Office of the Secretary Patient-Centered Outcome Research Trust Fund Project

(IAA #750120PE090051)

National Center for Health Statistics
Centers for Disease Control and Prevention

Augmenting the National Hospital Care Survey (NHCS) Data through Linkages
with Administrative Records: **Conduct Privacy Preserving Record Linkage (PPRL)**
Methodological Work and Link the National Hospital Care Survey Data to Centers for
Medicare & Medicaid Services T-MSIS data

FINAL REPORT

May 2024

Table of Contents

Team Members

Team Member Organizations	Team Members
Division of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention	Cordell Golden – Project Lead Crescent Martin Jessie Parker Lisa Mirel (former Project Lead)
Division of Health Care Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention	Carol DeFrances Geoff Jackson Donielle White
NORC	Ed Mulrow - Project Manager Dean Resnick Scott Campbell Shawn Linman Chris Cox Dave Baugh

***Augmenting the National Hospital Care Survey (NHCS) Data through Linkages
with Administrative Records:*** **Conduct Privacy Preserving Record Linkage (PPRL)
Methodological Work and Link the National Hospital Care Survey Data to Centers for
Medicare & Medicaid Services T-MSIS data**

FINAL REPORT

1. Executive Summary

The National Center for Health Statistics (NCHS) houses unique national data resources and expertise, including (1) the National Hospital Care Survey (NHCS), which provides statistics on health and health care utilization based on patient hospital inpatient and emergency department (ED) visits, through the collection of administrative claims records and electronic health records (EHR), as well as information on patient characteristics; and (2) the NCHS Data Linkage Program, which has substantial statistical and methodological expertise in creating, managing, and analyzing linked files. Although various organizations compile and disseminate data collected by hospitals, few integrate subsequent health care coverage information or contextual variables, like Medicaid claims and administrative data. Through data linkage, the NCHS Data Linkage Program creates rich new data resources that can be used to study the relationship between health insurance and health outcomes, outpatient care and inpatient care, and participation in federal housing programs and management of chronic health conditions. This project utilized the agreement between NCHS and the Centers for Medicare & Medicaid Services (CMS) to link survey data with claims records from CMS.

This project was three pronged. The first undertaking was to compare the results of a commercially available privacy preserving record linkage (PPRL) software to the results of a clear text matching approach to link survey and administrative records. The second undertaking was to link 2016 NHCS data to CMS Transformed Medicaid Statistical Information System (T-MSIS) administrative records. The third undertaking was to further expand the analytic utility of the previously linked 2014 and 2016 NHCS-CMS Medicare data by filling data gaps with additional Medicare program enrollment, health care utilization and expenditure data. The enhanced data linkage algorithm that was adopted to link the survey and T-MSIS data and the validation of commercially available PPRL software are critical additions to the portfolio of methods and institutional knowledge used by the NCHS Data Linkage Program that can be used to inform others interested in linkage methodology at the Department of Health and Human Services and beyond. Filling the data gaps in the linked NHCS-CMS Medicare data to enable trend analyses and the assessment of utilization and expenditures over time as well as the examination of health care utilization among patients enrolled in the rapidly expanding Medicare Advantage (MA) program.

2. Background

2.a Goal

This project aimed to conduct an evaluation of commercially available PPRL software; link inpatient and ED claims and EHR data from the NCHS' National Hospital Care Survey (NHCS) to administrative T-MSIS data from CMS, using clear text matching; and fill data gaps in the previously linked NHCS-CMS Medicare data.

2.b Data Sources

National Center of Health Statistics, National Hospital Care Survey
NHCS is an establishment survey that collects inpatient and ED, and outpatient department (OPD) visit level data from sampled hospitals. NHCS is one of the National Health Care Surveys, a family of surveys covering a wide spectrum of health care delivery settings from ambulatory and outpatient to hospital and long-term care providers (https://www.cdc.gov/nchs/dhcs/index.htm). The goal of the NHCS, when fully implemented, will be to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, health services utilization, and substance-use involved ED visits.

NHCS collects patient personally identifiable information (PII) such as name, date of birth, and Social Security Number (SSN), which allows for the linkage of episodes of care within a surveyed hospital as well as linkages to other external data sources, such as the CMS and HUD administrative data. The linkage described in this report includes inpatient and ED claims, as well as EHR data.

Eligible hospitals for NHCS are non-institutional, non-federal hospitals with six or more staffed inpatient beds. There are 6,622 hospitals which met these criteria as of 2013 to form the survey sampling frame. The hospital sample size for the 2014 and 2016 NHCS data collection (which re-use the 2013 sample) was 581 hospitals.

Data from all inpatient, ED, and outpatient visits occurring during the calendar year are collected from NHCS participating hospitals. In 2014, 95 NHCS participating hospitals provided these data in the form of UB-04 administrative claims records.

For 2016, in an effort to reduce the burden of transferring their hospital records, participating hospitals were given the option to submit their patient records using the UB-04 claims data format or to submit extracts from their EHR systems. Hospitals submitting EHR records did so in the form of a custom extract or Consolidated Clinical Document Architecture (C-CDA), which are a set of Health Level Seven International (HL7) clinical document architecture specifications and include templates such as Continuity of Care Documents (CCDs) (1). In 2016, 158 hospitals participated in the NHCS. Of these hospitals, 89 hospitals provided UB-04 administrative claim records and 47 provided EHR records – 16 as custom extract and 31 as CCD. The remaining 22 hospitals provided records submitted in the format to Vizient, a healthcare performance company, which are similar to UB-04 data but contain no patient PII so were not used in linkage.

National Center of Health Statistics, National Death Index (NDI) Data
The NDI is a centralized database of United States death record information on file in state vital statistics offices. Working with these state offices, NCHS established the NDI as a resource to aid epidemiologists and other health and medical investigators with their mortality ascertainment activities. The NDI became operational in 1981 and includes death record information for person dying in the U.S. or a U.S. territory from 1979 onward. The records, which are compiled annually into the longitudinal register, include detailed information on the underlying and multiple causes of death.

Centers for Medicare and Medicaid Services, Medicaid T-MSIS Data
Enacted in 1965 as Title XIX of the Social Security Act, Medicaid is a federal and state partnership to provide health insurance coverage to low-income individuals in the United States. The program has changed continuously since it was enacted through a series of legislative action. Medicaid is jointly financed with federal and state/local funds. States must meet federal requirements to receive federal funding. Management and oversight activities are shared by federal and state governments, with identified federal and state roles and responsibilities. There is significant variation among state Medicaid programs in both the eligible population groups and the covered services. For this reason, each state must develop and maintain a state Medicaid plan to assure that the state abides by federal requirements for administering its program and claiming federal matching funds.

Over 85.8 million individuals were enrolled in Medicaid and CHIP in the District of Columbia and the 50 states that reported enrollment data for November 2021. Among that total, over 78.9 million individuals were enrolled in Medicaid and nearly 6.9 million individuals were enrolled in CHIP. Enrollment in these programs represented over one quarter of the U.S. population, with over 90 million individuals enrolled for at least one day in 2018. Medicaid provided coverage for 42.3% of U.S. births in 2018.

There are five Medicaid/CHIP files available to analysts who have access to the linked NHCS/T-MSIS data. The Demographic and Eligibility (DE) file contains demographic and enrollment information on persons enrolled in Medicaid and/or CHIP. The remaining T-MSIS Analytic Files (TAFs) contain claims records for services provided under fee-for-service, premium payments to prepaid managed care plans, and encounters for services provided by managed care plans, and are organized into four claims files: inpatient hospital services (IP), long-term care services (LT), pharmacy services (RX), and all other services (OT) organized by date of service. Further, excluding the pharmacy (RX) which has only a header and line file, each of the claim's files has a header, line, and occurrence file. Additional information about what can be found in each of the files is detailed below:

Demographic and Eligibility (DE) File – This file provides demographic and program eligibility and enrollment information on each person who was enrolled for at least one day in the calendar year in Medicaid and/or CHIP. Demographic data elements in the DE file include race and ethnicity; primary language; and marital status. Eligibility data elements include Medicaid and CHIP enrollment days, eligibility group, CHIP program, dually eligible individual status,

restricted benefit status, and participation in managed care, for each month in the calendar year. The file also includes information about the enrollee's participation in other federal programs, such as Social Security Disability Insurance (SSDI) and SSI, and Temporary Assistance for Needy Families (TANF).

Inpatient Hospital (IP) File – This file includes records for inpatient hospital services for Medicaid and CHIP enrollees during the calendar year. Emergency room visits that result in an inpatient hospital admission are identified in Uniform Billing (UB-04) revenue codes (T-MSIS claim line-item data element REV_CNTR_CD).

Long-Term Care (LT) File – This file includes records for institutional long-term care services for Medicaid and CHIP enrollees during the calendar year. Records include claims for room and board, which may include prescribed drugs if they are included in the institution's per diem rate, which has historically been the case in only a small number of states. LT records also include ancillary services, such as speech therapy or specialized dietary services, if they are provided by the institution's staff. Otherwise, prescribed drugs and ancillary services are reported in the RX and OT files, respectively.

Pharmacy (RX) File – This file includes records for prescribed drugs, supplies and other items provided by a free-standing pharmacy, either directly to an enrollee or to a long-term facility for the enrollee's use. This includes prescribed and covered over-the-counter drugs, supplies, and durable equipment.

Other Services (OT) File – This file includes records for all other community-based services not reported in the IP, LT, and RX files. These services include physicians (including separately billed services provided to patients during inpatient hospital stays), clinic, laboratory, radiology, EPSDT, home health, dental, therapy, transportation, case management, family planning services, waiver services, and Home- and Community-Based Services (HCBS).

Centers for Medicare and Medicaid Services, Medicare Data
Medicare is the primary federal health insurance program for people age 65 or older, people under age 65 with qualifying disabilities, and people of all ages with End Stage Renal Disease (ESRD). In 2016–2017, nearly two-thirds of persons enrolled in Medicare, known as Medicare beneficiaries, were enrolled in original Medicare, also known as Medicare fee-for-service (FFS). Nearly all Medicare FFS beneficiaries receive Part A hospital insurance benefits, which help cover IP hospital care, Skilled Nursing Facility (SNF) stays (not custodial or long-term care), home health care, and hospice care. Most FFS beneficiaries also enroll in Medicare Part B medical insurance benefits, which help to cover physician services, OP care, durable medical equipment (DME), and some home health care services.

In 2016–2017, approximately one-third of Medicare beneficiaries received Medicare benefits through a Medicare Advantage (MA) plan, also known as Medicare Part C. MA plans are administered by approved Medicare Advantage Organizations (MAOs). MAOs sponsor privately managed care plans such as Health Maintenance Organization (HMOs), Preferred Provider

Organization (PPOs), and Special Needs Plans (SNPs) which provide, at a minimum, the same covered services provided in Medicare Parts A and B. MAOs may also elect to provide additional services not covered by Medicare Parts A and B such as dental and vision care. MAOs are responsible for providing Medicare benefits directly to enrollees through prior arrangements with providers or by paying for the benefits on behalf of enrollees.

In 2006, Medicare beneficiaries could begin to elect optional prescription drug coverage, known as Medicare Part D. Part D coverage can be obtained through Medicare approved Part D private plans, known as Prescription Drug Plans (PDPs) or through Medicare Advantage Prescription Drug Plans (MA-PDs). MA-PDs provide prescription drug coverage that is integrated with the health care coverage provided to Medicare beneficiaries enrolled in MA plans.

The CMS Medicare Data Files are comprised of Standard Analytic Files, or SAFs, containing standard format extracts of research-oriented Medicare program data. The CMS Medicare Data Files contain information on the enrollment status, health care utilization, and expenditures of Medicare-enrolled beneficiaries. The SAFs for Medicare beneficiaries enrolled in FFS Medicare contain final action health care claims submitted for payment by both institutional and noninstitutional health care providers. A final action claim contains all payment adjustments between Medicare and providers and represents Medicare's final payment action for a given health care claim. Medicare FFS SAFs are organized by seven health care settings: IP, SNF, institutional outpatient (OP), practitioner/provider services (Carrier), home health agency (HHA), DME, and hospice care.

The SAFs for MA-enrolled beneficiaries contain all health care encounter records submitted by MAOs for the given calendar year for each enrolled Medicare beneficiary. MA SAFs are organized by six health care settings: IP, SNFs, OP, Carrier, HHA, and DME. Hospice care services provided to Medicare beneficiaries enrolled in MA are paid under Medicare FFS rather than as part of the managed care plan.

The Medicare Part D Prescription Drug Event (PDE) File contains a summary of prescription drug costs and payment data used by CMS to administer benefits for all Medicare Part D enrollees including beneficiaries enrolled in both Medicare PDPs and MA-PDs.

In addition to the SAFs and the PDE Files, two assessments are also included in the linked dataset – the Home Health Outcome and Assessment Information Set (OASIS) and the LongTerm Care Minimum Data Set (MDS). The OASIS assessment contains data pertaining to patient outcomes and home health care. The OASIS assessments are required of all HHAs certified to accept Medicare and Medicaid payments. The MDS is a health status screening and assessment tool used for all residents of long-term care nursing facilities certified to participate in Medicare or Medicaid, regardless of payer. The MDS assessment is also required for Medicare payment of SNF stays.

2.c Tasks, Objectives, and Deliverables

This section outlines the tasks, the objectives, and the deliverables.

Table 1. Tasks, Objectives, and Deliverables.

Task	Objective	Deliverables
Task 1	• Assessment of Privacy Preserving Record Linkage Using the NHCS-NDI as the gold standard • Purchase of Datavant software license	• Peer reviewed journal article on the results of the privacy preserving record linkage validation: https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210891
Task 2	• Link the 2016 inpatient & ED claims and EHR data (2016 only) to CMS T-MSIS data • Create a written report on NHCS-T-MSIS data linkage methodology and analytic considerations to be published on the web	• New data file containing linked 2016 NHCS inpatient and ED claims and EHR to CMS T-MSIS data available to researchers through the NCHS and Federal Research Data Centers (RDC) [NCHS Data Linkage Restricted-Use Linked NHCS-CMS Medicaid Data (cdc.gov)] • Report on NHCS-T-MSIS data linkage methodology and analytic considerations published on the NCHS website. [The Linkage of the 2016 National Hospital Care Survey to 2015–2017 Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data: Matching Methodology and Analytic Considerations (cdc.gov)]
Task 3	• Research and Evaluation of the CMS T-MSIS Linked Data • Create NCHS publications and peer-reviewed articles detailing innovative aspects of the NHCS-T-MSIS linkage methods and assessments of linkage quality	• T-MSIS PII Analysis (The Linkage of the 2016 National Hospital Care Survey to 2015–2017 Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data: Matching Methodology and Analytic Considerations (cdc.gov)) • NHCS Race/Ethnicity Imputation Validation (D2Mirel_FCSM_Proceedings.pdf) • Citation list: https://www.cdc.gov/nchs/data/datalinkage/LinkedNHCSDataCitationList_508.pdf

Task	Objective	Deliverables
Task 4	• Add 2014 and 2015 Medicare claims, prescription drug, and assessment data to make the set of linked 2014 NHCS-Medicare files more comparable to the linked 2016 NHCS-Medicare data	• Additional linked data files containing 2014 NHCS inpatient and ED claims and EHR linked with 2014 and 2015 Medicare claims, prescription drug, and assessment data available to researchers through the NCHS and Federal Research Data Centers (RDC) [NCHS Data Linkage Restricted-Use Linked NHCS-CMS Medicare Data (cdc.gov)] • Report on 2014 NHCS-CMS Medicare data linkage methodology and analytic considerations updated and published on the NCHS website. [Linkage of the 2014 National Hospital Care Survey to 2014/2015 Medicare Data: Methodology Overview and Analytic Considerations (cdc.gov)]
Task 5	• Add 2017 Medicare Advantage (MA) encounter data to the set of linked 2016 NHCS-Medicare files to be consistent with data years for fee-for-service (FFS) claims	• Additional linked data files containing 2016 NHCS inpatient and ED claims and EHR linked with 2017 MA encounter data available to researchers through the NCHS and Federal Research Data Centers (RDC) [NCHS Data Linkage Restricted-Use Linked NHCS-CMS Medicare Data (cdc.gov)] • Report on 2016 NHCS-CMS Medicare data linkage methodology and analytic considerations updated and published on the NCHS website. [Linkage of 2016 National Hospital Care Survey to 2016/2017 CMS Medicare Data: Matching Methodology and Analytic Considerations (cdc.gov)]

3. Major Accomplishments

3a. Assessment of PPRL tools

The assessment of and subsequent publication on PPRL methods positioned the Data Linkage Program to serve as experts within CDC and the Department on deploying methods for PPRL. In fact, NCHS received funding through OS-PCORTF FY 22 to continue these assessments with different data sources and different PPRL tools. This work will have a lasting effect on agencies across government as they pursue linking data without sharing direct identifiers.

3a. Enhanced Linkage Algorithm

The linkage methodology utilized in this project includes technical enhancements not included in previous linkages and a robust analysis of linkage accuracy. The enhancement involved a new sequential coverage algorithm (SCA). The SCA is a machine learning technique that implements and tests a rule, evaluates its accuracy, then removes it from the linkage routine and tests a new rule. This approach was compared to a previously used linkage method that did not use the SCA. A comparison of Type I and Type II error rates determined that the SCA method was the most accurate.

Blocking is a key step in record linkage. It identifies potential candidate pairs without comparing every single pair in the Cartesian product. Blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)" (4). Rules can be used to define the blocking criteria however, for the CMS linkage, instead of rules, machine learning techniques were used to help create a set of blocks that would efficiently join the datasets together. By using the data to create the efficient block set, the number of false positive links were reduced while retaining a high percentage of true positive links. For the purpose of this linkage, the 'truth deck' was used as the training dataset. When the data are used in this manner, it is commonly referred to as a machine learning algorithm. For more detailed information on the method that was used please refer to "Learning Blocking Schemes for Record Linkage" (2) and the published methods and analytic guidelines (2, 3).

The new linkage algorithm also included an SSN odds adjustment to further improve the accuracy of the probabilistic matching within the linkage. Furthermore, the logistic regression step in the linkage algorithm was replaced with a partial expected-maximization (EM) model, which further improved the accuracy of the linkage. The linkage algorithms developed for this project were created and customized using SAS statistical software and saved for future use by the NCHS Data Linkage Program.

3b. Final Linkage Rate Table: 2016 NHCS-CMS T-MSIS Linkage

The tables below illustrate the linkage rates by demographic variables for the linked 2016 NHCS -2015-2017 T-MSIS claims records.

Table 3. Linked 2016 NHCS-CMS T-MSIS Claims Records: Sample Sizes and Percent Linked, by Age and Sex

	Sample Size			Percent Linked	
	Total Sample	**Eligible for Linkage[3]**	**Linked to CMS T-MSIS Claims Data[4]**	**Total Sample[5]**	**Eligible Sample[6]**
2016 NHCS					
Age[1]					
0-17	1,293,458	1,205,473	854,671	66.1	70.9
18-39	1,268,852	1,191,263	671,948	53.0	56.4
40-64	1,130,616	1,062,386	457,400	40.5	43.1
65 and over	762,766	717,624	178,472	23.4	24.9
Total	4,455,692	4,176,746	2,162,491	48.5	51.8
Sex[2]					
Male	2,597,453	1,851,201	915,274	35.2	49.4
Female	3,157,461	2,278,263	1,224,609	38.8	53.8
Total	5,754,914	4,129,464	2,139,883	37.2	51.8

NOTES: Data are presented at patient level.

[1] Age is as of final IP or ED encounter (date of last known contact). Age could not be determined for 1,367,473 patients in the 2016 NHCS due to missing data. Age is calculated by subtracting patient date of birth (DOB) from the final encounter date. When more than one DOB was present, the minimum of the non-missing DOB was selected.

[2] Sex could not be determined for 68,251 patients in the 2016 NHCS due to missing data.

[3] Eligibility for linkage is based upon having sufficient PII in at least two of three data element groups: SSN, name, and date of birth. 1,642,060 patients in the 2016 NHCS were missing all PII and were also considered ineligible for linkage.

[4] This group includes linkage-eligible patients who linked to CMS T-MSIS enrollment database at any time during the linkage interval (2016 NHCS: 2015 – 2017 CMS T-MSIS).

[5] This percentage is calculated by dividing the number of linked patients by the number of patients in the total sample.

[6] This percentage is calculated by dividing the number of linked patients by the total number of linkage-eligible patients.

3c. Linkage Error Estimation: 2016 NHCS-CMS T-MSIS Linkage

The estimation of Type I and Type II errors for the CMS linkage followed the methods described in the appendix of the methods and analytic guidelines report ([The Linkage of the 2016 National Hospital Care Survey to 2015–2017 Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data: Matching Methodology and Analytic Considerations (cdc.gov)](cdc.gov)) and are noted in Table 4.

Table 4. Algorithm Results for Total Selected Links by 2016 NHCS Data Source

Data Source	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
UB-04 Claims	0.92	1,690,830	309,062 (18.3%)	1,381,768 (81.7%)	0.04%	2.0%
EHR Custom Extract	0.92	265,694	158,041 (59.5%)	107,653 (40.5%)	0.01%	0.4%
CCD	0.92	207,062	0 (0%)	207,062 (100%)	0.07%	*

*Unable to estimate Type II linkage error due to no SSN information on CCD records.

4. Lessons Learned

Several critical lessons were learned from this project. The project illustrated that PPRL can work well when linking data with high quality PII and may support linkages in the future that may not be able to rely on linking with clear text matching. This project also highlighted the importance of creating high-quality linkage algorithms to combine data sources that can answer key policy and patient-centered research questions. The linkage from this project highlighted the importance of collecting patient personally identifiable information as part of the NHCS to enable such linkages and the resulting data files have been featured as important recruitment tool to encourage hospital participation. It also supported enhancements to linkage algorithms that will improve efficiency and accuracy for future linkage projects. This work aligns with the Evidence Act of 2018 and HHS Federal Data Strategy. The 2016 NHCS UB-04 administrative claims data, and EHR data were successfully linked to CMS T-MSIS data to produce linked files that are available to researchers through the NCHS and Federal RDCs. NCHS is accepting requests to access the linked files, and the data have been analyzed by internal NCHS staff.

5. Publications and Presentations

5a. Presentations:

- **May 19, 2021**, L. Mirel presented "Privacy Preserving Techniques: Case Studies from the Data Linkage Program" at the NCHS Board of Scientific Counselors Meeting. https://www.cdc.gov/nchs/data/bsc/bscpres-Mirel-Data-Linkage-20210519-508.pdf
- **May 21, 2021,** L. Mirel presented "Innovations in Data Integration" at the Data Foundation Data Symposium. https://www.datafoundation.org/events-list/data-symposium-2021-exploring-emerging-data-capabilities-in-government/2021
- **August 11, 2021,** L. Mirel, D. Resnick, C. Cox, and J. Aram presented "Expanding Data Linkage Opportunities and Measuring Quality" at the Joint Statistical Meetings (JSM). https://ww2.amstat.org/meetings/jsm/2021/onlineprogram/AbstractDetails.cfm?abstractid=317346
- **October 30, 2021,** S. Garcia, S. Jaeger, P. Keenan, S. Lumsden, and L. Mirel presented "Balancing Access to Federal Data Sets with Enhancing Privacy and Security" at the American Medical Informatics Association Annual Symposium https://s4.goeshow.com/amia/annual/2021/schedule_at_a_glance.cfm?session_key=A5F8FDA9-38D0-EB11-80EF-D1E7CDE2D2B0&session_date=Monday%2C%20Nov%2001%2C%202021
- **October 27, 2022**, J. Parker and C. Martin presented "Concordance Between Self-Report of Medicaid Enrollment in the National Health Interview Survey and Medicaid Administrative Records" at the Federal Committee on Statistical Methodology (FCSM) Research & Policy Conference. https://www.fcsm.gov/assets/files/docs/2022-conference-docs/H4.3_Parker.pdf
- **March 17, 2023**, J. Parker presented "National Center for Health Statistics Survey Data Linked to CMS Medicaid Data" at the Intellectual Disability/Developmental Disability Administrative Data Federal Workgroup Monthly Meeting
- **August 8, 2023,** C. Golden presented "Privacy Preserving Techniques: Case Studies from the National Center for Health Statistics" at the Joint Statistical Meetings (JSM).

5b. Publications:

- Ashman J, Cairns C, DeFrances C, Schwartzman A. Respiratory Illness Emergency Department Visits in the National Hospital Care Survey and the National Hospital Ambulatory Medical Care Survey. National Health Statistics Reports; no 151. Hyattsville, MD; National Center for Health Statistics. 2021. https://www.cdc.gov/nchs/data/nhsr/nhsr151-508.pdf

- Resnick D, Cox C, Mirel L. Using Synthetic Data to Replicate Linkage Derived Elements: A Case Study. Health Services and Outcomes Research Methodology; 21, 389-406. 2021. https://link.springer.com/article/10.1007/s10742-021-00241-z
- Campbell S, Resnick D, Cox C. Using Supervised Machine Learning to Identify Efficient Blocking Schemes for Record Linkage. Statistical Journal of the IAOS, vol 37, no. 2, pp673-680. 2021. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200779
- Mirel L, Resnick D, Aram J, and Cox S. A Methodological Assessment of Privacy Preserving Record Linkage Using Survey and Administrative Data. Statistical Journal of the IAOS, vol 38, no. 2, pp413-421. 2022. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210891
- Mirel, L.B., Resnick, D., Parker, J.L., Zhang, C. Cox, C. (2022). Race and Ethnicity Modeling Applied to Linked Health Care Data. In Proceedings of the Federal Committee on Statistical Methodology Research Conference, Session D-2. https://www.fcsm.gov/assets/files/docs/2022-conference-docs/D2Mirel_FCSM_Proceedings.pdf

6. Future Considerations

This project opened the door for many other opportunities, including PCORTF FY21 project that explores the creation of synthetic linked data files and PCORTF FY 22 project to continue the assessment of privacy preserving record linkage methodology with different sources of data and different PPRL tools.

7. Summary

This project accomplished a comprehensive data linkage to provide a new source of data for the PCOR community. The methods developed and subsequent improvements in linked data quality through this project will be utilized in future NCHS data linkage projects. NCHS will continue to monitor and promote the use of the linked NHCS data. As more years of the NHCS become available, the algorithms developed through this project will continue to be used to link new sources of data and create new resources.

Additionally, the project assessed the feasibility of utilizing commercially available PPRL solutions. The results of this validation will be used to inform NCHS's future decisions for conducting linkages utilizing PPRL tools.

8. How to Request Linked NHCS Data

The linked NHCS – T-MSIS and NHCS – Medicare files are made available through the NCHS Research Data Center (RDC) or Federal Statistical Research Data Centers (FSRDC). Researchers must submit a written proposal that will be reviewed by NCHS staff. For more information on RDC access, please see this link: https://www.cdc.gov/rdc/. Questions related to the linked files can be directed to the NCHS Data Linkage Team (datalinkage@cdc.gov).

9. References

1. National Center for Health Statistics. National Hospital Care Survey Research Data Center Documentation. Hyattsville, MD. 2019. Available from: https://www.cdc.gov/rdc/data/b1/NHCS_RDC_USERS_GUIDE.pdf

2. Michelson M, Knoblock CA. "Learning Blocking Schemes for Record Linkage." In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, 440–445. AAAI'06. Boston, Massachusetts: AAAI Press, 2006

3. National Center for Health Statistics. Division of Analysis and Epidemiology. *The Linkage of the 2014 National Hospital Care Survey to the 2014/2015 Centers for Medicare & Medicaid Services Master Beneficiary Summary File: Methodology Overview and Analytic Considerations*, August 2019. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/index.htm

4. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag, 2012. http://www.springer.com/us/book/9783642311635