N3C Privacy-Preserving Record Linkage and Linked Data Governance

Title: **N3C Privacy-Preserving Record Linkage and Linked Data Governance**	
Version No: 1.0	**Effective Date: 2021-08-04**
Point of Contact	Ken Gersing

Version history

Version	Description of changes
All versions	https://doi.org/10.5281/zenodo.5165212 This DOI represents all versions, and will always resolve to the latest one.
1.0 2021-08-04	Final Draft

Table of Contents

Background

This document is meant to give data participating sites greater visibility into and context around the types of linkages also known as **Privacy Preserving Record Linkage (PPRL)** (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), that will occur and how linked data will be used downstream. **Linkage** is defined here as any operation involving two or more datasets using cryptographic hash codes (tokens) as a way to match records associated with the same individuals anonymously, without ever using the individual true PII/PHI identifiers**.** Participating sites execute the linkage tool locally, PII is never sent to the honest broker and PII cannot be extracted from these hash codes, as they are strictly one-way hash algorithms.

Who must read this document: N3C data contributing site PIs interested in participating in Privacy Preserving Record Linkage (PPRL), the participating site's institutional signing official and other relevant decision makers for participation in PPRL linkage.

Who should read this document: All parties interested in the activities relevant to records linkage and records deduplication.

The Linkage Honest Broker Agreement does not change the terms of the N3C Data Transfer Agreement (DTA), Data Use Agreement (DUA) or the User Code of Conduct (CoC). It is a 3-way agreement that is intended for execution between N3C participating sites, NCATS as steward of the enclave securing the data and protecting privacy, and the Regenstrief Institute in its role as the Linkage Honest Broker using software vendor Datavant. A Linkage Honest Broker in the NCATS PPRL's infrastructure is a party that holds cryptographic hash codes (tokens) and operates a service that matches tokens generated across disparate datasets to formulate a single Match ID representing records that should be linked for a specific use case.

The concepts put forth within the Linkage Honest Broker Agreement outlines 3 types of data linkages (**deduplication, linking multiple-datasets, cohort discovery**) which require further data governance context to support N3C participating sites' decision making.

N3C Privacy Preserving Record Linkage Principles

The NCATS' Linkage Honest Broker Agreement includes a set of principles that unambiguously define the relationship between NCATS, Honest Data Broker, and Participating Sites.

N3C Participating Sites to the N3C Data Enclave

- Participation in the PPRL is voluntary
- Per existing procedures, Participating Sites must also have a signed Data Transfer Agreement (DTA) to transfer their electronic health record data (EHR) (in the form of a Limited Data Set) to the N3C Data Enclave

- Only the participating sites have access to and control of Personal Health Information and Personally Identifying Information (PHI/PII) (actual identifiers)
- Participation is not an all or none proposition for all linkage activities. Participation in PPRL obligates participating institutions to deduplication of duplicate records and is necessary for accuracy of counts and prevalence information. Linkage of multiple datasets and cohort discovery are optional activities.
- Participation is controlled (see below) and pre-determined by the participating sites.
- Participating sites may discontinue participation in the Linkage Honest Broker Agreement at any time (see below). However, if an investigator is actively using data for linkage at the time of discontinuance, that investigator will be allowed to complete their work.

Linkage Honest Broker (LHB)

- The Linkage Honest Broker for N3C is the Regenstrief Institute. The LHB is a neutral entity located outside of the N3C Data Enclave that serves as an escrow for the cryptographic hash codes (tokens), and operates the technology platform which facilitates PPRL using these tokens.
- The LHB does **NOT** receive, store, or process PHI/PII or clinical information. As aforementioned the PHI/PII is ONLY held by the data participating sites. For the avoidance of doubt, Regenstrief may utilize tokens and metadata at the request of a Participating Site and consistent with the NCATS N3C Data 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, e.g., EHR data, chest x-ray, viral variant data. The LHB is contracted by NCATS.

Exclusions of use:

- Linked datasets will be used for scientific research only. Uses for administrative and performance measurements/assessments are not permitted.
- Data participating sites within N3C are providing clinical data to advance research on SARS-CoV-2 and COVID-19. Data cannot be used for administrative or performance metrics such as quality, reimbursement, and medical errors.

PPRL Participation Options

Participation in PPRL is voluntary. Once participation in PPRL is active, participating sites have the option of choosing the type of data linkage they want to participate in. Participation in the LHBA is not an all or none proposition. Linkage to multiple datasets and for cohort discovery are optional. However, data participating sites that sign the LHBA are agreeing, at minimum, to participate in records deduplication. Participating sites at any time can change their participation in linkage of multiple datasets and/or cohort discovery by going to a secure website and setting institutional parameters. (**see Image: PPRL Options matrix**). If a data contributor discontinues participation no

new linkages will be allowed; however, any ongoing work will be allowed to continue to completion of the project described in the data use request.

Image: PPRL Options matrix. Signing the agreement obligates the participating site to records deduplication

Deduplication (required):

- Signing the LHB Agreement allows for patient deduplication in N3C

Linkage (optional):

- Sites control which types of data can be linked to the site data e.g., EHR and Imaging
- All linkage studies require local IRB determination

Cohort Discovery (optional):

- **Agree to point of contact being informed of potential clinical trial studies eligible for participation**
- Contact is always initiated by the Linkage Honest Broker platform

Category	Site 1	Site 2	Site 3	Site 4
Deduplication	Yes	Yes	Yes	Yes
Dataset Linkage	No	Yes	Yes	No
Cohort Discovery	Yes	Yes	No	Yes

Deduplication

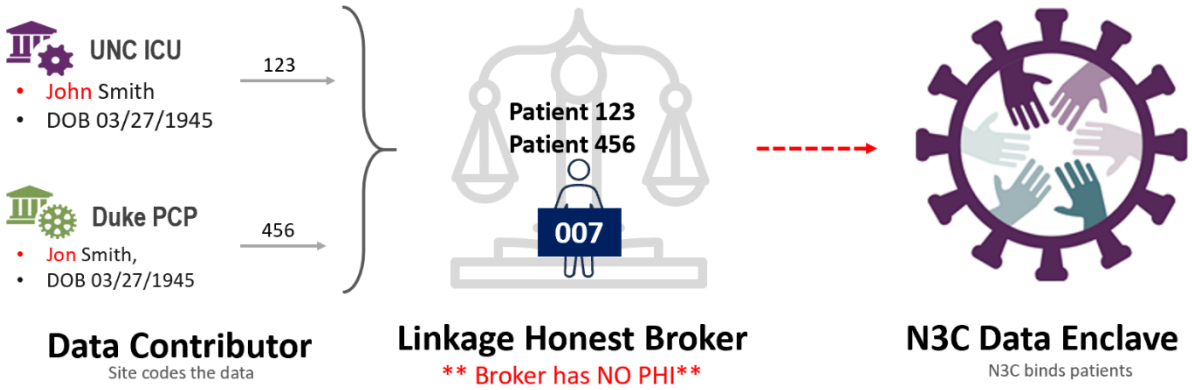
"**Deduplication**" means eliminating duplicate or redundant information within a data source and among two or more datasets. In the context of N3C, linkage with deduplication means combining the data from medical records of a unique patient that for some reason has repeated records (duplicates). Multiple records for a unique patient is surprisingly common and can happen for a variety of reasons such as being registered under different names (maiden and married surname), multiple registrations caused by misspellings, or the combining of multiple institutions records where the same patient may have gotten care. Deduplication takes place within the N3C Data Enclave (**see Image: Illustrates deduplication process from multiple institutions**), identifies data records associated with unique individuals with multiple records and allows an N3C Data user (investigator) to adjust for counts in their analysis.

Deduplication is a requirement for any institution that participates in the LHBA, because of its importance to the data quality of the N3C Data Enclave and its scientific mission.

Deduplication requires a data contributing site to:

1. Apply the cryptographic hash code (token) to relevant patient records.
2. Send the cryptographic hash code (token) with metadata to the linkage honest broker.
3. Include the cryptographic hash code (token) as part of the N3C data transfer payload.
4. Allow N3C to use the cryptographic hash code (token) to deduplicate redundant information.

Image: Illustrates deduplication process from multiple institutions



Linking Multiple Datasets

Though there are many types of datasets and ways to link to them, the Linkage Honest Broker Agreement applies only to datasets that are within N3CthN3C Data Enclave and requires linkage using the hash/PPRL. N3C has developed an external dataset classification system (See description below Multi-Dataset Linkage Classification) the LHBA only applies to datasets classified as class “0” and Class “2”. Linkages to external datasets that do not require the hash or PPRL are not covered by this agreement. The difference between Class 0 datasets and Class 2 datasets is Class 0 datasets originate from different enclaves and allows for a temporary extension of the N3C Data Enclave to accommodate this requirement. If additional computational resources are required for large datasets, the N3C Data Enclave will utilize NCATS High PC Performance Computing (HPC) services for data processing.

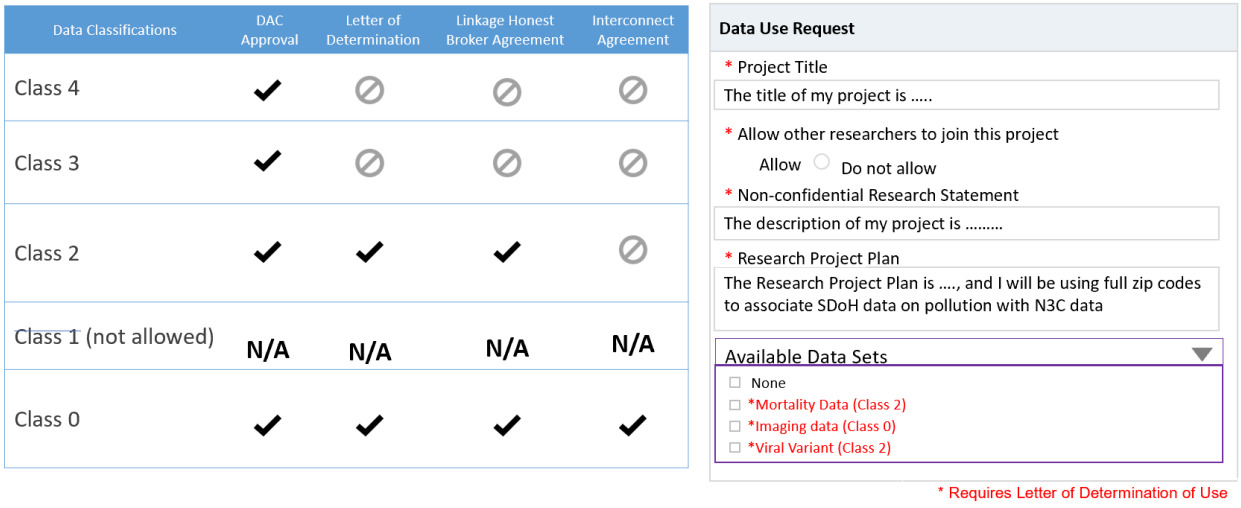
Multi-Dataset Linkage Classification Summary

- **Class 0**: Linkages using cryptographic hash codes (tokens) managed by a third-party linkage honest broker to connect multiple Enclaves.
- **Class 1**: Linkages leading to immediate re-identification of patients and is not permitted with the N3C
- **Class 2**: Linkages using cryptographic hash codes (tokens) within a single enclave leading to higher confidence re-identification of patients.
- **Class 3**: Linkages leading to data sufficiently aggregated to reasonably mitigate the risk of re-identification.

- **Class 4**: Linkages or use of data not involving individual persons.

Access to various classes of data require sets of agreements (**see Image: Proposed requirements for use of multiple datasets in research**) to be in place for the data requesters.
All datasets that use the PPRL (Class 0 and 2) are considered level 3 data and as such must work through their institutional policies and require a letter of determination when submitting an N3C Data Use Request (DUR). Additional information on terms of use for available linkable datasets can be found at https://discovery.biothings.io/dataset?guide=/guide/n3c/dataset .

Image: Proposed requirements for use of multiple datasets in research



Data Classifications	DAC Approval	Letter of Determination	Linkage Honest Broker Agreement	Interconnect Agreement
Class 4	✔	⊘	⊘	⊘
Class 3	✔	⊘	⊘	⊘
Class 2	✔	✔	✔	⊘
~~Class 1~~ (not allowed)	N/A	N/A	N/A	N/A
Class 0	✔	✔	✔	✔

Class 2 dataset linkages require existing institutional N3C Data Use Agreement, Dual authentication and authorization, a signed institutional linkage honest broker agreement for multiple datasets, an approved data use request (DUR) by the federally staffed data access committee (DAC), and local institutions IRB letter of determination. IRBs must clearly have reviewed the DURs proposed protocol and the specific use of multiple datasets beyond N3C EHR-derived data. Class 2 dataset linkage are contained within the single N3C Data Enclave. An example of a class 2 multiple linkage datasets would be if N3C data is linked to Mortality data that was sent to N3C.

For class 0 dataset linkages, that connect more than one enclave, an additional interconnect agreement will be in place. The interconnect agreement will be agreement between two trusted enclaves in order to instantiate what is referred to as a temporary virtual or ephemeral workbench. The workbench is ephemeral because it is short-lived for a specific task and then destroyed when an investigator's work is completed.

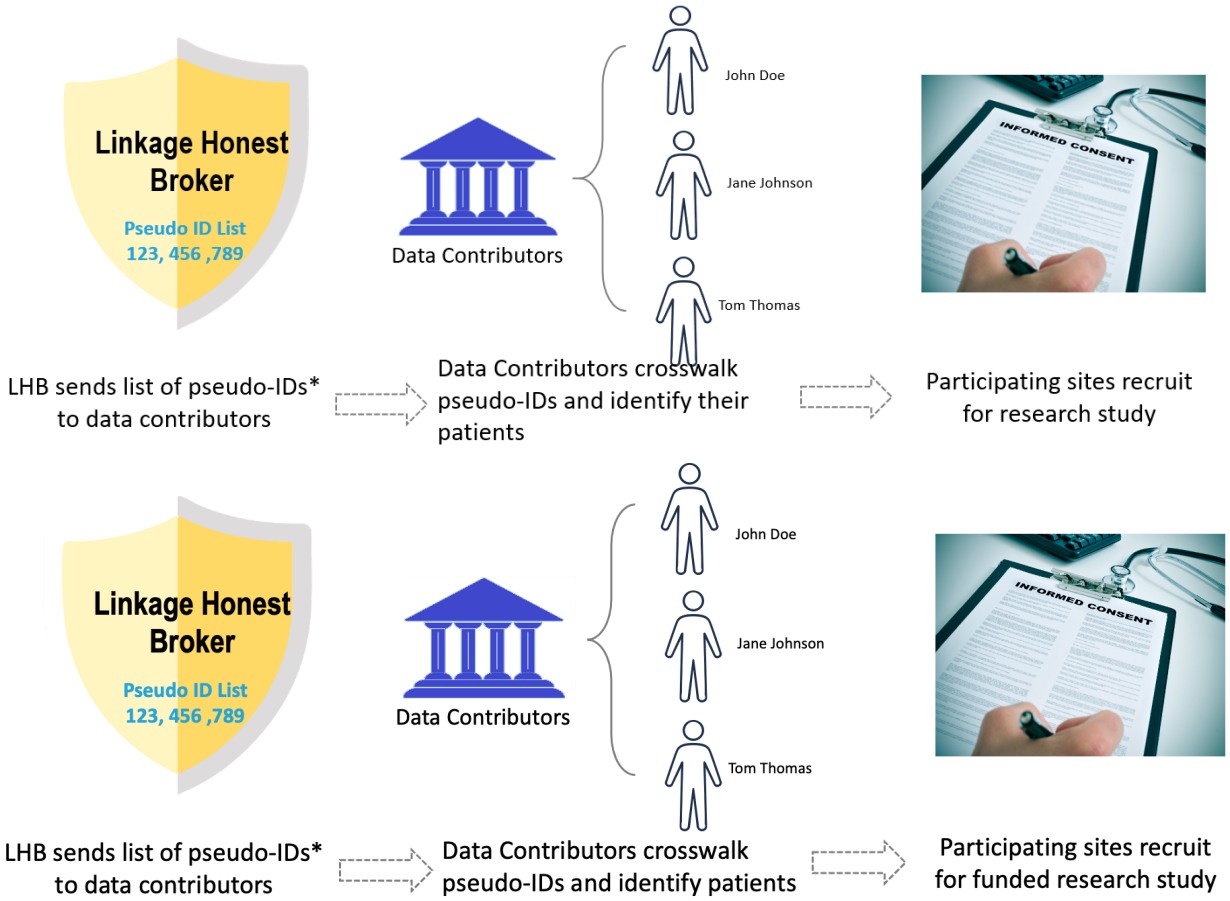
Classes 3 and 4 require a DUR for the study approved by the DAC.

Cohort Discovery for Research Studies

The third and final type of linkage is cohort discovery. Cohort Discovery is a web-based tool that allows researchers to discover research-specific population cohorts across multiple linked datasets within N3C. Cohort discovery is a common and familiar process to many data participating sites that participate in networks like TriNetX or Accrual to Clinical Trials (ACT) activity. In the ideal world, prior to asking sites to do cohort discovery for a clinical study, an organization has done feasibility research that confirms there is a large enough population exists to power a study. Once all of feasibility research is determined only then are sites contacted with potential list of de-identified patient keys sent to a data contributor for cohort discovery by the LHB. Interested sites can decide to participate or not in any particular study. Only the participating sites are technically able to re-identify patients. If a site chooses to participate in any given study, the recruitment process follows local institutional policy and procedures on contacting patients and consenting them for a study (**see Image: Cohort Discovery Prospective Study Process**). Cohort discovery is key to use-cases that require recruitment of patients from a de-identified (but PPRL-ready) cohort or when certain prospective data or local data augmentations is required. Those use-cases should follow both N3C and local institutional governance and be reviewed and approved by both IRBs of record.

It is very important that cohort discovery not be conflated with patient re-identification. In cohort discovery ONLY participating sites that have signed the LHBA and opted for cohort discovery can re-identify their own patients. The linkage honest broker list of de-identified patient keys does not have any Personally Identifying Information, (PII).

Image: Cohort Discovery Prospective Study Process



For cohort discovery, the requirements are as follows:

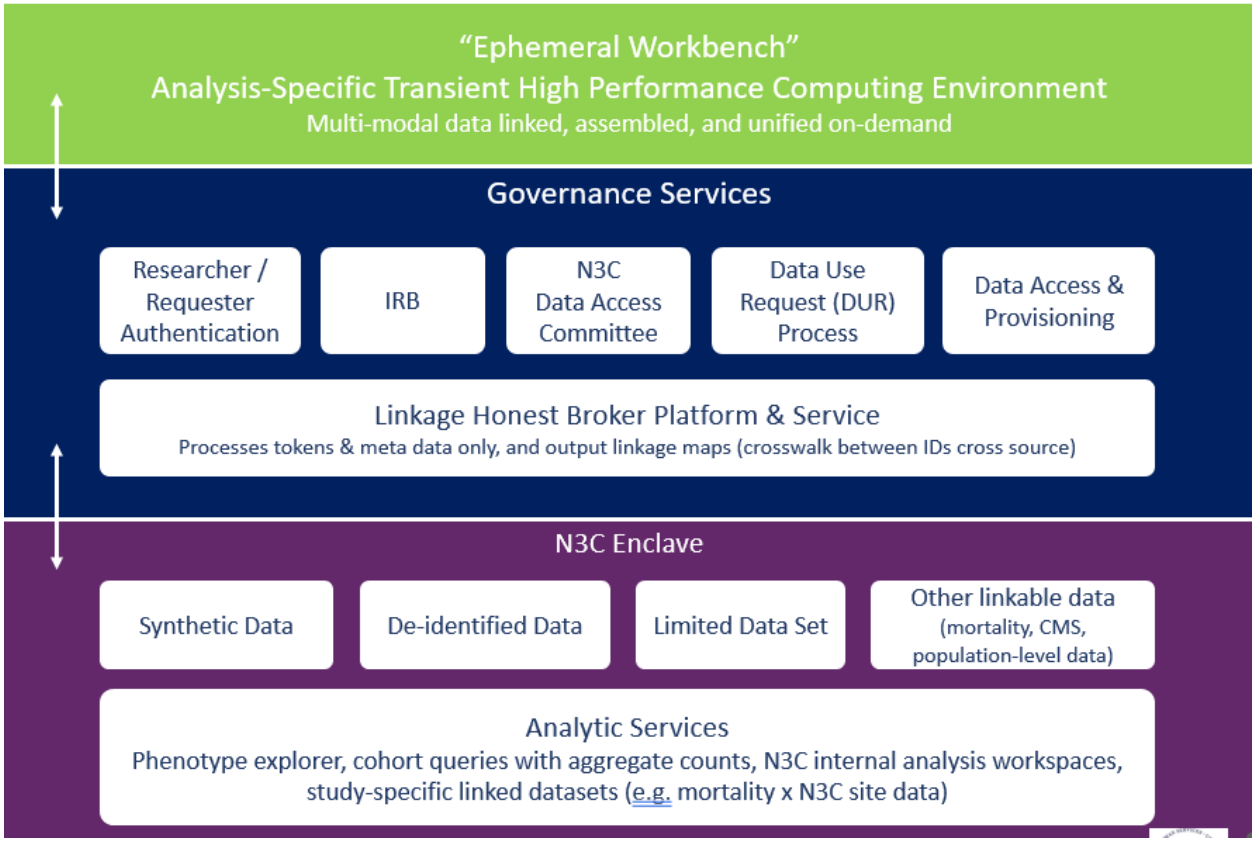
- Tokens from N3C participating sites will be linkable to other linkable datasets for the purpose of study feasibility assessments.
- Research feasibility requests may be initiated by authorized researchers, and NIH for its operational purposes.
- Compliance with N3C download policy or cell sizes < 20 and data release policies for aggregate counts, which are considered will apply.
- The Linkage Honest Broker results will only include aggregate counts and will not include row level data or participating site information.

Technical and Data Governance Architecture for N3C PPRL Linked Data Infrastructure

There are several considerations related to the technical and data governance architecture for the N3C PPRL linked data infrastructure:

1. Tokens only reside with the Linkage Honest Broker
2. Data resides and is unified in the authorized data enclaves: 1) N3C data and linkable datasets that will be available within the N3C environment will be available for authorized researchers within the N3C analysis workspace, 2) the ephemeral (Virtual Machine, (VM)) workbench connecting multiple enclaves is an extension of the N3C Data Enclave (**see Image: Multiple enclave process**), where datasets will be unified based on data governance approvals.
3. The Linkage Honest Broker platform produces (or will produce) a linkage dashboard depicting the linkages in records between disparate datasets.
4. An authentication and authorization system managed by the NIH will determine the nature of information that can be shared with the requesting party.

Image: Multiple enclave process. Note: **De-identified data** refers to level 2 requested access, where 17 of 18 HIPAA identifiers have been removed; longitudinal data are data-shifted to protect individual privacy. The **Limited Data Set (LDS)** available from N3C consists of health information from individuals who have received a COVID-19 test or whose symptoms are consistent with COVID-19. Data will also be collected from individuals infected with pathogens such as SARS 1, MERS, and H1N1 to support comparative studies. 16 of 18 HIPAA identifiers have been removed; data retain dates and zip codes.

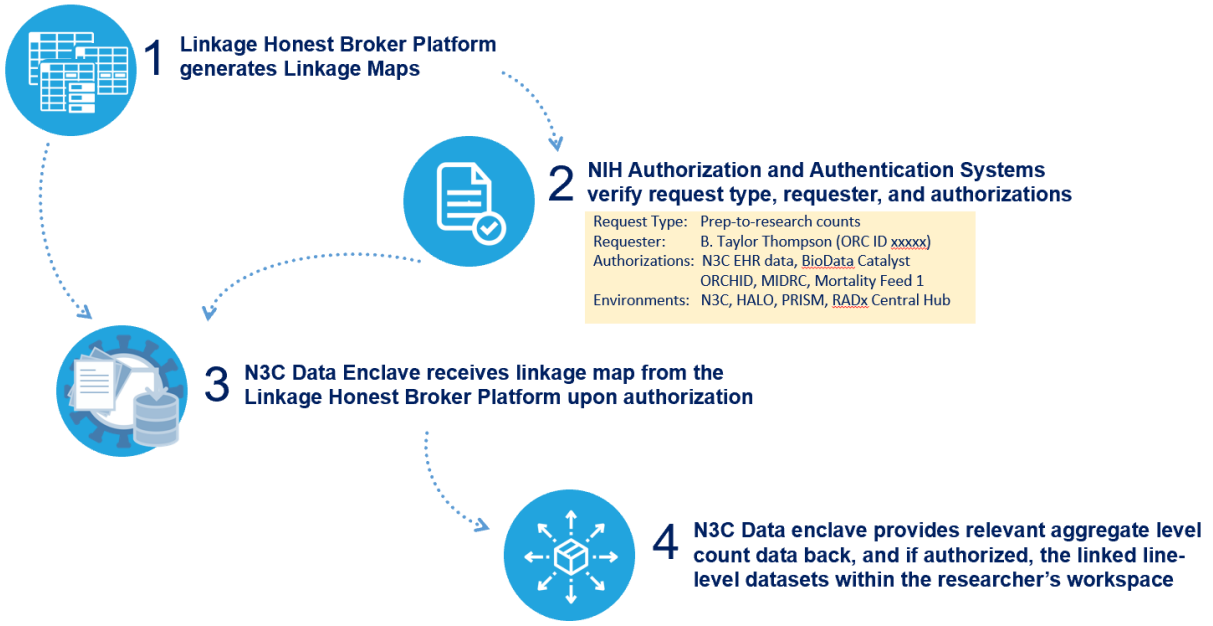


Linkage Honest Broker Platform and Service

The Linkage Honest Broker platform and service will interface with the relevant Researcher/Requester Authentication Systems, the N3C Data Enclave, and the ephemeral workbench environments.

1. NIH staff and registered N3C Data Enclave investigators will have access to the Linkage Honest Broker platform for an aggregate-level view of overlaps between disparate datasets and repositories.
2. The Linkage Honest Broker platform will hold all tokens centrally in its role as a privacy escrow for de-identified, linkable tokens.
3. The platform will ingest and run linkages on all tokens held centrally and sent to it by various participating sites and repositories. The platform will generate linkage maps that can only be accessed by other platforms based on data governance authorizations. See illustration below for the interfaces between the Linkage Honest Broker platform and the N3C Data Enclave.

Image: Interfaces between the Linkage Honest Broker platform and the N3C Data Enclave.


1 Linkage Honest Broker Platform generates Linkage Maps
2 NIH Authorization and Authentication Systems verify request type, requester, and authorizations
Request Type: Prep-to-research counts
Requester: B. Taylor Thompson (ORC ID xxxxx)
Authorizations: N3C EHR data, BioData Catalyst ORCHID, MIDRC, Mortality Feed 1
Environments: N3C, HALO, PRISM, RADx Central Hub
3 N3C Data Enclave receives linkage map from the Linkage Honest Broker Platform upon authorization
4 N3C Data enclave provides relevant aggregate level count data back, and if authorized, the linked line-level datasets within the researcher's workspace

Appendix A Glossary

- "**Cohort Discovery**" means the process for enabling authorized researchers to query the N3C Data Enclave for data records associated with de-identified individuals and meet specified inclusion and exclusion criteria.
- "**Deduplication**" means eliminating duplicate or redundant information within a data source.
- **De-identified Patient Keys** Encrypted strings that are processed through a cryptographic method called "hashing", with the resulting output referred to **hashes;** these are referred to as **tokens** more generally. Datavant de-identified patient keys are certified as de-identified per the HIPAA Privacy Rule. Note: Not all hashes and hashing are considered de-identified per HIPAA.
- "**Health Data**" means the Participating Site's information related to an individual's medical history, including but not limited to structured information such as demographics, vital signs, diagnoses, procedures, admission, discharge and transfer information and semi-structured information, including laboratory tests and results, medications, imaging, waveform, variants etc. Health Data that includes real dates and zip codes is a Limited Data Set as defined herein. Health Data is referenced in the Data Transfer Agreement as Data.
- **Investigator Data Access:** Investigators that meet requirement to use different types of data must include but not limited to a data use request, letter of determination, Data Access Committee approve and interconnect agreement.
- **Interconnect Agreement, (ICA) sharing agreements between enclaves: Enclave entities like** NIBIB Medical Imaging and Data Resource Center (MIDRC), and NCATS N3C.
- **Linked ID** When the linkage honest broker generates a link between two cryptographic hash codes (tokens), they also generate a new random ID that corresponds to the linkage itself; this is the Linked ID. Using unique Linked IDs ensures that the linkages can be used only where data governance restrictions allow.
- "**Limited Data Set**" is The Limited Data Set (LDS) available from N3C consists of health information from individuals who have received a COVID-19 test or whose symptoms are consistent with COVID-19. Data will also be collected from individuals infected with pathogens such as SARS 1, MERS, and H1N1 to support comparative studies. 16 of 18 HIPAA identifiers have been removed; Data retain dates and zip codes. of service and zip codes. "**Metadata**" means a set of data that provides a structural or administrative description about the Participating Site's Health Data. Metadata does not include Health Data.
- "**Privacy Preserving Record Linkage (PPRL)** or **Record Linkage**" 1) means 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. 2) Method to generate de-identified patient keys ("hashes") that enable data likability
- **Pseudo ID** Originating from an institution or source system, these are randomly generated IDs that accompany the de-identified patient keys / hashes when sent to the third-party linkage honest broker.

- "**Research**" means a systematic investigation, including research development, testing, and evaluation designed to develop or contribute to generalizable knowledge.
- **Site Permission for Linkage:** Can a site's data be included in linkage studies.
- "**Token**" or "**Hash**" mean an encrypted value created by an irreversible conversion algorithm and any underlying Protected Health Information that has been de-identified using the expert determination method as described under HIPAA regulations at 45 CFR 164.515(b)(1).
- **Authentication**: the act of proving an assertion, such as the identity of a computer system user.
- **Authorization:** the function of specifying access rights/privileges to resources

Appendix B Linkage Honest Broker Agreement

LINKAGE HONEST DATA BROKER AGREEMENT

https://ncats.nih.gov/files/NCATS_LHBA-508.pdf

Incorporated by reference

Appendix C Regulatory Considerations for Privacy-Preserving Record Linkage

NOTE: Covered entity in the document below refers to data participating sites.

Privacy Preserving Record Linkage De-identification and Re-identification Regulatory Considerations

V1.0, 16-Dec-2020, jas@datavant.com

Datavant tokens (also known as keyed secure hashes) are de-identified universal patient keys that are used to link records across datasets in a de-identified manner.

Datavant Tokens Are a 'Code' Using The NIST Definition

The HHS Office of Civil Rights uses the same guidance as the NIST definition of what constitutes a "code", which states:

> De-identified information can be re-identified (rendered distinguishable) by using a code, algorithm, or pseudonym that is assigned to individual records. The code, algorithm, or pseudonym should not be derived from other related information* about the individual, and the means of re-identification should only be known by authorized parties and not disclosed to anyone without the authority to re-identify records. A common de-identification technique for obscuring PII [Personally Identifiable Information] is to use a one-way cryptographic function, also known as a hash function, on the PII.
>
> *This is not intended to exclude the application of cryptographic hash functions to the information.

Datavant tokens are generated using *cryptographic hash functions*, specifically with SHA-256, one-way, irreversible cryptographic hash function from identifiable information.

Datavant Tokens Are Certified Under The Expert Determination Standard

The Expert Determination Standard at §164.514(b)(1) in the HIPAA Privacy Rule is one of 2 methods to achieve de-identification. Datavant tokens are certified through the Expert Determination standard.

> (b) Implementation specifications: requirements for de-identification of protected health information. A covered entity may determine that health information is not individually identifiable health information only if:
> (1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
>
> (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably

> available information, by an anticipated recipient to identify an individual who is a subject of the information; and
> (ii) Documents the methods and results of the analysis that justify such determination; or

Datavant's independent, third-party Expert Determination certification is provided by Dr. Patrick Baier from Scheuren-Ruffner Consultants. Dr. Baier is a cryptographic and statistical de-identification and re-identification risk expert.

Datavant Tokens Can Be Re-Identified by the Covered Entity
The HIPAA Privacy Rule provides the following direction with respect to "re-identification" by the covered entity in §164.514(c)

> (c) Implementation specifications: Re-identification. A covered entity may assign a code or other means of record identification to allow information de-identified under this section to be re-identified by the covered entity, provided that:
>
> (1) Derivation. The code or other means of record identification is not derived from or related to information about the individual and is not otherwise capable of being translated so as to identify the individual; and
>
> (2) Security. The covered entity does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification.

The re-identification provision permits assignment of a code or other means of record identification that is derived from identifying information by a covered entity so long as an *expert* determines the derived information meets the expert determination standard.

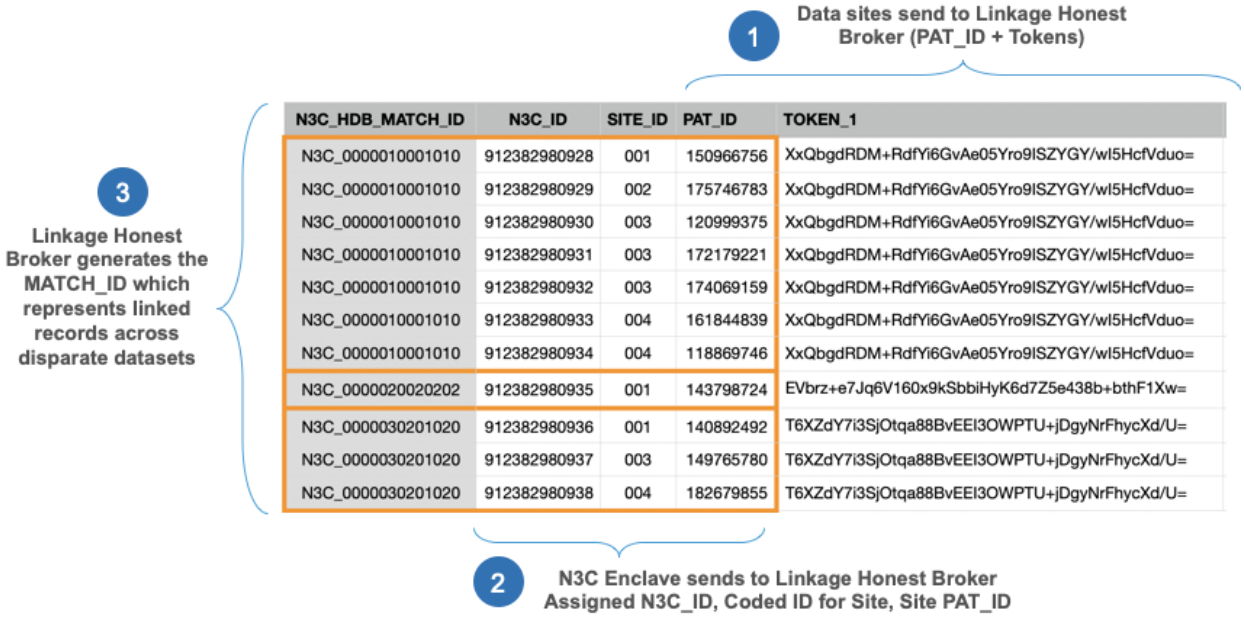
Therefore, codes derived from PHI as part of a de-identified data set can be disclosed if an expert determines that the data meets the de-identification requirements at §164.514(b)(1).

N3C as a Reference Implementation
Within the N3C data infrastructure, a separate entity is designated as the *linkage honest broker*. The role of this linkage honest broker is to serve as the neutral steward and processor of Datavant tokens that are generated by data participating sites (frequently covered entities).

The linkage honest broker (Regenstrief Institute) holds certain de-identified IDs, cryptographic hashes (codes/tokens) and generates a new ID representing the linked record. The various IDs and processes are illustrated in the diagram below.

1. Participating sites send a randomly generated ID (PAT_ID) and the cryptographic hashes to the linkage honest broker
2. When the N3C data enclave onboards clinical data from a site, an N3C_ID is assigned. The N3C_ID together with the site's randomly generated ID for a specific record (PAT_ID) are sent by the N3C data enclave to the linkage honest broker.

3. The linkage honest broker runs Datavant matching algorithms based on specific linked data use cases, and generates a MATCH_ID which represents the records that should be linked. In the example below, there are 3 individuals with data residing across multiple sites, and capturing duplicate records within a site.



N3C_HDB_MATCH_ID	N3C_ID	SITE_ID	PAT_ID	TOKEN_1
N3C_0000010001010	912382980928	001	150966756	XxQbgdRDM+RdfYi6GvAe05Yro9ISZYGY/wl5HcfVduo=
N3C_0000010001010	912382980929	002	175746783	XxQbgdRDM+RdfYi6GvAe05Yro9ISZYGY/wl5HcfVduo=
N3C_0000010001010	912382980930	003	120999375	XxQbgdRDM+RdfYi6GvAe05Yro9ISZYGY/wl5HcfVduo=
N3C_0000010001010	912382980931	003	172179221	XxQbgdRDM+RdfYi6GvAe05Yro9ISZYGY/wl5HcfVduo=
N3C_0000010001010	912382980932	003	174069159	XxQbgdRDM+RdfYi6GvAe05Yro9ISZYGY/wl5HcfVduo=
N3C_0000010001010	912382980933	004	161844839	XxQbgdRDM+RdfYi6GvAe05Yro9ISZYGY/wl5HcfVduo=
N3C_0000010001010	912382980934	004	118869746	XxQbgdRDM+RdfYi6GvAe05Yro9ISZYGY/wl5HcfVduo=
N3C_0000020020202	912382980935	001	143798724	EVbrz+e7Jq6V160x9kSbbiHyK6d7Z5e438b+bthF1Xw=
N3C_0000030201020	912382980936	001	140892492	T6XZdY7i3SjOtqa88BvEEI3OWPTU+jDgyNrFhycXd/U=
N3C_0000030201020	912382980937	003	149765780	T6XZdY7i3SjOtqa88BvEEI3OWPTU+jDgyNrFhycXd/U=
N3C_0000030201020	912382980938	004	182679855	T6XZdY7i3SjOtqa88BvEEI3OWPTU+jDgyNrFhycXd/U=

References

Office of Civil Rights, U.S. Department of Health and Human Services. *Guidance regarding methods of de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* November 2012.

Mccallister E, France T, Scarfone K. *Guide to protecting the confidentiality of personally identifiable information (PII): recommendations of the National Institute of Standards and Technology*. Special Publication 800-122, National Institute of Standards and Technology, 2010.