



Privacy Preserving Record Linkage (PPRL) Strategy and Recommendations

PPRL Linkage Strategies Report

Sponsor: National Institute on Aging
Contract No.: 75FCMC18D0047
Outcome No: 1

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

©2022 The MITRE Corporation.
All rights reserved.

McLean, VA

Stacy J. Petersen
Robert D. Lieberthal
Keith J. Miller
Niyati H. Vakil

January 2023

1 Executive Summary

This report provides an overview of Privacy Preserving Record Linkage (PPRL) technology as a strategy to securely link real-world data (RWD) to National Institute on Aging (NIA) funded studies. This report is part of a larger project to achieve impact in Alzheimer's Disease and Alzheimer's Disease Related Dementias (AD/ADRD) research to allow researchers to focus on a wider array of factors in AD/ADRD prevention, care, treatment, and other important areas of investigation.

This project has several components including this overview of PPRL strategies, a selection of NIA use cases well suited for implementation of PPRL, identification of common data elements (CDEs) to facilitate interoperability, and expert feedback on RWD sources and CDEs that can be linked. The selection of a PPRL technology and potential vendor heavily relies on the technical and practical requirements of a particular use case, which are described in the Use Case Report, however this report provides a general overview of the roles, methods, and limitations of the PPRL process.

Given the large number of people affected by AD/ADRD¹ and the significant amount of NIA-funded clinical trials, research projects, and real-world datasets (RWD)², it is ideal to be able to link data privately and securely across these resources that could lead to advances and discoveries in a range of outcomes in clinical, economic, and universal domains. However, the lack of a unique common identifier, inconsistent identifying information, and privacy concerns limits the ability to align data between data sources. Therefore, any linkage strategy would require (i) no PII sharing across data partners, (ii) scalability to many participants and (iii) flexibility regarding inclusion of certain data elements used for linkage.

PPRL is a cutting-edge technique that can match records from many data sources across institutions and jurisdictions without directly sharing personally identifiable information (PII). PPRL relies on data partners who use an encryption key, provided by a key escrow, to obfuscate their PII. The obfuscated data is then linked by a third-party linkage agent, who creates a unique identifier which can be used to link data from the same individuals across data partners. The PPRL method relies on trust between all parties and strict enforcement of regulation of the encryption key to mitigate adversarial attacks, which would result in breaches in privacy. This report discusses the types of attacks and common limitations of PPRL, as well as ways to avoid and overcome them.

The method of encryption can either enforce exact matches or allow for errors (such as misspellings, token inversions, etc.) in the underlying data elements. One common technique for probabilistic data matching is the use of Bloom Filters, which are data structures used to hold encoded personal identifiers. The technique will vary depending on the selected CDEs, selected PII / biographic/demographic data and use case. Considerations for setting up, testing, and evaluating PPRL in an NIA use case are detailed in this report.

As part of this report, we have performed a market scan and preliminary assessment of PPRL technologies to facilitate the further implementation of PPRL by NIA in its trials and aging data

¹ More than six million people in the United States currently live with AD/ADRD, a number projected to jump to 14 million by 2060. Alzheimer's Association. 2021. *Alzheimer's Disease Facts and Figures: Special Report*. 17:3. Available at <https://www.alz.org/alzheimers-dementia/facts-figures>. Accessed 2/8/2022.

² AD/ADRD studies include 442 active clinical trials, 5,356 actively funded NIA projects in AD/ADRD, and real-world data sets (RWD). See Use Case Report.

sets. This assessment, guided by the overall purpose of the project, is designed to provide actionable guidance to NIA in the use of PPRL for AD/ADRD research in the field. We highlight one specific technology, Bloom filter-based probabilistic matching, and three specific products / vendors, ANONLINK, Datavant, and Sensing, as having strong potential for implementation in aging research. These three options also provide options NIA may want to consider, as one (ANONLINK) is available as an open-source product, while the other two (Datavant and Sensing) are proprietary. We are aware of use in the health domain of ANONLINK and Datavant, but we were unable to obtain information to verify the use of Sensing in health. We will continue to refine these recommendations as we engage in the next steps of the current project.

An assessment also was performed of the ability to determine what mapping activities have occurred to date between different providers. The MITRE team was able to verify 1-1 and 1-2 mappings by specific vendors. In other words, current PPRL tools allow data providers to map individuals across the data sets they provide, and there are vendors that provide solutions that allow for mapping with data from other vendors (“1-2” mapping). Mapping across three or more data sources was not a solution that was identified, although there is no technical reason that such a mapping cannot be implemented using PPRL. Instead, proprietary, legal, and ethical concerns with wider forms of matching are the main impediments to applying PPRL across a wider array of data sets.

To ensure effectiveness of selection and implementation of any PPRL approach, there are several points which must be taken into consideration:

- Stakeholders must be willing and able to conduct the PPRL processes and transmit data to a third party.
- PPRL tools should require minimal IT support to implement and run.
- There must be strong coordination of activities across data partners.
- The process must be seamless and invisible to data users.

We will address these points in future parts of this project. We are planning a virtual expert panel to ensure stakeholder engagement and a “user-friendly” approach for researchers to integrate the assessment of PPRL with the harmonization of CDEs in that part of the project. Future considerations of cost and support requirements for NIA also will be delivered as the final part of this project. That report, along with transition activities designed to ensure the current project is implemented by the NIA, will incorporate the selection criteria for any PPRL strategy made as part of a final recommendation.

This page intentionally left blank.

Table of Contents

1	Executive Summary.....	1-1
2	Background.....	2-1
2.1	Privacy-Preserving Record Linkage	2-1
2.1.1	Matching Individuals Using PII.....	2-1
2.1.2	The PPRL Solution to the Matching Problem	2-1
2.1.3	Example Applications in the Field.....	2-2
2.2	Application of PPRL to AD/ADRD Research for NIA.....	2-3
2.2.1	Rationale for AD/ADRD Research.....	2-3
2.2.2	NIA Applications of PPRL	2-3
3	PPRL Roles, Process, and Matching Methods.....	3-1
3.1	Roles	3-1
3.1.1	Data Partners	3-1
3.1.2	Key Escrow	3-1
3.1.3	Linkage Agent.....	3-1
3.1.4	Data Users	3-1
3.2	Matching Process	3-2
3.2.1	Data Preprocessing and Selection of Schema	3-2
3.2.2	Configuration Information	3-2
3.2.3	Creating a Hashed Data Set	3-2
3.2.4	Creating and Sharing Unique Identifiers	3-2
3.2.5	Data Sharing and Transmission	3-3
3.2.6	Selecting a Linkage Agent	3-3
3.3	Matching Methods	3-4
3.3.1	Deterministic Matching.....	3-4
3.3.2	Probabilistic Matching	3-6
3.3.3	Comparing Bloom Filters.....	3-7
3.3.4	Combined Matching Approaches.....	3-8
3.4	Available tools for use in health and healthcare	3-9
3.4.1	Search Strategy	3-9
3.4.2	Commercial and Open-Source Options	3-9
3.5	Market Scan of PPRL Tools	3-11
4	Limitations of PPRL.....	4-1
4.1	Limitations of Specific Methods.....	4-1

4.1.1	Limitations of Deterministic Matching.....	4-1
4.1.2	Limitations of Probabilistic Matching	4-1
4.1.3	Limitations of a Combined Matching Approach	4-2
4.2	Privacy and Data Breaches	4-2
4.2.1	Types of attacks	4-2
4.2.2	Number of Parties	4-3
4.2.3	Record Linkage and Data Breaches	4-3
4.2.4	Privacy and Ethical Concerns	4-4
4.3	Linkage Quality	4-5
4.3.1	Assessment of Quality	4-5
4.3.2	Data Challenges	4-5
4.3.3	Quality Improvement	4-6
4.4	Computational Scalability.....	4-6
5	Conclusion	5-1
5.1	Current and Potential Future State of PPRL	5-1
5.1.1	Potential for Impact.....	5-1
5.1.2	Ensuring Effectiveness.....	5-1
5.1.3	Coordinating Key Destruction	5-2
5.2	Recommended Technologies and Vendors.....	5-2
5.2.1	Recommended Technologies	5-2
5.2.2	Recommended Vendors	5-3
5.3	Evaluating PPRL on a Pilot Basis for NIA.....	5-1
6	References	6-1
	Appendix A Glossary	A-1
	Appendix B Abbreviations and Acronyms	B-1

List of Figures

Figure 2-1: Process of PPRL using blind matching.....	2-2
Figure 3-1: PPRL Process Flow Chart.....	3-4
Figure 3-2: Hashing Example from a Sample Individual Record.....	3-5

List of Tables

Table 3-1: Example of an encryption key value	3-5
Table 3-2: Example of hashed values with an encryption key	3-6
Table 3-3: Example Bloom filter construction	3-7
Table 3-4: Dice coefficient	3-8
Table 3-5: Market scan of available PPRL tools	3-11
Table 5-1. Recommended Technology List.....	5-3
Table 5-2. Recommended Vendor List.....	5-1

This page intentionally left blank.

2 Background

2.1 Privacy-Preserving Record Linkage

2.1.1 Matching Individuals Using PII

The process of matching individuals in the absence of a shared, unique identifier often requires organizations to exchange information with each other or a third party. Matching occurs by comparing shared personal identifiable information (PII) to see if there are similarities in demographic attributes that appear in the form of *direct identifiers*, such as name and medical record number, as well as *quasi-identifiers*³, such as gender, date of birth, email address, telephone number, or zip code.

The impact of using an approach that can match individuals within and across longitudinal studies to link research efforts is likely to have high impact across several domains. Therefore, technologies and approaches that can solve the matching problem are key to enhancing the value of research and application of scientific studies in practice. An approach that conducts matching at scale and does not involve openly exchanging PII beyond organizational boundaries while still allowing for the linkage of individual records is likely to have high impact in health domains such as research and patient care.

2.1.2 The PPRL Solution to the Matching Problem

PPRL refers to a range of technologies and approaches that solve the privacy and scalability problems that arise when matching records across organizations and institutions. The key technological innovation underlying PPRL is *the ability to secure clear text PII* within the originating organization's system and behind the firewall. Clear text refers to a message or data that has not been subject to encryption; it is in a form that is immediately understandable and readable to a human being without additional processing. In the PPRL process, PII is obfuscated, or hashed in a series of prescribed steps prior to transmission beyond an organizational boundary for matching. The process of hashing results in information from which it is nearly impossible for an outside party to recover PII, but still allows for the establishment of links across organizations.

The PPRL allows for “blind” matching of individuals across data sets and research organizations. In this case, a third party, known as a “linkage agent”, is provided access to the hashed data, but is unable to view PII. The linkage agent then compares the obfuscated information to establish linkages. Figure 2-1 illustrates this process. The linkage agent that conducts the matching assigns an identifier when a linkage is found and communicates a unique identifier back to the participating organizations. This identifier can then be used when transmitting data through healthcare provider, research data, and public and private insurer systems.

³ Quasi-identifiers are variables which are not sufficient to identify an individual in isolation but may do so when combined.

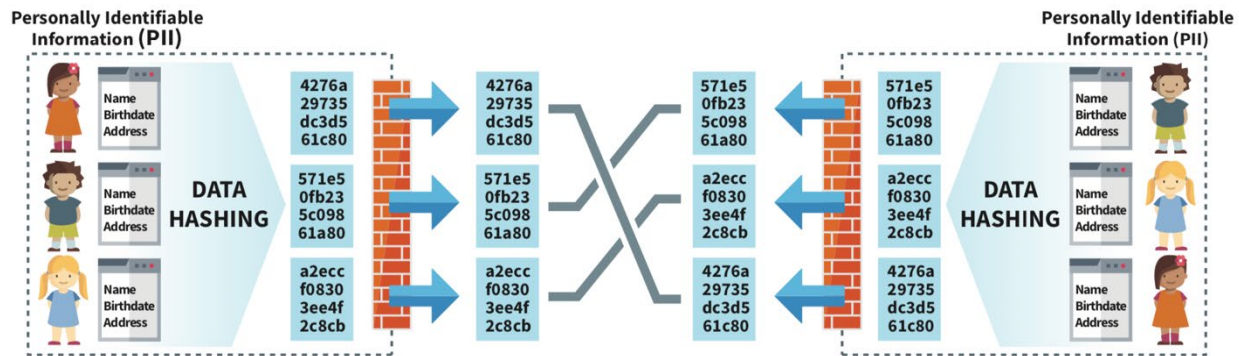


Figure 2-1: Process of PPRL using blind matching

While Figure 2-1 demonstrates PPRL as a simplified process, approaches to implementing PPRL can vary in sophistication and complexity. PPRL can use a variety of methods intended to minimize privacy breaches, prevent malicious attacks, improve efficiency, and provide partial matches that can account for variations in the data, such as data entry errors, name or address changes, or nicknames. These concerns all are highly relevant in the context of health and healthcare research. A selection of these methods is discussed in the following section.

2.1.3 Example Applications in the Field

While still cutting edge, the practical application of PPRL technologies is advancing beyond research laboratories and is being deployed in a range of industries. The following are some examples, and details on the specific technologies can be found in our Market Scan of PPRL Tools as described in Section 3.5:

- Datavant uses PPRL technology to protect, match, and share health data of its clients (Datavant, 2020). Datavant is part of a consortium that recently launched a pro-bono COVID-19 research database that “enables public health and policy researchers to use real-world data to better understand and combat the COVID-19 pandemic” (COVID-19 Research Database, 2020).
- The Australian Collaboration for Coordinated Enhanced Sentinel Surveillance of Sexually Transmissible Infections and Blood-borne Viruses (ACCESS) is using the PPRL software tool, GRHANITE®, to link de-identified patient records for public health surveillance such as blood-borne viruses and sexually transmissible infections (Nguyen et al., 2020).
- The core technology behind Enveil’s ZeroReveal® PPRL solution was created in the U.S. intelligence community and has been implemented at scale in security-sensitive environments (Enveil, 2020). It is now used independently outside of the intelligence community across a range of industries including financial services, healthcare, data monetization, audit and compliance, cloud security, and government.
- Finally, IBM’s Sensing PPRL solution may be applied to any industry including banking, securities, insurance, healthcare, manufacturing, marketing, telecommunications, energy, and government (IBM, 2019). Sensing supports real-time decision-making and is scalable “up to 400M records on a \$5,000 commodity server and performs millions of new entity

resolutions a day in real time with sub-second response rates”. What is not currently clear is whether scalability also applies to Senzing’s newest PPRL solutions.

2.2 Application of PPRL to AD/ADRD Research for NIA

2.2.1 Rationale for AD/ADRD Research

The above examples show that PPRL can be applied to a wide variety of domains. Our project specifically will focus on the application of PPRL to health and healthcare research, with a focus on Alzheimer’s Disease/Alzheimer’s Disease Related Dementias (AD/ADRD). Part of the rationale is pragmatic, as healthcare data providers and vendors such as Datavant and IQVIA already are applying PPRL technologies to their products. There also is a strong rationale related to the relatively siloed nature of healthcare data, which may be maintained in electronic health records (EHRs), health insurance company systems, clinical registries, public surveillance data, and personal health records.

The need to link data across these systems has become more urgent as the data locations and types of data proliferate. In addition, matching and assessing long-term outcomes is likely to be important for those disease states and domains where longitudinal data is needed. The domain of aging and disease states such as AD/ADRD provide a particularly important use case for the application of PPRL.

2.2.2 NIA Applications of PPRL

Currently, application of PPRL within the realm of AD/ADRD research is primarily ad-hoc. A wide array of data sets contains important information about longitudinal outcomes related to AD/ADRD. Some of those data sets can be linked using PPRL. However, clinical trials focused on AD/ADRD therapies generally are not designed for easy linkage with other data sets.

While it is common to link payer data with aging data sets, for example, linking the NIA funded Health and Retirement Study with Medicare claims data, those linkages are based on unique identifiers rather than a PPRL approach. Linkages for other NIA studies with Medicare data are less common or not currently undertaken. The need to obtain and manage those sensitive identifiers, typically social security numbers or Medicare identifiers, limits their use in practice and exposes beneficiaries to the possibility of exposure of their PII or personal health information (PHI). PPRL represents an alternative technology that can address these challenges as described in the remainder of this report.

3 PPRL Roles, Process, and Matching Methods

This section describes the roles involved in the PPRL process, the record linkage process, and the underlying matching methodology. Definitions and terminology are listed in the Glossary (Appendix A).

3.1 Roles

Key roles supporting the PPRL process include data partners, key escrow, linkage agents, and data users.

3.1.1 Data Partners

Data partners are the organizations that own the data that will be collected in real-world data (RWD) environments. Examples include healthcare providers, facilities, laboratories, and other organizations, EHR vendors, clinical and observational trial administrators, public and private insurers and other third-party payers, and vital records registrars. RWD also includes a wide array of data collection initiatives including social science data sets and community surveys. These organizations conduct hashing, share hashed data for record linkage, and incorporate the unique identifiers derived from the record linkage process into their data systems.

3.1.2 Key Escrow

The key escrow organization must be trusted by all data partners to act in their best interests. This relationship is usually supported through a legal agreement between the data partners and the key escrow. The key escrow shares the configuration files used in encryption and the encryption key with the data partners.

3.1.3 Linkage Agent

Data partners provide the linkage agent with encrypted hashed values to conduct the record linkage process. The linkage agent shares the output of the linkage process in the form of unique linked identifiers back to each data partner. To ensure the hashed values remain encrypted, the linkage agent must not have access to the encryption keys.

3.1.4 Data Users

While data users, primarily scientists and researchers, do not use and may be unaware of PPRL, they are a key stakeholder in PPRL implementation. Scientists determine the types of participants needed for specific research studies, as well as the types of data needed to answer research questions. These data users can play a key role in providing input into the types of data needed for patient records linked across different data sets, as well as the types of patients needed for inclusion (with a particular focus on populations highly burdened by disease or underrepresented in current trials).

Data users also play a role to ensure informed consent and authorization are enforced and properly conducted. The data users should protect against bias and misuse of patient information, especially those who interact directly with research participants.

3.2 Matching Process

The matching process typically involves the following steps and is depicted in Figure 3-1 below. This process consists of configuration information, creating a hashed data set, creating a unique identifier, sharing identifiers, and data sharing. Each step in the process is described in greater detail below:

3.2.1 Data Preprocessing and Selection of Schema

To ensure the best results from the PPRL process, the data elements from each dataset should be cleaned and preprocessed for possible errors by the data partners, especially if the linkage agent is performing exact matching (see Section 3.3.1). Additionally, PII data elements must meet agreed-upon specifications that are typically defined in a common data model or other specification document. The choice of PII attributes may depend on characteristics of the data population and will affect the level of linkage accuracy attainable. The selection process may involve data users as well as the data partners.

3.2.2 Configuration Information

As a first step in the PPRL process, a key escrow shares configuration information with each data partner. Configuration information includes data required to use the linkage method and the encryption key that is used by all the data partners. Given that the key escrow is a trusted organization, selection of this partner for AD/ADRD research likely would include a third-party vendor without a conflict of interest, including non-governmental organizations with the technical capabilities to safely collect and share information.

3.2.3 Creating a Hashed Data Set

The next step in the process is undertaken by the data partners. Each data partner creates a hashed dataset by extracting PII from its operational database and passing the PII through a hashing process. The key escrow is responsible for providing encryption keys to all data partners, ensuring that the process will obfuscate the information shared in a way that makes it usable by the linkage agent to perform matching as described in section 3.3. Sharing the hashed data with the linkage agent is the final step in creating a hashed data set. It is important that the data partners do not share the encryption key with the linkage agent.⁴ In particular, the linkage agent does not see either the clear text PII or the associated health-related information (clinical data and other CDEs). Only the data owner has PII and clinical/health related info in the same place.

3.2.4 Creating and Sharing Unique Identifiers

The third step is for the linkage agent to develop a unique identifier known as a LINKID.⁵ Determining which hashed values correspond to the same individual based on the data that is shared allows the linkage agent to establish a unique LINKID for each individual.

⁴ It is important that the linkage agent only performs matching and does not have access to the encryption keys, which would allow them to decrypt the data and access the PII. This arrangement prevents any party from having access to all PII used in the matching process.

⁵ Note that LINKID is not an acronym, but rather a name for the identifier.

Once the LINKIDs have been created, the linkage agent shares the LINKIDs with each data partner. The LINKIDs can be stored by the data partners for future queries, provided the data holdings of the data partners remain constant.⁶

3.2.5 Data Sharing and Transmission

After LINKIDs are shared with each data partner, data contained within the originating EHR, RWD, and other public and private data systems can be associated with a LINKID. The linkage agent can then match records within and across public health surveillance systems using LINKIDs.

3.2.6 Selecting a Linkage Agent

The role of the linkage agent could be served by NIA or a separate third party that is both trusted and technically capable of serving in this role. The benefit to NIA serving as the linkage agent is that the linkage agent must be a trusted entity between data partners. NIA already sits in a trusted position and provides significant research funding for a wide array of studies. In this case, NIA itself would receive hashed data from data partners, link data, and provide a LINKID back to the data partners. Data partners would then submit data through data systems associated with the correct LINKID. NIA would link data by matching LINKIDs across datasets.

Alternatively, the linkage agent role could be fulfilled by a separate third party; however, all data partners would have to trust this entity. In this case, NIA would not receive hashed data; it would instead be sent to the third-party linkage agent. The linkage agent would link the hashed data and provide a LINKID back to the data partners. The data partners would then each submit their data associated with these LINKIDs to NIA, where the data could be linked by matching LINKIDs across data sets.

⁶ More specifically, the results of linkage (i.e., the LINKIDs) remain valid as long as the values of the PII attributes in the data that were used for linking the data do not change. Modification of those data values or the addition of records with those PII attributes populated could change the results of linkage.

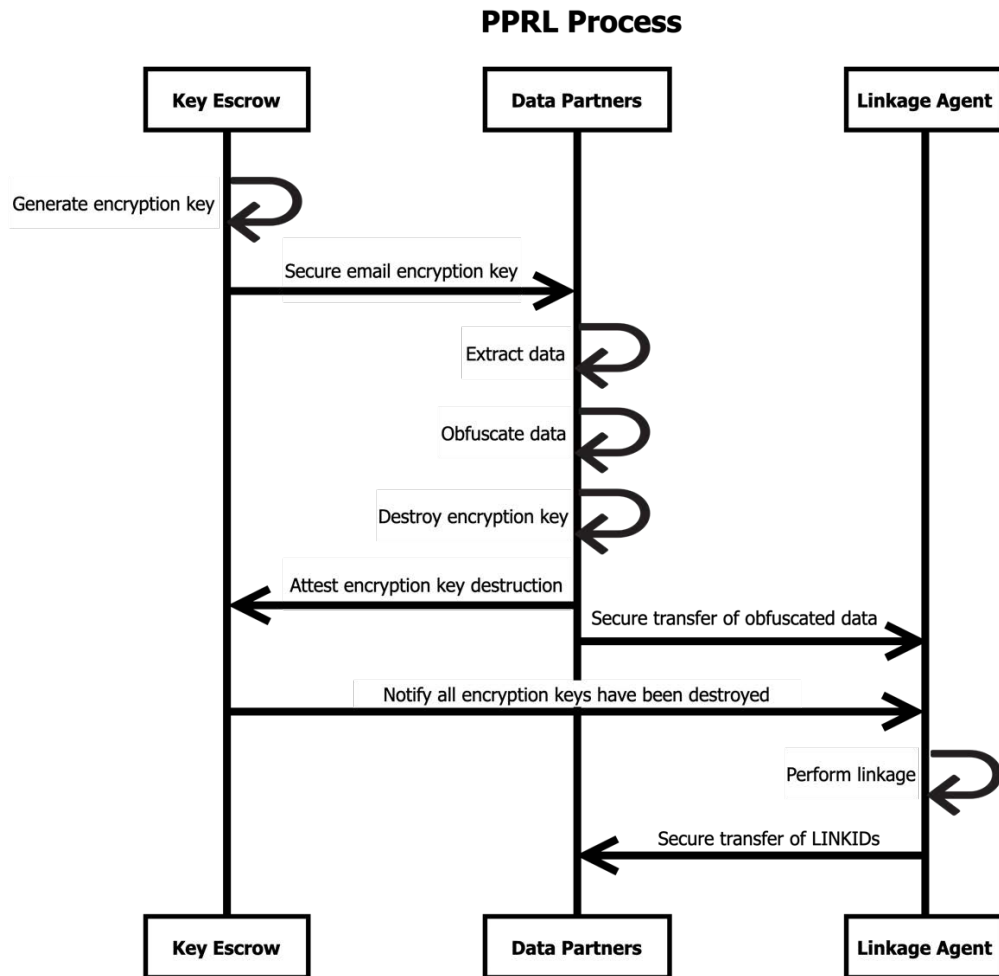


Figure 3-1: PPRL Process Flow Chart

3.3 Matching Methods

PPRL operates on the premise that PII is hashed prior to being sent to the linkage agent. There are several methods to accomplish hashing. The following section will outline three matching methodologies: deterministic, probabilistic, and a combined approach. A description of the matching method, the methodology used to perform matching, and the limitations of each method will be described.

3.3.1 Deterministic Matching

Deterministic matching relies on a linkage agent to perform exact matching—one piece of PII is an exact match to another piece of PII—based on information that is hashed by the data partner. The inputs to the hashing process are usually combined with an encryption key to prevent attacks that would reveal PII.

Hashing is an integral component of deterministic matching because it gives a way for values to be matched without revealing PII. If two hashed values are identical, then the inputs that produced

those hashed values must also be identical⁷. Thus, if two data partners separately have information about John Doe, the hashing process will result in the same value for John Doe.

Hashing is a mathematical function with two key properties. Hashing functions are:

- **Deterministic:** The same inputs always produce the same hashed output.
- **One-way:** Based on the output, it is nearly impossible to determine which inputs were used.

Deterministic matching generates multiple hashed values for the PII in each record. These hashed values are created by combining different data elements from the patient's record and hashing the resulting values. Figure 3-2 presents an example of hashed values derived from a synthetic individual record. In this example, three hashed values were generated for this record.

Sample Individual Record		
Data Element	Value	
First Name	John	
Last Name	Doe	
Date of Birth	7/4/2005	
State	MA	
Insurance Number	23-0009876	



Hash Combination Elements	Combination Value	Hashed Value
First Name + State + Insurance Number	JohnMA23-0009876	1bf6e725d482e28d774a987688c59e4faca0213f
Last Name + Date of Birth + State	Doe7/4/2005MA	ccc20f4d42210c781e1be805ce6162a8c718059c
First Name + Last Name + Date of Birth	JohnDoe7/4/2005	0a49bf0326dbdf3e29ec18a50bc71701c8c42d2

Figure 3-2: Hashing Example from a Sample Individual Record

Hashing obfuscates PII so it is less vulnerable to data attacks by adversaries. To further ensure PII is protected from adversaries, an encryption key is randomly generated by the key escrow and is provided to the data partners as an extra input to the hashing function (Table 3-1). The encryption key must be generated by the key escrow because it is critical to keep the encryption key a secret.

Table 3-1: Example of an encryption key value

Elements	Value
Encryption key	tm0eoRWdkW

Applying the encryption key to the hashed values displayed in Figure 3-2 would now yield the results presented in Table 3-2.

⁷ This is true except in the case of extremely unlikely hash collisions (two non-identical input values producing the same hashed output). In practice, hashing functions are normally designed to make collisions so rare as to be negligible.

Table 3-2: Example of hashed values with an encryption key

Data Elements	Value	Hashed Value
First Name + State + Insurance Number + Encryption Key	JohnMA23-0009876tm0eoRWdkW	99f04beb9494b66ba0530b1c6c4eae91c8ac45dc
Last Name + Date of Birth + State + Encryption Key	Doe7/4/2005MAtm0eoRWdkW	6759019832dbdfbe300fc257f0d0dbbc57378d5b
First Name + Last Name + Date of Birth + Encryption Key	JohnDoe7/4/2005tm0eoRWdkW	9b4d8ebae857caef82949a6fd87f4522afcb496b

The changed hashed values prevent an attacker from determining whether a patient has information stored at a given organization, based on the attacker’s own hash values generated from known PII (e.g., generating a hash with knowledge of John Doe’s first name, last name, and date of birth would yield a different hashed value than these same known PII values when combined with the secret encryption key).

3.3.2 Probabilistic Matching

Probabilistic matching offers a way to handle variations in demographic information by providing an alternate method to hashing data. This method uses a statistical approach to measure the probability that two records correspond to each other. Probabilistic matching does not require an exact match, which allows for variation in the underlying data (e.g., Jon Doe and John Doe).

There are several ways to implement probabilistic matching, but this report will focus on one method: Bloom filters. A Bloom filter is a data structure that offers efficient storage of information. Like deterministic matching, data partners still encrypt PII and include an encryption key in the probabilistic matching process. However, unlike deterministic matching, this approach to probabilistic matching fragments PII prior to the addition of the encryption key and input into the hashing process. The outputs of the encrypted and hashed fragments are then used to construct Bloom filters.

Each Bloom filter starts with a list of 0s that is a set length. When the encrypted fragments of PII mentioned above are added, a Bloom filter goes through a process where some of the 0s may be changed into 1s. This process starts by taking bits of data and passing it through a hashing function. This will result in a hash value for the item. The hash value is divided by the length of the Bloom filter and the remainder is calculated through the modulo operation. The remainder of this division is used to count to the position in Bloom filter’s list to see if it is a 0. If that value is still a 0, it will be changed to a 1. In some cases, the value in the list is already a 1. If that is the case, no change happens. This process is repeated multiple times for each fragment of PII, using different hashing algorithms or by using different encryption keys for the hashing process. This means that when adding an item, the item may set multiple places in the list to 1.

Using a Bloom filter, a data element, such as name, is broken into smaller segments that are hashed individually resulting in multiple output values (see Figure 3-2 for an example). For example, the name “John” could be fragmented into “Jo,” “oh,” and “hn.” The intent behind this process is to

increase the likelihood that variations in the data element (e.g., John vs. Johnny) will lead to a match across some portion of the various combinations. Table 3-3 presents an example of how a Bloom filter could be created for the name “John.” In this case, the process breaks the name into two-character segments, and hashes each using the same two encryption keys. The example Bloom filter is 64 bits long.

Table 3-3: Example Bloom filter construction

Name Fragment	Encryption Key	Combined Value	Hash Value	Bit to Set (Hash mod 64)
Jo	tm0eoRWdkW	Jotm0eoRWdkW	f8c6c76e3d4f69b42ed2d233591212fe0187c106	6
Jo	sLJp9wvfpY	JosLJp9wvfpY	177cfa71b1826df0968d343410bd88a199969731	49
oh	tm0eoRWdkW	ohtm0eoRWdkW	2b21bb44fff52320149bddd35266bfbbf1680ba6	38
oh	sLJp9wvfpY	ohsLJp9wvfpY	d7dd3104605c9680f6091c1c27f4736d5673fee6	38
hn	tm0eoRWdkW	hntm0eoRWdkW	3906a4eb6bbc4edd938e57895a657f5d39ba6c90	16
hn	sLJp9wvfpY	hnsLJp9wvfpY	be4d1c934e4d74b2139f8f4a55fc8ec0ec4e2689	9

It should be noted in the example calculations in Table 3-3 that two different encryption keys were used for the name fragment “oh,” which resulted in two different hash values. Passing these hash values through the modulo operation resulted in setting the same bit. While Bloom filters are space efficient, these situations in which different inputs result in the same bit being set are possible. When two different inputs generate the same output, it is called a collision.

The resultant Bloom filter from the Table 3-3 example is: 000000**1**00100000010000000000000000000000000010000000000100000000000000. This shows five distinct cases when the value for the filter has been set to “1,” as indicated in Table 3-3. In this example, the first name fragment will set the value at the index position of 6 (bolded), which is the 7th bit in the Bloom filter⁸.

3.3.3 Comparing Bloom Filters

Bloom filters are often used to check for probabilistic set membership. The Bloom filter generated in the previous section allows for checking for the presence of certain name fragments. As an example, the name “Johnathan” could be broken down into fragments, and following the same encryption key and hashing procedure, the presence of the “Jo,” “oh,” and “hn” fragments would be reported as true. However, fragments “na,” “at,” “th,” “ha,” and “an” will likely report as false.

⁸ As in many computational applications, counting of positions here starts with 0, so the first position in the Bloom filter is position 0, the second is position 1, ... and the seventh is position 6. This allows for setting the 0th position in the Bloom filter for data elements for which the length of the Bloom filter divides evenly into the hash value produced by that data element, i.e., the result of the modulo operation (remainder) is 0.

It is not possible to definitively state that these fragments will return false due to the possibility of collisions, as discussed in the previous section.

Instead of testing for the presence of name fragments for “Johnathan,” it is possible to construct a separate Bloom filter from this name using the same two encryption keys that result in two different 64-bit arrays. The array generated for “John” and the array generated for “Johnathan” can be compared by calculating a Dice–Sørensen coefficient (DSC), sometimes referred to as a Dice coefficient. Calculation of this metric starts by tabulating the values in Table 3-4.

Table 3-4: Dice coefficient

Value	Definition
True Positive (TP)	The bit at a given position in the first Bloom filter is set to 1 and the corresponding bit in the second Bloom filter is also set to 1
False Positive (FP)	The bit at a given position in the first Bloom filter is set to 0 and the corresponding bit in the second Bloom filter is set to 1
False Negative (FN)	The bit at a given position in the first Bloom filter is set to 1 and the corresponding bit in the second Bloom filter is set to 0

These terms can then be used in the equation in Equation 3-1

Equation 3-1: Dice Coefficient

$$DSC = \frac{2TP}{2TP + FP + FN}.$$

DSC provides a value between 0 and 1. Comparing Bloom filters with the exact same inputs would result in a coefficient of 1. Comparing filters created from increasingly dissimilar inputs will result in a value closer or equal to 0. Returning to the previous example, if a record has a given name of “John” and another record has a given name of “Johnathan,” Bloom filters can be created for these two separate records. These can be compared, and the resulting Dice coefficient value can be used to determine if the records match the same individual within a certain tolerance.

Using this approach, data partners can build Bloom filters based on individuals’ identity information. The Bloom filter approach can be accomplished through a single Bloom filter that includes all identity information or with multiple Bloom filters for each individual. In the latter approach, filters are constructed for a subset of data elements or even a single data element.

Data partners transmit the Bloom filters to the linkage agent. The linkage agent will compare these Bloom filters to determine if matches exist across organizations. Filters that have a Dice coefficient above a particular number established for the matching process are considered a match.

3.3.4 Combined Matching Approaches

A matching system can combine deterministic and probabilistic matching in successive rounds to improve match recall and precision.

A linkage agent may execute a deterministic matching phase to identify exact matches across data partners. The deterministic matching phase may incorporate multiple rounds of deterministic matching to compare hashes constructed from different data elements. Assuming each data partner has perfectly cleaned their data such that each has only one record for each unique individual in their data sets, the identities paired across data partners in these rounds can then be removed from the potential pool of matches in later rounds of the process.

After deterministic matches have been found, the linkage agent could employ a probabilistic approach to find non-exact matches. Like deterministic matching, there can be several rounds of probabilistic matching using Bloom filters that have been constructed from different combinations of data elements. Multiple Bloom filters may be considered in a matching process, combining the resulting Dice coefficients of comparisons through a weighting formula.

3.4 Available tools for use in health and healthcare

3.4.1 Search Strategy

MITRE performed a market scan of publicly available literature to find current identity resolution tools that might employ PPRL as a matching technique. The scan searched for any PPRL-like capability even if referenced using different terminology. The PPRL moniker was coined in recent years for research and functionality, but the technique has previously existed under other names. The search terms included “anonymous [entity/identity] resolution,” “record [linking/matching/fuzzy-matching] on cryptographically hashed attributes,” and “privacy-preserving record linkage.”

The list of tools presented in Table 3-5 builds on a prior evaluation of PPRL tools conducted by MITRE in late 2019 and 2020. The tools listed are both open source and commercially available and are described based on company brochures and technical papers. 15 tools were identified as using PPRL as a matching technique and are included below. Note that this market scan does not represent a complete list of all available PPRL tools and is instead a selection based on the search and prior resources as described here. It is also important to recognize that companies may offer PPRL or similar cryptographically hashed attribute matching, but do not include it as readily available public information; vendors of this type could not be included in the present market scan.

3.4.2 Commercial and Open-Source Options

The market scan includes both commercial and open-source tools. Open-source software is usually available free of charge. The benefit to consumers of open-source software is zero-to-low up-front cost, an ability to see and modify the source code, and availability to community-produced enhancements. However, there is risk to open-source software due to the potential for minimal support, software dependency on deprecated infrastructure, and abandonment by a once-flourishing maintenance and user community. These issues may lead to deferred, unplanned costs when implementing solutions with open-source software.

This market scan lists open-source software under two licenses: the General Public License (GPL) and the Apache License. The GNU⁹ GPL or GPL is a series of widely used free software licenses

⁹ GNU is a recursive acronym for “GNU’s Not Unix!” because of its Unix-like design but unlike Unix, it is free software and contains no Unix code.

that guarantee end users the freedom to run, study, share, and modify the software. The Apache License is an even more permissive free software license written by the Apache Software Foundation. GPL is a copyleft¹⁰ license, meaning the license agreement transfers to any derivative software. The Apache License on the other hand, does not require that any derivative software, or modifications to the original software, be distributed using the same license. The result is that Apache licensed software can be modified and repackaged in proprietary software.

¹⁰ Copyleft licenses for software require that information necessary for reproducing and modifying the work must be made available to recipients of the binaries.

3.5 Market Scan of PPRL Tools

Table 3-5: Market scan of available PPRL tools

Tool Name	Tool Description	License	Health Domain Use	Links
ANONLINK	Anonlink is an open source (Apache 2.0) suite of technologies developed by the Confidential Computing team at Data61 that allows organizations to carry out PPRL. It uses cryptographic hashes, blocking, and Bloom filters. Anonlink is written in C++ and provides an interface to Python. Anonlink is modular, consisting of different libraries for generating hashes, calculating similarity scores, and offering an entity service from which the clients can request mappings. Anonlink's entity service is designed for deployment to a cluster environment, and readily scales horizontally. The system has been designed to scale across multiple nodes and handles node failure without data loss. Anonlink also has several parameters for tuning to achieve better results.	Open	Yes	Information (Anonlink) Source Code (Anonlink)
CURL	Colorado University Record Linkage (CURL) system is a commercial (available for exclusive licensing, non-exclusive licensing, and co-development) platform developed by University of Colorado to perform record linkage operations, including linkage configuration and management, data normalization, data encryption and hashing, data linkage, data deduplication, and linked data dissemination. CURL uses SHA2 and Bloom filter-based hashing methods to support deterministic and probabilistic data blocking and linkage. CURL platform supports both centralized and distributed record linkage. CURL implements modular software architecture, which allows each module in CURL to be updated, added, or removed without the need to change the source code of the core platform.	Commercial	Yes	Information (CURL)
Datavant	Datavant is a commercial platform developed to allow customers to connect de-identified patient records across datasets. Datavant's tools de-identify data using HIPAA-certified, site-specific, Datavant Patient Keys (encrypted tokens) to reduce the risk of unauthorized access and patient re-identification. Datavant's sophisticated approach to matching is flexible and matches records across datasets with industry-leading precision and recall.	Commercial	Yes	Information (Datavant)

Tool Name	Tool Description	License	Health Domain Use	Links
GRLC /PPRL R Package	German Record Linkage Center (GRLC) PPRL R Package is an open source (available for free on CRAN, GPL-3) toolbox developed by GRLC for deterministic, probabilistic, and privacy-preserving record linkage techniques using R. It combines Merge ToolBox with current privacy-preserving techniques.	Open	No direct information	Information (GRLC) Source Code (GRCL)
GRHANITE®	GRHANITE® extracts and curates the delivery of sensitive data to secure research data storage facilities, with data custodian permission, and subject to ethics committee approval and legal agreement. It can extract from most data repositories and is independent of other health-related software. Since its development in 2007 GRHANITE® has been installed in over 1,000 general practices and other organizations in a variety of organizations around Australia. GRHANITE® enables PPRL, for example, linking GP, pathology, hospital, and registry data using irreversible ‘hashes’ or ‘signatures’ from PII before the data leaves the host computer. When information is extracted using GRHANITE® from multiple organizations, the signatures provide a mechanism to link records.	Commercial	Yes	Information (GRHANITE) Publication (GRHANITE)
FRIL + LinkIT	LinkIT is an open-source tool developed by Emory University that uses a secure transformation technique for linking sensitive string attributes. The Fine-Grained Records Integration and Linkage Tool (FRIL) is an open-source tool developed jointly by Emory University and CDC that performs the linkage on the attributes. LinkIT is implemented in Java and can run as a stand-alone software for linking string attributes or serve as a preprocessing module to FRIL to provide a complete record linkage toolkit for a variety of data attributes.	Open, Academic	No direct information	Information (FRIL) Information (LinkIT) Source Code (FRIL)
PRIVATEER	PRIVATEER is a PPRL research prototype developed by Hellenic Open University and University of Macedonia. This toolkit enables practitioners to compare various techniques involved in the PPRL process and determine the best technique for their needs. The toolkit is based on a simulator, designed to be highly configurable, modular, and extensible, allowing the user to test different configurations by combining several privacy-preserving blocking and matching methods with corresponding distance and similarity measures on their own or sample data.	Academic	N/A	Information (PRIVATEER)

Tool Name	Tool Description	License	Health Domain Use	Links
Senzing	Senzing is a commercial product that is a plug-and-play artificial intelligence module for entity resolution. The Senzing Application Programming Interface is designed for software and data engineers who need to quickly and affordably add highly accurate entity resolution to applications and services. It quickly matches and links data about people, organizations, and other entities. It supports full attribution, favors false negatives, self-corrects false positives, tethers data, and performs selective field hashing. IBM licenses Senzing for use in its big data analytics products. Senzing is also available as an application. By default, Senzing entity resolution (ER) uses an HMAC-SHA2-256 one-way hashing algorithm with a 1024-bit secret key.	Commercial: Free Download for up to 100K Records	No direct information	Information (Senzing)
PRIMAT	PRIMAT (Private Matching Toolbox) is an open source (Apache 2.0) toolbox developed by the Database Group of the University of Leipzig, Germany, for the definition and execution of PPRL workflows. It offers several components for data owners and the central linkage unit that provides state-of-the-art PPRL methods, including Bloom-filter-based encoding and hardening techniques, locality-sensitive hashing-based blocking, metric space filtering, post-processing, and more. PRIMAT provides two separate applications: one for data owners and one for the linkage unit. The data owner application consists of components for pre-processing (data cleaning and standardization) functions and Bloom-filter-based encoding of records containing person-related data. The linkage unit application provides linkage functionalities, including blocking, similarity calculation and classification, and post-processing. Furthermore, it consists of evaluation facilities to compare different PPRL workflows in terms of quality (recall, precision, f-measure) and scalability (runtime, reduction ratio).	Open	No direct information	Source Code (PRIMAT)
Enveil (Encrypted Veil)	Enveil's ZeroReveal® solution allows entities to securely derive insights, crossmatch, and search data assets without ever revealing the contents of the search itself or compromising the security or ownership of the underlying data. Organizations use ZeroReveal® to query data securely and privately across privacy jurisdictions while ensuring that the content of the interaction and the results are never exposed. By keeping sensitive search terms encrypted throughout the entire processing lifecycle, users can securely collaborate with multiple data sources, even when using highly sensitive or regulated indicators, to make improved intelligence-led decisions. Enveil uses Privacy Enhancing Technologies to deliver enterprise and consumer privacy and security across verticals. Users can perform operations such as searches and analytics on data without ever decrypting the content of that search or analytic.	Commercial	No direct information	Information (Enveil)

Tool Name	Tool Description	License	Health Domain Use	Links
Privitar Securelink	Privitar is a UK-based company focused on enterprise applications for privacy protection. Securelink is part of Privitar's data privacy platform, which allows organizations to create de-identified datasets via an API or user interface and focuses on privacy protection. It is a commercial tool which has been approved by the UK government for national health system data.	Commercial	Yes (NHS)	Information (Privitar)
AtyImo	AtyImo is an open-source record linkage application developed by the Federal University of Bahia in Brazil and University College London. AtyImo is implemented in Python and is aimed at linking very large databases via parallelization on GPUs. The software provides functionalities for Bloom filter encoding and comparison using both deterministic and probabilistic matching techniques. AtyImo was developed in part to link millions of individuals from the Brazil national health system and achieves high accuracy using probabilistic methods.	Open	Yes	Source Code (AtyImo)
HealthVerity	HealthVerity was founded with the goal of providing the healthcare industry a more accurate and reliable way to combine disparate datasets in a HIPAA-compliant manner to reveal patient insights and power innovation. HealthVerity is capable of matching very large datasets and claims high accuracy identity resolution. HealthVerity's Census technology supports and deploys privacy preserving record linkage at the patient level. They use cryptographic hashes (standard SHA-256 algorithm with salts) and Bloom filters for probabilistic matching.	Commercial	Yes	Information (HealthVerity)
BRICS	The Biomedical Research Informatics Computing System (BRICS) is a health records management ecosystem of tools, that includes a GUID-based PPRL solution. It was developed to support multiple disease-focused research programs and began as a joint effort between the Department of Defense (DOD), Department of Health and Human Services (HHS)/National Institutes of Health (NIH): National Institute of Neurological Disorders and Stroke (NINDS), and the Center for Information Technology (CIT). The BRICS Global Unique Identifier (GUID) client is a JavaScript application that runs via web-browser. PII or PHI is encrypted with a one-way hash, which is sent to the GUID linkage agent for matching. GUID client provides a set of necessary and optional data fields required to obtain a valid GUID.	Government	Yes	Information (BRICS)
IQVIA	IQVIA and IQVIA's product Linguamatics provide a suite of natural language processing machine learning technologies and pipelines for a range of areas including healthcare. IQVIA provides services including data cleaning, data extraction, and model training, evaluation, and deployment. IQVIA also works to structure, normalize, and de-identify/anonymize data to HIPAA standards. More information is needed to fully define IQVIA's PPRL capabilities.	Commercial	Yes	Information (IQVIA)

4 Limitations of PPRL

PPRL is intended to preserve the privacy of PII while enabling matching. However, like other record matching techniques, the PPRL process has several weaknesses that can impact the security and privacy of PII: (1) the PPRL process is vulnerable to adversary attacks resulting in breaches in privacy, (2) poor data quality and errors in matching can degrade **linkage quality**, and (3) there are computational limitations to **scale** the PPRL process. Data breaches are also discussed in this section.

4.1 Limitations of Specific Methods

4.1.1 Limitations of Deterministic Matching

Deterministic matching allows for identification of only exact matches in information. For example, if Organization A has a record for “Jon Doe” and Organization B has a record for “Jon Doe,” it will create a match. This approach does not, however, handle any variation in the information across organizations. Referring to the previous example, a subtle change in the name by one organization (e.g., adding an “h” to the first name for “John Doe” or including a middle name or middle initial for “Jon Q. Doe”) will cause the match to be missed. Because data in real-world information systems are rarely pristine, a more robust approach to matching is required.

Further, when using an encryption key, an additional level of coordination is required for two reasons. The first reason is the importance of all data partners using the same encryption key value as an input to the hashing function; if data partners used different encryption keys, matches would not be identified. The second reason is that the encryption key must also be kept a secret to prevent attacks and therefore must be destroyed after the data is hashed. This is a critical step in preserving the privacy of PII.

4.1.2 Limitations of Probabilistic Matching

Several issues can arise when using the approach to probabilistic matching described above, most of which are related to the use of Bloom filters. First, the storage space efficiency provided by Bloom filters is obtained through the loss of information. Secondly, the size of the Bloom filter has an impact on matching performance. Adding fragments into the Bloom filter changes bits from 0 to 1. There is no process that changes those bits back to 0. If too much information is placed into the filter, there is the potential that many or all bits will be set to 1. Referring to the example 64-bit Bloom filter created in Table 3-3, it is possible to also create fragments for a family name, date of birth, and address and pass them through the same encryption, hashing, and modulo operation. When these additional data elements are introduced to the Bloom filter, many more bits are set to 1. Given enough data, all bits for the Bloom filter will be 1.

When too many bits are set to 1 on a Bloom filter, there is an abundance of false positives, as the Bloom filters for all records appear similar. A solution to this issue is to increase the size of the Bloom filter. However, increasing the size of the Bloom filter increases (1) the size of information that must be transmitted between the data partner and linkage agent and (2) the computing resources needed by the linkage agent to compare the filters.

Finally, there is a trade-off in the combination of data elements used when building Bloom filters. Placing many data elements into a single Bloom filter increases system recall (the ratio that

provides the fraction of the correct possible answers that the linkage algorithm found) and reduces precision (the ratio that provides the fraction of the identified matches that are correct). An example of recall is when a Bloom filter is constructed based on given name plus a family name: a record that swaps those names so that the family name comes before the given name will still result in the same Bloom filter, and thus a match between the records. As an example of precision loss, the combination of city of residence with name information can lead to false positives because individuals living in Austin, TX, will have a greater likelihood of matching individuals with the given name Austin. Thus, it is important to carefully consider the selection of elements in construction of effective and efficient Bloom filters.

4.1.3 Limitations of a Combined Matching Approach

It should be noted that removing exact matches from consideration in later (probabilistic) rounds of the matching process can prove problematic in several circumstances. First, data providers may have multiple records for an individual in their data holdings, with one or more variations in data elements between records. If these records are not internally (probabilistically) matched prior to exact matching across data providers, records may be removed from consideration that should be considered for cross-provider probabilistic matching. Second, if the data elements chosen for exact matching are not sufficient to uniquely identify individuals, they may be erroneously matched and prevented from being paired with their actual matching record, which may contain some data errors.

4.2 Privacy and Data Breaches

One of the largest vulnerabilities when linking records is safeguarding privacy and confidentiality of PII throughout the process. The types of privacy attacks and the number of parties (e.g., data partners) involved in the PPRL process are variables that can influence privacy.

4.2.1 Types of attacks

There are different types of attacks that could make PII vulnerable despite the PPRL process (Vatsalan, Sehili, Christen, & Rahm, 2017).

A “dictionary attack” occurs when the attacker uses precomputed tables to transform the hashed data back into identifying data. An attacker could be an outside adversary or one of the data partners acting in bad faith. PPRL can largely overcome the risk of dictionary attacks by using an encryption key. As mentioned in Section 3.2, the encryption key is generated through a key escrow and shared through a secure protocol to data partners (Dong, Randolph, & Rajanna, 2020). The addition of the encryption key value causes a change in the output of the hashing function and prevents dictionary attacks because the adversary’s hashed lookup table values will no longer yield the same output as the value with the encryption key. It acts as a sort of password and should be protected as such; if an attacker learns the encryption key, a dictionary attack again becomes possible.

Another common attack is a “frequency attack” where an attacker matches the frequency distribution of a set of masked values with the distribution of known unmasked values to infer the original values of the masked values. This attack can be overcome by adding random records or noise to distort the frequency values, or by selecting combinations of data elements to hash for which the corresponding frequency values are not generally widely available.

As the number of parties involved in the PPRL process increases, PII is at higher risk for a “collusion attack.” In this type of attack, data partners could work together, or a data partner could collude with the linkage agent, to gain information on another data partner’s data.

A taxonomy of attacks on PPRL including adversary aspects, technical aspects, and practical aspects is provided in detail in Vidanage, Ranbaduge, Christen & Schnell (2022).

4.2.2 Number of Parties

In addition to the increased risk of a collusion attack, the risk of the misuse of the encryption key also increases with the number of parties (e.g., data partners) involved for two reasons. First, the encryption key must be common between all data partners as they hash their data, otherwise no matches would be identified. As the number of data partners increases, this level of coordination may become difficult, especially if data is collected from data partners in an ongoing manner. The encryption key would have to be provided and destroyed continuously in a coordinated way. Second, it is critical to keep the encryption key a secret, otherwise it is possible to trace hashed data back to its associated PII.

Increasing the number of data partners increases the risk that the encryption key is inappropriately revealed as more data partners need to keep the encryption key secret. Appropriately managing the encryption key in the PPRL process cannot be overstated. Without its proper use, destruction, and secrecy, the entire purpose of PPRL to preserve the privacy of PII becomes annihilated. While this coordination is easier when data collection and linkage are conducted in discrete phases, careful thought and specific processes would need to be established if there is a need to have data continually linked over an ongoing period.

4.2.3 Record Linkage and Data Breaches

MITRE reviewed the lay press and academic literature for evidence of PPRL involvement in data breaches. We searched for evidence of either failures of PPRL or instances where the use of PPRL helped to prevent the exposure of PII during an unrelated data breach.

The search terms used focused on the same areas covered by the market scan in section 3.4.1 (“anonymous [entity/identity] resolution,” and “record [linking/matching/fuzzy-matching] on cryptographically hashed attributes” along with “data breach”). Additional searches looked for evidence of data breaches related to any of the tools identified in the market scan. MITRE also searched for instances of “encryption key leak” or “dictionary attack” in recent articles.

These searches yielded no evidence of PPRL involvement in data breaches. One limitation with this search process, however, is that it is difficult to show that a breach was prevented or thwarted because it can be difficult to prove that an event did not happen. Additionally, most reports of data breaches provide very little detail of the details of a data breach. Reports also show that many countries are under-reporting health data breaches, partially because many institutions are unaware that they have been breached (Birmingham, 2020). MITRE noted that the Patient-Centered Outcomes Research Institute (PCORI) requires reporting of data breaches by funded institutions but does not provide any publicly reported information about any reported breaches (Patient-Centered Outcome Research Institute, 2016).

The search did reveal informative articles listing recent data breaches and an analysis of the domains and causes of recent data leaks (“Data breach,” n.d.; Cheng, 2017). Dictionary attacks resulting in data breaches are on the rise, especially during COVID-19 due to the increase in remote

work (Venkat, 2020; Eddy, 2020). In one example, attackers organized a brute-force dictionary attack on administration account passwords of storage vendor Synology, resulting in ransomware demands by the hackers (Tung, 2019). As discussed in Section 4.2, dictionary attacks can be mitigated with the addition of an encryption key. However, if the encryption key is not properly managed and secured, data breaches leveraging encryption keys, both malicious and accidental, have resulted in situations where PII was leaked and where breaches were mitigated before PII was exposed (Goodin, 2019; Krebs, 2020).

In one malicious breach that involved collusion, coordinated insiders copied a master encryption key in South African Post Office Bank (Postbank) that resulted in the compromise of millions of accounts, costing the bank several millions in fraud and damages (Muhlberg, 2020). In an accidental breach, an Amazon engineer committed nearly a gigabyte's worth of data, including personal documents, passwords, and cryptographic keys, to a personal GitHub repository; the leak was caught in roughly 30 minutes by a third-party security firm and no sensitive information was exposed (Cameron, 2020).

The examples above are not directly related to PPRL but are provided to show that data breaches have significant consequences and to help emphasize the need to protect and manage encryption keys to mitigate attacks.

4.2.4 Privacy and Ethical Concerns

It is important to recognize that the act of linking records pertaining to the same individual may make them easier to identify, and as such corresponds with a boost in adversaries' motivation to access the more valuable data sets (Baker, Knoppers, Phillips, Enckevort, Kaufmann, Lochmuller & Taruscio 2019). There is additional concern regarding the ethical issues that may arise when linking biomedical research.

A recent NIH workshop on the policy and ethics of record linkage (National Institutes of Health, 2021) highlighted concerns with the loss of autonomy, dignity, sovereignty, privacy, and researcher trustworthiness. It was noted that not all groups and communities equally experience privacy and security issues, and that incorrect and invalid linkages can disproportionately affect certain populations. Medical experts at this workshop provided the following insights regarding these concerns:

- Obtain informed consent for record linkages from the beginning of research participant engagement and perform record linkage in ways that protect privacy.
- Articulate the research purposes for data uses and return results to participants.
- Provide transparency to participants about the value and risks of data linkage and how their data may be used, including dissemination of information useful to participants.
- Research participants should be allowed to express their permission to modify their personal preferences for record linkage over time.
- Give research participants the ability to opt-in and opt-out of record linkage.
- There is a need for clear principles and detailed ways for identifying and measuring benefits and risks for different communities, especially those who are underserved.

An additional concern with data linkage is the possibility of inadvertent discovery of a previously unknown health-risk factor that affects a participant. Baker et al. 2019 note that while participant

re-identification for return-of-results may be prohibited, or the participant’s “right not to know” may need to be enforced, in other ethico-legal contexts, the ability to return results to participants in situations where a serious, preventable condition is discovered may be mandatory. If these are legal concerns for NIA, the implemented PPRL system would need to accommodate the ability of optional re-identification.

4.3 Linkage Quality

Overall linkage quality can be influenced by the PPRL method’s ability to accurately match records and the quality of the input data. Like a diagnostic test, the validity (sensitivity and specificity, among other measures) of a test can be determined, but the quality of the specimen that is tested will also influence the test result. Implementation of PPRL will have to consider the quality of the PPRL method to perform linkages and the overall quality of the data.

4.3.1 Assessment of Quality

To address the PPRL tool’s ability to accurately match records, methods can be used to perform a quantitative assessment of the quality of linkage performed by the linkage algorithm. This is typically done through an analysis of the linkage algorithm using metrics such as precision (positive predictive value) and recall (sensitivity).

4.3.2 Data Challenges

In terms of data quality, real-world data introduces a set of challenges for record linking including typographical errors, missing entries, differing coding schemes, and many other variations—foreseen or otherwise. These data challenges result from several sources. A non-exhaustive representative sample includes:

- Legitimate alternate representations for data items, for example:
 - “Street” vs. “St.” or apartment numbers in addresses: “no. 123,” “#123,” “Unit 123,” or simply “123”
 - Separators or other characters in (international) phone numbers: +1-202-555-1212, (202) 555-1212, 12025551212
 - Spellings or punctuation in names: Wm A Smith, Wm. A. Smith, William A Smith, William A. Smith, Bill Smith
- Differences in data models or standards used by data partners, for example:
 - Full Name vs. Given Name and Surname in separate fields; differing order for full name; inclusion or exclusion of middle name; preference for middle name or initial (these differences applied to the “William A Smith” example above would yield another entire set of valid varying name attributes)
 - Character set standards: UTF 8 (e.g., Nuñez) vs. lower-ASCII-only (e.g., Nunez)
 - Standards for date representation: 10/12/72 vs. October 12, 1972
 - Differing field length limits leading to truncation of data items by some but not all linkage partners
- Data errors:
 - Typographical errors

-
- Missing data or placeholder data: e.g., UNK, N/A, NULL

4.3.3 Quality Improvement

Linkage quality can be improved if data partners take steps to prepare their data, such as resolving data entry errors and using standardized data fields across data partners, before hashing occurs.

Ideally, record linking models should have capabilities beyond exact matching, which could otherwise miss matches due to such “messy” data. Models should be both flexible enough to deal with real-world data and effective enough to achieve high accuracy matching. In practice no measures taken will completely resolve all data issues. Furthermore, it is important to be cognizant that “over cleaning” data can result in signal being removed along with the noise, thus adversely affecting linkage quality. Real-world implementations must balance these considerations.

4.3.4 Linkage Across Multiple Providers

An assessment also was performed of the ability to determine what mapping activities have occurred to date between different providers. The MITRE team was able to verify 1-1 and 1-2 mappings by specific vendors. For example, IQVIA, a large data vendor, verified their ability to “bridge” between tokens. In brief, they represent that they can link between their token and the token developed by Datavant. They also expressed having the capability to develop bridging between other tokens but would explore those scenarios through specific use cases rather than providing this as a general service.¹¹

In general, nothing in the PPRL technologies available from the vendors identified either in Table 3-5 or Table 5-2 would prevent mappings beyond the 1-2 mappings available in the market. Instead, proprietary and ethical considerations, such as the possibility of re-identification especially for vendors who do not vary the “salt”, or random string used in their underlying hashing technique (see section 3.2.3), in order to build large internal data sets.¹² Additional research may be needed to verify that any selected PPRL technology that implements mappings across different data providers is indeed privacy-preserving.

4.4 Computational Scalability

The volume of data and the number of parties involved in linking data creates a computational challenge—comparing all records to all other records becomes unrealistic when the number of records per database and the number of databases increase. A naïve comparison of all record pairs in two databases where one database has m records, and the other database has n records would result in $m \times n$ comparisons. This number quickly becomes intractable when m and n are large numbers. To address this limitation, blocking and filtering techniques can be introduced to improve scalability by reducing the number of comparisons between non-matching data records.

Blocking and filtering processes are standard techniques that work to reduce the number of comparisons that need to be conducted between records by pruning as many pairs or sets of records as possible that are unlikely correspond to matches (Vatsalan, Sehili, Christen, & Rahm, 2017). Blocking will group similar data together according to one or more blocking criteria. Comparisons are then performed only for records that are in the same block. Filters reduce comparisons by

¹¹ This information was based on an informal discussion with IQVIA.

¹² This concern was raised by MITRE internal experts about at least one of the recommended vendors, Datavant.

removing potential non-matches based on their properties. This step can be done either by the data partners prior to hashing data or on the hashed records.

Limited literature has shown that PPRL can sufficiently address the limitations in computational scalability and be applied to large population-level datasets. Randall et al. reported on an Australian effort involving two databases with approximately twenty million and seven million records, respectively. Randall et al. claimed their study was the “first time that privacy-preserving linkage using the Bloom filter method for approximate string comparison has been applied to large, population-level data collections” (Randall, Ferrante, Boyd, Bauer, & Semmens, 2014). Additionally, commercial software may be engineered to handle large numbers of participants and records;¹³ however, commercial organizations tend to publish fewer implementation details than researchers.

¹³ For example, Datavant claims that “Datavant's technology has been successfully used since 2014. The Datavant ecosystem consists of over 200 data partners. In the last year, [Datavant] processed well over 100 billion records.” <https://datavant.com/how-we-do-it/> (accessed 6/17/2020)

5 Conclusion

5.1 Current and Potential Future State of PPRL

PPRL is a cutting-edge technique that can match records from many data sources across institutions and jurisdictions without relying on PII and has been integrated into many areas within healthcare and public health. However, MITRE was able to find two examples in its research of PPRL employed in ongoing public health domains. The first is the ACCESS program based in Australia (Nguyen et al., 2020), in which the scope was limited to sexually transmitted infections and blood-borne viruses and only pulled datasets housed in the ACCESS database. The second is Privitar, which is a UK-based company that links NHS healthcare data between millions of patients (Privitar, 2022).

5.1.1 Potential for Impact

While PPRL has significant theoretical potential to link data across NIA funded studies and data assets, further research is needed to determine PPRL will function within the AD/ADRD research infrastructure. PPRL could serve as a universal tool to link clinical, process, outcomes, universal, economic, and other types of data. The ongoing need to collect data from a diverse set of nationwide data sources to inform evidence-based healthcare interventions makes this a challenge for NIA, for NIH, and for the health research community. There are several barriers to the successful use of PPRL that need to be addressed.

5.1.2 Ensuring Effectiveness

- **Stakeholders must be willing and able to conduct the PPRL processes and share data with a third party.** Each data partner must prepare its PII data for PPRL processing according to a standard protocol, hash the data to generate obfuscated files, and transmit the data to a linkage agent for record linkage. Prior to data transmission, data partners will likely need to enter into a data sharing agreement with the key escrow. These processes take time and technical and human resources to complete.
- **PPRL tools should require minimal IT support to implement and run.** Data partners may not have IT resources or staff to accommodate a complex installation process or subsequent maintenance. Data partners must ensure their data information systems can consistently extract and process the data in the form required by the PPRL tool.
- **There must be strong coordination of activities across data partners.** An essential part of the PPRL process is using a shared encryption key across all data partners. Before the hashing process begins, the encryption key is added to ensure PII cannot be traced back to the originating data source if an adversary gains access to the hashed data. The encryption key is provided by a key escrow, and the same encryption key must be used by all data partners as an input to the hashing functions. If different encryption keys are used, matches would not be identified. Further, the encryption key must be kept confidential; otherwise, an adversary may be able to re-identify the data. Careful coordination is necessary to ensure (1) all data partners receive the same key; (2) while active and available, the encryption key is kept safe and secure from potential adversaries; and (3) the encryption key is destroyed by all data partners upon completion of the hashing process to ensure

confidentiality of the data. Failure to complete any of these coordination steps increases the risk that individuals' privacy and confidentiality is breached.

- **The process must be seamless and invisible to data users.** Data users are typically experts in clinical practice, trials, or health research. They likely lack the technical background or resources to implement any aspect of the PPRL approach other than sending and receiving data. Any implementation of PPRL needs to be designed with these users in mind to lower the burden on researchers, allowing them to focus on scientific discovery and advancement in their fields.

5.1.3 Coordinating Key Destruction

Additionally, and perhaps most significantly, coordinated destruction of the encryption key could be challenging. Destroying the encryption key after hashing data is critical to the overall process and security of PPRL as shown in Figure 3-1. In the ongoing collection of data for public health surveillance, managing the ongoing destruction of the encryption key across many health and healthcare data partners will require close coordination from NIA.

If an encryption key is not properly destroyed by all data partners, the nation's public health surveillance systems are vulnerable to breaches that could lead to identifying individuals. The consequences of a breach of PII on a national scale are significant and devastating. A potential solution to this issue is for the key escrow to create a new encryption key on a pre-determined (e.g., weekly) basis.

At the end of this pre-determined time, the encryption key is destroyed by all data partners, and a new key is generated. New hashed values and any corresponding LINKID for that record would be provided to the linkage agent to link records. It is important to note that this solution has not been tested in real time using real data; therefore, the feasibility, maintenance of privacy, linkage quality, and potential to scale have not been assessed. NIA would need to consider these types of encryption key management mechanisms when developing use cases and pilot studies. The challenge of key destruction is particularly relevant to the NIA's approach to this technology given the diversity of data sets funded by NIA and used in NIA funded clinical trials. The destruction of data sets, such as Medicare claims data, at the conclusion of studies is an example of the type of data disposition burden that researchers currently face.

5.2 Recommended Technologies and Vendors

5.2.1 Recommended Technologies

Based on our assessment of PPRL technologies and application to NIA research, we have provided a preliminary list of recommended PPRL technologies in Table 5-1 below. These recommendations may change based on the selected use cases, data elements, and other implementation considerations relevant to NIA.

Table 5-1. Recommended Technology List

Technology Name	Rationale for Inclusion	Potential Limitations	Application to NIA Activities
Bloom filter probabilistic matching	Bloom filter technology provides the ability to match data with a variety of elements. It can also account for variation and errors in the underlying data elements.	The size of the data elements and Bloom filter may lead to false positives (too many or all 1's).	Data elements across datasets, studies, and clinical trials may not be consistent. Bloom filters provide a probabilistic matching approach to account for these variations.

5.2.2 Recommended Vendors

Based on our assessment of PPRL technologies and application to NIA research, we have provided a preliminary list of recommended PPRL vendors who may be able to implement our recommended technologies in Table 5-2 below. Note that these recommendations may change depending on the selected use cases, data elements, and other implementation considerations relevant to NIA. A thorough market survey and full evaluation on a set of established requirements would be required to provide a more in-depth set of recommended vendors.

Table 5-2. Recommended Vendor List

Technology Name	Rationale for Inclusion	Potential Limitations	Potential Roles	Application to NIA Activities
ANONLINK: Open-source option for PPRL implementation	Open-source tools have little-to-no upfront cost and the ability to see and modify the existing code. ANONLINK is widely used and well supported in the open-source community.	Open-source tools have more limited support than commercial tools and may be decommissioned or unsupported in the future.	ANONLINK acts as a linkage agent and computes similarity scores between sets of hashed entity records. NIA or a third party can use ANONLINK to perform the linkage.	Wide array of tools may facilitate different types of linkage (claims and EHR data, clinical trials) May be combined with other open-source tools for best results.
Datavant: Commercial platform developed to allow customers to connect de-identified patient records across datasets	Datavant's tools de-identify data using HIPAA-certified, site-specific, Datavant Patient Keys (encrypted tokens). Previously part of COVID-19 consortium.	May not be able to provide tokens that link across multiple data sets due to proprietary implementation.	Datavant works with the data partners to both encrypt the data and link it. In this way, Datavant is key escrow and linkage agent. Data partners use site-specific hash keys as well as a 'master key'.	Currently engaged in discussions with NIA staff about use in NIA research activities Partners include Regenstreif, Palantir
Senzing: Commercial option for PPRL implementation	Commercial tools come with full technical support and potential customization by the vendor. Senzing is well-known and backed by IBM, a large vendor with plentiful resources.	Commercial tools come with up-front costs and may offer fewer customization options than open-source tools (i.e., black box). We were unable to verify the use of Senzing in health applications.	Senzing is an API that a third-party or NIA would use to perform the linkage. Senzing has no access to PII data.	May not relate to current activities. Consider as new vendor or partner for project transition plan.

5.3 Evaluating PPRL on a Pilot Basis for NIA

Until a mechanism to continuously manage an encryption key is determined and tested, NIA should evaluate the utility and feasibility PPRL on a limited and controlled scale. As an example, this could involve developing a small number of limited-scope use cases and pilot studies with a few data partners that evaluates one type of dataset such as NIA funded clinical trials or EHR data from a single health system. Use cases and pilot studies will help NIA understand the necessary processes, metrics, challenges, and institutional and jurisdictional barriers to evaluate whether PPRL adds value to the practice of public health and its potential to scale to more complex processes.

6 References

1. Baker, D. B., Knoppers, B. M., Phillips, M., van Enckevort, D., Kaufmann, P., Lochmuller, H., & Taruscio, D. (2018). Privacy-preserving linkage of genomic and clinical data sets. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4), 1342-1348.
2. Birmingham, A. (2020). *Australia's healthcare sector may be under reporting PII data breaches: ForgeRock MD - Which-50*. Which-50. Retrieved 2 July 2020, from <https://which-50.com/australias-healthcare-sector-may-be-under-reporting-pii-data-breaches-forgerock-md/>.
3. Cameron, D. (2020). *Amazon engineer leaked private encryption keys. Outside analysts discovered them in minutes*. Gizmodo. Retrieved 2 July 2020, from <https://gizmodo.com/amazon-engineer-leaked-private-encryption-keys-outside-1841160934>.
4. Centers for Disease Control and Prevention. (2020). *NNDSS Supports the COVID-19 Response*. Retrieved 15 June 2020, from <https://wwwn.cdc.gov/nndss/covid-19-response.html>.
5. Cheng, L., Liu, F., & Yao, D. (2017) Enterprise data breach: causes, challenges, prevention, and future directions. *WIREs Data Mining and Knowledge Discovery*, 7(e1211). <https://doi.org/10.1002/widm.1211>.
6. U.S. Department Of Health And Human Services. Retrieved 25 October 2022, from <https://www.acf.hhs.gov/sites/default/files/documents/cb/im1504.pdf> .
7. Datavant. (2020). *Datavant*. Retrieved 1 July 2020, from <https://datavant.com/>.
8. Datavant. (2020). *Leading Healthcare Companies Announce COVID-19 Research Database - Datavant*. Retrieved 30 June 2020, from <https://datavant.com/news/leading-healthcare-companies-announce-covid-19-research-database/>.
9. Dong, X., Randolph, D. A., & Rajanna, S. K. (2020). Enabling Privacy Preserving Record Linkage Systems Using Asymmetric Key Cryptography. *AMIA Annual Symposium Proceedings*, 380-388.
10. Eddy, N. (2020). *Cyberattacks continue to mount during COVID-19 pandemic*. Healthcare IT News. Retrieved 2 July 2020, from <https://www.healthcareitnews.com/news/cyberattacks-continue-mount-during-covid-19-pandemic>.
11. Enveil. (2022). *Enveil*. Retrieved 20 October 2022, from <https://www.enveil.com/>.
12. Goodin, D. (2019). *Hackers steal secret crypto keys for NordVPN. Here's what we know so far*. ARS Technica. Retrieved 2 July 2020, from <https://arstechnica.com/information-technology/2019/10/hackers-steal-secret-crypto-keys-for-nordvpn-heres-what-we-know-so-far/>.
13. Krebs, B. (2020). *'BlueLeaks' exposes files from hundreds of police departments*. Krebs On Security: In-Depth Security News and Investigation. Retrieved 2 July 2020, from <https://krebsonsecurity.com/2020/06/blueleaks-exposes-files-from-hundreds-of-police-departments/>.

-
14. Muhlberg, B. (2020). *South Africa's Postbank replaces 12 million bank cards after internal security breach exposes master key*. CPO Magazine. Retrieved 2 July 2020, from <https://www.cpomagazine.com/cyber-security/south-africas-postbank-replaces-12-million-bank-cards-after-internal-security-breach-exposes-master-key/>.
 15. National Institutes of Health, Office of Data Science Strategy. (2021). NIH Workshop on the Policy and Ethics of Record Linkage: Workshop Summary. June 29-30. Retrieved 26 October 2022 from <https://datascience.nih.gov/news/nih-policy-and-ethics-of-record-linkage-workshop>
 16. Nguyen, L., Stoové, M., Boyle, D., Callander, D., McManus, H., Asselin, J., Guy, R., Donovan, B., Hellard, M., & El-Hayek, C. (2020). Privacy-Preserving Record Linkage of Deidentified Records Within a Public Health Surveillance System: Evaluation Study. *Journal of Medical Internet Research*, 22(6):e16757. URL: <https://www.jmir.org/2020/6/e16757/>. DOI: 10.2196/16757
 17. Patient-Centered Research Outcomes Institute. (2016). *PCORI policy on Data and Safety Monitoring Plans for PCORI-Funded Research*. Retrieved 2 July 2020, from <https://www.pcori.org/sites/default/files/PCORI-Policy-Data-Safety-Monitoring-Plans.pdf>.
 18. Privitar. (2022) *Unlocking Siloed Data For the NHS: Leveraging the Full Power of Sensitive Healthcare Data*. Retrieved 20 October 2022, from <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.privitar.com/app/uploads/2022/07/P1026-CS-NHS-Digital-DP.pdf>.
 19. Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). Privacy-preserving record linkage on large real-world datasets. *Journal of Biomedical Informatics*, 50, 205-212.
 20. Tung, L. (2019). *Ransomware crooks hit Synology NAS devices with brute-force password attacks*. ZDNet. Retrieved 2 July 2020, from <https://www.zdnet.com/article/ransomware-crooks-hit-synology-nas-devices-with-brute-force-password-attacks/>.
 21. Vatsalan, D., Christen, P., O'Keefe, C. M., & Verykios, V. S. (2014). An Evaluation Framework for Privacy-Preserving Record Linkage. *Journal of Privacy and Confidentiality*, 6 (1). <https://doi.org/10.29012/jpc.v6i1.636>.
 22. Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In A. Zomaya, & S. Sakr, *Handbook of Big Data Technologies* (pp. 851-895). Springer International Publishing.
 23. Venkat, A. (2020). *RDP Brute-force attacks rise during COVID-19 crisis: report*. Bank Info Security. Retrieved 2 July 2020, from <https://www.bankinfosecurity.com/rdp-brute-force-attacks-rise-during-covid-19-crisis-report-a-14206>.
 24. Vidanage, A., Ranbaduge, T., Christen, P., & Schnell, R. (2022). A Taxonomy of Attacks on Privacy-Preserving Record Linkage. *Journal of Privacy and Confidentiality*, 12(1).

Appendix A Glossary

Term	Definition
Bloom Filter	A data structure that is often used to probabilistically test the presence of an element within a set. Bloom filters are space efficient, meaning that they allow for the testing of presence in a set without requiring access to the entire set. This space efficiency is achieved by a process that can allow for false positives to be provided when testing for element presence.
Configuration File	A computer file used to define parameters and settings for computer applications.
Data Partners	Data partners are the organizations that own the data that will be shared through public health surveillance systems. Data partners can include public health departments, laboratories, healthcare organizations, vital records registrars, and immunization information systems.
Encryption Key	An encryption key is typically a random string of bits generated specifically to scramble data. Encryption keys are created with algorithms designed to ensure that each key is unique and unpredictable.
Hashing	Hashing is a mathematical function with two key properties. First, the same inputs always produce the same output. Second, given the output, it is nearly impossible to determine which inputs were used. Hashing transforms input data by shuffling and mixing up the information it is given.
Key Escrow	The key escrow, sometimes referred to as a trusted third party, is an organization that is part of the PPRL process. The key escrow provides the encryption key and configuration files that are used by all parties. This organization must be trusted by all data partners to act in their best interests; this is usually supported through a legal agreement between the data partners and the key escrow.
Linkage Agent	The linkage agent is an organization that is part of the PPRL process. Data partners provide the linkage agent with encrypted hashed values to conduct the record linkage process. The linkage agent shares the output of the linkage process in the form of unique linked identifiers back to each data partner.

Term	Definition
Linkage Quality	Linkage quality measures the degree to which records in different databases for the same individual can be correctly identified.
LINKID	A unique identifier that is assigned to an individual.
Modulo	A mathematical operation that provides the remainder after division of one number by another.
Obfuscated Data	Data that has been rendered unreadable.
Personally Identifiable Information	Any information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual.
Positive Predictive Value	See Precision.
Precision	A ratio that provides the fraction of the identified matches that are correct.
Privacy	Not revealing patient information to entities who are not authorized to view that information.
Recall	A ratio that provides the fraction of the correct possible answers that the system found.
Scalability	The ability of a process or computer system to grow to accommodate additional participants or data records.
Sensitivity	See Recall.

Appendix B Abbreviations and Acronyms

Term	Definition
ACCESS	Australian Collaboration for Coordinated Enhanced Sentinel Surveillance of Sexually Transmissible Infections and Blood-borne Viruses
AD/ADRD	Alzheimer's Disease/Alzheimer's Disease Related Dementias
BRICS	Biomedical Research Informatics Computing System
CIT	Center for Information Technology
CURL	Colorado University Record Linkage
DOD	Department of Defense
DSC	Dice-Sørensen Coefficient
EHR	Electronic Health Record
ER	Entity Resolution
FFRDC	Federally Funded Research and Development Center
FRIL	Fine-grained Records Integration and Linkage tool
GNU	GNU's Not Unix
GPL	General Public License
GPU	Graphics Processing Unit
GRLC	German Record Linkage Center
GUID	Global Unique Identifier
HHS	Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act of 1996
ML	Machine Learning
NIA	National Institute on Aging
NIH	National Institutes of Health
NINDS	National Institute of Neurological Disorders and Stroke
PCORI	Patient-Centered Outcomes Research Institute
PHI	Personal Health Information
PII	Personal Identifiable Information
PPRL	Privacy-Preserving Record Linkage
PRIMAT	Private Matching Toolbox
RWD	Real World Data
SHA	Secure Hashing Algorithm