

Data Linkage with an Establishment Survey

Jennifer Sayers¹, Scott Campbell², Clinton Thompson¹, Geoffrey Jackson¹

¹Centers for Disease Control and Prevention, National Center for Health Statistics, ²NORC at the University of Chicago

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference

Background

The National Center for Health Statistics (NCHS) Data Linkage Program links NCHS survey data with vital and administrative records, such as the National Death Index (NDI), Centers for Medicare and Medicaid Services (CMS) enrollment and claims data, and the Housing and Urban Development (HUD) housing assistance program data [1]. NCHS surveys include both household and establishment surveys.

One of the establishment surveys conducted by NCHS is the National Hospital Care Survey (NHCS). NHCS was designed to provide accurate health care statistics that answer key questions of interest to health care and public health professionals and researchers. A sample of 581 hospitals across the United States were selected and data was collected from the participating hospitals [2]. In 2014, 95 hospitals provided data. NHCS provides patient and encounter-level information on patient health characteristics and health care utilization from hospital inpatient stays and outpatient and emergency department visits, along with provider characteristics. 2014 NHCS data were collected using elements of the Uniform Bill (UB-04), a standard claims form [3]. These elements include personally identifiable information (PII), demographic information, encounter dates, and codes for diagnosis, procedures, and revenue. In NHCS, PII such as name, address, date of birth, and social security number are key factors for data linkage. Since NHCS collects PII of patients seen in the sampled hospitals, it presents a unique opportunity for entity to entity data linkage to administrative and vital records. Although NHCS is not currently nationally representative due to low response rates, 95/581=16% in 2014, linking NHCS with other data sources will allow for new analyses, such as studying mortality post hospital discharge, along with specific causes of death.

With support from the Office of the Secretary Patient Centered Outcomes Research Trust Fund (OS-PCORTF), the 2014 NHCS was recently linked with death records from NDI occurring in 2014-2015. NDI is a centralized database of death record information on file from state vital statistics offices in the United States and is housed at NCHS[4]. Death records have been added to the database annually since 1979, and records contain information on the state of death, date of death, death certificate number, and cause of death.

Given that NHCS is an establishment survey, data processing was distinct from the household surveys typically linked within the NCHS Data Linkage Program. The processing of the NHCS data included developing new approaches for standardizing the data and collating information collected at more than one visit to facilitate record linkage. This paper presents the four main objectives and accompanying solutions for standardizing the NHCS data: 1) processing the data, 2) cleaning the data, 3) accounting for incomplete records, and 4) storing the data. In addition, a brief overview of the modified linkage algorithm used to link NHCS to NDI is explained, along with the total number of matches.

Objectives

Objective 1: Processing the data

The 2014 NHCS data were distinct from most of the household survey data since they may contain multiple records for each encounter the patient had in the hospital or inpatient record (Figure 1). For a single patient, it was possible to have different names, addresses, dates of birth, and other PII on different encounters (see Objective 2: “Cleaning the data” for a description of how multiple records were handled). A key element to processing the data was understanding the timing of the encounters. For each encounter, there was a date of admission and discharge. Processing the data required setting one beginning and one end date while maintaining PII for all encounters. This process insured that each alternate name, date of birth, etc. had the same beginning and end date. To increase the likelihood of finding a match, multiple or alternate records were retained for each linkage-eligible patient. Having a consistent beginning and end date on the alternate records establishes the time frame the patient could technically match to NDI.

PATIENT_ID	Beginning date of encounter	End date of encounter	NAME (Last, First)	DOB	ADDRESS	SSN
Patient_A	02/15/2014	02/16/2014	Wane, Jalisa	05/18/1962	Prisk Place	SSN 1
Patient_A	03/05/2014	03/10/2014	Whane, Jalissa	05/01/1962	Prisk PL	SSN 2
Patient_A	06/10/2014	06/12/2014	Wane, Jalisa	01/01/1960	Prisc Place	SSN 2
Patient_B	01/30/2014	02/05/2014	Madison Solomon	12/05/1990	Tazwell Circuit	SSN 1
Patient_B	09/15/2014	09/15/2014	Madiso Solomon	12/05/1990	Tazewell Circ	SSN 1

Figure 1. Sample of NHCS 2014 original survey data format. Names and addresses from Christen, 2012 [5] and not actual NHCS data.

Objective 2: Cleaning the data

The survey data included PII from multiple encounter records. In order to increase the efficiency of the linkage process, the data needed to be cleaned to remove all duplicate records in the data for each patient. The goal was to maintain all unique PII record information, by de-duplicating alternate records. First, middle, and last names, addresses, dates of birth, and other demographic variables were cleaned of duplicates. It was then revealed that the resulting alternate records for some patients were not unique. These often included name misspellings and spacing differences, and address abbreviations. Further steps were needed to mitigate all abbreviations and spacing before de-duplication. In order to maximize the number of name variations for linkage, last names with multiple parts, such as “De La Garza”, did not have spaces removed. This allowed an increase in name combinations for linkage, and thereby an increased likelihood of getting a match, as records could now be matched based on either “Garza” or “De La Garza”, and even “DeLaGarza”. Names with apostrophes connecting multiple parts, such as “O’Henry”, had the apostrophe removed, resulting in a record reading “OHenry”. Abbreviated street names, such as “St” and “Ave”, were fully spelled out to “Street” and “Avenue”. Once corrected, all the records were de-duplicated again, and different records that remained were kept and prepared for linkage. Records that included different name spellings, such as “John” and “Jon” were all maintained since it cannot be confirmed which record lists the correct form of the patient’s name. The NHCS also included many general names that were assigned to babies born in the sampled hospitals. Names such as “babygirl”, “babyboy”, and “babyJohnson” were removed, but all other PII for the patient’s record was retained.

Objective 3: Accounting for incomplete records

Part of the data standardization process included accounting for incomplete Social Security Number (SSN) values in the data. SSN is a key identifier for patients. A valid SSN has 9 digits, without repeating numbers for the entire SSN, such as 888-88-8888, and conforms to several other formatting standards. The original 2014 NHCS data included numbers in the SSN field with less than 9 digits, and some with invalid formats. These values were considered invalid and were not kept as the patient’s SSN. Additional checks were performed to validate SSN, which are outlined at <https://secure.ssa.gov/apps10/poms.nsf/lx/0110201035>. The number of invalid SSNs, combined with missing SSN values for many patient records, led to a low number of valid SSNs being available for linkage. In order to maximize the availability of SSN for linkage, we also extracted SSN from the patient’s Medicare Health Insurance Claim (HIC) numbers, because they often contain the patients’ SSN within them. Since HIC numbers can be linked to the primary beneficiary’s SSN, Insurance ID numbers in the NHCS data were screened for numbers matching the current HIC format. This method increased the population of unique patient records with a valid 9 digit SSN by 91% compared to using the information from the SSN field alone (Table 1).

Table 1. Number of unique patient records with SSN

Record type	Number of Patients
Valid SSN	474,675
Valid SSN derived from HIC	433,273
Total records with SSN	907,948

Objective 4: Storing the data

The NCHS Data Linkage Program has developed a process for storing survey data collected by NCHS in a standardized format in preparation for linkage to vital and various administrative record databases. The standardized data are stored in the Record Linkage Repository (RLR) in preparation for linkage. In RLR, survey information, including name, date of birth, address, and SSN (Figure 2), is stored in different tables. Storage of the 2014 NHCS data in the RLR required altering the tables to account for an increased number of alternate records within some RLR tables. The RLR subject table contains one record per unique patient identifier and includes information on the first and last contact dates, sex, and marital status. The earliest encounter date for each patient was assigned as the date of first contact and latest encounter date was set as the date last known alive. The latest known marital status was recorded for each patient, regardless if there were multiple statuses. Patients with multiple sexes reported were designated as having more than one sex, keeping only one record for the patient. Other tables such as name and date of birth had up to 8 records per patient ID.

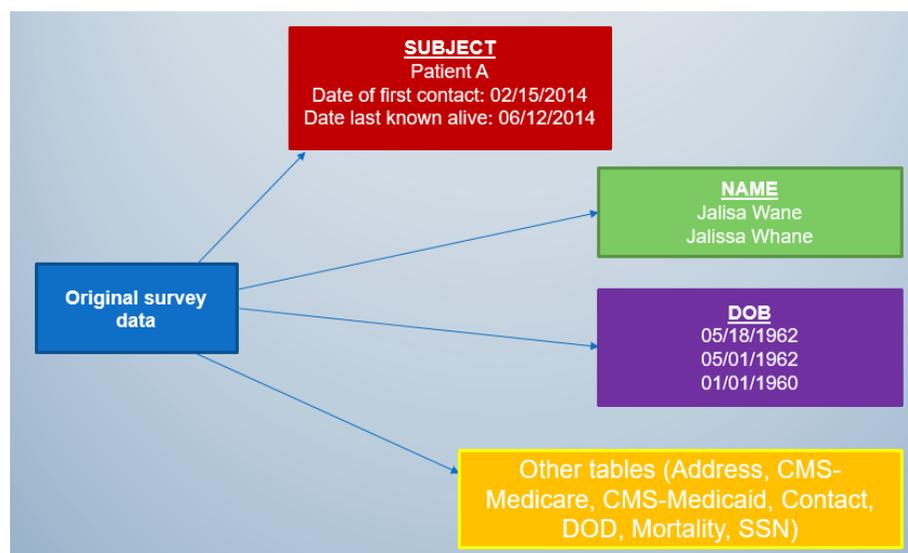


Figure 2. Sample of Record Linkage Repository format. Names and addresses from Christen, 2012 [5] and not actual NCHS data.

Linkage approach

Once the survey data were standardized and loaded into the RLR, a modified linkage algorithm was used to link the patient level data collected in NHCS to NDI. The modified algorithm incorporated both a deterministic and probabilistic component to identify matches, versus the primarily probabilistic approach applied for standard linkages of household survey data to NDI. The standard linkage algorithm also needed to be modified to accommodate the fact that race, ethnicity, and state of birth were not collected for all patient data in NHCS. Lastly, the modified algorithm incorporated date of death using patient discharge status. A two-part process was used for linking the 2014 NHCS data with the 2014-2015 NDI (details on this linkage process can be found at <https://www.cdc.gov/nchs/data-linkage/nhcs-ndi.htm> [6]). First, a deterministic match using SSN was completed, with name, state of residence, and date of birth used to validate matches. Second, a probabilistic match was conducted using other non-SSN PII. Of the 3,244,917 eligibleⁱ patient records in the 2014 NHCS, 168,253 records were linked to NDI in 2014-2015. Sensitivity analysis comparing the hospital discharge status as deceased with the linked NDI data showed that 98% of patients with a discharge status indicating they died in the hospital were linked

to NDI (data not shown). The linked file is currently available, through the NCHS Research Data Center, to researchers with an approved submitted proposal [7, 8].

Discussion and Conclusion

Processing establishment survey data for linkage purposes can differ from processing cross sectional household survey data due to the volume of encounters collected per patient. There is not a “one size fits all” approach since these data are captured in multiple encounters, requiring iterative de-duplication and cleaning. Taking the time to process the data in a unique manner helps to increase the pool of records that can be used in the linkage.

The processes for standardizing the 2014 NHCS data prior to conducting the linkage will permit future linkages to other administrative records. For instance, the 2014 NHCS data are currently being linked with CMS Medicare administrative records. The standardization process used for the 2014 NHCS will also be applied to future NHCS years, including the 2016 NHCS, which will contain data collected using electronic health records.

The resulting NHCS-NDI linked file is limited by the fact that 2014 NHCS is not nationally representative. Any analyses using the file cannot be generalized to the population of U.S. hospitals. However, the linked data augment the survey data and allow for analyses that would not be possible with each source alone. The NHCS-NDI linked data can be analyzed for 30-, 60-, and 90-day mortality post hospital care, along with causes of death.

References

1. CDC - NCHS Data Linkage. November 9, 2017 [cited 2018 April 9]; Available from: <https://www.cdc.gov/nchs/data-linkage/index.htm>.
2. National Hospital Care Survey. February 16, 2017 [cited 2018 April 9]; Available from: <https://www.cdc.gov/nchs/nhcs/index.htm>.
3. Official UB-04 Data Specifications Manual 2016. 2015, American Hospital Association.
4. National Death Index. March 7, 2017 [cited 2018 April 9]; Available from: <https://www.cdc.gov/nchs/ndi/index.htm>.
5. Christen, P., *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. 2012: Springer Publishing Company, Incorporated. 289.
6. National Center for Health Statistics. Office of Analysis and Epidemiology. *The Linkage of the 2014 National Hospital Care Survey to the 2014/2015 National Death Index: Methodology Overview and Analytic Considerations*, April 2018. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data/datalinkage/NHCS14_NDI14_15_Methodology_Analytic_Consider.pdf
7. National Hospital Care Survey Data Linkages. February 27, 2018 [cited 2018 April 9]; Available from: https://www.cdc.gov/nchs/data-linkage/nhcs_linkage.htm.
8. NCHS Research Data Center. December 16, 2015 [cited 2018 April 9]; Available from: <https://www.cdc.gov/rdc/index.htm>.

ⁱ For NHCS 2014, a patient record must contain at least two of the following: SSN; two parts of the name (first, middle, last); and/or two parts of the date of birth (day, month, year) to be considered eligible for linkage.