Use of large-scale privacy protecting record linkages for data sharing to enhance the value of clinical data for research

Shaun Grannis, MD, MS[1,2]; Chris Beesley[2]; Jeff Warvel[2]; Brandy Phalora[2]; Umberto Tachinardi, MD, MS, FAMIA, FIAHSI, FACMI[1,2]; Stephanie Hong[3]; Sara Rogovin[4]; Amin Manna[5], MEng; Benjamin Amor, PhD[5]; Nabeel Qureshi[5]; Leonie Misquitta, PhD[6]; Samuel Michael[6]; Melissa Haendel, PhD, FACMI[7]; Christopher Chute, MD, DrPH, MPH[3]; Kenneth Gersing, MD[6]

[1]Indiana University School of Medicine; [2]Regenstrief Institute; [3]John Hopkins University; [4]Datavant; [5]Palantir Technologies, Inc;
[6]National Center for Advancing Translational Sciences (NIH/NCATS); [7]University of Colorado

Regenstrief Institute
Better Care. Better Health.

Palantir

SCHOOL OF MEDICINE

Background

The National COVID Cohort Collaborative (N3C) is an open science effort led by the NIH National Center for Advancing Translational Sciences aimed at accelerating COVID-19 research. Backed by a secure research platform, the N3C Data Enclave offers 2,773 researchers 19.4B rows of EHR data on 6.7M patients from 76 sites.

The linkage honest broker (LHB) is a neutral, vendor-agnostic entity, external to the N3C data enclave, which serves as an escrow for encrypted identity tokens and operates the platform that facilitates PPRL using these tokens. The LHB does not receive, store, or process PHI/PII or clinical information, which is only held by the data participating sites. (Figure 1) 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. The LHB 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). Finally, the LHB can respond and adapt to emerging use cases: different match algorithms can be applied to different data sources and use cases.

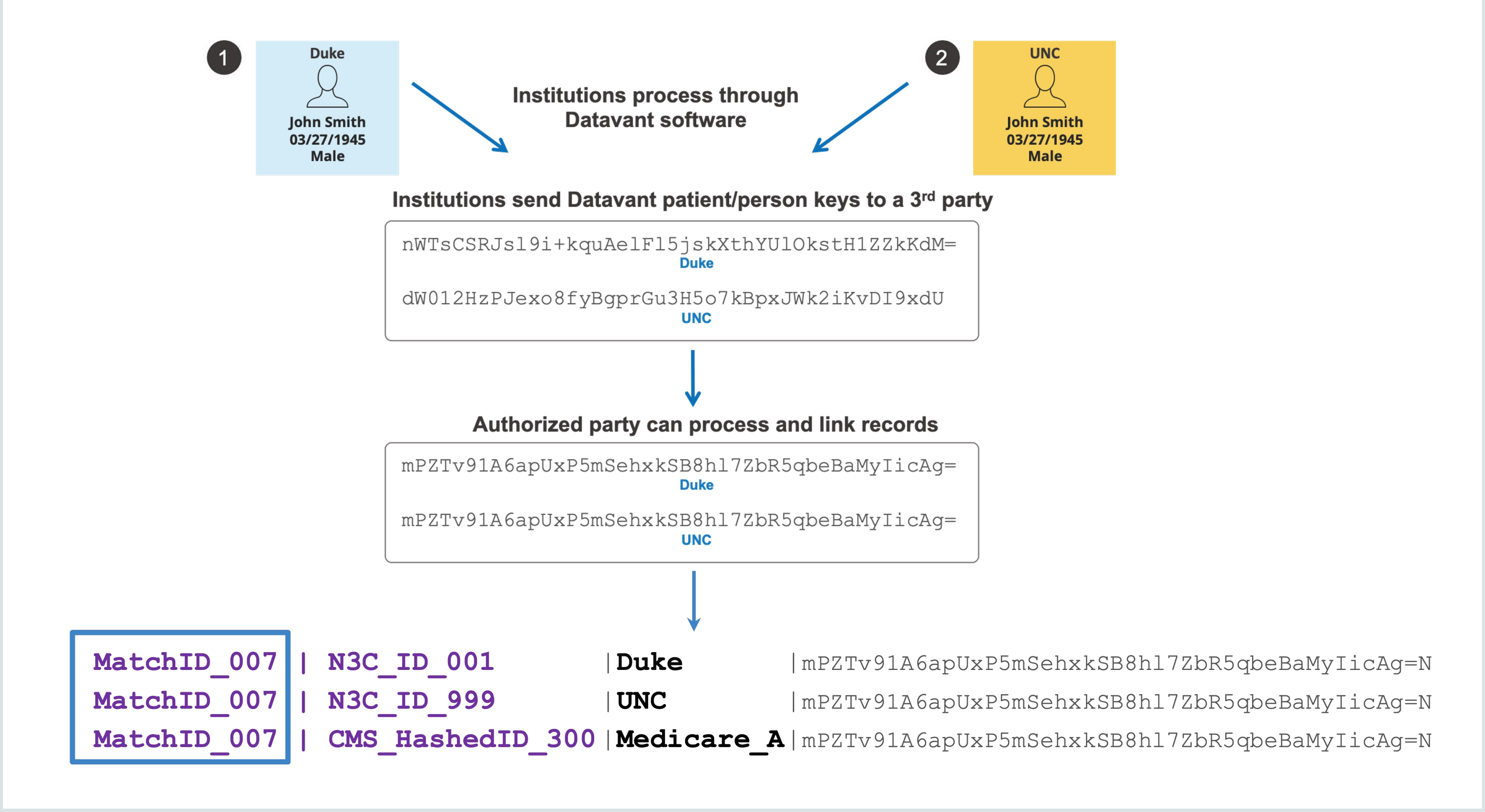
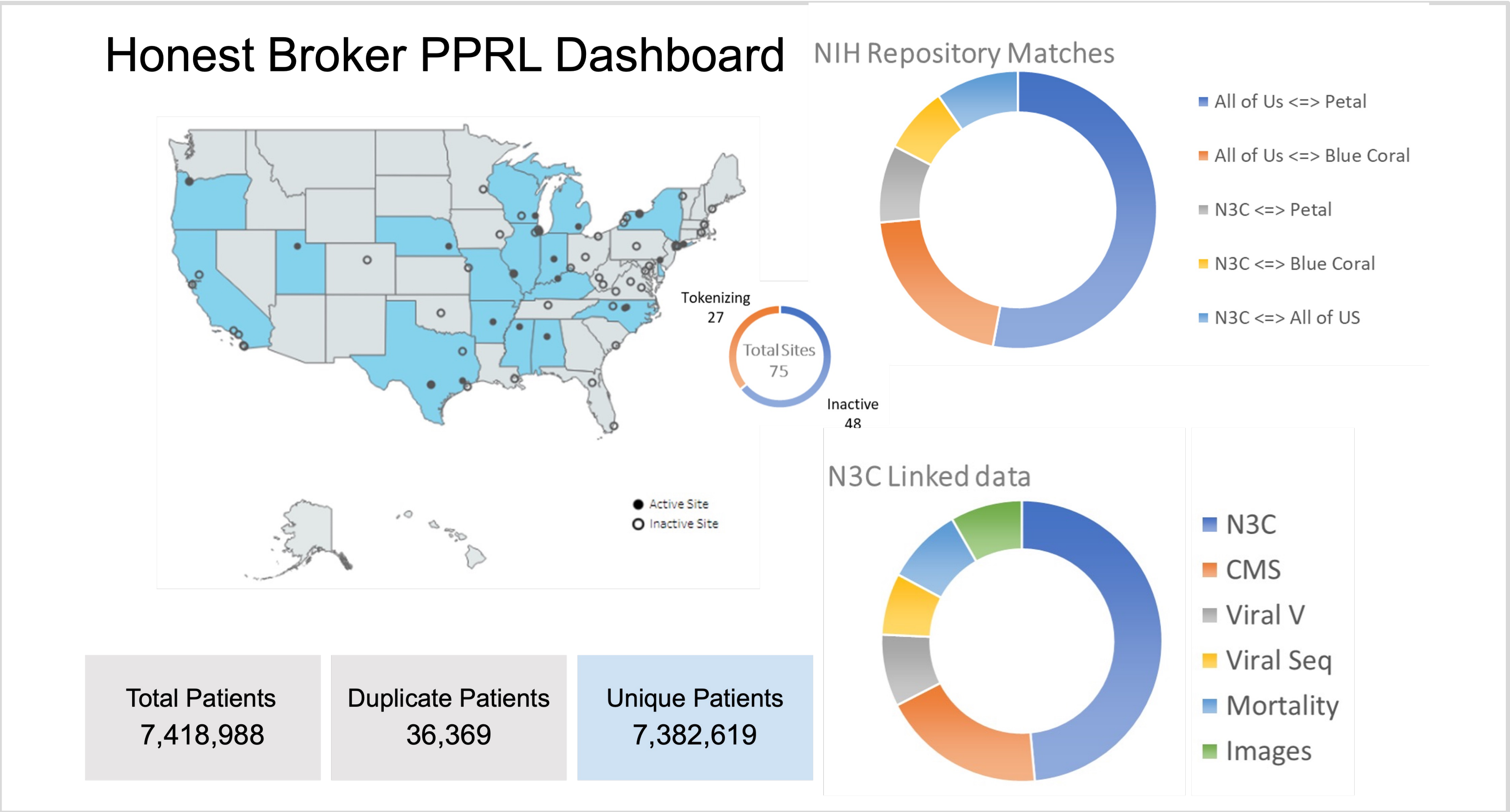

1
Duke
John Smith
03/27/1945
Male
2
UNC
John Smith
03/27/1945
Male
Institutions process through Datavant software
Institutions send Datavant patient/person keys to a 3rd party
nWTsCSRJsl9i+kquAelFl5jskXthYUlOkstH1ZZkKdM=
Duke
dW012HzPJexo8fyBgprGu3H5o7kBpxJWk2iKvDI9xdU
UNC
Authorized party can process and link records
mPZTv91A6apUxP5mSehxkSB8hl7ZbR5qbeBaMyIicAg=
Duke
mPZTv91A6apUxP5mSehxkSB8hl7ZbR5qbeBaMyIicAg=
UNC
MatchID_007 | N3C_ID_001 |Duke |mPZTv91A6apUxP5mSehxkSB8hl7ZbR5qbeBaMyIicAg=N
MatchID_007 | N3C_ID_999 |UNC |mPZTv91A6apUxP5mSehxkSB8hl7ZbR5qbeBaMyIicAg=N
MatchID_007 | CMS_HashedID_300 |Medicare_A |mPZTv91A6apUxP5mSehxkSB8hl7ZbR5qbeBaMyIicAg=N

Figure 1: Conceptual overview of the PPRL hashing process. Note that hash values (tokens) for the same individual at different sites are identical, allowing a linkage to be made without identifying the individual.

Methods

Since capture of mortality data within health systems may not fully represent all-cause mortality during the COVID-19 pandemic, we deployed a privacy-preserving record linkage (PPRL) algorithm to associate N3C patient data with external mortality data. We applied the gold standard validated using a software package to process identifiable demographic elements and formulate HIPAA-compliant deidentified tokens. The tokens enable linkage between patient and mortality data, which contains deidentified US death from multiple sources (covering ~90% historical and ~75% ongoing deaths) and includes data source, birth/death dates, state, and sex.

A key PPRL benefit is that protected health information (PHI) never leaves contributing sites. Instead, sites send tokens to a Linkage Honest Broker (LHB), which constructs a linkage map that the Enclave uses to link EHR data to the mortality records. At no point do either the LHB or Enclave receive PHI. The LHB does not hold data.



Total Patients	Duplicate Patients	Unique Patients
7,418,988	36,369	7,382,619

Figure 2: Geographical representation of site locations. The current statistics for N3C are included. Note that the PPRL process identified over 47,000 links within and among participating N3C sites.

Results

To date, mortality data PPRL has been performed for 27 contributing sites, with 47,100 linkages found within and across these sites. Nearly 6.7 million unique patents have been identified (Figure 2). PPRL linkage is performed on a weekly basis incorporating approximately 20% improvement in providing mortality evidence over EHR data alone.

Summary

Dataset linkage to obituary and SSA records is a particularly impactful PPRL application due to high rates of missingness in EHR death data.

Conclusion

In the future, we will further develop the PPRL algorithm to allow association with additional datasets, such as CMS claims, imaging, genomic, or clinical trial data, for an even richer research data asset.

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. 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.
4. N3C Dataset Registry. https://discovery.biothings.io/dataset?guide=/guide/n3c/dataset. Accessed November 19, 2021.
5. 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.