had more than one NDI record as a potential match. The remaining potential matches were then ranked by highest score. The NDI match with the highest score was selected as the single best record match. Within each class, matches were determined to be true or false using the cut-off scores developed by NDI. Matches with a score greater than or equal to the cut-off score were considered true matches, while records with a score less than the cut-off were considered false matches. The cut-off scores[2] for Classes 2, 3, and 4 were 44.5, 37.5, and 32.5, respectively.

At the end of the match selection process, the NHCS “known dead” had a true match rate with the NDI of 94 percent (= 7,630 matched / 8,153 submitted), with an accuracy of 90 percent (=. [(4,947 Class 2 matches * 93.4% accuracy) + (1 Class 3 matches * 96.9% accuracy) + (2,682 Class 4 matches * 84.6% accuracy)] / 7,630 matched). The NDI User’s Guide provides recommended cut-off scores and accuracy of match data based on analyses of two calibration samples, with consideration given to maximizing the proportion of records correctly classified while at the same time minimizing the number of records incorrectly classified.

3.1.1 Unmatched Records

The records for the six percent (n=532) of the “known dead” individuals that did not have a true match to a record in the NDI were manually reviewed. The majority of discharge records with no true match were records that did not have any potential match in the NDI (57%). There were 358 records for 303 individuals (identified by PATIENT_ID) that did not have any potential matches in the NDI; 96% of those records were missing SSN and 27% had an obviously invalid first name (i.e., some form of “baby,” “boy,” “girl,” or “infant” was included in the first name). However, all discharge records had a birth date and state of residence.

The second group of discharge records with no true match is made up of individuals that had a potential match in the NDI but that were deemed false matches in the elimination

[1] The NHCS matching methodology was based on the methodology in *The Second National Health and Nutrition Examination Survey (NHANES II) Linked Mortality File, mortality follow-up through 2006: Matching Methodology*.

[2] A detailed description of how NDI developed the cut-off scores can be found in *National Death Index User’s Guide*.

process described above. Ten unique individuals were deleted because they had a score less than or equal to zero: five were missing SSN and one had a first name of “infant.” For all ten records, the date of birth, name, and SSN did not agree with the potential NDI match. Applying the date of death boundary resulted in a loss of 63 unique individuals: 44 were missing SSN, one was missing sex, and six included some form of “baby” in the first name. This is the only elimination step where potential true matches were also eliminated, based on the date of death on the NDI record not being within plus or minus one day of the date of discharge. There were 28 individuals that were eliminated that would have been included on the basis of class and cut-off score alone. The difference between date of death and date of discharge ranged from (-30 to -2) and (2 to 9). Finally, 147 individuals were eliminated using the class and cut-off scores provided by NDI: 97 were missing SSN. A review of these records also indicated that when SSN matched, the name on the NHCS and NDI records did not match, and vice versa. Unlike in the rejected records manual review, no one hospital appears to have a disproportionate share of no match or false match records.

The review emphasizes the importance of SSN as a variable in the matching process. Overall, almost 94% of the records that were not matched were missing SSN. Another common issue was the unnamed baby. The baby records consist of first name entries ranging from “BABY A” and “INFANT” to “BABY” followed by the mother’s name (e.g., a boy baby with the first name of “BABY JANE”).

3.2 Mortality Indicators

The match selection process for the “discharged alive” subsample is adapted from the “known dead” process and results in the 30-, 60-, and 90-day mortality indicators. First, NHCS-NDI potential match records with a score of less than or equal to zero are considered false matches and eliminated from the pool of potential matches. Next, potential match records that have a date of death on the NDI record less than minus one day to the date of discharge are considered false matches and eliminated (because they were known to be alive on the date of discharge). Among the remaining pool of potential matches, duplicate death certificate numbers (i.e. match records that referred to the same death certificate) are eliminated within likely individuals, retaining the match with the highest score. Of the remaining potential matches, the NDI match with the highest score is selected as the single best record match. Within each class, matches are determined to be true or false using the cut-off scores developed by NDI. Overall, 14,785 patients (5 percent) in the preliminary 2011 NHCS data are confirmed to have died after discharge within calendar year 2011.

At the end of the match selection process, 30-, 60-, and 90-day mortality indicators were calculated (Table 3). Of the patients that died after discharge, 52 percent died within 30 days, 69 percent died within 60 days, and 79 percent died within 90 days of discharge.

Table 3: Mortality indicators

	Percent	*Number*
30-day	52	7,634
60-day	69	10,186
90-day	79	11,651
Greater than 90 days	100	14,785

4. Conclusions

Based on the high match rate and match quality of this preliminary linkage of the NHCS and NDI, the Center is optimistic that in the future linked mortality data will contribute in meaningful ways to health services research. However, steps are being taken to increase the match rate. In the next round of linking to the NDI, newborn infant records will have state of birth imputed using the hospital's state, since state of birth is not a discrete data element collected through NHCS. Additionally, the results from this preliminary analysis will be shared with participating hospitals that do not provide SSN, as well as hospitals considering participation in the NHCS, to encourage the submission of SSN on their data files.

Acknowledgements

The authors would like to acknowledge the James Brittain and other analysts in the Office of Analysis and Epidemiology at the National Center for Health Statistics and the work that they did to clean the hospital discharge file for submission to the National Death Index.

References

National Center for Health Statistics. National Death Index user's guide. Hyattsville, MD. 2013.

Winglee, M., Valliant, R., and Scheuren, F. (2005) A case study in record linkage. *Survey Methodology*, 31(1), 3-11.