National Center for Health Statistics
Data Linkage

Data Linkage with an Establishment Survey

Jennifer Sayers[1], Scott Campbell[2], Clinton Thompson[1], Geoff Jackson[1]

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

Federal Committee on Statistical Methodology
Research and Policy Conference
March 9, 2018

Outline

1. Introduction: NCHS data linkage program
2. Background on National Hospital Care Survey (NHCS) and National Death Index (NDI)
3. Data processing objectives & solutions
4. NHCS-NDI linkage overview
5. Conclusion and future directions

Introduction

- The NCHS data linkage program links survey data with the NDI and other administrative data on a recurring basis

- Recently the NHCS data was linked with the NDI
  - The linkage was conducted with support from the Office of the Secretary Patient Centered Outcomes Research Trust Fund (OS-PCORTF)

- This linkage presented new challenges
  - NHCS is an establishment survey

National Hospital Care Survey

- Collects health care statistics from 581 non-institutional, non-federal U.S. hospitals with six or more staffed inpatient beds
- Not currently nationally representative
- Data collected as elements of UB-04:
  - Personally identifiable information
  - Demographic information
  - Encounter dates
  - Diagnoses and procedures
  - Revenue codes

https://www.cdc.gov/nchs/dhcs/index.htm

National Death Index (NDI)

- Centralized database of death record information on file in state vital statistics offices in the U.S.
- Death records added annually from 1979-2016
- Includes information on:
  - State of death
  - Date of death
  - Death certificate number
  - Cause of death (through NDI Plus service)

https://www.cdc.gov/nchs/ndi/index.htm

Objective 1: Processing the data

- For most linkages conducted in the NCHS Data Linkage Program
  - Typically one record per linkage eligible participant
  - Some alternate records are created based on nicknames or parsing out a last name with a hyphen

- However, the NHCS data are different
  - NHCS data are at the patient level, but not collected in the same way as other NCHS household surveys
  - Multiple encounters per patient
  - Multiple names, dates of birth, even at times different sexes reported

- Solution:
  - Transposed records to do a "roll-up" of alternate records per patient

Objective 2: Storing the data

- NCHS Data Linkage Program developed a process to store survey data in an standardized format for linkages

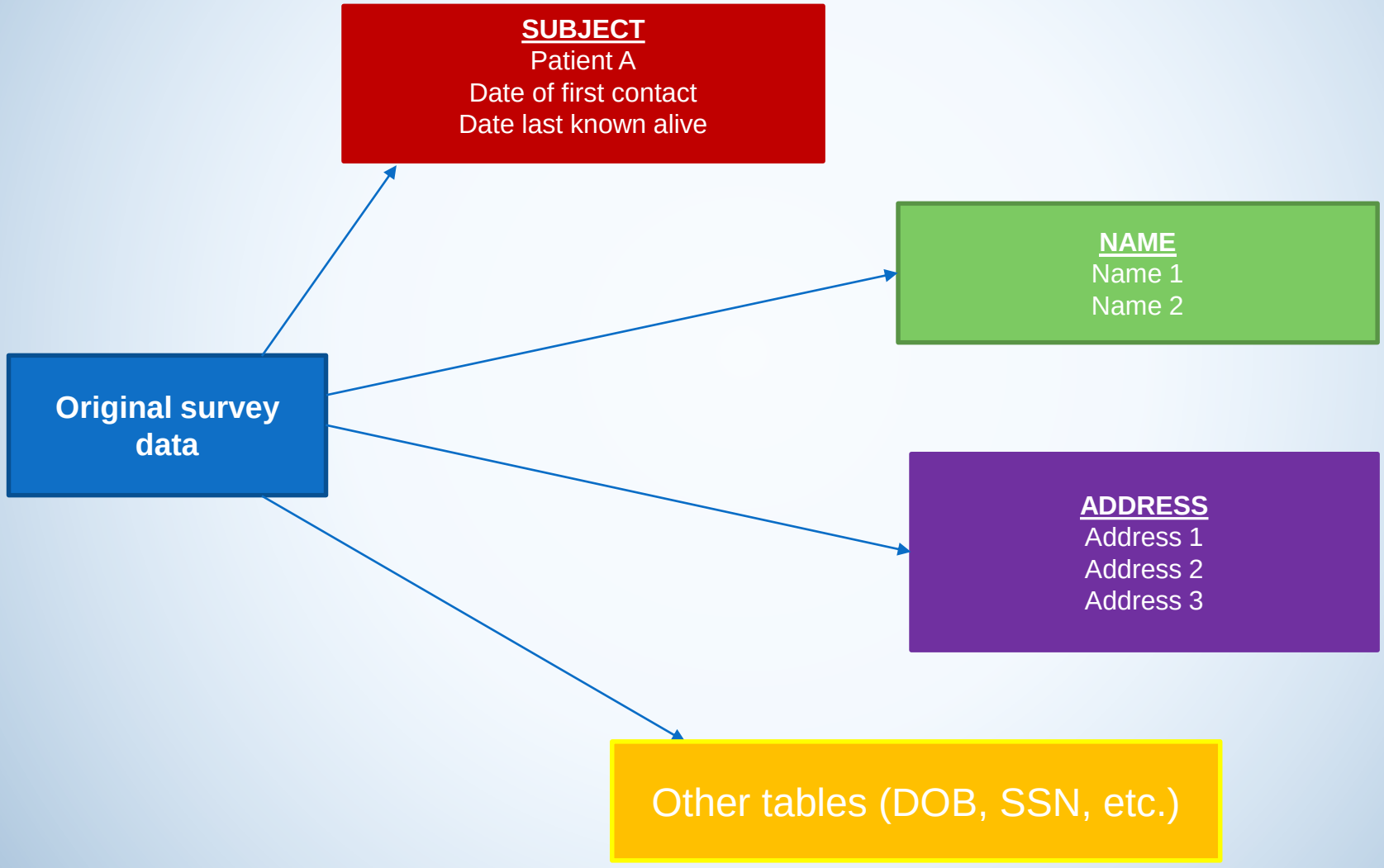
- Different pieces of information are broken up into different tables and stored in the record linkage repository (RLR)

- Solution:
  - NHCS posed unique storage challenges
  - New coding for multiple alternate records
  - New methods for ranking records

Original survey data

PATIENT_ID	Date of Encounter	NAME	ADDRESS	DOB	SSN
Patient_A	02/15/2014	Name 1	Address 1	DOB 1	SSN 1
Patient_A	03/05/2014	Name 2	Address 2	DOB 2	SSN 2
Patient_A	06/10/2014	Name 1	Address 3	DOB 2	SSN 2
Patient_B	01/30/2014	Name 1	Address 1	DOB 1	SSN 1

Record Linkage Repository (RLR)



Objective 3: Cleaning up the data

- Multiple records per patient were not always due to actual different records
  - Includes misspellings and abbreviations
    - Pear St./Peer St., John/Jon
    - St/Street, Ave/Avenue
  - Solution:
    - Spell out all abbreviations before deduplication
    - Collect all records with different spellings

- "Baby" names found in inpatient records
  - Eg. "Babygirl", "Babyboy", "BabyAnderson"
  - Solution: removed these names, but kept other valid PII for the patient

Objective 4: Accounting for incomplete records

- Missing information:
  - Race/ethnicity
    - Information not loaded
  - Contact dates
    - Used discharge date and discharge status to determine date of death or last encounter data from the survey
  - SSN
    - Medicare Health Insurance Claim (HIC) numbers were derived from insurance ID numbers following the CMS format

      #########$ | #########$$ | #########$#

Valid SSN	474,675
Valid SSN derived from HIC	433,273
Total records with SSN	908,916

** Valid SSNs are those with full 9 digits and correctly formatted*

Standardized data ready for linkage

- How to link NHCS 2014 inpatient and emergency department (ED) data to 2014-2015 NDI

- The standard linkage algorithm was enhanced to accommodate the structure for the NHCS
  - Certain fields were not available (race, state of birth)
  - Other fields were part of the data that could be used in the algorithm (date of discharge where status=dead)

Linkage Approach

- Enhanced linkage algorithm conducted in two passes:
  1. Deterministic match using SSN from hospital records
     - Identifier fields: name, state of residence, and date of birth are compared to validate
  2. Probabilistic matching techniques used to identify likely pairs using other identifiers (not SSN)
     - SSN is not used to create the match pool so it is used to measure linkage accuracy

Preliminary Match Rates

- In the file, out of **3,244,917** eligible patient records, **168,253** linked to the NDI in 2014/2015
  - In order to be considered eligible for linkage, the patient's record must contain at least two of the following: SSN, name, birth date

- This file is currently available for researchers through the NCHS Research Data Center with approved proposals

  https://www.cdc.gov/rdc/index.htm

Future Steps

- Use linked data to study mortality post hospital discharge (30-, 60-, 90-day mortality) and cause of death
- Electronic Health Records – NHCS 2016
- Link to other administrative data
  - CMS Medicare
- Link later years of NHCS and NDI

Conclusion

- Processing of establishment survey data differs from other survey data used in the Data Linkage Program
- Not one size fits all
- Taking the time to process the data helps increase pool of records that can be used in the linkage

Acknowledgements

Special thanks to Dean Resnick and the Division of Health Care Statistics

Co-authors:

Scott Campbell | Clinton Thompson | Geoff Jackson

Data Linkage Team:

James Brittain | Eileen Call| Adam Fedorowicz | Cordell Golden | Dedun Ingram | Lisa Mirel | Yu Sun | Clinton Thompson | Yeats Ye | Keith Zevallos | Cindy Zhang

Questions? Comments?

Thank you!

Email: JSayers@cdc.gov

https://www.cdc.gov/nchs/data-linkage/index.htm