

Public Health Sector Practice

The Missing Link

Solving patient privacy protection and data linkage challenges for the National Public Health Data System



By Linda Hermer, Ph.D., Andrew Goldberg, Kenona H. Southwell, Ph.D., and Emery Niemiec

Introduction

Major gaps in U.S. health data systems were exposed during the COVID-19 pandemic. Among the foremost gaps is the fact that the nation's health information exists in siloed databases. Across federal, state, and local public health agencies alone, there are more than 2,000 separate databases. Beyond health agencies, numerous other entities maintain U.S. health data for their independent purposes, including other federal agencies, public- and private-sector payers, physician groups, hospitals, and other providers, pharmacies, and clinical lab companies among others. The data exists in myriad different formats and is not readily analyzable as unified, longitudinal patient journeys through the healthcare system.

These siloed, fragmented, and inconsistently formatted data sets have adversely impacted the country.

For example, answering numerous policy questions during the pandemic took much longer than needed. The question of whether COVID-19 vaccines

were safe for pregnant women and their unborn children took almost a year to answer because there was no centralized database jointly containing data on vaccinations, pregnancies, births, deaths, and other postnatal outcomes. As a result, policymakers were forced to delay issuing critical guidance to pregnant women—as with many other subpopulations who were counting on the CDC, FDA, and other federal agencies for directives.

More generally, it became clear that the current system hinders scientific advancements and interferes with the prediction and improvement of public health. As a result, there is a newfound resolve among federal decision-makers to integrate the nation's health data starting with federal databases. The creation of the proposed **National Public Health Data System (NPHDS)** will be invaluable in equipping policymakers and public health researchers with the representative and reliable data sets needed to make impactful public health decisions.

A central challenge in creating the NPHDS is linking patient identities

across disparate systems with high accuracy while also safeguarding patients' health information in compliance with the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The patient identity resolution issue is especially acute for the health data in federal databases, almost none of which contain unique patient identifiers such as Social Security numbers and presenting a plethora of other linkage challenges.

Eagle Technologies, Inc., and HealthVerity, Inc., are two leaders in the health IT arena who are providing thought leadership as the government contemplates different solutions for the NPHDS. For example, HealthVerity has accurately linked private- and public-sector health data from more than 75 national data sources covering 330 million unique patients, comprising data on most of the current U.S. patient population. **Here we present a tested, proven solution to linking patient records with high accuracy in the NPHDS while complying with HIPAA regulations by not transferring and centrally storing any**

of the patients' primary personally identifiable information (PII) in the NPHDS central repository. Rather, the Eagle Technologies and HealthVerity solution relies on replacing PII with a unique but persistent universal patient enumerator—enabling the integration of health data on a fully interoperable but privacy-protected basis.


Our recommendations represent industry best practices for solving the above challenge that federal agencies and contractors awarded procurements should follow. The approach we present has demonstrated *an accuracy rate 10 times better than legacy industry alternatives have previously achieved*. Concurrently, it safeguards patients' identities throughout the data acquisition, linkage, and storage processes while successfully managing the increasing regulations around privacy and security.

The Data Landscape

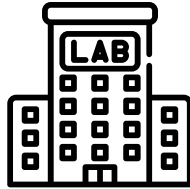
All the data sets to be ingested by NPHDS' extract, transform, and load (ETL) processes include patient PII, and this data can be used directly or


Duke General Hospital



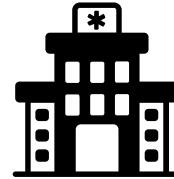
 **Name:** Edward D. Michael
Birth Date: 10/12/1965
Address: 423 Jade Dr.
Springs, IN 47909
Admission Date: June 5, 2019

Rainbow Laboratory, Inc.



 **Name:** Ed Michael
Birth Date: 10/12/1964
Address: 2315 Cumberland Ave.
Lafayette, Indiana 47907
Service Date: December 12, 2020

Pines Psychiatric Hospital




 **Name:** Edward Michael
Birth Date: 10/12/1965
Address: 412 Brea Lane
St. Paul, Minnesota 55106
Admission Date: September 15, 2021

Figure 1. Illustration of some of the complexities of matching patients. The first two records are from the same individual, whereas the third record is from a different person.

indirectly for record linkage. Patients are typically linked on information derived from string, numeric, and modulus PII fields such as names, addresses, and birthdates. Legacy approaches rely on purely deterministic or exact matching on such data that seldom yields consistently reliable matching statistics. Many of the available data fields overlap extensively across individuals (e.g., metropolitan areas contain many individuals with the same first and last names). In addition, real-world data such as these are typically rife with errors such as misspellings and transposed digits; variations such as inconsistent use of middle names, initials, and nicknames across multiple occurrences of the same individual; missing values; and

out-of-date information such as what might occur following a patient's move to a new address. **Figure 1** presents a case study illustrating several of these challenges for two individuals with the same first and last names who live in relatively close proximity.

Matching Statistics

In the face of these challenges, most private sector data vendors offer data sets with estimated 3–5% false positive and 9–42% false-negative match rates. Such poor accuracy rates are highly problematic. High false positive, incorrect matches may facilitate policymakers' and researchers' drawing incorrect inferences from the data. Similarly, high false negative rates

(incorrectly unmatched cases) result in fragmented patient journeys and suboptimal knowledge generation. The success of the NPHDS is dependent on an efficient and highly accurate approach to linking hundreds of millions of patients across billions of fragmented medical events spread over hundreds or thousands of data sources. The technologies and methods recommended below for the NPHDS are characterized by false positive and false negative rates lower than typical industry rates by a factor of 10.

Introduction to Patient Identity Resolution and Data Linking Solution for the NPHDS

The proposed NPHDS patient identity resolution and data linkage system features innovative technologies and solutions to enhance the utility, transparency, availability, and cost efficiency of traditional and emerging health care data. Overall, the system will enable processing of PII and transactional records for highly accurate de-identification and matching to a unique but persistent

universal patient identity, enabling interoperability across siloed data. This industry- best solution leverages advanced techniques including probabilistic matching and machine learning algorithms based on over six billion variations of U.S. PII to unite disparate patient records with *ten times the accuracy* of comparative deterministic methods. Here we describe the steps, methods, and technologies that will generate accurate, longitudinal, health data sets without compromising patients' PII. The system leverages advanced cryptographic techniques and centralized matching technology to optimize privacy-preserving patient entity resolution.

Protecting Patients' Personal Information

Meeting the need for extensive longitudinal linking of personal data in a way that preserves privacy is critical. One of the distinguishing features of the proposed system is that protection of patients' PII and protected health information (PHI) is at the forefront of the process and protection-related

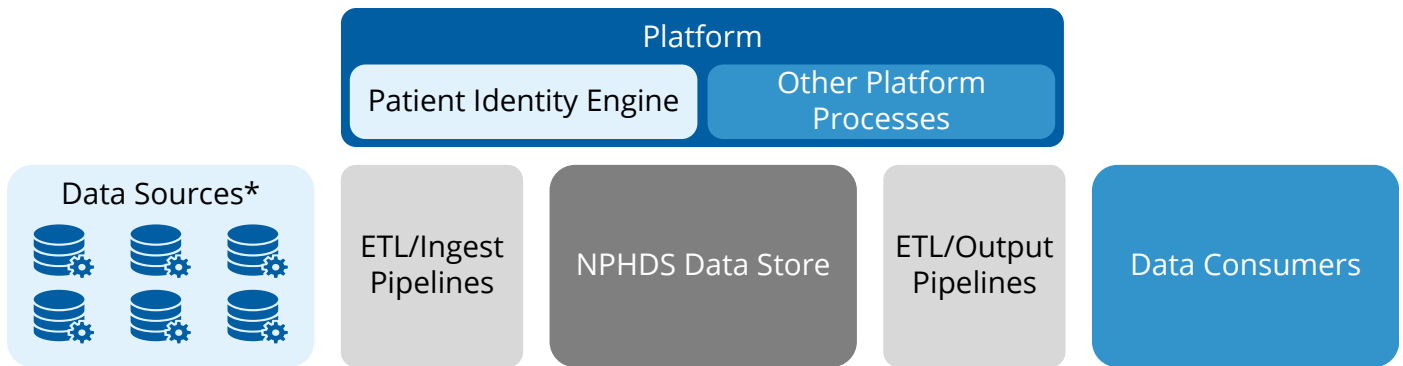
procedures and mechanisms are strategically intertwined throughout data acquisition, processing, storage, analysis, and sharing.

First, the NPHDS would never directly receive any PII, nor even require access to the data owners' network. All data transmitted to NPHDS would be de-identified and encrypted, with patient identity resolution taking place with data encoded with a secure hashing algorithm that generates a universal, de-identified patient ID. This technology has been certified as meeting the de-identification requirements of HIPAA using expert determination across multiple leading third-party privacy experts. Second, the solution will rely on privacy and governance techniques developed to ensure that broadscale clinical transaction data linked to patients' de-identified identity is also consistent with a master HIPAA certification and manages this health data to exclude or transform any data fields that would otherwise leak quasi-identifiers.

Local Engines for Data Hashing, Encryption, and Transfer

To transfer data to the NPHDS, we recommend that the system utilizes a lightweight de-identification engine that will be installed locally on each data owner's server. While safeguarded by the data owner's firewall, all protected health information is removed by the engine and replaced using sophisticated algorithms that generate an encrypted hash of the original PII. The technique leverages a one-way hash function meaning that it is statistically impossible to reverse the computation and discover the original PII values. Consequently, all health data leaving each data owners' site would be considered de-identified under HIPAA prior to transmission. This method would be certified as being in full accordance with HIPAA expert determination provisions.

To further minimize risk during data acquisition, all patient hashes should be further encrypted and sent to the sequestered, centralized NPHDS server for automated matching. This



*Each data source has its own de-identification and transfer engine.

Figure 2. NPHDS notional architecture with systems for data de-identification and transfer highlighted.

procedure protects each input system from both other input systems and the NPHDS, mitigating the risk of a successful hacking occurrence. It also enables each data owner system to continue maintaining its own PII independently. **Figure 2** illustrates the separation of the engines for data de-identification and transfer, on the one hand, and centralized matching, on the other, in the NPHDS architecture.

First Stage of Matching

The initial encrypted hashes of PII should then be used to accomplish patient identity resolution by assigning the correct patient identity from a central master database of patient identities. In most cases, the patient hash from a particular data set will

be assigned an ID from a national database of such IDs, although a new ID may be assigned in the rare case that the system believes the individual to be a new patient. In cases of slight variations with a patient's identifying information (such as misspelled names or a new address), if the system were to rely on purely deterministic matching, new patient identities would be created even though the data represent a single, existing patient. The probabilistic matching employed by this solution, however, accounts for variations in patient identity by comparing several field values to determine an accurate match, resulting in the lowest false positive and false negative rates attainable.

To complete the first stage of

matching, building on multiple personal identifiers and the associated secure chasing algorithm, probabilistic matching accumulates the probability of each piece of evidence, seamlessly accounting for larger probabilities relating to missing data, address changes, and ambiguities in the data. Each candidate match is evaluated based upon the evidence present in the observed tokens.

Final Stage of Matching

In the final stage of matching, machine learning should be used to account for additional patterns in the data. These models can adjust matching probabilities and assign final match

results based on further information on which the models should be trained, such as:

- The frequency of each name, location, and phone number
- Relocation patterns (e.g., patients moving between partially masked, three-digit zip codes such as from Long Island, NY, to West Palm Beach, FL)
- Rates of typos, swaps, and other errors in data entry

Figure 3 depicts all the stages of data de-identification, acquisition, and matching discussed above.

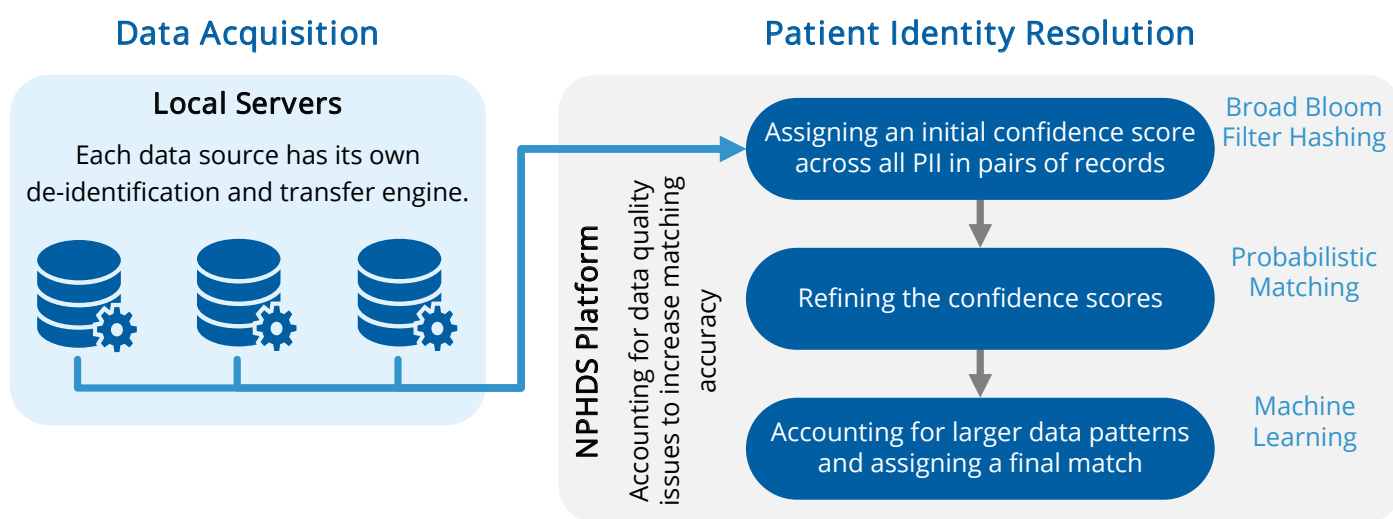


Figure 3. The stages of data de-identification, acquisition, and multi-stage matching in the proposed system.

Application of the Technologies and Methods

The proposed system has been used by HealthVerity in numerous large-scale, longitudinal healthcare and consumer data projects from more than 75 national data partners, covering over 150 billion de-identified health and consumer transactions across more than 330 million individuals.

The system has been used to process medical claims, pharmacy claims, EMR data, hospital or chargemaster data, lab data, and social determinants of health data—all of which were rendered completely interoperable yet privacy protected by our methods.

Working with HealthVerity, federal agencies have already leveraged these methods to accomplish the following:

- **Support for the COVID-19 vaccine program:** The Health and Human Services (HHS) contract entails enabling a Privacy-Preserving Record Linkage (PPRL) in support of the national COVID-19 vaccine program. The PPRL

solution delivers a de-identified linkage of individual patient records across jurisdictions and settings over time. With the additional use of longitudinally linked patient data, the U.S. government and jurisdictions are better able to estimate vaccine administration rates within its population, understand the degree of immunization among the population, securely share information with appropriate clinical staff about individuals' vaccination status, appropriately target vaccine acquisition and distribution, and identify areas that require targeted communication to promote vaccination.

- **Support for COVID-19 research:** As an extension to the vaccine project, the patient identity resolution and data linkage system was used to support the CDC's need to combine multiple data sources into a longitudinal and interoperable real-world data view to provide insights into new and emerging research related

to COVID variants, vaccination, testing, re-infection, health impact, natural history, and long-term effects of COVID-19. In addition, the data linkage system will be used to connect genomic sequencing data to real-world data in order to better understand demographic and clinical characteristics and outcomes associated with different COVID-19 variants. The award was the largest licensing of real-world data by the CDC in its history.

- **Connecting cancer patients to a real-world data ecosystem:** This National Institute of Health project utilized the de-identification and matching technology of the system. Across two cancer registries at the National Cancer Institute (NCI), cancer patients are being connected to a real-world data ecosystem for enhanced insight (e.g., to determine the impact of the pandemic on cancer outcomes).

Moving Forward with the NPHDS Data Linkage System

The COVID-19 pandemic underscored the need for a centralized public health data repository—the NPHDS—that can be used for fast and accurate analyses to support public health decision-making. In the process, however, patients’ PII should not be transferred across data systems. Here, we recommend the approach of privacy protection through initial patient de-identification and data encryption, followed by multi-stage matching to accurately resolve patient entities.

After linkage and conversion to longitudinal format have taken place, advanced data analytics can be performed in a HIPAA-compliant manner to support population health prediction, outbreak detection and other disease surveillance, monitoring of the healthcare system, assessing efforts to achieve health equity, and a wealth of additional purposes. NPHDS data will be updated regularly, facilitating real-time insight into problems of a range of scopes, scales,

and complexities limited only by the imaginations of researchers and policymakers. In addition, the data will be in a format conducive to merging with private-sector data ecosystems, allowing the NPHDS to eventually incorporate genomic and other data for advanced analyses. Finally, fully in accord with HIPAA regulations, the federal government will be able to share the data across federal agencies and with research institutions, non-profits, for-profit healthcare industry partners, and a wide variety of other stakeholders conceiving of new uses for the data. The possibilities are limitless.

About the Authors



Linda Hermer, Ph.D., is Senior Project Director for public health and health IT research at Eagle Technologies. She is a thought leader in the research and health management potential of integrated data systems in areas such as COVID-19, behavioral health, aging and dementia, and child development.



Andrew Goldberg is Co-Founder and COO at HealthVerity, Inc. He has more than 30 years of cross-sector experience focused on creating network effects through data connectivity. In addition, Mr. Goldberg has led corporate growth and development across various IT communications and cloud solutions companies.



Kenona Southwell, Ph.D., is a Senior Researcher at Eagle Technologies where she uses utilizes big data to understand emerging patterns in pressing current public health issues. She has generated deep insights into opioid and mental health treatment, telemedicine, and mental health diagnosis.



Emery Niemiec is Director, Partners & Alliances, at HealthVerity, Inc. She has more than 10 years of experience in data-driven transformations that enable partners to uncover enhanced insights and track outcomes. Ms. Niemiec supported various Fortune 500 companies through digital transformation efforts with a public health focus.

Point of Contact



Paul Garcia serves as Director of Business Development at Eagle Technologies. He is a former Government Contracting Officer with more than 30 years of experience supporting Government and Commercial sectors. Mr. Garcia leverages decades of management expertise to drive business development resulting in corporate growth.

E: paul.garcia@eagletechva.com | P: (571) 275-8894

About the Companies

Eagle Technologies, Inc., is a Technical and Systems Integrator delivering effective health IT and grants management solutions transcending the complex requirements of government and business clients nationwide. Eagle leverages its depth of experience delivering end-to-end solutions that rely on advanced cybersecurity and privacy systems, enterprise architecture, cloud-based services, mobile computing, and business intelligence services.

HealthVerity, Inc., Pharmaceutical manufacturers, payers and government organizations have partnered with HealthVerity to solve some of their most complicated use cases through transformative technologies and real-world data infrastructure. The HealthVerity IPGE platform, based on the foundational elements of Identity, Privacy, Governance and Exchange, enables the discovery of RWD across the broadest healthcare data ecosystem, the building of more complete and accurate patient journeys and the ability to power best-in-class analytics and applications with flexibility and ease.