Received: 22 July 2023 | Revised: 6 December 2023 | Accepted: 6 December 2023

DOI: 10.1002/lrh2.10404

LEARNING FROM DATA

Learning Health Systems

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

Umberto Tachinardi[1] | Shaun J. Grannis[2] | Sam G. Michael[3] | Leonie Misquitta[3] | Jayme Dahlin[3] | Usman Sheikh[3] | Abel Kho[4,5] | Jasmin Phua[5] | Sara S. Rogovin[5] | Benjamin Amor[6] | Maya Choudhury[6] | Philip Sparks[6] | Amin Mannaa[6] | Saad Ljazouli[6] | Joel Saltz[7] | Fred Prior[8] | Ahmen Baghal[8] | Kenneth Gersing[3] | Peter J. Embi[9]

[1]Department of Biomedical Informatics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

[2]Center for Biomedical Informatics, Regenstrief Institute, Department of Family Medicine, IU School of Medicine, Regenstrief Institute, Inc. and Indiana University School of Medicine, Indianapolis, Indiana, USA

[3]National Center for Advancing Translational Science, NIH, Bethesda, Maryland, USA

[4]Department of Medicine, Northwestern University, Feinberg School of Medicine, Chicago, Illinois, USA

[5]Public Sector, Datavant, Inc, San Francisco, California, USA

[6]Federal Health, Palantir Technologies, Denver, Colorado, USA

[7]School of Medicine, Stony Brook University, Stony Brook, New York, USA

[8]COM Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

[9]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Correspondence
Shaun J. Grannis, 1101 W. 10th Street, Indianapolis, IN 46202, USA.
Email: sgrannis@regenstrief.org

Funding information
National Center for Advancing Translational Sciences, Grant/Award Number: 75N95021D00028

Abstract

Introduction: 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. However, making this data most useful for research requires linking it with information such as mortality data, images, and viral variants. The objective of this project was to establish privacy-preserving record linkage (PPRL) methods to ensure that patient-level EHR data remains secure and private when governance-approved linkages with other datasets occur.

Methods: Separate agreements and approval processes govern N3C data contribution and data access. The Linkage Honest Broker (LHB), an independent neutral party (the Regenstrief Institute), ensures data linkages are robust and secure by adding an extra layer of separation between protected health information and clinical data. The

LHB's PPRL methods (including algorithms, processes, and governance) match patient records using "deidentified tokens," which are hashed combinations of identifier fields that define a match across data repositories without using patients' clear-text identifiers.

Results: These methods enable three linkage functions: Deduplication, Linking Multiple Datasets, and Cohort Discovery. To date, two external repositories have been cross-linked. As of March 1, 2023, 43 sites have signed the LHB Agreement; 35 sites have sent tokens generated for 9 528 998 patients. In this initial cohort, the LHB identified 135 037 matches and 68 596 duplicates.

Conclusion: This large-scale linkage study using deidentified datasets of varying characteristics established secure methods for protecting the privacy of N3C patient data when linked for research purposes. This technology has potential for use with registries for other diseases and conditions.

KEYWORDS
clinical research, data privacy, learning health systems, record linkage, translational research

1 | BACKGROUND AND SIGNIFICANCE

Having reliable and diverse healthcare data is necessary to conduct research, generate evidence, and advance a learning health system. However, it is challenging to utilize such data due to the fragmentation of information across various health systems, hindering our ability to develop evidence and make discoveries that can ultimately improve health. Integrating data from multiple health systems is essential to realize the vision of a data-driven learning health system. To achieve this, protecting patient privacy and ensuring patient data is secure and safeguarded against unauthorized access is critical to building trust in a data-driven learning health system. Additionally, it is vital to enable integration across sites to maximize the value of health data for improving patient care. Achieving both objectives—ensuring privacy and enabling integration—is a significant challenge in developing a data-driven learning health system. It requires careful consideration and planning to achieve both goals effectively.

The COVID-19 pandemic has highlighted the importance of abundant, secure, and accessible patient data from various sources for research and guiding public policy and clinical care. However, assembling and harmonizing real-world data from different sites and formats into a functional data repository poses significant challenges. Challenges of using RWD include duplicate, incomplete, and siloed patient records; patient privacy concerns; and barriers to sharing data across institutions and to linking both similar and different types of data.[1-4] To overcome these challenges with COVID-19 data and allow investigators to access it quickly and securely, the National Center for Advancing Translational Sciences (NCATS) of the United States National Institutes of Health (NIH) established the National COVID Cohort Collaborative (N3C).[5] A primary focus of N3C is creating and managing the N3C Data Enclave, a centralized repository of harmonized and deidentified COVID-19 electronic health record (EHR) patient data collected from institutions across the United States and held in a secure, cloud-based platform.

As of March 2023, 77 of the 93 N3C partner institutions have completed data submission requirements and have transferred their COVID-19 EHR data to the N3C Data Enclave. That data is accessible to 356 institutions that have applied and been approved to use it for research. The data enclave, hosted and managed by NCATS, is the largest repository of longitudinal row-level COVID-19 EHR data available for research in the United States. This data begins in January 2018 and includes clinical and demographic characteristics of patients tested for or diagnosed with COVID-19, as well as treatment information for those confirmed or suspected to have the virus. At this writing, the enclave contains records for 7.1 million COVID-positive patients and 18.1 million total patients from 48 of 50 states, encompassing 22.7 billion rows of data, 900 million procedures, 10.8 billion lab results, and 1.5 billion drug exposures (Figure 1). This collection continually expands as current N3C partners update their data and new partners join the collaborative.

Regardless of the size of an EHR repository such as the N3C enclave, it is limited to the healthcare encounters captured in the electronic record. Other health-related data, including imaging like chest x-rays, CT scans, MRIs, or genomic information, is often absent from the EHR. Additionally, contextual information that is not typically part of medical care, such as social determinants of health (SDOH), is also absent.[6] Obtaining a more comprehensive understanding of an individual's health is necessary for making high-impact scientific discoveries that can inform clinical decision-making. To that end, we have so far imported over 60 publicly available external datasets into the N3C Data Enclave from sources such as the US Census, US Postal Service, Environmental Quality Index, American Community Survey, Food Access Research Atlas, and Centers for Disease Control and Prevention pediatric growth data.[7] These datasets are available as resources

FIGURE 1 Patient linkage statistics for the N3C linkage honest broker process as of March 2023. Privacy preserving record linkage (PPRL) is used to identify and link multiple records for the same person within and across data contributing sites.

within the enclave and can be linked broadly to a patient by zip code or county. But if the N3C is to meet its full potential, it needs to also allow data linkages that are patient-specific. Adding a person's COVID viral variant type, for instance, or one's ability to return to work after a COVID hospitalization would be invaluable.

To maintain patient privacy and secure data linkage, robust systems are necessary to uphold public trust and meet the privacy safeguards mandated by the 1996 Health Information Portability and Accountability Act (HIPAA). Soon after HIPAA's release, members of our team began developing and evaluating some of the earliest techniques for securely matching patients across multiple organizations using secure hashing methods.[8,9] The objective of this project was to operationalize privacy-preserving record linkage (PPRL) methods to ensure that patient-level EHR data remains secure and private when it is linked within the N3C enclave and with external datasets for use in research. In this paper, we detail the secure methods and infrastructure developed for linking N3C data. We outline the three key functions enabled by this system: deduplication, multi-cohort dataset linking, and cohort discovery. Further, we describe the status of these linkages and discuss the insights gained from our experiences.

2 | MATERIALS AND METHODS

Based on best practices from established data-sharing models, N3C created a data governance framework consisting of agreements and practices that provide general data security and quality control.[5,10] A central Institutional Review Board (IRB) was established at Johns Hopkins University School of Medicine to cover contribution of data and provide ongoing oversight. Sites' data transfer can also be governed by site institutions' IRBs. The N3C enclave is covered by a Certificate of Confidentiality, which the limits release of sensitive data, and Community Guiding Principles describing ethical expectations. Beyond these general provisions, the transfer and use of N3C data require participants to sign specialized agreements. Because linking health data for research purposes has additional distinct requirements related to patient privacy, we also developed specially tailored PPRL methods implemented by a neutral party, the Linkage Honest Broker (LHB).

2.1 | Data transfer and data use agreements

The N3C partnership between data contributors and NCATS as the steward of this data is based on a Data Transfer Agreement (DTA) [https://ncats.nih.gov/n3c/resources/data-contribution] that defines each party's responsibilities including what will be sent to NCATS permitted data use and obligations to protect clinical information. All institutions contributing data must sign the DTA, agreeing to send a limited data set with true dates and full zip codes among other requirements.

Access to the N3C data is governed separately. Organizations that wish to access N3C data and related datasets for COVID-19 research purposes must sign an institution-wide Data Use Agreement (DUA) with NCATS. Haendel et al. emphasize that "the decision to cover data transfer and data use as separate agreements was intentional, as it allows organizations to access data even if they do not contribute data".[5] The DUA is part of the NCATS Data Use Request (DUR) framework, designed to protect patient data while ensuring it is only used as approved. Data access must be project-based and meet all use requirements. Requirements vary depending on level of access sought, but all requests for data to be used in a specific study must include a DUR specifying the intended use of the data and a signed User Code of Conduct as well as security and human subjects training. To access true dates of service as well as patient complete zip code information, investigators must also have project-specific IRB reviews from their institutions. The N3C Data Access Committee is composed of federal employees and is evaluates and adjudicates investigators' requests for permitted uses of the data.

Approved research projects are given a secure virtual workspace with access to the COVID-19 EHR records, the publicly available datasets, and over 3000 knowledge objects developed by the N3C community. The knowledge objects can be thought of as reusable codes and range from definitions of diseases to standard methods for defining a hospitalization. Sharing knowledge objects enables the research community to build upon others' work, facilitates reproducibility, improves efficiency, and increases collective understanding of COVID-19. For data security and oversight, researchers are not able to download the data to their institutional platforms, accessing resources only within the N3C Data Enclave.

2.2 | Additional requirements for linking health care data

Beyond these general governance agreements, additional provisions are needed when conducting research using linked patient data. As mentioned above, linking patients' EHR data with other datasets such as imaging, genomic data, viral variants, and information related to SDOH is important because it can yield rich findings that support high-impact decision-making in health policy and clinical care. Secure

and robust linkages of health care data, however, face two fundamental types of challenges.

First, health care data itself is fragmented.[11-13] Each time a patient visits a hospital, health system, clinic, pharmacy, long-term care facility, or public health agency, new information is created. Unfortunately, this information is stored in many data repositories without a single unique identifier, meaning clinicians and scientists cannot easily create a connected, complete record of each patient's information. Without a complete record, physicians (and in the case of research, data scientists) cannot see the full extent of the care received by patients, patient safety risks increase, public health reporting is weakened, and patient information for research is limited. Therefore, accurate data linkage, which is defined as identifying records for the same person across separate datasets, is necessary to deliver safe and effective health care and to realize the nation's cost and quality improvement goals.

The second challenge involves the identifiability of patient data used for matching. Most prior research on patient data linkage has been conducted on identified datasets, which contain information that ideally uniquely identifies individual patients. However, to minimize privacy risks, growing numbers of researchers need to link *deidentified* data, which lacks patient identifiers. In cases where business associate relationships do not exist, federal law prevents sharing of non-consented Personally Identifiable Information (PII) and Protected Health Information (PHI). Further, local regulations, business processes, and social expectations often limit the use of identified data. Linking data using only deidentified data provides optimal protection of patients' privacy, which is of paramount concern of N3C.

2.3 | Linkage honest broker

Trust and flexibility are essential for these linkages to occur. Trust is needed because data providers and users must have high levels of confidence that the system is reliable, effective, safe, and secure. Flexibility is necessary because the linkage system must accommodate various use cases (designing a method for each use would be unwieldly, inefficient, and cost-prohibitive). To meet the trust and flexibility criteria, we established the role of the LHB, an independent neutral party serving as an intermediary to facilitate data linkages between data contributors and data users. Unlike other types of honest brokers that process PHI by removing identifying information to generate deidentified or limited datasets, the LHB for the N3C Data Enclave does not receive, store, process, or access patients' PHI. Rather, PHI is held only by data-contributing sites, which process their identified data using PPRL software to generate limited datasets for ingestion into the N3C enclave. The LHB's overall purpose is to enable data linkages and enforce appropriate access to linkable data.

The Regenstrief Institute services as the LHB for the N3C initiative under contract with NCATS. N3C participants agree to allow their data to be linked by signing the LHB Agreement. This three-way agreement is between the data contributor, NCATS as steward of the N3C data enclave, and Regenstrief as the LHB using Datavant software (with Datavant PPRL licenses provided to data-contributing sites). While the system leverage is Datavant software, the LHB architecture design is vendor neutral, enabling other encryption and hash-generating software tools to be used.

The LHB Agreement includes a set of principles that unambiguously define the relationship among NCATS, the LHB, PPRL software vendor, and participating sites. First, participation in data linkages among N3C Data Enclave facilitated via PPRL is voluntary. Second, per existing procedures, sites must have a signed DTA to transfer their EHR data (in the form of a Limited Data Set) to the N3C Data Enclave. Third, only the participating sites (not the LHB or NCATS) have access to and control of actual patient identifiers (PHI and personally identifying information, PII). Fourth, participation is not an all-or-none proposition for linkage activities. Institutions are required to permit deduplication of redundant patient records for the accuracy of counts and prevalence information but can set data use preferences for other types of data linkages. Fifth, participation is controlled and predetermined by the participating sites, and sites may discontinue participation in the LHB Agreement at any time. However, if an investigator is actively using data for linkage at the time of discontinuance, that investigator will be allowed to complete the work. The N3C Governance Committee, made up of representatives from NCATS and Regenstrief as the LHB, meets weekly to address issues, review policy, and establish standard operating procedures on topics ranging from publications to codes of conduct.

2.4 | N3C privacy-preserving record linkage methods

PPRL methods typically match patient records with the use of "deidentified tokens"—hashed combinations of demographic and identifier fields that define a match across multiple data repositories without exposing patients' clear-text PHI or PII identifiers, and are deidentified based on the Expert Determination Standard of the HIPAA Privacy Rule.[14-17] In the N3C initiative, those clear-text identifiers remain only with the patient's home health care organization; home organizations convert the identifiers to tokens for its own patient data using hashing software and guidelines provided by the LHB and Datavant. Each token consists of multiple features chosen from a list of identifiers and demographic characteristics. Examples of tokens are {last name + first initial of first name+gender+date of birth} and {last name+first name+date of birth+zip5}. In the N3C's PPRL infrastructure, the LHB holds the deidentified tokens from the contributing sites and matches tokens generated across sites to formulate a single Match ID representing records that should be linked for a specific use case.

An important first step in building the N3C PPRL solution was selecting the most effective set of tokens from 35 pre-certified tokens based on Regenstrief, Northwestern, and Datavant researchers' prior work in developing innovative PPRL techniques and experience matching billions of clinical and claims records across a wide variety of use cases and datasets. Using a dataset from the Indiana Network of Patient Care, we tested a combination of tokens assessed in prior

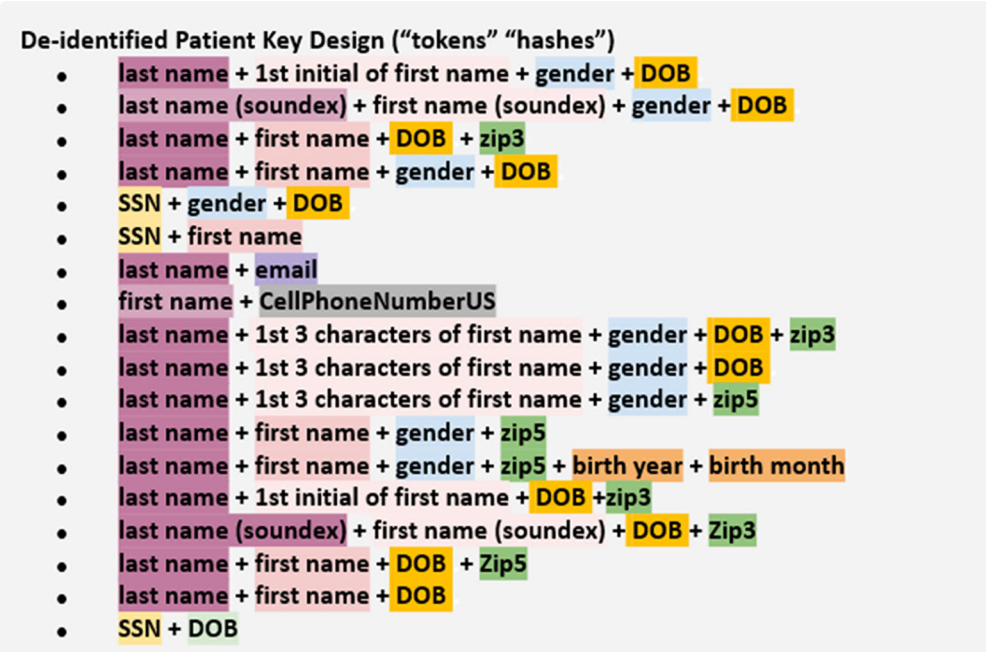

FIGURE 2 Eighteen tokens selected for use with N3C datasets. DOB, date of birth; SSN, social security number; zip3, first 3 numbers of zip code; Zip5, first 5 numbers of zip code.

studies. Out of this available set, we selected those with the best performance and availability that were also scalable and generalizable. After extensive evaluation using all possible permutations, we selected 18 for the N3C project that performed above the others (Figure 2).

In practice, other provisions for securing N3C data linkages will come into play. As needed, Regenstrief as the LHB may utilize tokens and metadata at the request of a participating site and consistent with the N3C enclave rules and policies for possible follow-on clinical research. The LHB will hold certain metadata such as the originating data contributor/data source and the nature of data associated with the received tokens (eg, EHR data, chest x-ray, viral variant data).

The LHB platform and service are designed to streamline interactions between the relevant researcher/requester authentication systems, the N3C Data Enclave, and the ephemeral workbench environments. The LHB platform holds all tokens in its role as a privacy escrow for deidentified, linkable tokens. The platform ingests and processes tokens using the PPRL software, formulating interoperable and linkable tokens. A series of matching algorithms are then applied to link the tokens. The platform generates linkage maps that provide a crosswalk between records that should be deduplicated and linked. Other platforms may access the linkage maps under authorized uses defined by the governance process.

Multiple security assurances are built into the N3C's technical and data governance architecture for the PPRL data. First, tokens reside only with the LHB, while data resides and is unified only in the authorized data enclaves. Second, N3C data and linkable datasets are available for authorized researchers only within the N3C analysis workspace, and the (virtual machine) workbench connecting multiple enclaves and is an extension of the N3C enclave. Third, an authentication and authorization system managed by NIH determines the nature of information that can be shared with the requesting party. Fourth, linked datasets must be used for scientific research only; uses for administrative and performance measurements such as quality or reimbursement are not permitted.

2.5 | Assessment of linkage methodology utility

The LHB platform is now operational. As of March 1, 2023 a total of 43 sites have signed the LHB Agreement, and 35 sites have sent tokens (Table 1). Those tokens have been generated for more than 9 528 998 patients. In this initial cohort, 135 037 matches have been identified, and 68 596 duplicates were identified. Among the duplicate records, 7% are identified within the same institution, while 93% reflect linkages across institutions, most of which are geographically proximate and are thus likely to share patients.

These LHB and PPRL methods have been designed to support three linkage functions: deduplication, linkage of multiple datasets, and cohort discovery. It is with these three forms of linkage that the N3C data becomes most widely useful for research purposes.

2.6 | Deduplication

Deduplication eliminates duplicate or redundant information (1) within a single data source and (2) among two or more datasets. In the first instance, even when a patient's data comes from multiple providers (clinicians, pharmacists, etc.), includes different kinds of data (clinical records, lab results, etc.), and is from different branches of a multisite organization, it is collated within the institution's EHR as a single data source. Multiple datasets, on the other hand, are those generated when separate health care institutions merge or when a unique patient receives care from neighboring but separate institutions that are N3C data-contributing partners (Figure 3). Situations in which a

TABLE 1 Status of Linkage Honest Broker platform (as of March, 2023).

Total Sites Engaged by NCATS	Total Sites Opt-In	LHBA Executed	Institutions Actively Sending Tokens	Number of Token Sets Sent to the LHB	Total Number of Matches
93	69	43	35	9 528 998	135 037

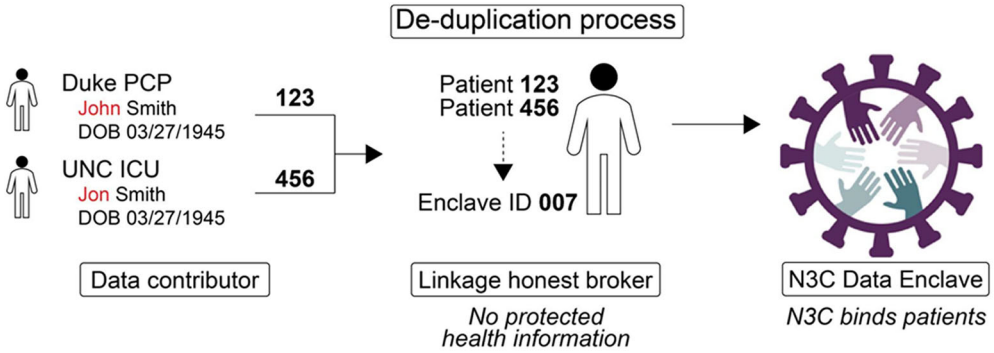

FIGURE 3 Deduplication process. Patients with multiple records, both within a single site and across different sites, are identified and linked.

unique patient has multiple patient records (in either or both single-source and multiple datasets) are common and can happen for a variety of reasons. Typically, duplicates occur when a patient is registered under different names (eg, by different surnames at different stages of life), when misspellings of names lead to multiple registrations, or if multiple institutions where a patient received care combine records.

Because deduplication is essential for the accuracy of counts and prevalence information, all N3C institutions that have signed the DTA and become N3C data contributors are required to participate in the deduplication process with the LHB. In the N3C, linkage with deduplication means grouping data from the medical records of a unique patient who has multiple records and concealing the duplicates so that the single patient is identified only once across all records. Deduplication takes place within the N3C Data Enclave, where the linkage based on PPRL tokens identifies unique individuals with multiple records and uses an additional adjudication process defined and operated by the N3C scientific community to combine information into a unified set of data for the investigator.

2.7 | Linking multiple datasets

A second form of linkage involves combining multiple datasets to fill knowledge gaps, commonly referred to as data enrichment or data augmentation—this involves linking the N3C EHR data with other types of data (eg, mortality) that is stored centrally in the N3C Data Enclave or decentralized in external repositories (eg, imaging). Allowing their contributed data to be used in this way is encouraged but not required for N3C partners.

To enable linkages between two different data types (multimodal data linkage), we developed an external dataset classification, by which all non-N3C EHR datasets are given a classification number (0 to 4) based on their contents (Figure 4). The process of classification, as well as the individual requirements to use the datasets, were designed to ensure that patient privacy was preserved after linkage. The PPRL process applies *only* to datasets in class 0 and class 2, as linkages to publicly available external datasets (classes 3 and 4) do not require PPRL. The difference between classes 0 and 2 is that class 0 datasets originate from different enclaves and allow for a temporary extension of the N3C enclave to accommodate this requirement.

Currently, the N3C initiative has successfully implemented tokenized linkages between the N3C EHR data and mortality, viral variant sequences, and viral variant data, as well as imaging data stored in two external repositories of special relevance for a pulmonary illness like COVID. One of these imaging linkages is with The Cancer Imaging Archive (TCIA) is a large, open-source archive of oncology medical images funded by the Cancer Imaging Program, a part of the US National Cancer Institute (NCI) and managed by the Frederick National Laboratory for Cancer Research.[18]

2.8 | Cohort discovery

The final type of linkage is cohort discovery. Like data enrichment linkages, an N3C data-contributing site may opt-in to allow its data to be used to discover cohorts of its patients meeting criteria for an observational study or clinical trial. Using PPRL for cohort discovery is like the process in previously established networks like TriNetX, PCORNet, or Accrual to Clinical Trials (ACT). Before initiating cohort discovery with local contributing sites, researchers with access to N3C use this form of linkage to determine if a sufficient population exists to power their proposed study (Figure 4). After establishing feasibility, the researchers' organization then works through the LHB to contact sites that indicate interest in potential participation in joint research studies. Interested sites will be contacted by the LHB and

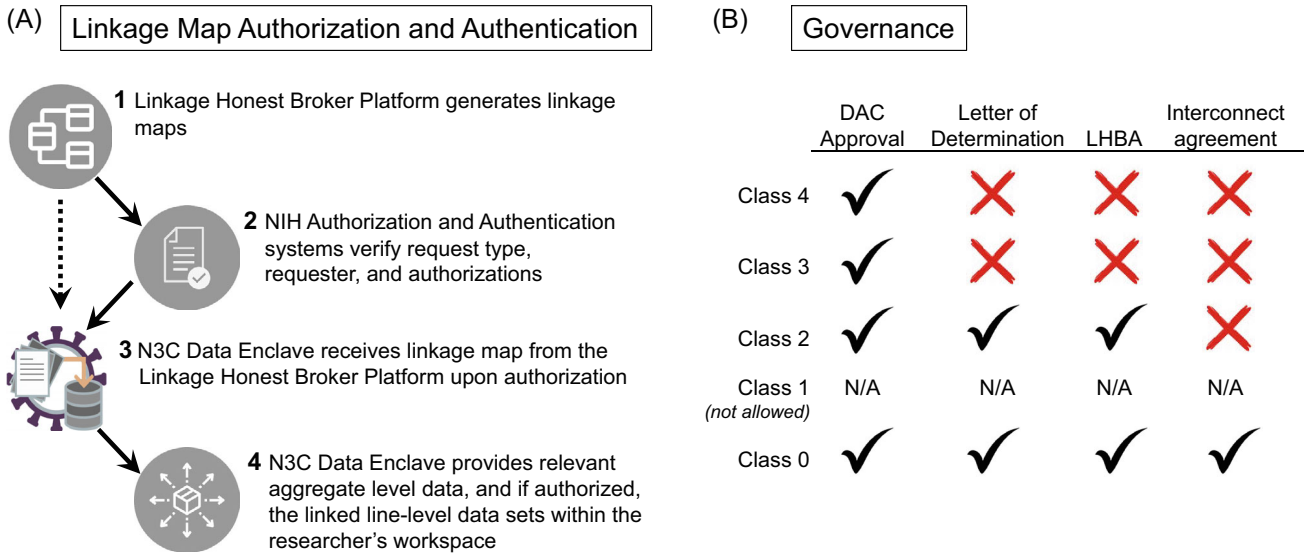

FIGURE 4 Linkage map process (A) and governance classes (B). Dataset classifications are defined as follows: class 0 is linkages using hashed identifiers managed by the third-party LHB to connect multiple enclaves; class 1 is linkages leading to immediate re-identification of patients; class 2 is linkages using hashed identifiers within a single enclave leading to higher-confidence re-identification of patients; class 3 is linkages leading to data sufficiently aggregated to reasonably mitigate the risk of re-identification; and class 4 is linkages or use of data not involving individual persons. DAC, Data Access Committee; LHBA, Linkage Honest Broker Agreement.

given an encrypted list of deidentified Pseudo-IDs. A Pseudo-ID is an ID generated solely for research; it is neither the patient's medical record number nor any other PHI.

It is important that cohort discovery not be confused with patient re-identification. In cohort discovery, the data-contributing sites control all aspects of participation from patient identification and methods of patient contact to consenting. Only the data-contributing sites can map back to the list of deidentified Pseudo-IDs and re-identify their own patients. The LHB list of deidentified IDs used in cohort discovery does not contain any PHI or PII.

3 | DISCUSSION

This project contributes to the growing body of literature on health record linkage research using RWD. Whereas significant earlier research on record linkage used identified data,[19-21] our project demonstrates the viability of PPRL methods on deidentified data. Our study, though not the largest, contributes significant evidence for the viability of deidentified matching systems, adding to the limited number of national-scale PPRL linkage studies using RWD datasets with varying characteristics.[15,22,23] This is important since multisource and multimodal data enable more comprehensive findings, but institutions typically mask patient identity in datasets to be shared, making it hard to produce those data integrations. Our study also expands the range and complexity of data being linked for research purposes and supports two core linkage functions beyond the more commonly used deduplication. Prior nationwide studies have had benefits ours lacked: being conducted, for example, within a national health service (Scotland) in which all patients have a unique identifier across datasets[24] or in a country (Switzerland) where patients agreed to use their social security numbers in record linkages.[25] Other studies have used PPRL to link datasets in specific parts of the US: within one county's services, for instance,[26] or in a state-based clinical research network,[27] or focused on a single disease in one state[28] or on deduplication across multiple sites in a large metropolitan area[29] or between national government indices and national surveys.[30] Because our project, with its broad and diverse national scope, has established and validated ways to link EHR data from many healthcare institutions across the United States and to link widely varying datasets (from EHR data to images to population-based data), these PPRL methods hold great potential for use beyond the N3C initiative.

Beyond the project's contribution to knowledge, we have learned valuable lessons from the collaboration. First, our experience shows that shared infrastructure works. The N3C partnership demonstrates the value of careful and thoughtful collaboration between data contributors, data users, institutions, governmental agencies, and investigators. Team science is required for projects like this, and NIH's role as a partner is to facilitate science. Second, common data model harmonization is possible, and sites can safely share data if agreed-upon governance procedures are established and followed. In data governance, trust is key—but trust is earned through the kind of data use and DTAs, LHB platform, and PPRL provisions developed for this project. Third, in studying diseases like COVID-19, improved data quality will enhance harmonization for research purposes. At the same time, EHR data alone is often not sufficient and must be linked to other forms of data to support rigorous and trustworthy scientific conclusions. Fourth, there is an acute shortage of resources available in clinical informatics, so projects like this one are essential to expand data analysis tools available to biomedical researchers. The research workforce needs professional support and ongoing training to take advantage of modern analytics techniques. Fifth, this project affirmed the

need for trust in collaborations like this. For a collaborative community like N3C to be effective requires a shared vision and the trust and freedom for its members to do their work.

This project and its results have some limitations. Although COVID-19 is global in its nature, the datasets in this study are only from the United States, so these approaches may not fully translate to other contexts. Our experience linking N3C and other data from sources across the United States, however, can meaningfully inform any opportunities to link US data with that from other countries. In addition, since the participation of N3C data-contributing sites in the LHB model is voluntary, its continuation into the longer-term future cannot be guaranteed. Transitioning the N3C initiative and data linkages to sustain the collaboration into the later stages of the COVID-19 pandemic and beyond will be needed in a future phase. Finally, the N3C collaborative was designed to address one specific disease and pandemic, so PPRL methods designed for it have so far been tested only for COVID-19 and related data. Future research will explore potential PPRL method optimizations and the application of methods developed in this project to support the creation of collaborative datasets for other diseases and conditions.

4 | CONCLUSION

This project contributes to the ongoing search for ways to protect patient privacy when linking information from different datasets in research to guide public policy and clinical care. As part of the NCATS National COVID Cohort Collaborative (N3C), we have established PPRL systems and procedures to ensure that patient-level data in the N3C Data Enclave remains secure and private when linked with internal and external datasets. The framework includes agreements and approval processes for both data contribution and data access; a LHB serving as a privacy-protecting intermediary between data contributors and data users; and PPRL methods that use "deidentified tokens" to define a match across data repositories without using patients' true identifiers. These methods support three linkage functions needed for research purposes: deduplication, linkage among multiple datasets, and cohort discovery. This data enrichment process helps COVID-19 research benefit from additional information such as mortality data, images, and viral variants and has the potential for future adaptations with registries for other diseases and conditions.

AUTHOR CONTRIBUTIONS

All co-authors substantially contributed to the conception or design of the work to the writing, and/or to significant review of this manuscript, and all who qualify for authorship are listed as co-authors. Dr. Grannis revised the manuscript in response to reviewer comments.

ACKNOWLEDGEMENTS

The LHB is funded by the NCATS/NIH through Contract 75N95021D00028. The authors would like to acknowledge the following individuals and groups for their guidance and support of this manuscript: Dr. Patricia Brennan, Dr. Susan Gregurick, Dr. Lynn Whittaker, The Regenstrief Institute Data Services team, and the team at Pintail Solutions, Inc., and the N3C community. The following NIH Institutes also contributed direct or indirect funding support: NCATS, NLM, OD, NIGMS, and NIAIDS. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

CONFLICT OF INTEREST STATEMENT

Jasmin Phua and Sara Rogovin are employed by Datavant, Inc. and Dr. Abel Kho has a financial interest in Datavant, Inc, a for-profit company that was contracted to provided software, expertise, and services for the creation and operation of the PPRL approach described in this manuscript. Benjamin Amor, Maya Choudhury, Philip Sparks, Amin Mannaa, and Saad Ljazouli are employed by the for-profit Palantir Technologies, which was contracted to provided software, expertise, and services for the creation and operation of the PPRL approach described in this manuscript. At the time that this work was performed, Drs. Peter Embi, Umberto Tachinardi, and Shaun Grannis all worked for and/or served as officers of Regenstrief Institute, Inc. a non-profit research institute contracted by the NIH to manage and oversee the PPRL and LHB work described in this manuscript.

REFERENCES

1. Harris S, Houser SH. Double trouble: using health informatics to tackle duplicate medical record issue. *J AHIMA*. 2018;89(8):20-23.
2. Lippi G, Mattiuzzi C, Bovo C, Favaloro EJ. Managing the patient identification crisis in healthcare and laboratory medicine. *Clin Biochem*. 2017;50(10–11):562-567.
3. Sragow HM, Bidell E, Mager D, Grannis S. Universal patient identifier and interoperability for detection of serious drug interactions: retrospective study. *JMIR Med Inform*. 2020;8(11):e23353.
4. Fahr P, Buchanan J, Wordsworth S. A review of the challenges of using biomedical big data for economic evaluations of precision medicine. *Appl Health Econ Health Policy*. 2019;17(4):443-452. doi:10.1007/s40258-019-00474-7
5. Haendel MA, Chute CG, Bennett TD, et al. The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28(3):427-443.
6. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep*. 2014;129(Suppl 2):19-31. doi:10.1177/00333549141291S206
7. N3C Dataset Registry. https://discovery.biothings.io/dataset?guide=/guide/n3c/dataset. Accessed November 19, 2021.
8. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc AMIA Symp*. 2002; 305. PMID: 12463836; PMCID: PMC2244404-309.
9. Schadow G, Grannis S, McDonald C. Privacy-preserving distributed queries for a clinical case research network. *CRPIT 14: Proceedings of the IEEE International Conference on Privacy, Security and Data Mining—Volume 14*. Maebashi City, Japan: Australian Computer Society; 2002: 55-65. doi:10.5555/850782
10. Pfaff ER, Girvin AT, Gabriel DL, et al. Synergies between centralized and federated approaches to data quality: a report from the National COVID Cohort Collaborative. *J Am Med Inform Assoc*. 2022;29(4): 609-618. doi:10.1093/jamia/ocab217

11. Mays JA, Jackson KL, Derby TA, et al. An evaluation of recurrent diabetic ketoacidosis, fragmentation of care, and mortality across Chicago, Illinois. *Diabetes Care*. 2016;39(10):1671-1676.
12. Walunas TL, Jackson KL, Chung AH, et al. Disease outcomes and care fragmentation among patients with systemic lupus erythematosus. *Arthritis Care Res*. 2017;69(9):1369-1376.
13. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. *AMIA Annu Symp Proc*. 2011;2011:409-416.
14. Trick WE, Hill JC, Toepfer P, Rachman F, Horwitz B, Kho A. Joining health care and homeless data systems using privacy-preserving record-linkage software. *Am J Public Health*. 2021;111(8):1400-1403.
15. Kho AN, Yu J, Bryan MS, et al. Privacy-preserving record linkage to identify fragmented electronic medical Records in the all of us research program. In: Cellier P, Driessens K, eds. *Machine Learning and Knowledge Discovery in Databases*. Cham: Springer International Publishing; 2020:79-87 (Communications in Computer and Information Science).
16. Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-preserving record linkage for big data: current approaches and research challenges. In: Zomaya AY, Sakr S, eds. *Handbook of Big Data Technologies*. Cham: Springer; 2017. doi:10.1007/978-3-319-49340-4_25
17. Office of Civil Rights, U.S. Department of Health and Human Services. Guidance regarding methods of deidentification of protected health information in accordance with the health insurance portability and accountability act (HIPAA). *Privacy Rule*. 2012. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed January 4, 2024.
18. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045-1057.
19. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol*. 2002;31(6):1246-1252.
20. Lui S. Development of record linkage of hospital discharge data for the study of neonatal readmission. *Chronic Dis Can*. 1999;20(2):77-81.
21. Daggy JK, Xu H, Hui SL, Gamache RE, Grannis SJ. A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Med Inform Decis Mak*. 2013;13:97.
22. Kiernan D, Carton T, Toh S, Phua J, Zirkle M, Louzao D, Haynes K, Weiner M, Angulo F, Bailey C, Bian J, Fort D, Grannis S, Krishnamurthy AK, Nair V, Rivera P, Silverstein J, Marsolo K Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet, the National Patient-Centered Clinical Research Network. *BMC Res Notes* 2022 31;15(1):337.
23. Marsolo K, Kiernan D, Toh S, et al. Assessing the impact of privacy-preserving record linkage on record overlap and patient demographic and clinical characteristics in PCORnet®, the National Patient-Centered Clinical Research Network. *J Am Med Inform Assoc*. 2023;30(3):447-455.
24. Hanna CR, Lemmon E, Ennis H, et al. Creation of the first national linked colorectal cancer dataset in Scotland: prospects for future research and a reflection on lessons learned. *Int J Popul Data Sci*. 2021;6(1):1654. doi:10.23889/ijpds.v6i1.1654
25. Branca-Dragan S, Koller MT, Danuser B, Kunz R, Steiger J, Hug BL. Evolution of disability pension after renal transplantation: methods and results of a database linkage study of the Swiss transplant cohort study and Swiss disability insurance. *Swiss Med Wkly*. 2021 Sep;23(151):w30027. doi:10.4414/smw.2021.w30027
26. Trick WE, Rachman F, Hinami K, et al. Variability in comorbidities and health services use across homeless typologies: multicenter data linkage between healthcare and homeless systems. *BMC Public Health*. 2021;21(1):917. doi:10.1186/s12889-021-10958-8
27. Bian J, Loiacono A, Sura A, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. *JAMIA Open*. 2019;2(4):562-569. doi:10.1093/jamiaopen/ooz050
28. Olatosi B, Zhang J, Weissman S, Hu J, Haider MR, Li X. Using big data analytics to improve HIV medical care utilisation in South Carolina: a study protocol. *BMJ Open*. 2019;9(7):e027688. doi:10.1136/bmjopen-2018-027688
29. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *J Am Med Inform Assoc*. 2015;22(5):1072-1080. doi:10.1093/jamia/ocv038
30. Mirel LB, Resnick DM, Aram J, Cox CS. A methodological assessment of privacy preserving record linkage using survey and administrative data. *Stat J IAOS*. 2022;38(2):413-421. doi:10.3233/sji-210891

How to cite this article: Tachinardi U, Grannis SJ, Michael SG, et al. Privacy-preserving record linkage across disparate institutions and datasets to enable a learning health system: The national COVID cohort collaborative (N3C) experience. *Learn Health Sys*. 2024;8(1):e10404. doi:10.1002/lrh2.10404