Final Report

National COVID-19 Longitudinal Research Database Linked to Medicare and Medicaid Data

Dr. Kenneth Gersing
Director of Informatics
NCATS, Division of Clinical Innovation

November 2024

TABLE OF CONTENTS

LIST OF FIGURES

PCOR Final Report:
CMS-N3C Linked Dataset and Use Cases

Executive Summary

The Patient-Centered Outcomes Research Trust Fund (PCORTF) was established in December 2010 through the Patient Protection and Affordable Care Act to expand comparative effectiveness research through patient-centered outcomes research (PCOR). More specifically, the Affordable Care Act of 2010 (ACA), as amended, authorized the Secretary of the U.S. Department of Health and Human Services (HHS) to provide for the coordination of relevant Federal health programs to build data capacity for patient-centered outcomes research.

National COVID Cohort Collaborative (N3C).

The overall goal of this project is to enhance the COVID-19 data capacity for patient-centered outcomes researchers by linking claims data from the Centers for Medicare and Medicaid (CMS) with clinical electronic health record (EHR) data in the National Center for Advancing Translational Sciences (NCATS) National COVID Cohort Collaborative (N3C). NCATS, along with its government, academic, and industry partners created N3C as an open science collaborative analytics resource starting in May of 2020. In the subsequent 4 years N3C has become the largest centralized longitudinal COVID-19 repository in the US, at the time of this writing, over 240 participating care delivery organizations have contributed the de-identified medical records of over 23 million individuals from all 50 states.

Over 4100 investigators from more than over 400 organizations currently have access to NCATS's secure FISMA moderate government cloud collaborative analytics infrastructure of over 33 billion rows of data and have conducted over 580 unique clinical studies. A list of all N3C COVID studies is available and consists of work ranging from Long Covid outcomes to the impact of Social Determinants of Health, SDOH, on COVID outcomes.

Rationale for Linked Data

Despite its size, N3C has limitations. The data in N3C comes from de-identified data from Electronic Health Records, EHR at specific health care systems. Because of the fractured nature of the US health system patients often get their care from multiple health systems and therefore the clinical information pulled from one institution is often incomplete. The N3C PPRL Enrichment Dashboard quantifies the amount of missing data when only looking at EHR without the addition of claims data. The impact on missing medical data is documented in the results of an analysis, by Makkar, Estepp, and Sidky in the preprint *Combining EHR and claims data in N3C increases scientific validity*, which concludes that multi-model data used in N3C COVID lead to greater scientific validity and without it incorrect conclusions are likely

Even if the health care data was comprehensive the information is limited. De-identified EHR data by itself is necessary but not sufficient, because the information only represents a small window of time in a human's life when they are being seen by their health care providers and does not capture the breadth of activity nor the level of functionality needed to fully assess a person.

To produce a more accurate and complete view of an individual, RWD needs to incorporate information from multiple sources. It should include SDOH information to include a person's financial status, housing and or environment, to functional data that compares the pre and post COVID data to their ability to hold a job or go to school.

PCOR Final Report:
CMS-N3C Linked Dataset and Use Cases

The National COVID-19 Longitudinal Research Database Linked to Medicare and Medicaid Data is part of the N3C EHR data enhancement initiative to design a reproducible scalable model that can be used to fill in some of the missing gaps in EHR data

Project Summary

The COVID-19 pandemic precipitated an urgent need to have a near real-time centralized research dataset of real-world data, RWD and a collaborative secure analytics platform to conduct patient-centered outcomes research on COVID-19 and generate evidence on effective interventions. As part of the development of a national resource for COVID-19 research, the Assistant Secretary for Planning and Evaluation (ASPE) OS-Patient-Centered Outcomes Research Trust Fund (OS-PCORTF), through a generous grant, funded the acquisition, linkage, harmonization, and integration of CMS Medicare and Medicaid claims data with National COVID Cohort Collaborative (N3C) de-identified clinical data. The project had seven objectives: four primary objects for linking CMS Medicare Data to N3C clinical data and three contingent objectives on linking CMS Medicaid Data to N3C Clinical Data if initial Medicare objects were met.

Medicare Objectives

Objective 1: Demonstrate the feasibility of linking clinical EHR data with Medicare claims data using the proposed N3C data linkage strategy and engage PCOR researchers in using the linked research dataset and providing input on the project use cases.

Objective 2: Prepare and Link Medicare claims data to N3C clinical EHR Data to be Used by PCOR Researchers.

Objective 3: Produce PCOR COVID use cases demonstrating the utility of the linked Medicare claims-N3C clinical data to conduct patient-centered outcomes research on COVID-19, including potential evaluation of economic outcomes.

Objective 4: Support the joint activities of the OS-PCORTF COVID Collaborative and overall project management, communications, and governance activities.

Medicaid Objectives

Objectives 1-4, linking CMS Medicaid claims data to N3C clinical data were successful, the Medicaid Linkage objective 5-7 were funded and implemented as part of this project. Though CMS Medicare and Medicaid Data have different structures, files and content, because both were transformed into the common data model OMOP the work was combined into a single data workstream.

Objective 5: Assess the feasibility of linking clinical EHR data with Medicaid claims data using the proposed N3C data linkage strategy.; and,

Objective 6: Prepare and link Medicaid claims data to N3C clinical EHR data to be used by PCOR researchers.

Objective 7: Produce PCOR covid use cases demonstrating the utility of the linked Medicaid claims-N3C clinical data to conduct patient-centered outcomes research on COVID-19.

Infrastructure Assets:

On going and available assets for investigators: Linked CMS Claims and N3C Clinical Data, Educational Material and Governance

- CMS Medicare-N3C linked national longitudinal COVID research dataset - available to researchers via the N3C data enclave,
- The N3C PPRL Enrichment Dashboard displays key statistics on the CMS Medicare-N3C linked data including numbers, characteristics, representativeness of the sample compared to the CMS
- Public Information on linking data using Privacy Preserving Record Linkage, PPRL
- N3C Privacy-Preserving Record Linkage and Linked Data Governance

N3C Publications, Preprints and Presentations

A complete list of all 296 N3C publications, preprints and presentation

Select Clinical Studies using linked CMS MEDICARE AND MEDICAID and N3C Clinical

The subset of studies listed below were selected because they used the linked CMS and N3C data

[RP-7B659A] Statistical and Machine learning method development for LONG COVID research

With our existing experience with the LONG COVID analyses. We found several issues from the N3C data. For example, the heterogeneity of LONG COVID reporting across different sites, the scarcity of the LONG COVID labeling and the complexity of the risk factors. In this project, we aim to develop novel data analysis and informatic tools to help overcome the deficiencies from the data, including missingness with complex mechanisms, high-dimensionality, and large scales. *Application Submission Date 12/19/22*

[RP-BBB544] Use of Evusheld and descriptive analysis of utilization

Retrospective study to examine the current pattern of use within the populations who are indicated for EVUSHELD use. EVUSHELD is an antibody treatment for immunocompromised and high-risk patients that provides pre-exposure prophylactic protection against COVID-19. EVUSHELD has been approved for emergency use but not yet with a full marketing authorization from the FDA. The research will describe baseline characteristics of patients, COVID-19 exposure, and all-cause outcomes for the population of eligible patients expected to receive benefit from EVUSHELD as defined under the terms of authorization. The study will examine both patients who have and have not received EVUSHELD. Areas of analysis include the volume of use within the indicated populations, how EVUSHELD is being used including prescribing patterns, and to evaluate the feasibility to study advanced questions such as efficacy, variation relative to COVID strain, and use in post-exposure populations. *Submission Date 12/14/22*

[RP-7767AC] Analytic methods for investigating effects of COVID-19 during pregnancy on birth outcomes.

Research on COVID-19 during pregnancy has suggested that the disease may increase the risk of adverse birth outcomes such as preterm birth. However, it is difficult to estimate the effects of infections and other exposures during pregnancy because the magnitude of those effects may differ depending on gestational age. In addition, people with shorter pregnancies are less likely to be infected while pregnant, making infection appear protective against time-dependent outcomes like

preterm birth. Finally, severe COVID-19 in a pregnant person can be an indication for delivery even if the pregnancy is not yet at term. When an indicated preterm delivery occurs, such as by Cesarean section, we don't know when labor would have begun spontaneously, making it difficult to tease apart physiologic effects on preterm birth vs. iatrogenic effects, and the magnitude of the latter may have changed throughout the pandemic as treatments were established. In this project, we aim to develop methods that allow researchers to ask and answer more specific research questions about effects of exposures during pregnancy and to use those methods to assess the impact of COVID-19 on birth outcomes. *Date 11/02/22*

[RP-0E7622] Neurological Complications Post-Covid-19 in Patients with Diabetes

Diabetes is associated with a wide range of neurological complications, of which, the most common is symmetric diabetic sensorimotor polyneuropathy (DSPN). This occurs in approximately 55% in people with type 1 diabetes mellitus (T1DM) and over 45% in those with type 2 diabetes (T2DM). Importantly, DSPN is a risk factor for mortality in diabetes. Therefore, prevention and treatment of DSPN are important not only in reducing the patient burden from this common complication but also in reducing overall morbidity and mortality in diabetes. Patients with diabetes mellitus (DM) are more likely to have severe complications with COVID-19 (SARS-CoV-2). In addition, recovering COVID-19 patients develop several neurological complications including peripheral neuropathy. The current study will determine if patients with concomitant COVID-19 infection and diabetes are at higher risk of neurological complications compared to those with diabetes alone. *Date 01/30/23*

[RP-BD23A5] Inflammation, thrombogenesis and myocardial injury and Covid-19 outcomes

To investigate the phenotype and outcomes in individuals with activation of one or more of the pathophysiologic pathways of inflammation, thrombosis, and myocardial injury during Covid-19 hospitalization. Sex and race differences will be assessed. Outcomes will adjust for baseline characteristics (demographics and clinical factors). In-hospital outcomes will include need for ICU admission, ventilation, cardiovascular events (MI, stroke, thrombotic events) and death. *Date 12/13/22*

[RP-42D046] NCATS Investigation into drug efficacy of disease pathology and post-acute COVID-19 syndrome

NOTE: This is upgrade to LDS from an existing project "NCATS Investigation into drug efficacy of disease pathology and post-acute COVID-19 syndrome"

Answers to several important questions related to COVID-19 treatment and long-term outcomes remain unclear. The overall scope or statistical power of existing studies has largely been limited by small cohorts. Access to the de-identified (Level 2) N3C data will allow us to revisit these questions on an unprecedented scale and better understand rarer long-term consequences of the disease. Both hypothesis and data-driven approaches will be used to conduct investigations into three key areas. The first involves the effect of drugs and drug classes in treating different stages and/or cohorts of COVID-19 including but not limited to SSRIs, ACE inhibitors, corticosteroids, antivirals, and antihistamines. Preliminary studies have indicated that at least some of these therapies may decrease overall mortality or reduce the likelihood of clinical deterioration, although some data are conflicting. The second relates to the quantification and prediction of the relative risk, and the factors contributing to that risk, for COVID-19 patients across the severity spectrum. These models

will be analyzed across various vulnerable patient cohorts including but not limited to diabetics, COPD, kidney disease, immunocompromised patients, etc. Also, assessing the risk of specific complications over a larger population to determine outcome rates and comorbidities more comprehensively will be performed with the N3C dataset. Finally, the third involves assessing the long-term consequences of COVID-19. A more precise definition of "post-acute COVID-19 syndrome" is needed, and patient phenotyping has yet to be conducted. Thus, both identification, stratification, and prediction of these patients will be performed using the N3C dataset. *Date 01/17/23*

[RP-924490] Comorbidity Comparisons, Including Patients with Immune Dysfunction, Between COVID-19 Negative and COVID-19 Positive Cohorts in N3C.

This project will compare comorbidity rates across both the COVID-19 negative and COVID-19 positive patients in N3C. The utility of this analysis is to determine if comorbidity incidence, which will facilitate more trips to the hospital for routine care and condition management, is indicative of a higher or lower incidence of COVID-19 in the N3C population. Preliminary evidence suggests that patients with comorbidities, particularly those with immunosuppressive and/or low survival over 10 years, may have a lower incidence of COVID-19 due to better adherence to social distancing practices. *Date 11/14/22*

[RP-E8FD97] Association Between COVID-19 infection and fracture morbidity and mortality in the geriatric population in the National COVID Cohort Collaborative (N3C)

Fragility fractures are an age-related disease among the most common injuries sustained by patients over 50 years of age, with an overall incidence of 1.1% in the United States. Patients are usually elderly, with limited physiological reserves and multiple comorbidities. With the emergence of the COVID-19 pandemic in early 2020, COVID-19 has become the third leading cause of death in individuals ages 65 and older. Elderly individuals are particularly vulnerable during pandemics. Fewer visits from relatives and caregivers result in a significant increase in the number of elderly living alone, which can increase the risk of falls. Moreover, elderly individuals have waning immune responses due to aging and chronic comorbidity that makes them more susceptible to infection. COVID-19 atypical symptoms such as falls, delirium, confusion, dizziness, and unusual fatigue in older patients could potentially increase the risk of fracture. Therefore, this project aims to explore the association between COVID-19 infection and fracture incidence and mortality in the elderly aged 50 years and above using the Limited data set. *Date 03/21/23*

[RP-059EDE] Disparities in COVID-19 treatment

Our project examines racial and ethnic disparities in Paxlovid treatment rates for COVID-19 positive patients within the NCATS N3C cohort. There is aggregate evidence that racial or ethnic minority patients receive such resources at lower rates. However, comparatively little scholarship elucidates the contributing factors driving these disparities. Are differences in resource allocations explained by observable characteristics of the patient and setting at the time a clinical decision is made? Or are different allocations made to patients who appear equally "at-risk" ex ante? Different diagnoses concerning the cause of disparities might yield different implications about the appropriate

mechanism for remedying the disparities, e.g., changes to guidelines or methods for assessing risk vs reducing segregation of minority patients into facilities. *Date 04/06/23*

[RP-2AD3C8] Immune response in preexisting autoimmune patients to SARS-CoV-2 variants: A retrospective study focuses on severity, long COVID and vaccination status based on N3C data.

Our preliminary study demonstrated that pre-existing autoimmunity is associated with increased severity in COVID-19 patients. There are several unanswered questions (e.g., association of pre-existing autoimmunity for different variants severity, impact of vaccination status on different variants, its relationship with long COVID etc.). In this project, we hope to validate our pre-existing autoimmune and patients on immunosuppressant cohorts for different variants, vaccination status, long COVID and usage of antiviral treatments. Functional data analysis and stats model will be used to answer relevant questions associated with different variants and autoimmune/immunosuppressants patients. We need PPRL data access to answer these questions. *Date 07/05/23*

[RP-9A216C] Assessing Severity of Disease based on Immune History and Variant

As the COVID-19 pandemic unfolded the timing of variant waves is necessarily correlated with immune histories. This study will attempt to control for the comorbidities of individual's and their medically recorded immune histories (bolstered by external records of vaccination rates and model estimate levels of infections) to estimate the relative differences in severity (and differential presentations) of the different waves of variants and sub-variants. Additionally, this may provide a basis for defining cohorts of the population based on their immune histories and variant exposures that may have implications for long covid or other associated co-morbidities in the future. *Date 04/04/23*

[RP-E4988E] Impact of the COVID-19 Pandemic on Treatments and Outcomes for Cancer Patients Using the N3C Limited Data Set

Compared to non-cancer patients, cancer patients with COVID-19 infections usually suffer more from the symptoms and experience compromised delivery and quality of health care service. However, only very few studies focused on this topic and there is an emerging need to establish a large-scale "real-world" study to comprehensively explore the impact of COVID-19 infection on cancer treatments and outcomes. We propose to conduct a retrospective cohort study for cancer patients and valid COVID-19 test results in the N3C dataset (Limited Data Set) and investigate the impact of the COVID-19 pandemic on cancer treatments and outcomes. *Date 05/18/23*

[RP-E8DA40] Study effects of COVID-19 Infection on cancer patients

Recent studies highlight the importance of COVID-19-cancer precision medicine. Published on Lancet, Lee and colleagues (Lee, Cazier et al. 2020) find that the only primary cancer type that demonstrates statistically significant mortality changes due to COVID-19, after adjusted for age and sex, is leukemia (OR: 2.25; 95% CI: 1.13 – 4.57; p-value: 0.023; ICD10: C51-C58. COVID-19. Another

study published on Nature Medicine, Robilotti et al. (Robilotti, Babady et al. 2020) found that, in the cancer cohort of Memorial Sloan Kettering Cancer Center (423 symptomatic COVID-19 cases out of 2,035 cancer patients tested between 3/10/2020 and 4/7/2020), immune checkpoint inhibitors and age > 65 together are predictors for hospitalization and severe disease of COVID-19 among cancer patients. These studies prove the concept of the heterogeneity of COVID-19's impact on cancer patients. Systematic data mining on a more representative, broader cohort with rich clinical, demographic, and socioeconomic features will provide more insights on the risk population and actionable knowledge. Correspondingly, our Specific Aims are:

1- Identify risks factors for cancer patients based on primary site and morphology on a larger scale due to varying results (e.g. Gustave Roussy and CCC19): Death and severity of COVID-19 (e.g. ICU use, WHO Ordinal Scale) will be used for outcome,

2- Development of cancer phenotype definitions on Palantir to be used for this study. Atlas has already 96 predefined concept sets for carcinoma.

3- Study of outcomes in cancer patient from healthcare utilization effect perspective: including delays in treatment due to COVID-19 and worsening effects including AE (i.e. CTCAE Diagnosis initially and NLP generated AE symptoms). *Date 02/12/24*

Select List of Methodology Assets using N3C Clinical and CMS Claims data

Methods Harmonization of N3C clinical data and CMS claims data developing a method to construct the CMS claims data as clinical encounters to complement and augment the 'OMOP-ified' N3C datasets.

- Augmenting the National COVID Cohort Collaborative (Paper)
- Augmenting the National COVID Cohort Collaborative (Slides)
- Augmenting the National COVID Cohort Collaborative (N3C) Dataset -Medicare and Medicaid (CMS) Data, Secure & Deidentified Clinical (Presentation)

Privacy-preserving record linkage across disparate institutions and datasets to enable a learning health system: The national COVID cohort collaborative (N3C) experience

Research driven by real-world clinical data is increasingly vital to enabling learning health systems, but integrating such data from across disparate health systems is challenging. As part of the NCATS National COVID Cohort Collaborative (N3C), the N3C Data Enclave was established as a centralized repository of deidentified and harmonized COVID-19 patient data from institutions across the US. **- Wiley Online Learning; Umberto Tachinardi, et. al; January 11, 2024**

This study established the possibility of secure large-scale health data linkage using deidentified datasets. This provides methods to ensure data quality necessary for research including deduplication, multi-dataset linkage, and cohort discovery and potentiates future scalability to other diseases.

The National COVID Cohort Collaborative (N3C) Privacy-Preserving Record Linkage Powered by Datavant and Regenstrief Wins 2022 FedHealthIT Innovation Award

SAN FRANCISCO, May 24, 2022 (GLOBE NEWSWIRE) -- Datavant, the leader in helping organizations securely connect health data today announced that the National COVID Cohort Collaborative Privacy-Preserving Record Linkage (N3C PPRL), powered by Datavant technology and Regenstrief Institute's Linkage Honest Broker services, has been recognized with a 2022 FedHealthIT Innovation Award.

Utilizing Datavant's privacy-first linking technology and Regenstrief Institute's Linkage Honest Broker data governance model together provides a synergistic approach that handles data deduplication and enrichment across different data types to improve data quality.

Regenstrief, Indiana CTSI, Datavant honored for support of NIH project with 2022 FedHealthIT Innovation Award

Regenstrief Institute, Indiana Clinical and Translational Sciences Institute (CTSI), the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) and Datavant have been honored with a 2022 FedHealthIT Innovation Award for the National COVID Cohort Collaborative (N3C) Privacy Preserving Record Linkage (PPRL).

Regenstrief Institute and Datavant have worked collaboratively to develop processes in HIPAA, gold-standard PPRL environment ensuring patient privacy while enabling research to address COVID-19 public health impacts, treatments, and prevention.

NIH's COVID-19 data enclave continues to evolve with the virus

The National Center for Advancing Translational Sciences within NIH launched the largest COVID-19 dataset in the U.S., the **National COVID Cohort Collaborative (N3C) Data Enclave** , in April. And now NCATS wants to use privacy-preserving record linkage (PPRL) to link data from its enclave with medical images, omics tools, electronic health records (EHRs), and social determinants of health to answer researchers' lingering questions like why COVID-19 symptoms linger in some patients.

Utilizing the established, secure PPRL process to link various data types will empower researchers to ask more complex health-related questions and improve clinical research utilizing a holistic approach.

Regenstrief, Indiana CTSI, Datavant partner on NIH national COVID-19 data effort

Regenstrief Institute, Indiana Clinical and Translational Sciences Institute (CTSI) and Datavant are supporting the National Institutes of Health (NIH) in a national effort to securely gather data to help scientists understand and develop treatments for COVID-19. Supported by a contract from the NIH, Regenstrief will serve as the national project's Honest Data Broker, using specialized technologies and processes to create more complete and informative data sets. Specifically, the Honest Data Broker will handle requests for data and manage a process referred to as "privacy-preserving record linkage" (PPRL) using technologies and approaches that help ensure **N3C** data are shared safely, securely and privately, all in compliance with HIPAA standards.

This technology will provide a secure PPRL environment with a multitude of health data to propel the researcher community towards new public health insights and discoveries.

Preprint Combining EHR and claims data in N3C increases scientific validity.
Quantifies the scientific benefits of using multi-model linked data (CMS claims and N3C Clinical) improves our ability to conduct valid research by enriching capture of exposures, outcomes, and comorbidities.

Educational Material for N3C Investigators (see Appendix)

N3C Background

The National COVID Cohort Collaborative (N3C) is a partnership to include three funding organizations, NIH, NIGMS, and NCATS, 77 Data contributing organizations including Community Health Systems, FQHCs, Citizen Scientists, and Academic Medical Centers, four distributed clinical data networks (PCORnet, OHDSI, ACT, TriNetX), and the more than 3400 data users representing over 360 organizations.

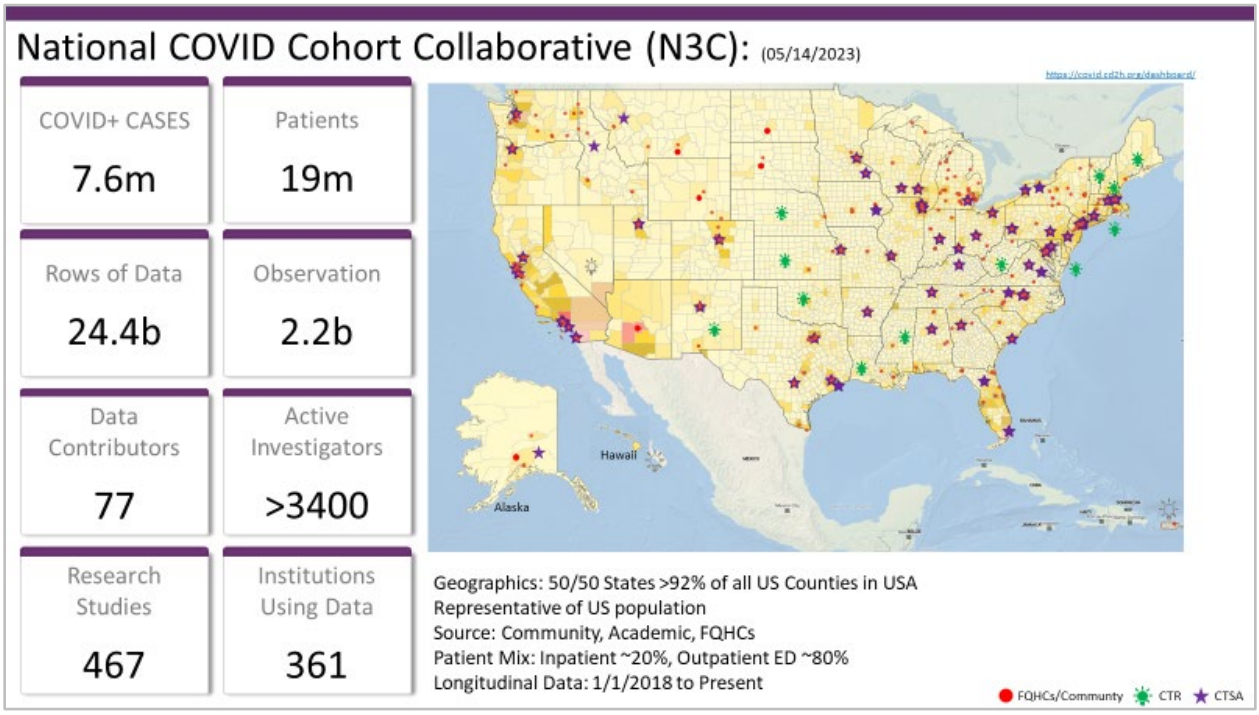


Figure 1 N3C Characterization: Includes enclave counts, studies, utilization, sources of data.

Because of the urgency of the pandemic, NCATS partnered with the common data model (CDM) communities (PCORnet, OHDSI, ACT, TriNetX), to create a centralized repository. The unknown unknowns of COVID-19 necessitated open access to row-level data so investigators could iterate over the data quickly and use techniques like machine learning for non-hypothesis driven discoveries.

To make this happen, N3C harmonized the four data models which required over 2 million data transformations and was made possible by leveraging the work done for the PCOR-TF Common Data Model Harmonization, CDMH project.

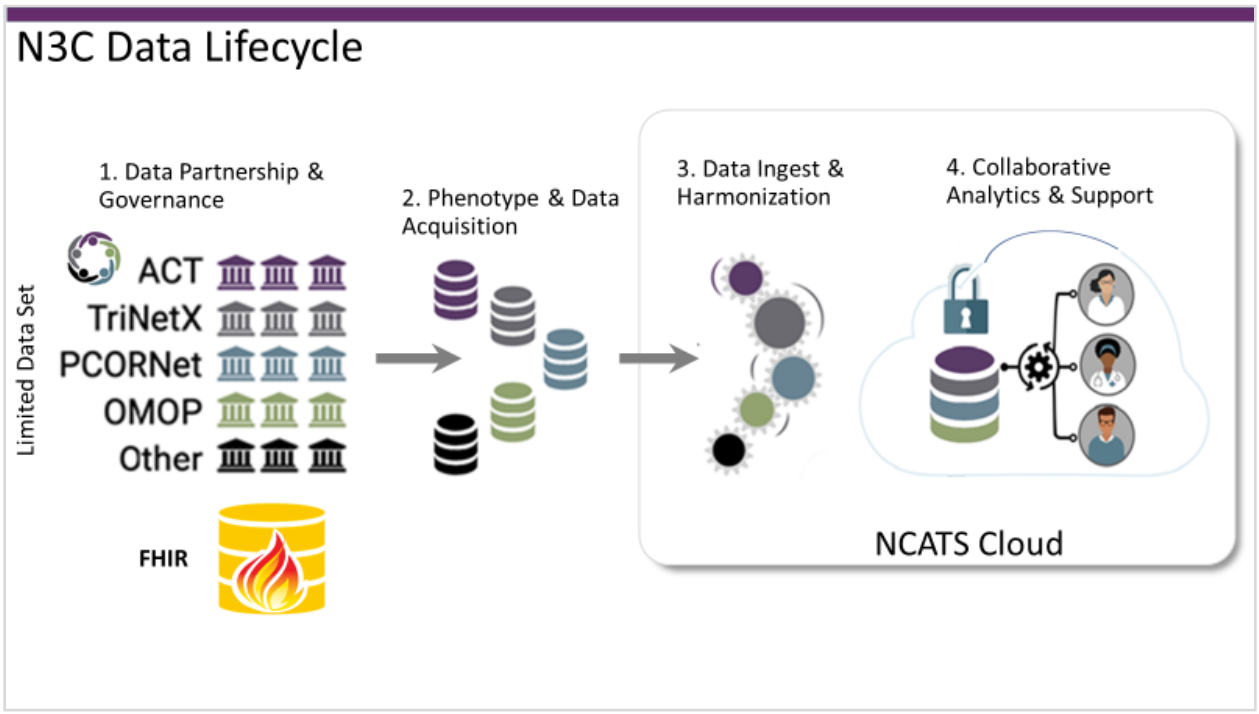


Figure 2: N3C Data Lifecycle is a 4-step Common Data Model harmonization process

All 24 billion rows of data are hosted by NCATS, using the N3C Data Enclave's secure, cloud-based environment certified through the Federal Risk and Authorization Management Program (FedRAMP) which provides standardized assessment, authorization and continuous monitoring for cloud products and services, ensuring the validity of the data while protecting patient privacy.

The N3C is longitudinal in nature and contains pre-covid data starting in 1/1/2018 and will continue to be updated on a regular interval until 10/2024. The data set includes such information as demographics, symptoms, lab test results, procedures, medications, medical conditions, physical measurements and more.

Rationale: The Need for Data Enhancements

De-identified EHR data by itself is necessary but not sufficient for a comprehensive real-world data warehouse that can produce impactful science. Although EHR uniquely contains outcomes and free text, suffers from quality issues and missingness and only represents a small slice of a human being. To produce a more accurate and complete view of an individual, RWD needs to incorporate information from multiple sources; from SDOH information on a person's financial status, housing, or environment to functional data that compares the pre and post COVID data to their ability to hold a job or go to school.

As part of the N3C EHR data enhancement initiative, NCATS identified CMS Medicare/Medicaid Claims data, along with national mortality data, viral sequence data, imaging data and over 60 public health datasets (e.g., pollution index, census data, RUCA code) as high priorities.

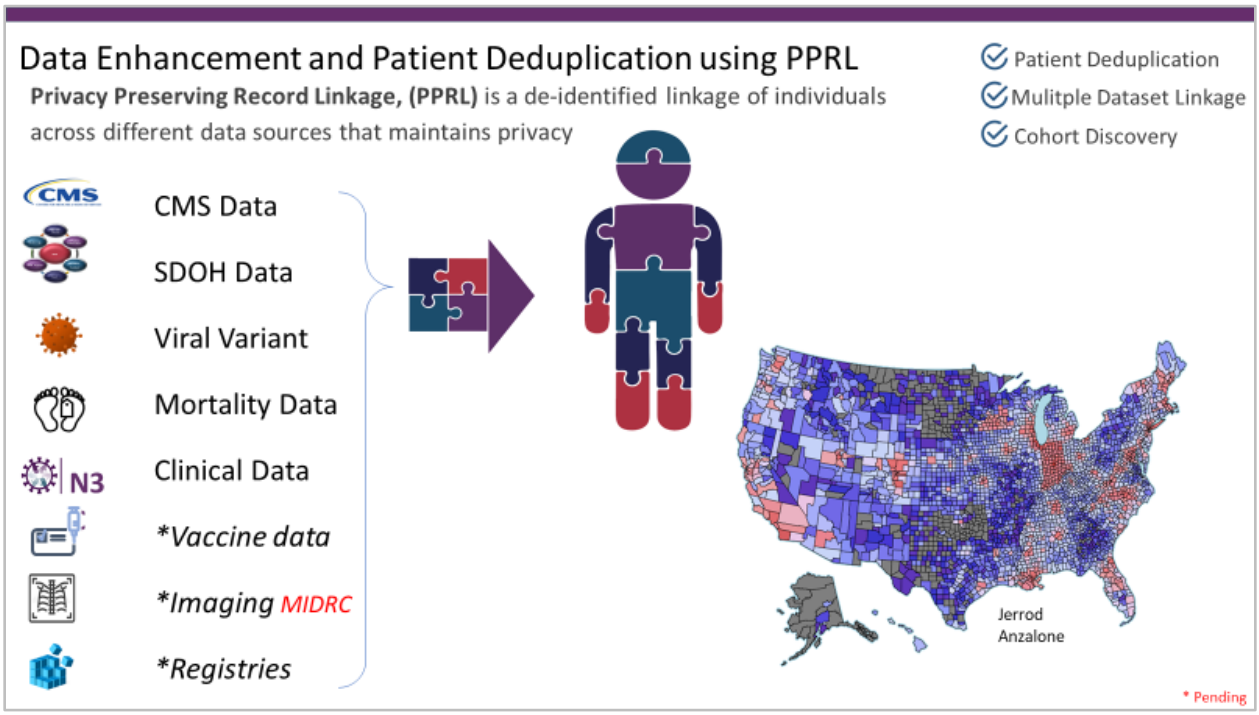


Figure 3 Available Linked Data in N3C:

N3C and CMS Medicare and Medicaid Data linkage

CMS Medicare and Medicaid was identified for data enhancement because, unlike de-identified EHR data, claims data contains any visit where a financial charge was filed. The addition of CMS Medicare claims data to the N3C enclave is an exponential leap in the scientific potential of N3C. The collaboration between N3C and CMS not only strengthens the overall capacity of the Enclave, but also largely resolves the issue of missing data from visits for patients that get their care from multiple providers.

Linking CMS Medicare and Medicaid claims data with clinical EHR data contributed to the N3C Data enclave will supplement this COVID-19 dataset to address key patient-centered outcomes research questions by:

- Improving the comprehensiveness of the clinical care history and longitudinal outcomes of patients in the N3C cohort (as EHR clinical data may be limited to only the health care received in N3C participating health care institutions) using CMS claims data on therapeutics, comorbid diagnoses, vaccinations, health care utilization and hospital data on deaths, as well as to evaluate post-COVID “long-haulers.”
- Incorporating aggregated provider characteristics and health systems affiliation using Agency for Healthcare Research and Quality’s (AHRQ) Compendium of U.S. Health Systems and CMS administrative data on health care providers which will allow researchers to evaluate system interventions on COVID-19 outcomes and system/provider characteristics on COVID-19 treatments and outcomes.
- Incorporating community characteristics via linkages to area or social deprivation indices to improve understanding of disparities in COVID-19 treatment and outcomes.

- Providing N3C researchers access to a CMS 5% sample to assess the representativeness of the N3C cohort and construct sampling weights and/or control groups to make generalizable inferences to the US population.

Accommplishments-N3C and CMS Medicare and Medicaid Data linkage

CMS-N3C Linked Data Sets Project Objectives-Approaches and Methods

Except for the ongoing scientific results, **all the objectives in the statement of work linking both Medicare and Medicaid data to N3C clinical data have been accomplished. The N3C and Medicare linked data sets were released to the public in November of 2022 and Medicaid and N3C linked data sets were released in April of 2023.**

Below are examples active scientific studies using the linked datasets. All investigators are granted 12 months to complete a study thus the first results are expected in November of 2023.

See Appendix 2 for more details.

Use Cases

1. Use of Evusheld and descriptive analysis of utilization
2. Statistical and Machine learning method development for LONG COVID research
3. Analytic methods for investigating effects of COVID-19 during pregnancy on birth outcomes.
4. Inflammation, thrombogenesis and myocardial injury and Covid-19 outcomes
5. Neurological Complications Post-Covid-19 in Patients with Diabetes
6. Comorbidity Comparisons, Including Patients with Immune Dysfunction, Between COVID-19 Negative and COVID-19 Positive Cohorts in N3C

Objective 1. Linking Clinical EHR Data with Medicare Claims Data

The purpose of Objective 1 and 4 weres to demonstrate the feasibility of linking clinical EHR data with Medicare and Medicaid claims data respectively using the proposed N3C data linkage strategy and engage PCOR researchers in using the linked research dataset and providing input on the project use cases. To demonstrate the feasibility of linking N3C de-identified clinical data with CMS Medicare and Medicaid data a significant amount of both technical as well as governance work was required prior to the investigators being given access. This included changes in the data access process and permissions, new vendor contracts, security enhancement, and technical development.

Approach and Method Used to Link CMS Claims Data to N3C De-identified Clinical Data

The linkage of CMS data and NCATS-funded N3C Data Enclave data was completed using the services of three entities:

(1) the Tokenization Contractor secures custody of CMS data; applies unique tokens; removes PII; and uploads the data to the N3C Data Enclave,

(2) the Linkage Honest Broker (LHB) receives and maintains custody of the tokens but not the underlying data from the Tokenization Contractor; and

(3) the Data Analytics Contractor hosts the data in the secure N3C Data Enclave.

First, the tokenization contractor applies privacy-preserving tokenization/hash to allow data with direct identifiers removed from the CMS Medicare and Medicaid claims data to be linked to the N3C Data Enclave. Second, the linkage/honest broker uses the token or hash to link data across participating providers for each individual patient ("mapping"); and sends the "mapped" CMS data to the data analytics contractor, who receives the mapped (hashed data with direct identifiers removed) data, links it to the existing data within the N3C Data Enclave and maintains the linked data. NCATS, as the steward of the combined CMS-NCATS linked data, maintains a FISMA-compliant FedRAMP moderate instance of an Amazon Web Services (AWS) GovCloud.

Data Contributor Modifications for Linkage to CMS Data

The 84 health systems that have contributed limited data sets to N3C have all signed the NCATS data transfer agreement with NCATS that stipulates how and what their data can be used for. As part of the implementation of PPRL linkage, a new and separate agreement was developed called the linkage honest broker agreement, LHBA, which lays out the terms and conditions on when, and how a health systems data can and cannot be linked to other data sets. Only after an institution has signed the LHBA and is specifically given permission to link their data, will it be made available to investigators who have met requirements to use the linked data.

As part of the implementation of CMS data, the site permission portal was modified to all sites control to if their clinical data could be linked to the CMS claims data (Left side of image below)

Data User modifications for investigators to get access to CMS data.

The requirements for an investigator to be granted access to CMS data include:

- Registration with N3C
- IT Security Training
- Human Subject training
- Must be from an institution that has signed the NCATS data use agreement
- local letter of determination on use of N3C/CMS data
- Attestation to the code of conduct
- Data User Request (DUR) proposal defines the research study data requested.
- Approval of DUR by federally staff data access committee

As part of the DUR process, an investigator must request the use of CMS data (see Figure 4. - below right side) and submit a letter of determination that addresses the need for CMS data. The letter of determination is then reviewed, along with the DUR proposal, by the NCATS Data Access Committee (DAC) for approval. Figure 4. Site permission portal (left side) allows each site to manage which data set/s are allowed to be linked to the institution's EHR data. The Investigator request for Access is a generalized process to review proposals and link them to the DUA, Institutional IRB, and NCATS DAC.

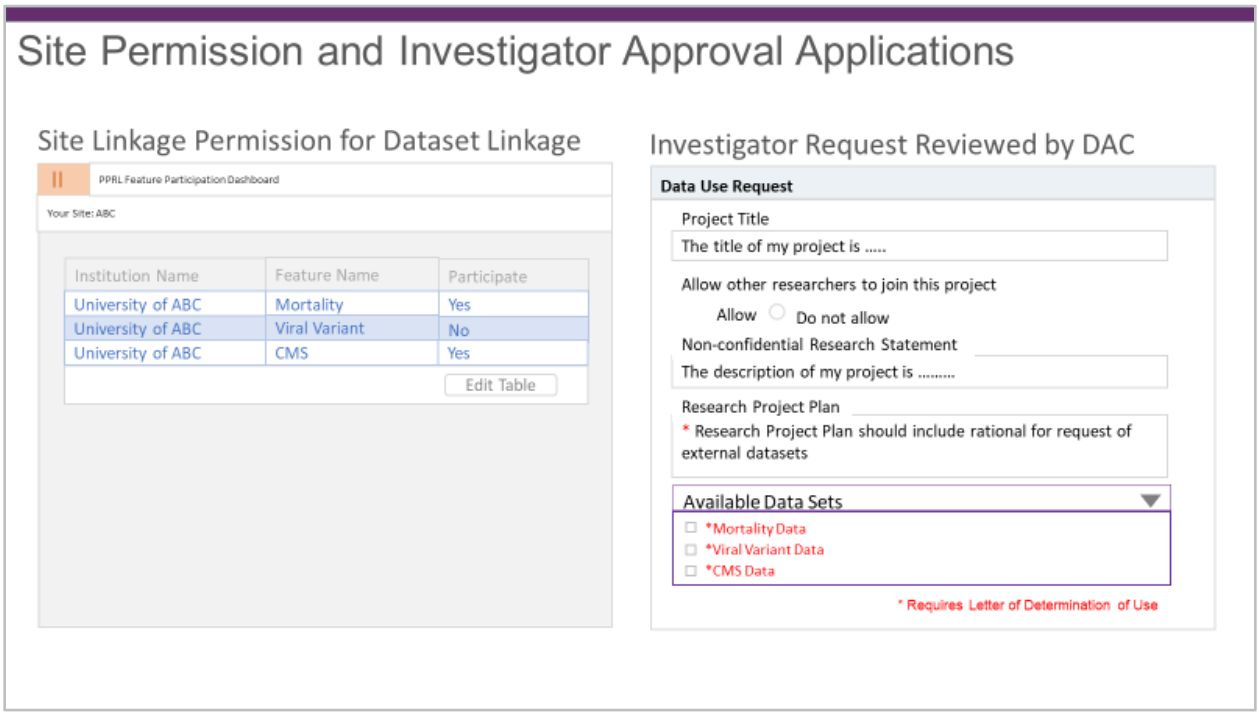


Figure 4: Site Permission and Investigator Approval Applications Portal

Objective 2 and 4: Prepare and Link Medicare and Medicaid Claims Data to N3C Clinical EHR Data to be Used by PCOR Researchers.

Data Harmonization and Curation.

The process of integration of CMS Medicare and Medicaid using the OMOP common data model was an arduous and complicated task and details were presented at AMIA meeting 2022. A preprint of the process Titled: Augmenting Medicare and Medicaid Services (CMS) Data into The National COVID Cohort Collaborative (N3C), Secure and Deidentified Clinical Dataset can be found in Appendix: "CMS Medicare and Medicaid data harmonization using the OMOP common data model" or on-line using the link above. Included below is a short synopsis of the process.

https://docs.google.com/document/d/1ZU2qHsriehLt88BHn-UP3O3isHQzef_JxYXSjt8kwpo/edit

One of the goals include demonstrating that it is possible to link CMS data securely and safely, can be harmonized with the common data model, and finally CMS data can be harmonized with a common data model OMOP to lower the investigative burden on using RWD.

- OMOP harmonization and transforming of CMS data into the OMOP to facilitate investigators.
- Semantic and syntactic harmonization of Ontological concepts from both CMS and OMOP values sets and vocabularies. This included a large team of Ontologists, CMS Subject Matter Experts and Clinicians.
- Quality Assurance (QA) testing of CMS transformation through a series of quality assessment checks

We worked with Acumen to identify data of the original code and code system in CMS claim files. The source files contained data elements in some cases over four thousand columns and we identified the columns where ICD10CM, ICD10 PCS, HCPCS, CPT4, and NDC terminology codes were found. We generated a code map crosswalk table to translate the native terminology code values found in the claim files to OMOP concepts ids. The data elements needed to be ingested and then created the mapping from resulting 271 thousand native CMS data elements into the 190 columns of the OMOP data model. We created a Code Map Service and a Data Transformation Pipeline, which involved combining sometimes up to 45 columns of CMS data into a single patient column record and creating visit constructs within a claim and macro visit constructs across multiple claims. The pipeline also required deidentification of patient and provider identifiers.

Linkage Honest Broker (LHB)

One of the fundamental challenges in gaining access to a researcher's RWD for research is trust that their data will be secure, and private. NCATS has gone through extraordinary efforts to win this trust and as part of this effort was separating the linking of data from holders of the data. NCATS contracts with a 3rd party called the linkage honest broker, LHB that acts as an independent party between the data contribution and the data users.

The linkage honest broker:

- is a neutral, vendor-agnostic entity, external to the N3C data enclave, which serves as an escrow for the encrypted identity tokens and operates the platform that facilitates PPRL using these tokens.
- does not receive, store, or process PHI/PII or clinical information, which is only held by the data participating sites. For validation purposes, the LHB may utilize tokens and metadata at the request of a participating site, consistent with the NCATS N3C Data Enclave rules and policies for possible follow-on clinical research.
- holds useful specific metadata such as the originating data contributor, data source, and the nature of data associated with the received tokens (e.g., EHR data, chest x-ray, viral variant data).
- includes capability to respond and adapt to emerging use cases: different match algorithms can be applied to different data sources and use cases.

The N3C linkage honest broker model disaggregates tokenization matching from data linkage. This allows enclave to enclave matching as well only the data contained within N3C. In the image below see in the upper right corner match counts are available between 3 separate enclaves ALL of US, NHLBI clinical trials, NIBIB MIDRC Images, as well as N3C (lower right pie chart)

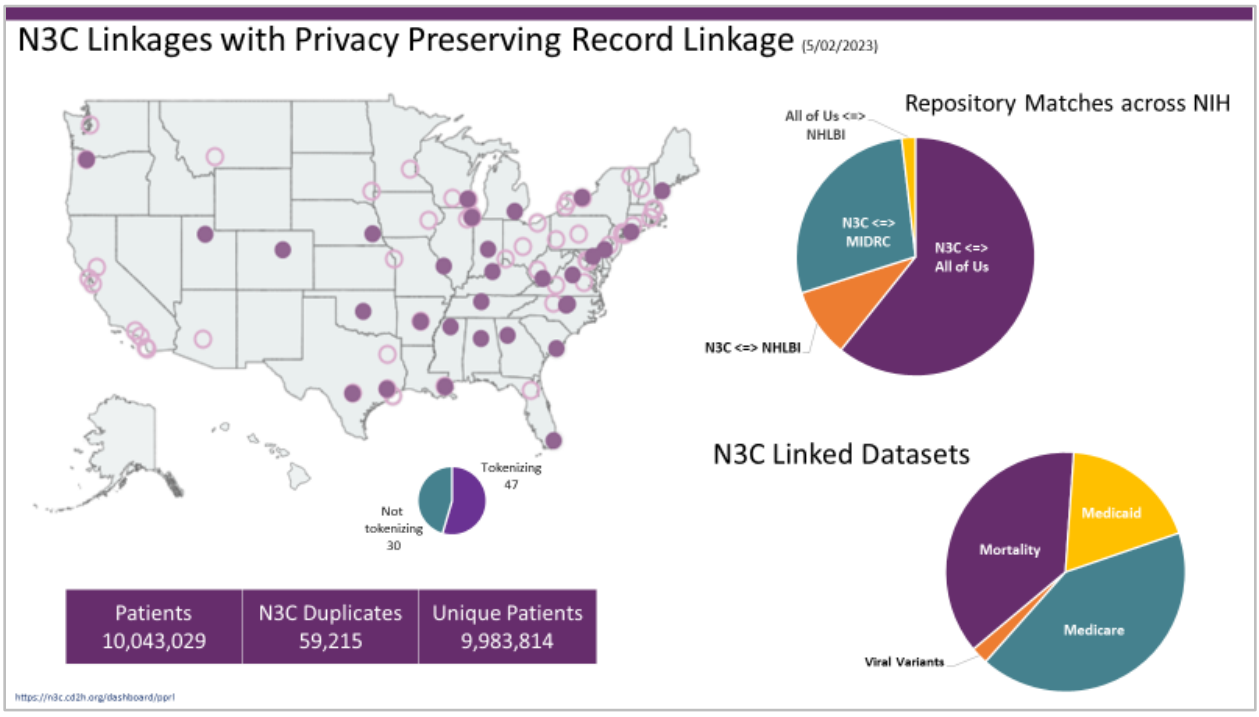


Figure 5 N3C Linkages with Privacy Preserving Record Linkage:

Data Linkage Enhancement

As previously mentioned, to get a complete picture of a person and how they are truly functioning, you need access to multiple types of information including a person's social determinants of health, level of functioning, as well as a person's medical history. If privacy was not a concern, the easiest way to do this would be to have all clinical data sources, including a person's PII, and then combine the files.

Thankfully this is not an option, because privacy concerns are real, and any data linkages must be de-identified and contain no PII. Privacy preserving record linkage (PPRL) allows a de-identify linkage. PPRL uses tokenization to insure a person's privacy is protected, while simultaneously accommodating data variability like spelling errors in a person's name or the fact that individuals may have multiple addresses.

Privacy Preserving Records Linkage Validation

While PPRL guarantees privacy because it contains no PII, it raises other concerns, e.g., if you don't know the identity of the individual, how do you validate the accuracy of the matching.

Using a non-N3C data set, N3C's honest broker, Regenstrief Institute, validated PPRL under an IRB, by comparing re-identified hashed patients at Indiana University (IU) with identified patients in the state's CMS Health information exchange, HIE.

1. Data from Indiana University Health
    - Compared matched tokens with data in IU Health records.
    - Confirmed age ≥ 65 years old for Medicare eligible population.

- Confirmed IU Health record contained mention of Medicare for matched population.

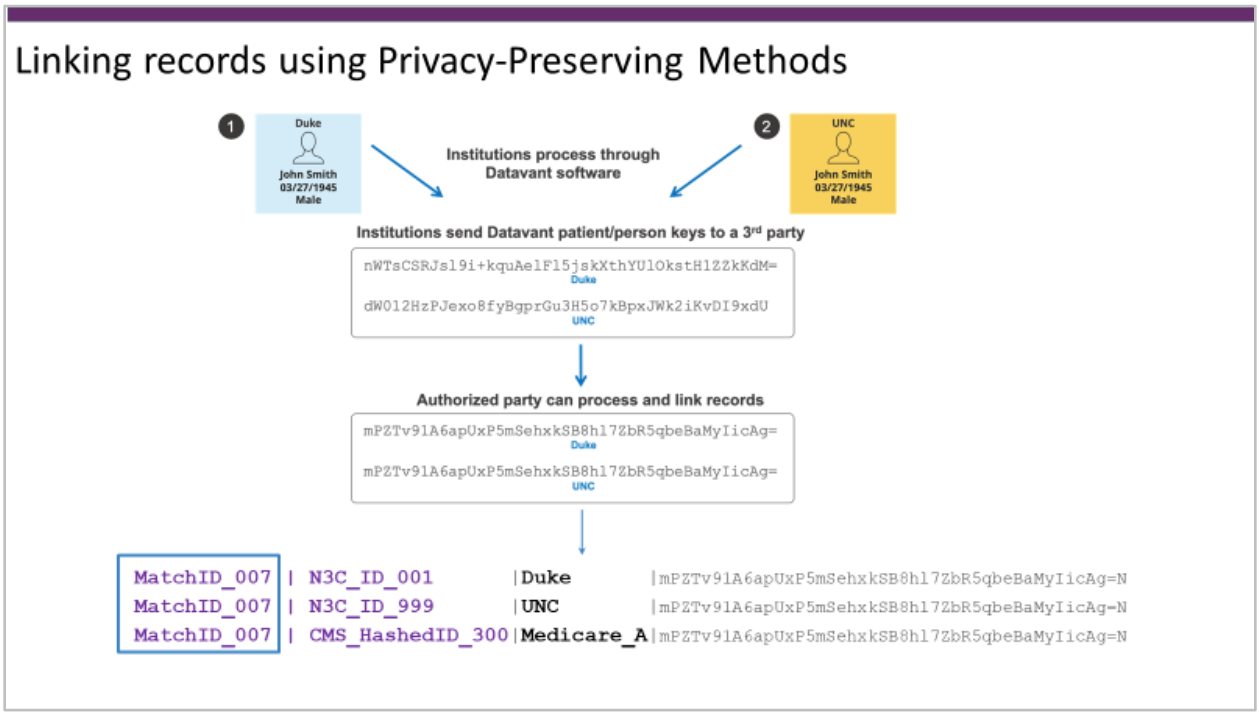


Figure 6: PPRL Patient matching and patient deduplication of clinical and claims data.

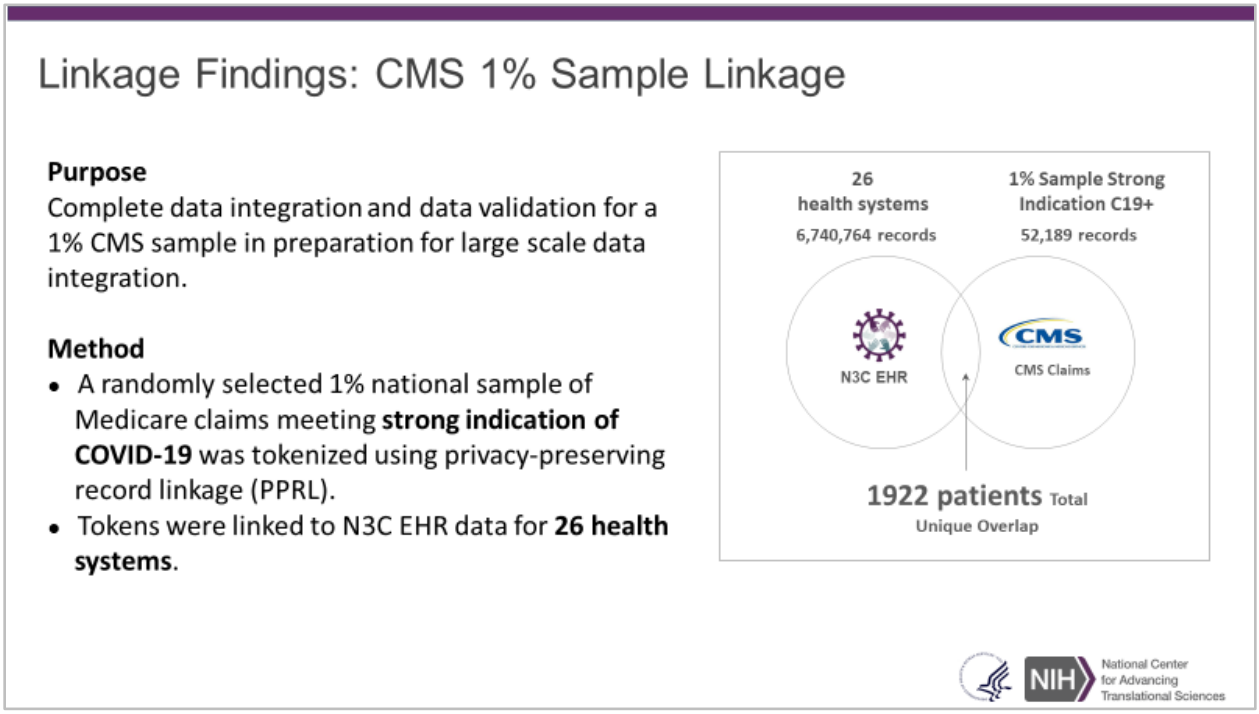


Figure 7 Linkage Findings: CMS 1% Sample Linkage

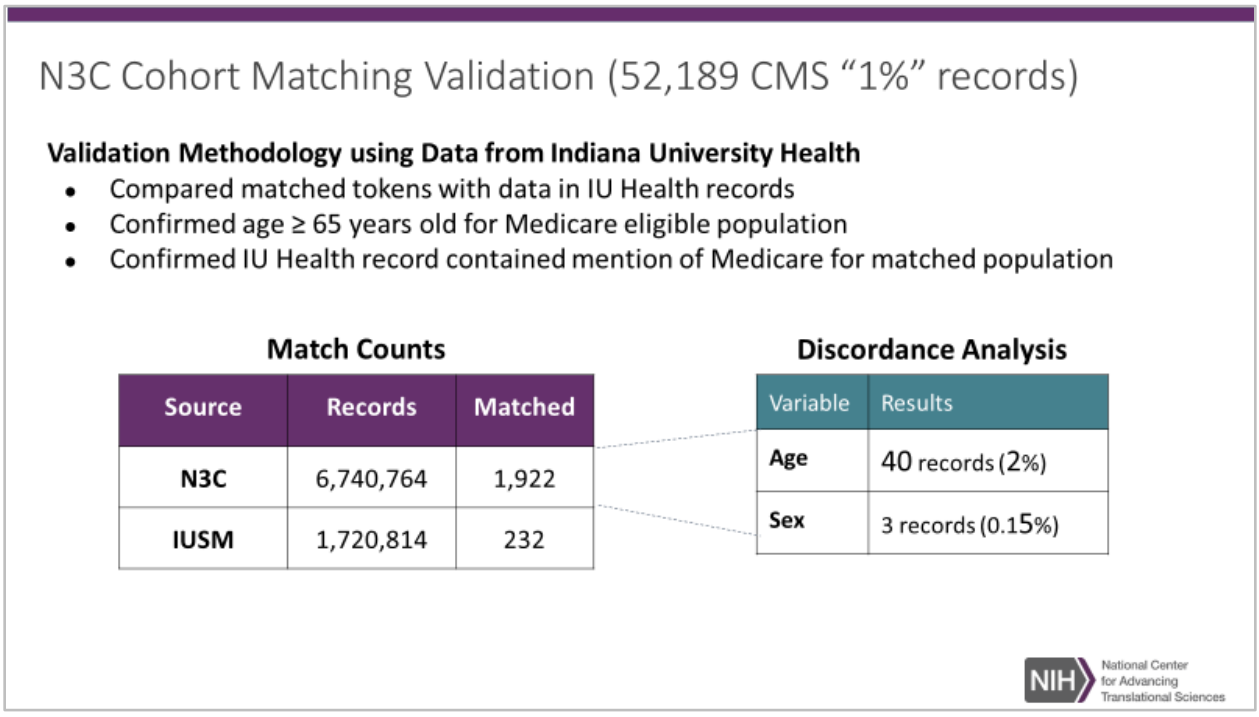


Figure 8: PPRL Validation: Indiana University/Regenstrief Institute

Privacy Preserving Records Linkage Clinical Validation.

In addition to validation of patient matching, a second level of validation was performed that compared the clinical information in the N3C with a national mortality database. Mortality was chosen for validation because EHR, though it has a high degree of missingness, does record date and reason for death and because of the importance of the death as a primary outcome. N3C used PPRL to link multiple types of missing data (see illustration below) to an individual's medical record with commercially available obituary data. The metrics used to assess the quality of the privacy preserving data linkage (PPRL) included collisions in age, gender, and date of death; within-site uniqueness; and cluster sizes.

Age and Gender:

An age collision is defined as a set of linked records in which the year of birth of the linked records differs. The percentage of age collisions was very low: 0.11% for linked records between EHR data and 0.15% for linked records between EHR and CMS data. Some age collisions may be due to EHR sources date-shifting a date of birth, as most collisions were apart by only 1 year.

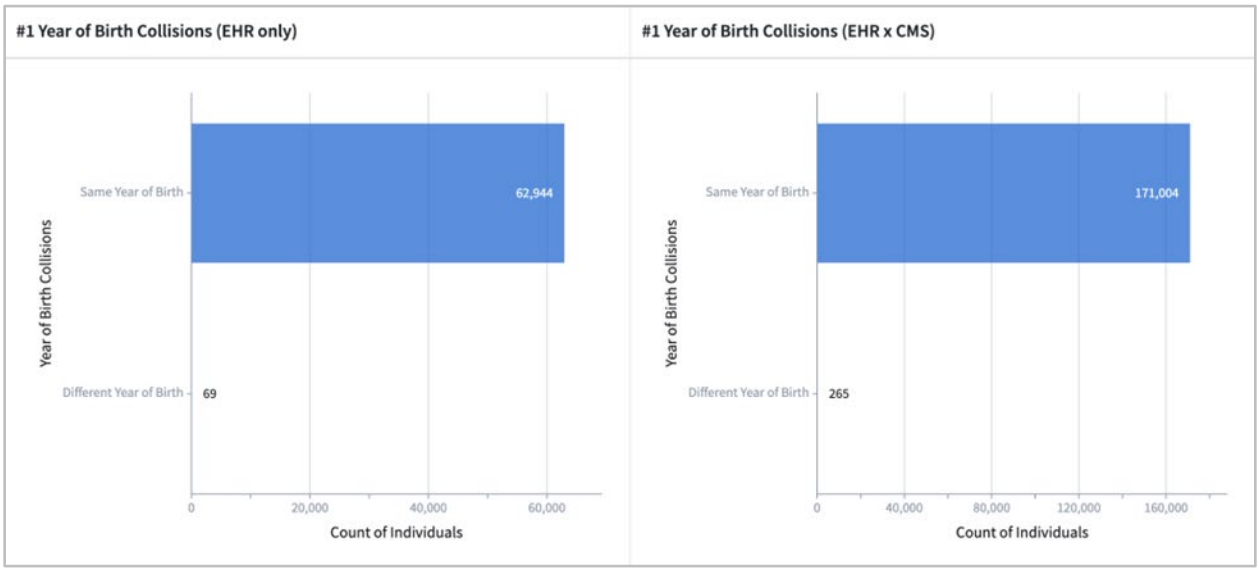


Figure 9: Year of Birth Collisions

Similarly, a gender collision is defined as a set of linked records in which the gender of the linked records differs. The percentage of gender collisions was also very low: 0.42% for linked records between EHR data and 0.08% for linked records between EHR and CMS data.

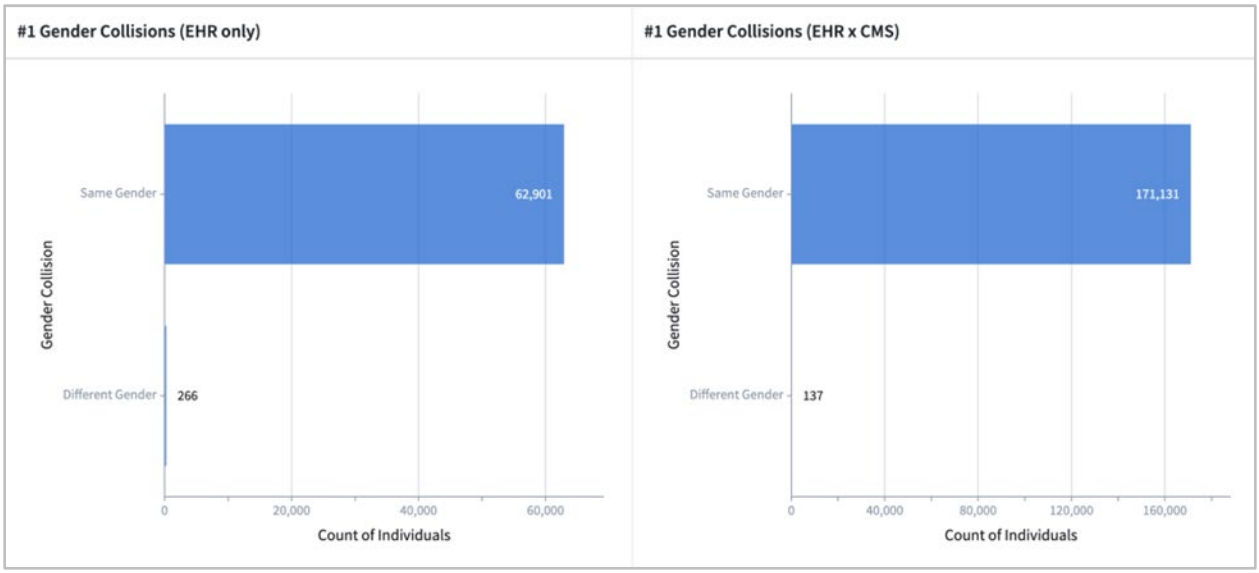


Figure 10: Gender Collisions

Date of Death:

To assess the mortality date of death collisions, the date of death was compared to the last known visit date, with the assumption that the last known visit date should come before the date of death. Around 20% of linked individuals have a date of death 30 or more days after the date of death. Many administrative encounters (e.g., autopsy, legal, or financial encounters) are documented on deceased patients, which may explain the high number of post-death encounters. As a more accurate proxy for mortality collisions, the date of death is being compared to a last known measurement date in an upcoming update.

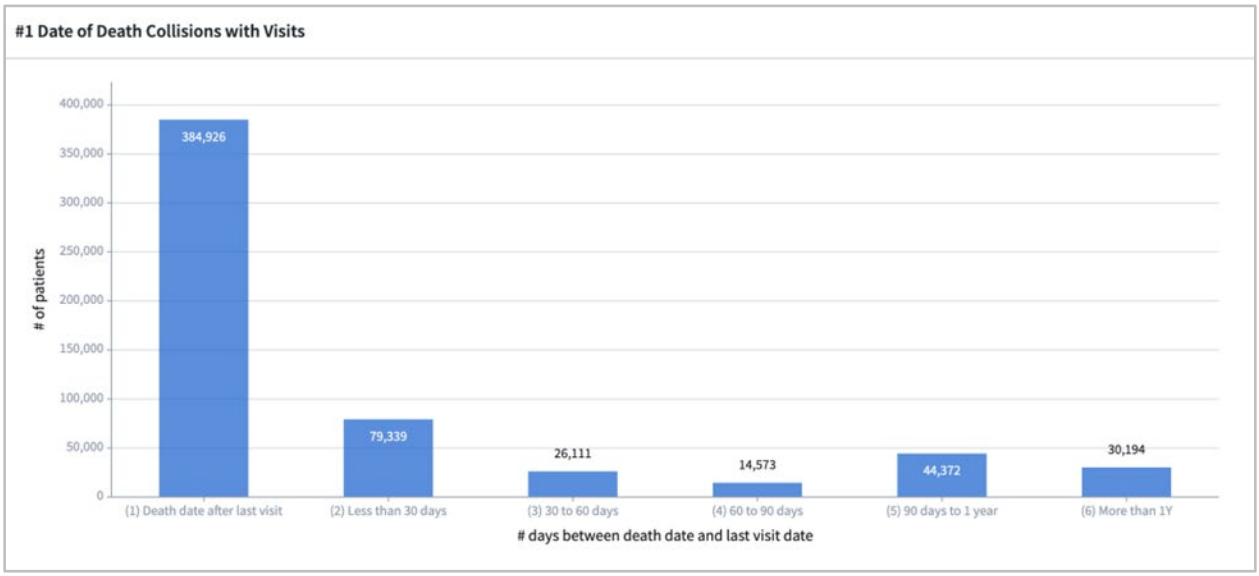


Figure 11: Date of Death Collisions

Within-site Uniqueness and Cluster Size

Within-site uniqueness, or the rate of duplicates found from a single submitting site, was evaluated to assess the quality of data submitted by individual submitting sites. The median duplicate rate was 0.0034%, indicating nearly all sites are submitting unique records.

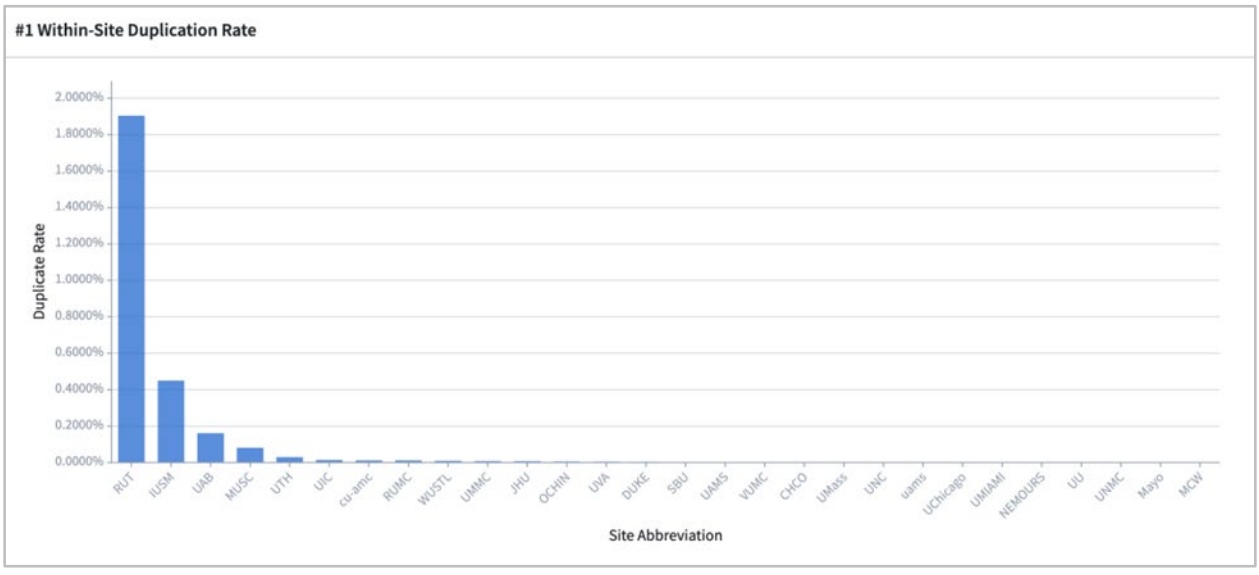


Figure 12 Within-Site Duplication Rate

Finally, cluster size frequency was evaluated. Cluster size indicates the number of records that are linked together, either within a single site or from multiple sites. For example, a triplicate is a cluster of size 3, indicating three separate records have been linked to the same individual. The distribution of cluster size followed the expected distribution: as cluster size increases, the frequency drops: 0.03% of linked records are in a cluster of size 4 or greater.

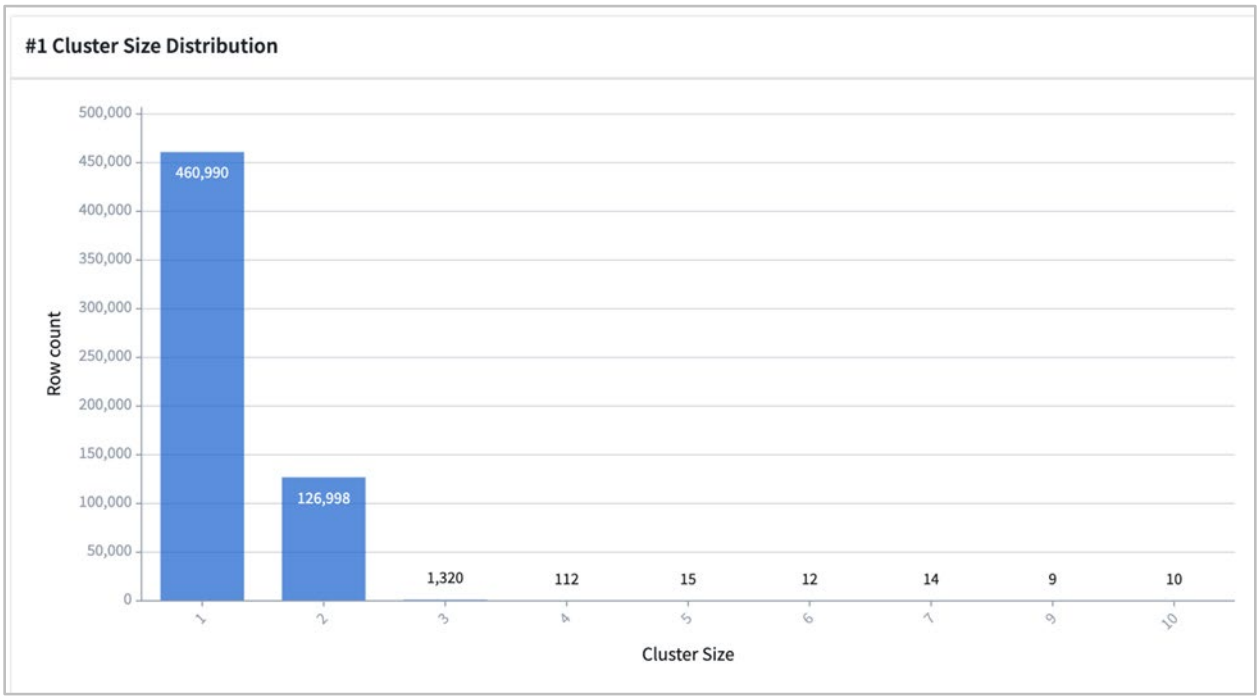


Figure 13: Cluster Size Distribution

CMS Linkage: Bidirectional Information Exchange, Honest Broker, & Tokenization Contractor

To link CMS claims data to N3C Clinical data, NCATS contracted with Acumen, as the tokenization contractor. Unlike most linkage efforts which only require a one-way interface, CMS linkages require a two-way interface because the tokenization contractor only sends CMS data of people with clinical data in N3C. This required the tokenization contactor to send a large file of COVID patients to the honest broker, the honest broker to identify matches and Acumen to use this subset of tokens to create the claims payload that was sent to N3C.

Tokenization Process Requirements (see image below)

1. De-identify the CMS claims data and place a Datavant PPRL hash.
2. Implement a bidirectional information exchange between tokenization contractor and linkage honest broker. To identify patients who are both in CMS and N3C. The additional functionality is required because the clinical files that CMS will allow to go to N3C is not all CMS data but only patients who have a matching LDS from one of the 77 data contributors.
3. Create of a de-identified claims payload of matched N3C and Claims data patients
4. Securely transfer the de-identified payload to the analytic contractor

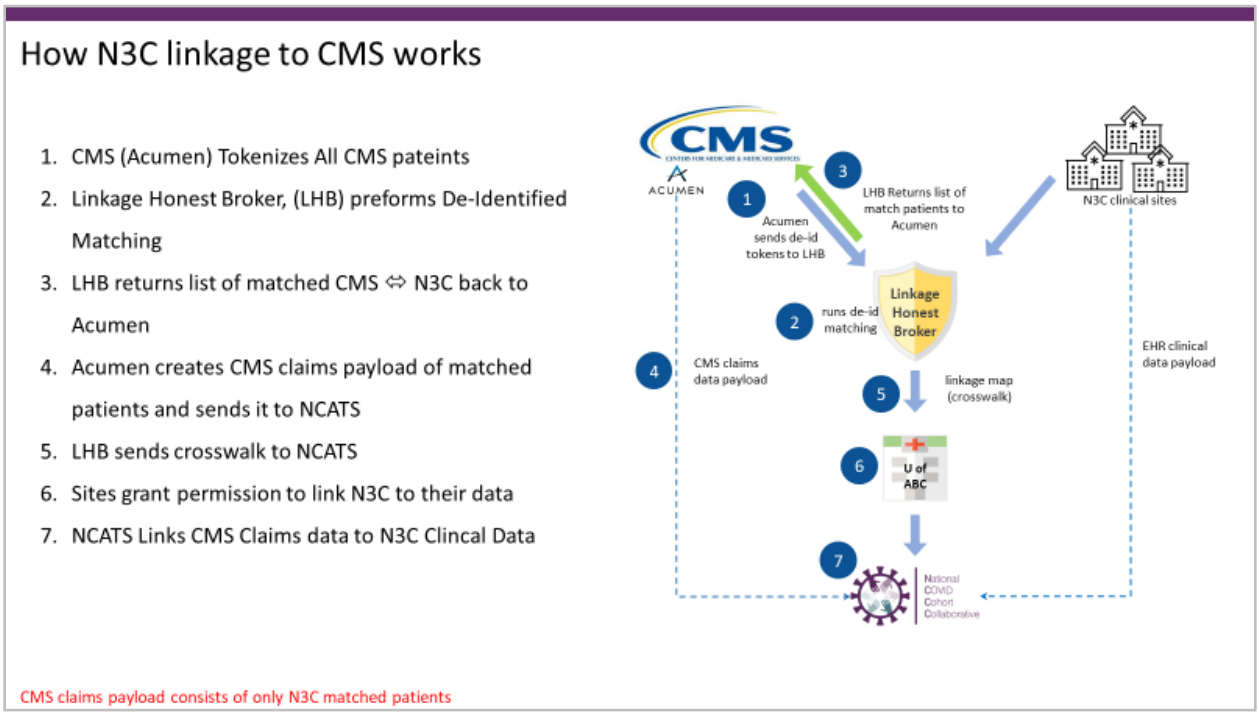


Figure 14: CMS Data Process: Source data, Linkage Honest Broker, Tokenization Contractors and N3C.

Patient Deduplication (See Appendix 1 for more details of using N3C deduplication)

The addition of CMS data to the N3C enclave required a refactoring and implementation of the patient deduplication functionality. "Deduplication" in the context of N3C refers to an individual from the same data contributor or multiple data contributors who refer to the same patient. Multiple patient records of a unique patient are surprisingly common and can happen for a variety of reasons, from being registered under different names (maiden and married surname), multiple registrations caused by misspellings, the combining of multiple institutions records where the same patient may have gotten care, or as in CMS, combining of different types of information.

N3C only acquired CMS data for patients with an existing clinical record from at least one health care system in N3C. This means that every CMS patient in N3C has multiple identifiers. For instance, if John Doe was a CMS Medicare recipient and passed away from COVID during his care at the University of ABC he will have at least 3 unique identifiers in N3C (U of ABC, Medicare, and Mortality). For investigators the presence of duplicates will be the presence of a globally unique identifier that links all 3 source identifiers to their respective unique information.

Objective 3 and 6: Produce PCOR COVID Use Cases Demonstrating the Utility of the Linked Medicare Claims-N3C Clinical Data and Medicaid claims-N3 Clinical Data to Conduct Patient-Centered Outcomes Research On COVID-19, Including Potential Evaluation of Economic Outcomes.

Use Cases and studies Deliverable of Linking CMS Claims-N3C EHR Data

Use cases serve many essential functions such as technical validation of a proposal that includes testing user's experience, debugging software, or performance metrics. More importantly, use cases should serve to test the utility, or value add, that linking claims and EHR data brings to either improving public health directly or indirectly facilitates better and more impactful science. Though it is our belief that having access to a more comprehensive CMS claims-N3C EHR data set will be more impactful confirming it is important.

The use cases from this project will help demonstrate the utility of the national COVID longitudinal research dataset to PCOR researchers on how claims data can augment and/or validate information from EHRs and produce unbiased evidence to yield a more comprehensive picture of an individual's past and current comorbid conditions and care patterns, including provider and community characteristics. Validation analyses will also evaluate accuracy of the linkages using the token/hash and to improve understanding of how adding CMS payer claims augments clinical data in the N3C Data Enclave by assessing potential concordance of information from different data sources and where they may diverge.

Origin of N3C Use Cases

The use cases for testing the utility of CMS claims-N3C EHR data will be derived from the normal course of operation by the 3400 N3C investigators. NCATS supports the infrastructure of N3C but does not fund investigators to use N3C. Despite the lack of funding, at the present time there are over 470 current studies in N3C. Access to the linked CMS claims-N3C EHR data requires an investigator to submit a proposal that is reviewed and approved by a Data Access Committee (DAC).

The advantage to using the N3C community test the utility of CMS claims-N3C EHR data is that questions are driven by the science because all the work is self-motivated, and users are not unduly influenced to use the linked data because of remunerations.

Currently the total number and type of use cases continues to grow with new projects being approved on a rolling basis. See Appendix 2 for a list of a few Real-World Data (RWD) and the Need for User Support

Effectively using RWD is challenging for inexperienced investigators. There can be a misperception that using RWD requires the same skills required for a randomly controlled trial (RCT). Conceptually they are very different and fit for different purposes. Where a RTC data collection was thoughtfully planned to answer a research question, RWD's uniqueness is its breadth and variability, which is a better representation of the population. Where RCT data is narrow and deep with a defined set of permissible values, RDW is broad and shallow with no predefined acceptable clinical parameters. Where an RCT data is 100% complete RWD has missingness. Where RCT

patients are a narrowly matched cohort, RWD patients represent the heterogeneous variability of the populations.

If you approach RWD as a poor man's RCT the first step is to filter 90% of the patients to remove as much variability of the data collected and create a pseudo-RCT value set. This certainly is a possible use of RWD and given that we will never be able to do a RCT on every disorder, a legitimate use of the data. However ultimately the conclusion of this approach will inevitably result in some variation of the author stating, "because of the limitation of the way the data has been collected an RCT is needed to definitively answer the question." To put this in context a fast majority of the "art of medicine" is done with no evidence so any evidence would still be welcomed.

Randomly Controlled Trial and Real-World Data

RWD is optimized for non-hypothesis driven data research. In the past, non-hypothesis driven research was often disparaged as a "fishing exercise." This description may have had merit in the past but severely underestimates the advances made over the last 30 years in computer science like machine learning, and large language models. All the while overestimating the applicability of conclusions being drawn from the biased populations found in an RCT that fails to account for the infinite variability of the human condition.

To address this, N3C is organized into multi-disciplinary Domain Teams. Domain teams enable a group of like-minded individuals to work together on scientific analysis, collect pilot data for grant submissions, train algorithms on larger datasets, inform clinical trial design, learn how to use tools for large scale COVID-19 data, and validate results. At the present time there are 33 N3C domain teams ranging on subject matter from Cancer and COVID to Machine learning best practices for N3C

Understanding the Difference Clinical Data and Claims Data

EHR were not designed for research; they were initially designed for clinical documentation but have evolved into regulatory platforms to enforce compliance and maximize charges. Investigators expecting clean data that is documented in a uniform manner, or a comprehensive picture of a patient will be disappointed to find out that EHR documentation is in fact idiosyncratic and that much of the information, like a person's sequence or imaging data, is not in the EHR but stored in a parallel application. In addition, because the US healthcare system is fragmented and patients often receive their care from multiple provider networks, the EHR data is not comprehensive and basic assumptions of a complete problem list, or a list of medications is misguided.

Investigators frustrated by the poor data quality and missingness of EHR data may feel using claims data is preferable because all encounters have a claim. Unfortunately claims data has its own challenges, for instance the diagnosis found in claims does not match the clinicians' clinical notes. This is a result of hospital coders whose job it is to maximize revenue by adding diagnosis and services codes that the clinical documentation supports rather than supporting accurate clinical notes. For instance, a physician might record a diagnosis of diabetes, but the claims data might contain a diabetes primary diagnosis but also new secondary disorders such as diabetes induced loss of vision or renal insufficiency.

In addition to conflicting information, claims data does not contain the results of procedure, labs, or diagnostic studies which limits its utility if you are interested in the efficacy of different interventions. Finally, private claims data is from private insurers and is generally timely but if you are using CMS Medicare or Medicaid data, the files can be years old before they are made available.

Less obvious challenges are that investigators who tend to use claims data are not the same investigators who use clinical data. EHR data users are accustomed to clinical data being available in a common data model such as OMOP, or PCORNet where services researchers who use claims data expect a format consistent with a billing file. This is much more problematic than at first blush as the semantic meaning of like-named information is different. For example, a unit of care, a visit, even what encompasses a hospital stay, are different and collaboration between such teams is difficult.

Investigator Education and Support Using N3C

One way to conceptualize the use of RWD is to think of a wet lab that requires a wide variety of skill sets, from expertise in chemistry, cell biology to methodological rigor. RWD is analogous to the wet lab, it requires a variety of skills each requires training, and support to be successful. Below is a short description of the education and support afforded N3C uses.

CMS User support: N3C was initially an EHR-based data warehouse but with the addition of other data, including claims data, we have developed an extensive support structure. All support services are cross disciplinary and can be accessed using multiple tools including a ticketing system, training videos, Slack, newsletter, websites, video, and the textbook of N3C.

Data Liaisons: Data Liaisons' primary purpose is data curation including data harmonization, data quality and semantics equivalence. A few brief examples will be illustrative: harmonious includes things like all temperatures are represented in centigrade, where semantic equivalence focuses on the meaning of a concept for instance is male gender and male sex the same.

Logic Liaisons: Logic Liaisons on the other hand focus more on the use of the enclave and scientific support. For example, N3C has concept sets that represent a group of medications like antibiotics. The logic liaisons validate such concepts and dissemination them to improve reproducibility and the quality of research.

Data Science Support:
Clinical informaticians skilled in modern data sciences, like machine learning are rare. Multiple reviews of conclusions reached by RWD investigation in health care have shown most are poorly executed, did not consider bias, and generalizable. Within the Logic liaison team there is also a consultant service we call Good Algorithmic Practice, GAP. GAPs support focusing on non-hypothesis driven data science such as machine learning. Unlike most RWD research networks N3C data is centralized, and investigators have access to row level data. The unique rescue makes ML possible but also presents a challenge in the need to make sure algorithms are not biased, transparent, and generalizable.

Education, and User Support

As previously mentioned, N3C has an extensive support system from office hours to an online ticketing system. In addition, there is a team completely staffed to user training. The Education branch has created hundreds of training resources including a textbook on the use of N3C, 30 training modules, and training webinars. In addition, N3C has now added subject matter expertise on the use of claims data for services researchers who want to continue to use Claims in its native format or use Claiming data that has been transformed in the OMOP.

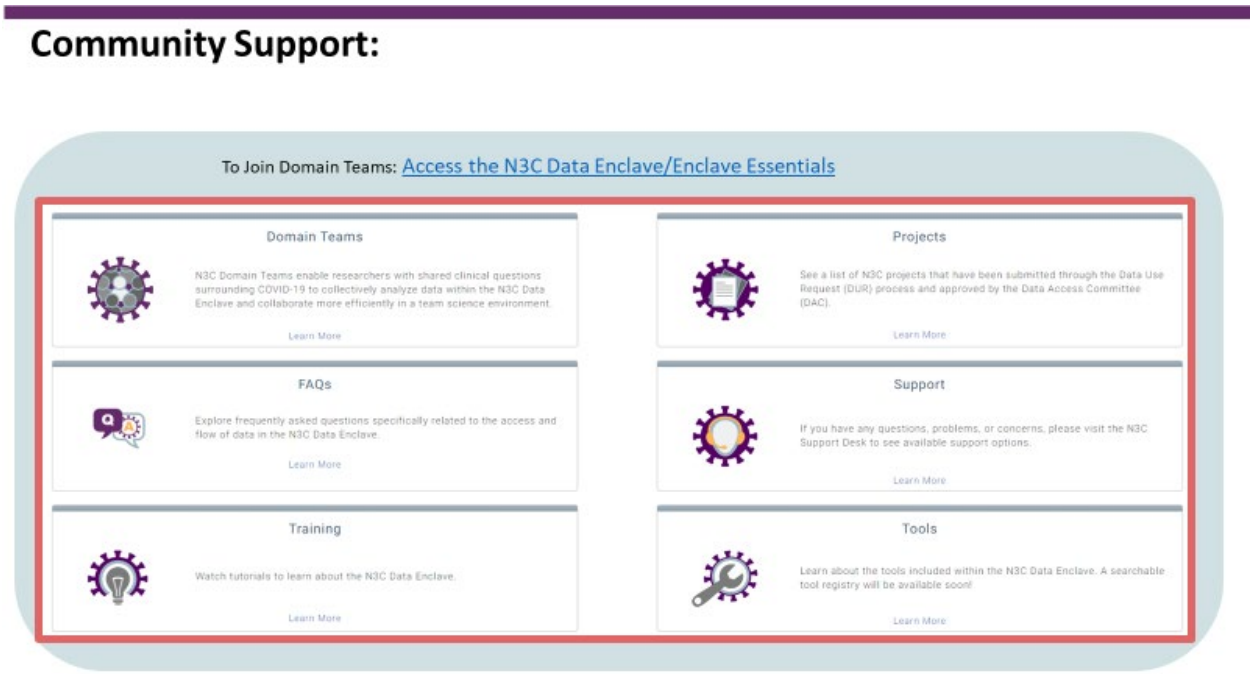


Figure 15: Available Investigator Community Support

Challenges Encountered

By most conventional measures of success N3C has excelled (see list below), and yet its impact has been limited by many challenges that must be addressed to maximize the use of RWD for the benefit of public health in the future

N3C Measures of Success

The following is the list of the success achieved by the N3C:

- **Linked Data Sets:** N3C has successful linked Nation Mortality, CMS Medicare/Medicaid data, Viral sequence data and imaging data.
- **Team Science:** > 4100 Users, > 580 studies, N3C Leadership is predominantly Women and Minority Leadership
- **Citations: :** 4845 Citations, H-index 33, 126 Press Articles, 79 Publications, 2023 "article of the year" The Journal of Rural Health.
- **Largess:** Largest COVID repository in the USA >23 million patient, 33 billion rows of data, 84 health systems

- **Data Quality:** Score Card, Data Quality Checks, Unit harmonization
- **Inclusive Networks:** Only Network that includes: PCORNET, OMOP, ACT, TriNetX
- **Education/support:** 763 training resources, personal help, office hours, best practice, tickets, website, newsletter, video, office hours, Domain Team, Forum
- **Recognition:** Biden administration, senate, and governor requests; Dataworks! Grand prize, NIH director's blog, NPR
- **SDOH:** AI/AN, 80+ public data sets, CMS Medicare, and Medicaid data
- **Organizational Users and Partners:** ONC, FDA, NCI, ASPE, ASPR, AHRQ, NIBIB, All of Us, NHLBI
- Funding: 72 awards, $> 109,000,000 from CLC, NCATS, NHGRI, NIA, NIAID, NICHD, NIDA, NIDCD, NIDDK, NIGMS, NIMH, NIMHD, NLM

N3C Challenges to Having Greater Impact

- **Access to HHS data (CMS Claims, CDC Vaccine)**
  - To address the pandemic, the US needed access to data from a variety of sources. One of the most important was CMS claims data and vaccine data. Logically HHS sister agencies such as NIH, CDC, and CMS as part of HHS should have shared their respective resources. But in reality, access to CMS claims data was a difficult task that took over a year of effort and the PCORT-TF funding to purchase. Linkage to vaccine data, despite funding and verbal agreement to collaboration, was blocked by multiple HHS agencies and never was realized.
- **NIH resistance for the use on waiver of consent data**
  - HIPAA, the common rule anticipated the use of RWD for public health. The FDA has a RWD initiative, ONC from meaningful use to the TEFCA support the use of RWD. Nonetheless the largest research health institute in the world, the NIH, does not support its use. Misplaced fear and understanding of RWD threatens future and present access to RWD not only for COVID research but other national issues such as gun violence, opioid epidemic, and maternal health.
- **Claims Data**
  - Claims data is not available in near real-time which is needed for public health response and the data is complicated to use and formats are designed for interoperability.
  - Access to CMS claims data is very delayed. In the emergency of a worldwide pandemic, access to current data from both Medicare and Medicaid was not available. New COVID variants are evolving weekly yet claims data in the case of Medicaid is years old before it is made available for research making evaluation for certain underserved populations impossible.
  - N3C purchased the use of Clams data from the vendor Acumen but this data is available in a proprietary format and different from other vendors such as Mathematica. During the worst public health pandemic since the 1918 flu, N3C spent precious months harmonizing Acumen files with the common data model OMOP instead of doing science. To make matters worse, the harmonization effort only

applies to a signal prior model and is not generalizable to CMS data supplied by another vendor.

- **Open Science Collaboration of RWD**
  - N3C is the largest open science repository of COVID data in the US. NCATS is a small institute among the 27 institutes and centers at the NIH. During the pandemic, when NHLBI received over a billion dollars of support, NCATS received less than 3 million. NCATS N3C has become the de facto national public health COVID repository of RWD in the US yet is financially not supported by NIH or HHS. The lack of funding and a coordinated effort between HHS agencies is a risk to future public health needs.
- **Lack of a skilled Data Science Workforce**
  - The US has a shortage of available government and academic investigators with the skills needed to effectively use RWD. This lack of qualified investigators is a significant impediment to the country's ability to respond to both a future pandemic as well as our on-going public health crises such as teen suicides, opioid addition, and the decline in maternal health.

Outcomes

The N3C centralized dataset is built on the success of a PCORTF-funded project –*Harmonization of Various Common Data Models and Open Standards for Evidence Generation* that established the foundation for harmonizing disparate clinical data across health care institutions using different common data models.

Although we continue to refine and improve the integration of the N3C clinical and CMS claims data, at this point they are available and being used in our production environment. Investigator support for the use of the linked N3C clinical and CMS claims data is extensive and includes on-going training, data harmonization, data use governance, publication support, and collaborative analytic environment for research.

The scientific impact of having access to linked N3C clinical and CMS claims data is still pending. All N3C projects have at minimum a year of access to the data. Since the CMS Medicare data became available late fall of 2022, we do not expect results until late 2024 or early 2025.

On going efforts to expand CMS Linkage Capacity for investigators

As part of a continuing efforts to make linked CMS claims available. NCATS though funding from the Assistant Secretary for Planning and Evaluation, ASPE's Office of the Secretary Patient-Centered Outcomes Research Trust Fund, OS-PCOR-TF, has funded three National Clinical Cohort Collaborative, N3C studies on, Cancer, Renal Disorders and Long COVID that will leverage infrastructure developed in present project to development a process that can serve as a reference architecture whose goals and objectives include

Goal

1. Accommodation Centers for Medicare & Medicaid Services new policy for investigators CMS data acquisition linkage exception
2. Create a process that is mutually beneficial, efficient, and cost-effective for investigators to get linked claims data for research

PCOR Final Report:
CMS-N3C Linked Dataset and Use Cases

Objectives

1. Work towards a “master agreement (DUA) between CMS and NIH
2. Identify a set of standard files that will be automatically supplied as part of an approved CMS/NIH linkage project/study
3. Phenotypes will be performed within CMS VRDC
4. Tokenization will take place within CMS AWS enclave

November 15, 2024

Appendices: N3C Intra Enclave PPRL Educational Material

Introduction to Privacy-Preserving Record Linkage (PPRL) for N3C Researchers

Definitions

This document is aimed at researchers already familiar with the National COVID Cohort Collaborative (N3C), especially N3C's data partner/Common Data Model (CDM) harmonization framework, the Data Use Request (DUR) process, and the three available 'levels' of N3C data. As such we use the following phrases with these specific meanings:
Clinical data partner or data partner: A site contributing Electronic Health Record (EHR) data via a source CDM for harmonization into N3C's OMOP CDM. These are represented by data_partner_id entries in N3C data.
EHR records or EHR data: Records provided by data partners via source-to-target CDM harmonization.
PPRL supplemental data: Data (i.e. mortality, viral variant data, etc.) linked to an individual using the de-identified PPRL technology.
De-identified Data: Data (whether EHR or otherwise) that has had individually identifying information removed or irreversibly encrypted. In the case of EHR data we use this term to refer to either de-identified or limited data according to HIPAA definitions (for definitions of these see this Johns Hopkins IRB page).

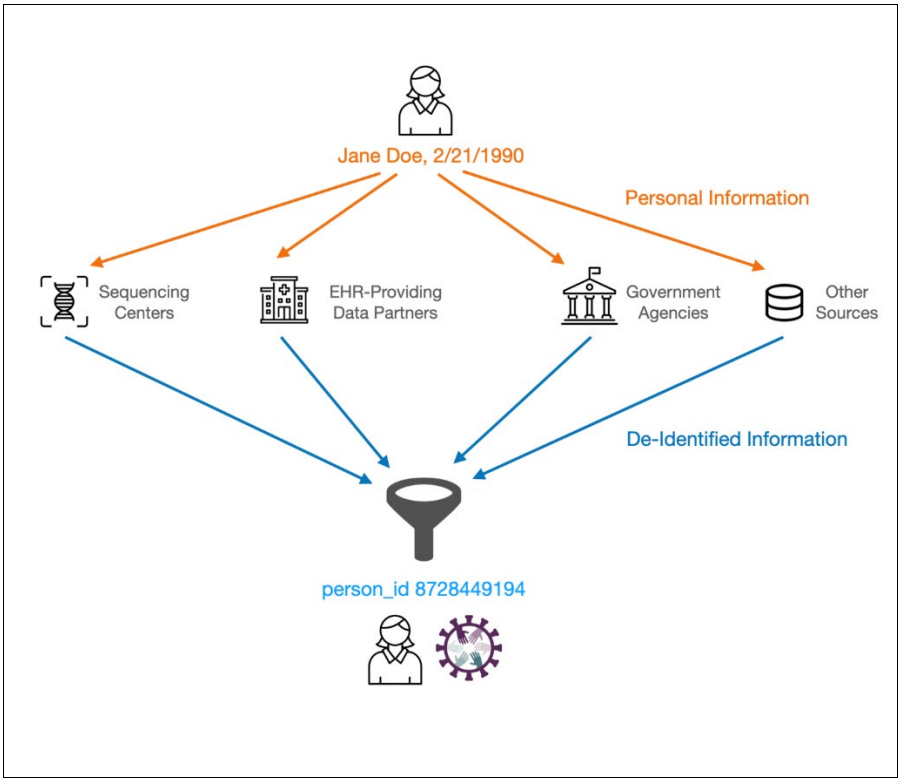
Summary

PPRL (Privacy Preserving Record Linkage) is a cryptographically secure method to link records about de-identified individuals from multiple data sources. PPRL currently supports linking EHR records to data from supplemental data sources (e.g. mortality information from government agencies or viral variant information from sequencing centers). Eventually, PPRL will also support "de-duplicating" individuals known to multiple clinical data partners. Each N3C clinical data partner must opt in to allow their EHR data to be linked to supplemental data of different types. Linked supplemental PPRL data are only available for use alongside the Limited Data Set (a.k.a. Level 3, or LDS). The DUR form has been updated to include a PPRL data request section, and the DUR's project rationale must include a justification for the request if used.

Introduction

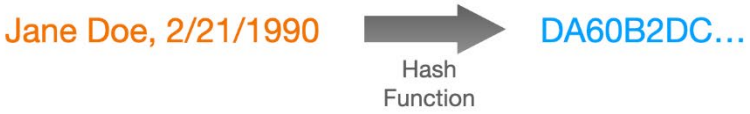
If you've worked with health-related data in any capacity, you are likely familiar with the restrictions on sharing Protected Health Information (PHI) defined by HIPAA, especially identifying information like names and birth dates. To avoid unnecessary risks to patient privacy, research consortia like N3C utilize de-identified or limited datasets (for definitions of these see this Johns Hopkins IRB page).
What happens, however, when a single patient's data comes from multiple data sources, to be

later harmonized in the N3C Data Enclave? In these cases, the identifying information is removed prior to data sharing, making it impossible to know when two data records from different sources refer to the same person.

PPRL is a cryptographically secure method to link information about individuals from different sources of data, without revealing personal information about those individuals. PPRL enables N3C to collect information such as EHR data from clinical data partners, viral variant information from sequencing centers, and mortality information from government and private sources, all into a single unified database for researcher use.	

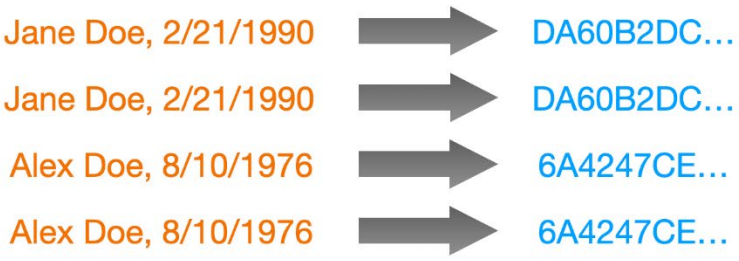
How it Works

To get a sense of how PPRL works, it helps to be familiar with what is known as a *hash function* in cryptographic applications. A hash function maps a set of data to a long string known as a *hash*:



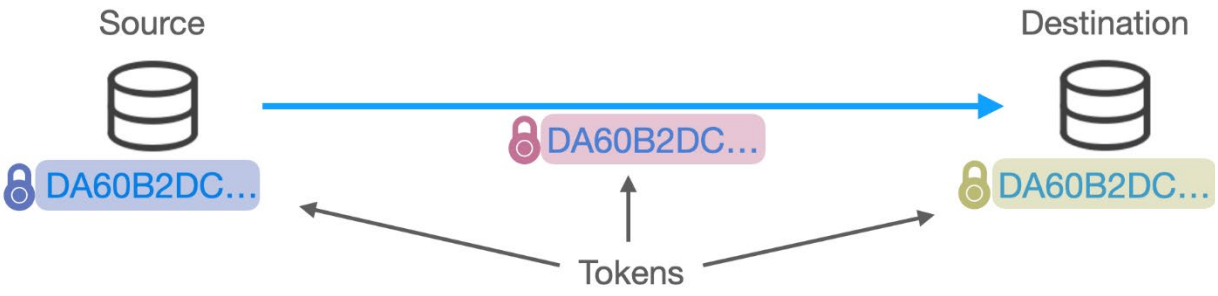
Hash functions are widely used in cryptography, including in digital signatures and banking and finance applications. Cryptographic hash functions have the following important properties for PPRL:

First, the same input data will always result in the same hash:



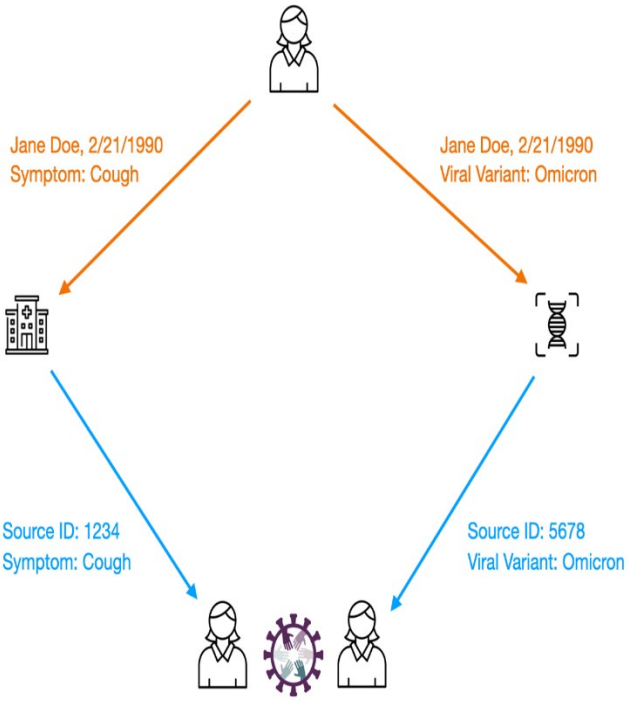
Second, it is not possible to extract the input from the hash; the hash function is "one way" or "irreversible":

How does this work in the context of data sharing between two entities involved in N3C? Although hash functions are a part of the process, the hashed information is also *encrypted* in several ways. First, when hashes are stored by a data source producing them or a destination receiving them, they are encrypted "at rest" using secret encryption keys unique to the source or destination. Second, when hashes are transferred between two entities, they are encrypted "in transit" using a third key unique to the source/destination pair. As a result, a data breach at any given source or destination (or anywhere in between) protects information *even from other participating entities*. Once encrypted, the hashes are referred to as *tokens*.

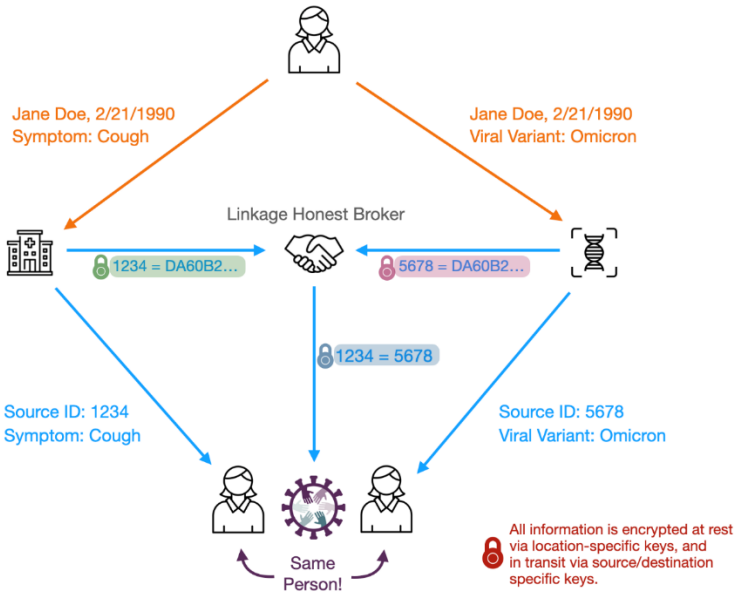


Locked boxes of different colors indicate unique encryptions of data they cover. The transferred value of DA60B2DC… would thus be something like 3DAF258…, while the stored value of DA60B2DC… would be C18D56A….

But that's just for a single source/destination pair. To understand PPRL as used by N3C, consider an example of two sources of data about a patient named Jane Doe. One is an N3C clinical data partner with EHR data about her, and the other is a sequencing center with information about her viral variant. Because information is de-identified at the source, each data source creates a random "source ID" for everyone in their payload. Because N3C doesn't know which source IDs refer to the same individuals, these two records for Jane are assumed to be about two different individuals (right).



With PPRL, data sources can send to a trusted third party the source IDs they are using, and the tokens securely encoding Jane's identifying information. This third party is known as the Linkage Honest Broker, who can in turn alert N3C which source IDs refer to the same person.



What It All Means

Now, there are a lot of identifiers floating around, but importantly, **protected health information only travels between the patient and their healthcare providers**. As a researcher you needn't worry about matching these various identifiers, either: N3C handles unifying them so that the same individual is always represented by the same person_id in the OMOP-formatted data and any

supplemental data available. (With the caveat that using PPRL to de-duplicate patient_ids across clinical data partners is not yet live.)

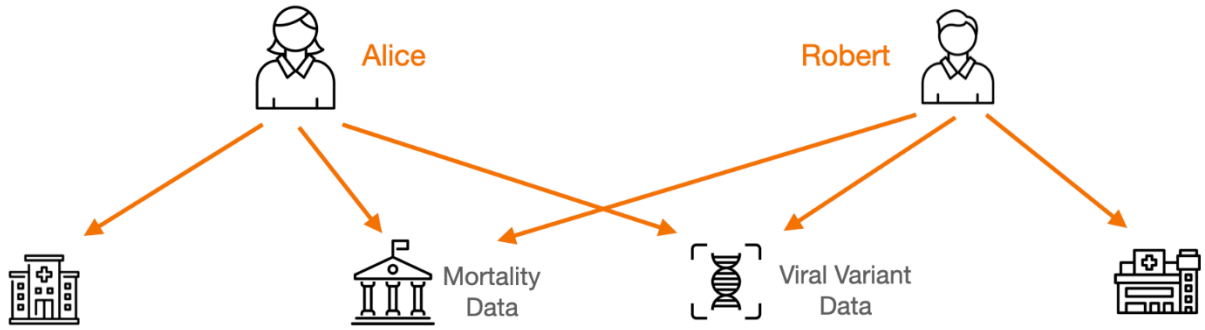
You may be wondering: why the need for an honest linkage broker at all? Couldn't data sources send N3C tokens directly for matching? Technically, yes. However, although N3C is the driving force behind PPRL, using a linkage honest broker establishes a hub-and-spoke model that can benefit multiple NIH-supported efforts, and opens the door to linking with future enclaves like N3C. This model also provides important privacy and governance features–**the broker only has access to tokens (which are de-identified at the source), and N3C only has access to the specific linkage information deemed appropriate by the broker**.

If all of this seems complicated, it is! Consider that every data flow represented requires legal agreements and risk analysis, and N3C coordinates over 70 data sources as of early 2022. In addition, we've simplified the picture, since we are naïvely assuming that name and date of birth uniquely identify an individual. In reality more information is used to produce the tokens (including information like patient sex, social security number, and zip code), but the process must also account for missingness of information, which means that multiple "versions" of tokens (securely encoding different subsets of information) are shared with the linkage broker for matching purposes, who uses specific rules to produce high-quality matches depending on the token versions available.

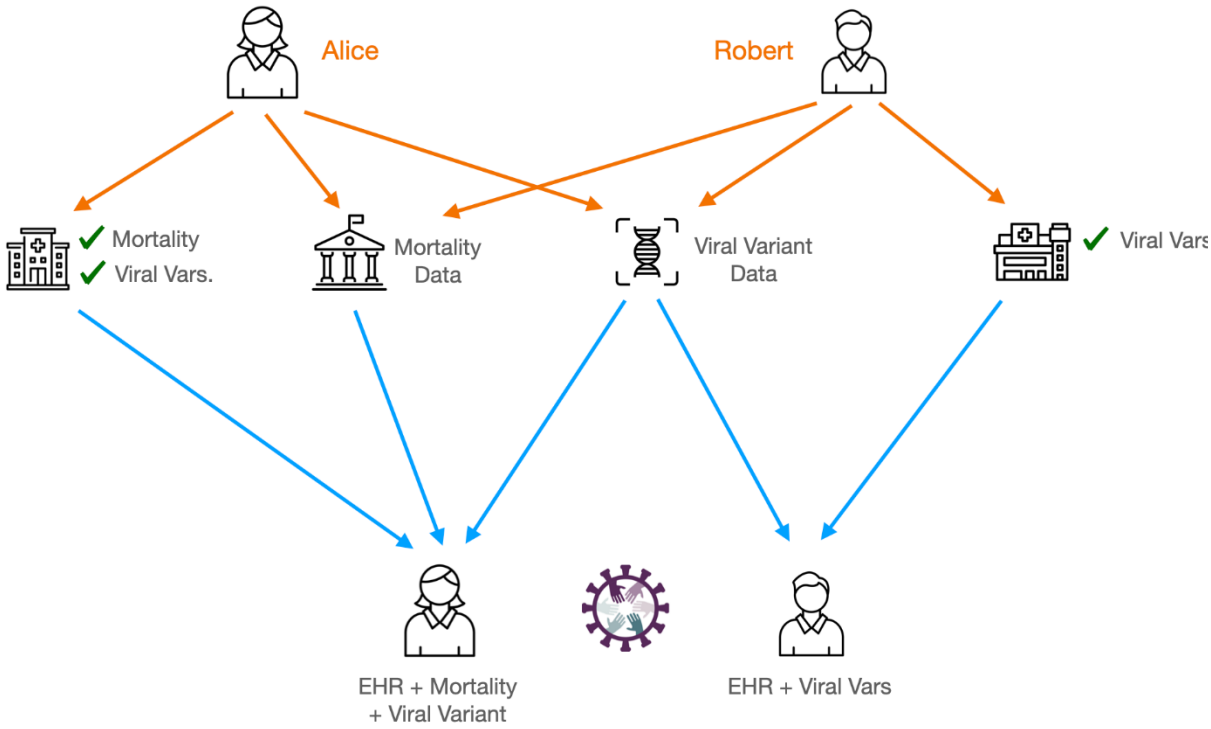
Supplemental Data Completeness

As mentioned in the summary, PPRL may serve two purposes for N3C: first, providing linkage between EHR records provided by clinical data partners and supplemental data from PPRL data sources, and second, de-duplicating individuals known to multiple data partners. As of February 2022, the de-duplication feature is not yet enabled, so we'll focus on supplemental PPRL data.
As a researcher, it is important to note that supplemental PPRL data will only be available for use alongside Level 3 (LDS) data, and require justification for the PPRL data as part of the DUR submission (we'll discuss access in more detail below).
Even with approved supplemental PPRL data access, however, only a subset of linked records will be available to researchers, because individual data partners opt in (or out) of linking their data to individual types of supplemental data. To illustrate this more concretely, let's consider two hypothetical patients, Alice and Robert, who's EHR records are provided by two different data partners:

As it so happens, both Alice and Robert had their viral variants sequenced. Both also subsequently passed away outside of the healthcare system, so their mortality information is only available via supplemental data:



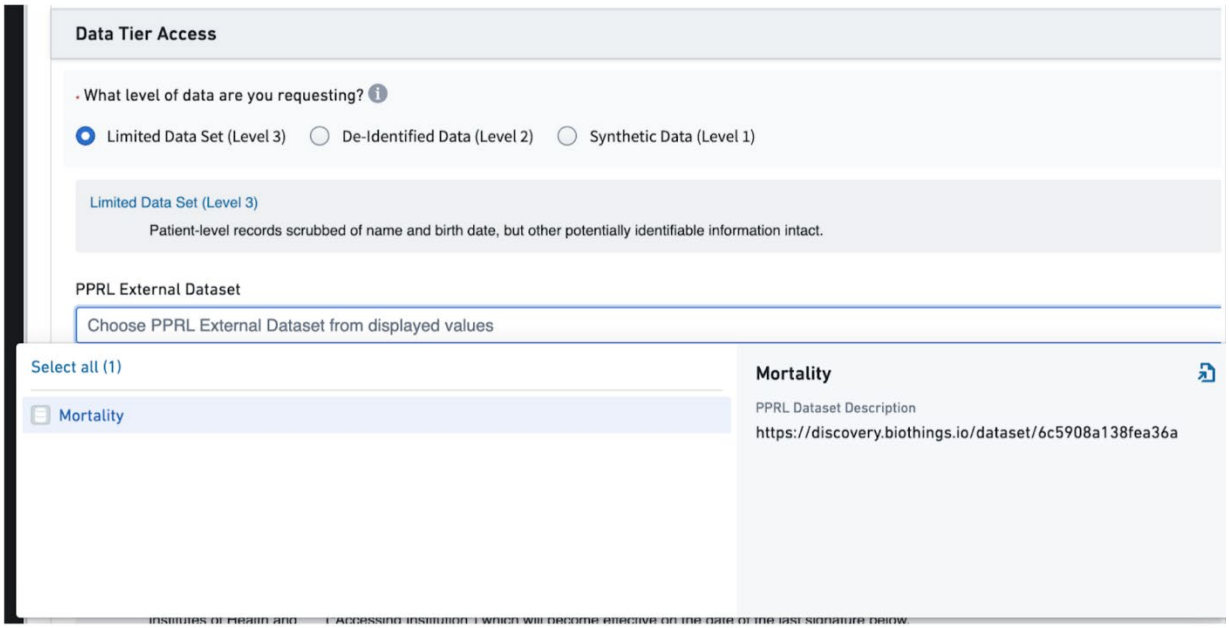
These four data sources are available to N3C via PPRL. However, while Alice's EHR-providing data partner has consented to linkage with both mortality and viral variant data, Roberts has only consented to viral variants linkage. As a result, in the N3C enclave information about Alice includes both viral variants and mortality, but for Robert only viral variant information is available.



Besides data partner linkage permission, each supplemental data type comes with its own completeness considerations. Mortality data for example lags actual mortality by days to weeks, and viral variant information is only available for a subset of the population.
Further details about specific supplemental PPRL data will be available in other resources as they are made available to researchers.

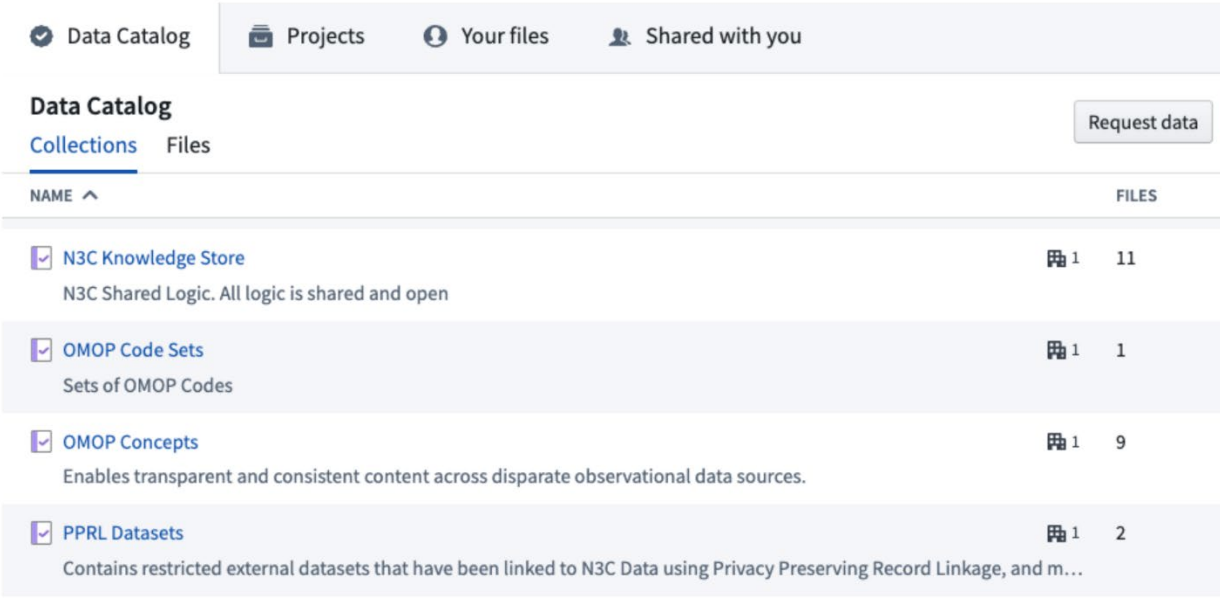
Accessing Supplemental PPRL Data

PPRL data is only available alongside Level 3 (LDS) data, and must be explicitly requested per data type, either as part of a new DUR submission, or as part of a DUR upgrade request (even for projects that are already approved for Level 3 access). The DUR form now includes a segment where project leads can specify which PPRL datasets they would like access to:



Important: When requesting access to supplemental PPRL data as part of a Level 3 DUR, the submitting lead must include as part of the research project rationale why the selected datasets are required for the project. Failure to do so will result in rejection of the DUR.

Once the DUR is approved and data access has been granted, the datasets will appear in the PPRL Datasets entry in the Data Catalog:



Additional Information

More information about PPRL can be found at the following links:

- N3C PPRL Public Page: https://covid.cd2h.org/pprl
- Regenstrief (Linkage Honest Broker) Page: https://www.regenstrief.org/n3c-lhb/
- NCATS Program FAQ (mentioned PPRL under "about the data"): https://ncats.nih.gov/n3c/about/program-faq
- NCATS Data Contribution Forms and Resources Page (for data sources and partners): https://ncats.nih.gov/n3c/resources/data-contribution
- N3C Biothings Catalog for external & supplemental data: https://discovery.biothings.io/faq/n3c

N3C PPRL CMS Data Guide

Definitions

This document is aimed at researchers already familiar with the National COVID Cohort Collaborative (N3C), especially N3C's data partner/Common Data Model (CDM) harmonization framework, the Data Use Request (DUR) process, and the three available 'levels' of N3C data. As such we use the following phrases with these specific meanings:

Clinical data partner or **data partner**: A site contributing Electronic Health Record (EHR) data via a source CDM for harmonization into N3C's OMOP CDM.

EHR records or **EHR data**: Records provided by data partners via source-to-target CDM harmonization.

PPRL supplemental data: Data (i.e. mortality, CMS, etc.) linked to an individual using the de-identified PPRL technology

De-identified Data: Data (whether EHR or otherwise) that has had individually identifying information removed or irreversibly encrypted. In the case of EHR data we use this term to refer to either de-identified or limited data according to HIPAA definitions (for definitions of these see this Johns Hopkins IRB page).

Summary

Before getting started, you may wish to review the Introduction to PPRL documentation to understand what PPRL (Privacy Preserving Record Linkage) is, how it is used in N3C, and how to get access to supplemental PPRL data–like the CMS data discussed here. As a reminder, supplemental PPRL CMS data is only available with a Level 3 DUR specifically requesting access, and the project rationale section must justify the additional request. Existing Level 3 projects that wish to add access to PPRL supplemental data will need to submit a revised DUR for approval.

CMS PPRL data provide an additional source of health information from billing claims submitted to the Centers for Medicare & Medicaid Services. These claims data are rich, and cover may of the same domains as other EHR data sources, including conditions, drug exposures, procedures, and so on. Claims data also include information about providers and visits, and while the source format

for CMS data is complex, N3C has transformed these records into near-standard OMOP domain tables with minimal structural (schema) differences compared to primary N3C EHR OMOP tables. CMS data contains records originating not only from participating N3C data partners, but all entities submitting Medicare or Medicaid claims for a given patient. Due to the way the PPRL linkage process works, CMS patients are given CMS-unique identifiers (person_ids), and a lookup table is provided to map these to N3C EHR identifiers (person_ids) and their corresponding data partners (data_partner_ids). Like all PPRL datasets, this mapping is only provided for N3C patient IDs associated with data partners who have opted in to linkage against the CMS data, but for included patients CMS records are provided from all CMS-billing providers. Researchers should thus consider carefully which EHR records, if any, to include alongside CMS data for research.

See the Dec. 12, 2022, CMS Data Webinar at https://youtu.be/fs0tM7RnL54.

Introduction

Medicare and Medicaid are federally run programs that entirely or partially cover medical costs for eligible populations. As described on the HHS website:

Medicare is an insurance program. Medical bills are paid from trust funds which those covered have paid into. It serves people over 65 primarily, whatever their income; and serves younger disabled people and dialysis patients. Patients pay part of costs through deductibles for hospital and other costs. Small monthly premiums are required for non-hospital coverage. Medicare is a federal program. It is basically the same everywhere in the United States and is run by the Centers for Medicare & Medicaid Services, an agency of the federal government.

Medicaid is an assistance program. It serves low-income people of every age. Patients usually pay no part of costs for covered medical expenses. A small co-payment is sometimes required. Medicaid is a federal-state program. It varies from state to state. It is run by state and local governments within federal guidelines.

Eligible patients may enroll for these services, and medical providers (including pharmacies) submit claims for those services or drugs for reimbursement. As claims data, CMS information provides slightly different information compared to EHR data; for example, rather than providing medication prescription information, CMS provides information on medication dispensation from the pharmacy. Enrollment in Medicare and Medicaid are well defined and thoroughly tracked, providing insight into patient's healthcare utilization vs eligibility that may be lacking in EHR data. For a more detailed investigation of the usefulness of EHR vs administrative claims data, see Comparing Population-based Risk-stratification Model Performance Using Demographic, Diagnosis and Medication Data Extracted From Outpatient Electronic Health Records Versus Administrative Claims by Kharrazi et al., 2017.

Date Range and Freshness

CMS records available cover the period from Jan 1, 2017, to the most recently available, which may lag by up to 3 months. CMS-derived OMOP tables are updated on schedules specific to the source CMS data type, future releases are planned to include a manifest providing data freshness

information.

Data Completeness

In addition to the requirement that data partners opt-in to PPRL data linkage as described in the Introduction to PPRL documentation (though see Patient ID Duplication below), CMS records are currently only available for patients who have a reported COVID-19 *diagnosis* in the CMS data; data about other N3C-included patients are not available (including matched 'control' patients and those with only positive lab tests or weak positive indicators as described by the N3C phenotype). Currently data are also not available for sites that pre-date-shift prior to submitting to N3C (these are indicated in the Level 3 and Level 2 manifest table). N3C is working to expand the set of included patients, and we will update this documentation when more information is available.

OMOP Domain Mapping

Available CMS records are mapped to their closest-match OMOP domain table, with CMS-defined medical concepts mapped to OMOP standard concepts. The following OMOP tables are currently available:

- care_site
- condition_era
- condition_occurrence
- death
- device_exposure
- drug_era
- drug_exposure
- observation
- observation_period
- person
- procedure_occurrence
- provider
- visit_occurrence

These tables generally match the schema of their counterparts in the N3C LDS (Level 3) OMOP tables, with a few notable exceptions.

1. The person_id fields are formatted to contain CMS-specific IDs of the format CMS61313455, which emphasizes that these IDs are not directly compatible with person_ids in other N3C OMOP tables (a mapping file is provided, see below), but violates the OMOP standard assumption that person_ids should be integers. You may thus need to modify these identifiers when using some OHDSI-provided tools that require integer IDs.
2. CMS OMOP tables also include a data_partner_id column, but all values are set to the string CMS to highlight their origin. This differs from the data_partner_id columns in other N3C OMOP tables in that it is a string value.
3. CMS information provides race and ethnicity information in a single column, prioritizing ethnicity over race when available, whereas OMOP records this information in two

independent columns. As a result, in the CMS OMOP person table when ethnicity information is available race information will be missing.

Other differences and similarities will be noted in a data dictionary (still in development).

Patient ID Duplication

CMS records are sourced independently of the EHR data provided by N3C data partners. N3C's PPRL process (in conjunction with the Linkage Honest Broker as summarized in the Introduction to PPRL documentation) starts by identifying CMS patients matching patients described by one or more data partners who have agreed to linkage with the PPRL CMS data. **All** CMS records from these matched patients are included in the CMS data, but the linkage between CMS patient identifiers and N3C person_id identifiers is only provided for those data partners who have opted in to this linkage.

To make this clear, consider the following two patients with records in the N3C LDS data; only one has records associated with a data partner who has opted into CMS linkage:

In the CMS data, you will find a CMS condition_occurrence table, as well as a mapping file linking CMS person_ids to N3C person IDs and their corresponding data partner IDs:

CMS condition_occurrence Table

person_id	data_partner_id	condition_concept_id	...
CMS1000000	CMS	2155156	...
CMS1000000	CMS	10589	...

CMS person_id Mapping Table

CMS person_id	N3C person_id	N3C data_partner_id
CMS1000000	11111111	732

This does *not* mean that all CMS records are sourced from EHR-providing data partner 732. Rather, the CMS data contain information from all CMS-billing sources, and that this patient is additionally known as person 11111111 to data partner 732.

In some cases, a single individual will be linked to multiple N3C person_ids, indicating this is the same individual known to multiple data partners. (In the initial release, approximately 1% of CMS

person_ids are associated with more than one N3C person_id.) Here's an example where an individual is known to two data partners who have both agreed to CMS linkage:

CMS condition_occurrence Table

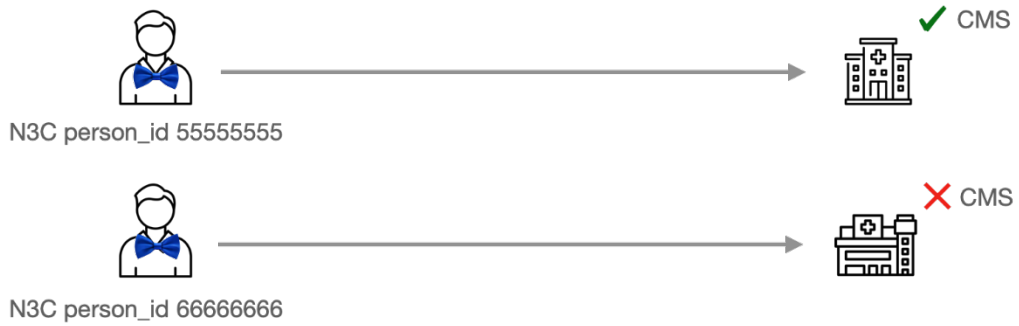
person_id	data_partner_id	condition_concept_id	...
CMS2000000	CMS	2155156	...
CMS2000000	CMS	10589	...

CMS person_id Mapping Table

CMS person_id	N3C person_id	N3C data_partner_id
CMS2000000	33333333	732
CMS2000000	44444444	119

The CMS data do not indicate which data partner, if any, the records are sourced from. Indeed, CMS records will include data from a variety of sources, including N3C data partners and other providers not otherwise affiliated with N3C.

Because CMS data include records from all CMS-billing sources, these may include records originating from data partners not opting in to CMS PPRL data linkage. In the figure below, we suppose N3C person_ids 55555555 and 66666666 are in fact the same person known to multiple data partners, but only one data partner has opted in to this linkage. The person ID mapping table will contain linkage information only for the opted-in data partner, but the CMS data itself will likely contain records originating from both data partners.



CMS condition_occurrence Table

person_id	data_partner_id	condition_concept_id	...
CMS3000000	CMS	2155156	...
CMS3000000	CMS	10589	...

CMS person_id Mapping Table

CMS person_id	N3C person_id	N3C data_partner_id
CMS3000000	55555555	732

Remember, it is against N3C policy to attempt to re-identify patients or data partners. Researchers should also be aware of this fact and be thoughtful about what datasets (EHR, CMS, or other) are included in analyses (see Record Duplication, below).

EHR/CMS Record Duplication

It is important to understand that CMS data contains an overlapping set of information provided by N3C EHR data. A Medicare-eligible patient may, for example, have a procedure recorded in the standard LDS procedure table, and the same procedure recorded in CMS procedure table. Coding differences and date discrepancies between EHR and claims data make it non-trivial to perform automatic de-duplication of such records, and N3C has not attempted to do so on behalf of researchers.

Patient Deduplication

N3C Patient Deduplication Announcement:

N3C is announcing national de-duplication functionality to all N3C investigators in all future data releases. Deduplication on a national scale has enormous scientific implications and is increasingly important as we link different types of data like CMS Medicare and Medicaid.
However, though we are excited to expose this powerful new functionality, we also want to caution investigators of the importance of understanding the ramifications of deduplication before incorporating it into their analysis. Over the coming weeks more information on best practices will be rolled out and incorporated in release notes as well the manifest tables.

What is deduplication and how does it work?

N3C deduplication uses Privacy Preserving Record Linkage, (PPRL) which is a means of connecting records using secure, pseudonymization processes in a data set that refer to the same individual across different data sources while maintaining the individuals' privacy.
The process of de-duplication is beyond the scope of this document. It is important to understand that PPRL is has contains not PII ONLY data contributing sites have access to PII. The image below gives a high-level overview of the process, for more information can be found at N3C Privacy-Preserving Record Linkage

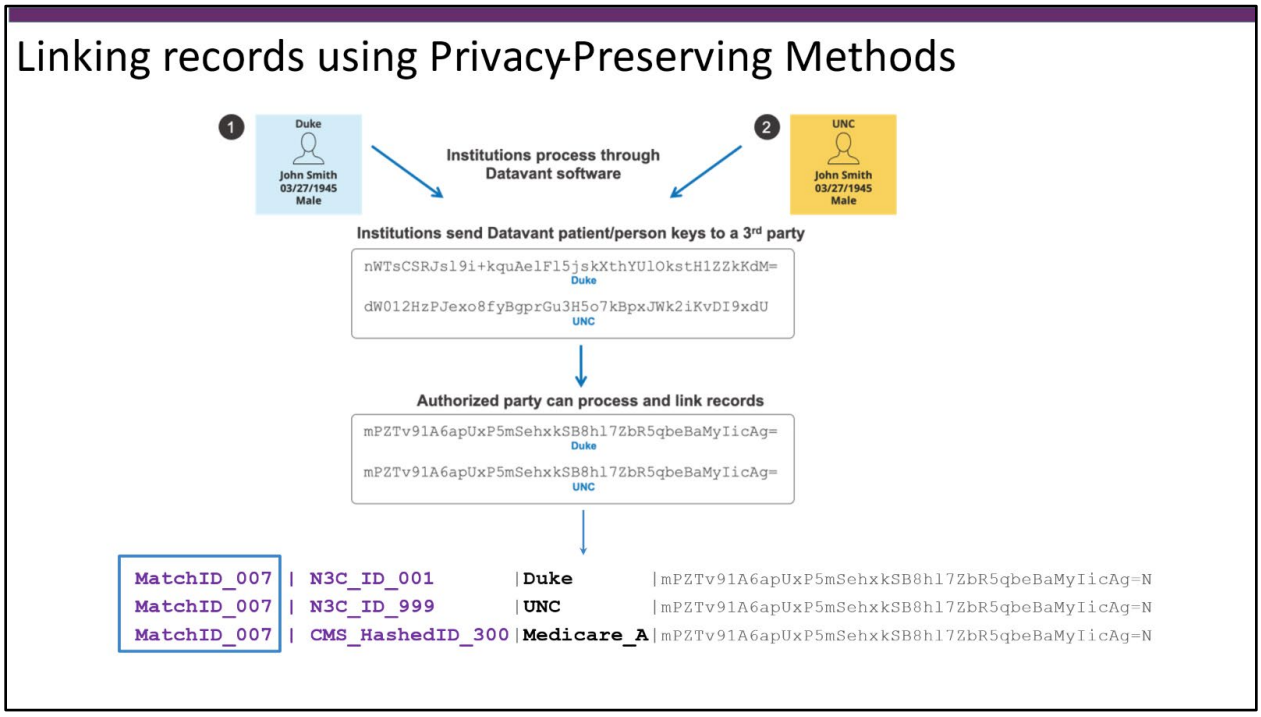


Figure 16: PPRL Patient Matching Process

How will this impact my ongoing analysis?

There will be no impact to your existing analysis. Implementing deduplication is controlled by investigators and your data will not be impacted.

How will I know if a patient has a duplicate?

The information shown is only informational, and any changes requires investigators to alter their code to implement deduplication.

The OMOP Person Table will have two additional columns, Global_ID and Duplication type.

Global ID	Duplication Types		Site ID
	NULL	111	AB
123	Intra-Site Duplicate	111	CD
123	Intra-Site Duplicate	111	EF
456	Inter-Site Duplicate	111	GH
456	Inter-Site Duplicate	222	IJ
	NULL	222	KL
	Unknown	333	MN

2 Person Table:

Global_ID = Is an identifier present in the Global_ID column when there are two or more person IDs that point to the same individual. A blank column indicates that no duplicates for the is person exists and
Duplication type = Defines 4 unique types of duplicates from different sources of information. This includes.

1. **Intra-Site Duplicate** = indicates duplicate patients from the same data source for example
    a. John Doe at University ABC has 2 MRN.
2. **Inter-Site Duplicate** = indicates duplicate patients from different data sources for example
    a. John Doe at University ABC <=> John Doe at University of XYZ
    b. John Doe at University ABC <=> John Doe from Medicare
3. **NULL** = Data Contributing site is participating in PPRL and the patient has NO duplicates.
4. **UNKNOWN** = Data contributing Site indicates site is **NOT** participating in PPRL and duplicate status cannot be determined. It is important to understand, a status of "Unknown" does not indicate there are no duplicates.

How do I know if a site is participating in de-duplication

As part of the implementation of de-duplication N3C has added the status of each site's participation to the manifest table.

Site_ID	Clinical Linkage	Mortality Linkage	Medicaid Linkage	Medicare Linkage
111	Yes	Yes	Yes	Yes
222	Yes	Yes	No	No
333	No	No	No	No

3 Manifest Table: Data Contributor Sites and Data Type Linkage Status

Why can't I just merge all duplicate patients?

Merging each type of de-duplication has different implications and depending on your analysis you may or may not want to deduplication patient, and

1. **Intra-Site Duplicate** = indicates duplicate patients from the same data source for example
    a. John Doe at University ABC has 2 MRN.
2. **Inter-Site Duplicate** = indicates duplicate patients from different data sources for example
    a. John Doe at University ABC <=> John Doe at University of XYZ
    b. John Doe at University ABC <=> John Doe from Medicare

3. **NULL** = Data Contributing site is participating in PPRL and the patient has NO duplicates.
4. **UNKNOWN** = Data contributing Site indicates site is **NOT** participating in PPRL and duplicate status cannot be determined. It is important to understand, a status of "Unknown" does not indicate there are no duplicates.

This is so complicated. Why don't I just drop all duplicate patients.

This is not an unreasonable option. For instance, if your population has many patients the removal of a small percentage will probably not impact your results. However, if you are investigating a rare disease with very small numbers of patients, knowing which patients are the same is essential.

Does N3C have a best practice?

Duplication is complicated and will differ by deduplication type. At the present time we encourage investigators to deduplicate all **"intra-site"** duplications. The logic being the data contributor, data site, common data model and providers from this institution are the same and therefore the impact on merging will enhance the data quality of your analysis.

Global Id	Duplicate Status	Pt_ID	Site ID
+1234	ABC	ABC	111
*NULL	DEF	DEF	111
*NULL	GHI	GHI	123
*NULL	JKL	JKL	123
+ 1234	MNO	MNO	123
++5678	PQR	PQR	155
*NULL	STU	STU	155
++5678	VWX	VWX	155
^ Unknown	YZ	YZ	175

Site ID	Clinical Linkage	Mortality Linkage	Medicare Linkage	Medicare Linkage
111	Yes	Yes	Yes	Yes
123	Yes	Yes	Yes	Yes
155	Yes	Yes	No	No
175	No	No	No	No

Global Id
* Null = indicates site is participating in PPRL and there are no duplicates with other data/sites using PPRL
^ Unknown = indicates site is NOT participating in PPRL and duplicate status is undetermined
Global Id is an integer and
+ Same site IDs = indicates duplicate patients from the same data source
++ Different Site IDs = indicates duplicate patients from different data sources i.e., N3C and CMS

*** put in manifest same site dup two columns

Figure 17: Person Table

Global Id	Duplicate Status	Pt_ID	Site ID
+1234	*dup	ABC	111
	*NULL	DEF	111
	*NULL	GHI	123
	*NULL	JKL	123
+ 1234	*dup	MNO	123
++5678	share	PQR	155
	*NULL	STU	155
++5678	share	VWX	155
	Unknown	YZ	175

Figure 18: Person ID

Table Key:

Global Id

1. Null = indicates site is participating in PPRL and there are no duplicates with other data/sites using PPRL
2. ^ Unknown = indicates site is NOT participating in PPRL and duplicate status is undetermined
3. Global Id is an integer and
4. + Same site IDs = indicates duplicate patients from the same data source
5. ++ Different Site IDs = indicates duplicate patients from different data sources i.e., N3C and CMS
6. *** put in manifest same site dup two columns