Journal of the American Medical Informatics Association, 2024, **31(11)**, 2605–2612
https://doi.org/10.1093/jamia/ocae196
Advance access publication 24 July 2024
Research and Applications

Research and Applications

Privacy preserving record linkage for public health action: opportunities and challenges

Aditi Pathak, PhD[1], Laina Serrer, MS[1], Daniela Zapata, PhD[1], Raymond King, PhD[2], Lisa B. Mirel, MS[3], Thomas Sukalac[4], Arunkumar Srinivasan, PhD[5], Patrick Baier, DPhil[1], Meera Bhalla, BS[1], Corinne David-Ferdon, PhD[6], Steven Luxenberg, PhD[6], Adi V. Gundlapalli, MD, PhD*,[6]

[1]American Institutes for Research, Arlington, VA 22202, United States, [2]National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA 30341, United States, [3]National Center for Science and Engineering Statistics, National Science Foundation, Alexandria, VA 22314, United States, [4]Center for Forecasting and Outbreak Analytics, Centers for Disease Control and Prevention, Atlanta, GA 30333, United States, [5]National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333, United States, [6]Office of Public Health Data, Surveillance, and Technology, Centers for Disease Control and Prevention, Atlanta, GA 30333, United States

*Corresponding author: Adi V. Gundlapalli MD, PhD, 1600 Clifton Rd, Office of Public Health Data, Surveillance, and Technology, Centers for Disease Control and Prevention, Atlanta, GA 30333, United States (agundlapalli@cdc.gov)

Abstract

Objectives: To understand the landscape of privacy preserving record linkage (PPRL) applications in public health, assess estimates of PPRL accuracy and privacy, and evaluate factors for PPRL adoption.

Materials and Methods: A literature scan examined the accuracy, data privacy, and scalability of PPRL in public health. Twelve interviews with subject matter experts were conducted and coded using an inductive approach to identify factors related to PPRL adoption.

Results: PPRL has a high level of linkage quality and accuracy. PPRL linkage quality was comparable to that of clear text linkage methods (requiring direct personally identifiable information [PII]) for linkage across various settings and research questions. Accuracy of PPRL depended on several components, such as PPRL technique, and the proportion of missingness and errors in underlying data. Strategies to increase adoption include increasing understanding of PPRL, improving data owner buy-in, establishing governance structure and oversight, and developing a public health implementation strategy for PPRL.

Discussion: PPRL protects privacy by eliminating the need to share PII for linkage, but the accuracy and linkage quality depend on factors including the choice of PPRL technique and specific PII used to create encrypted identifiers. Large-scale implementations of PPRL linking millions of observations—including PCORnet, National Institutes for Health N3C, and the Centers for Disease Control and Prevention COVID-19 project have demonstrated the scalability of PPRL for public health applications.

Conclusions: Applications of PPRL in public health have demonstrated their value for the public health community. Although gaps must be addressed before wide implementation, PPRL is a promising solution to data linkage challenges faced by the public health ecosystem.

Key words: privacy-preserving record linkage; PPRL; public health; data linkage.

Introduction

Effective public health action depends on linking information about elements of prevention and care, including clinical interventions and outcomes, disease screening and testing, vaccine administration, and social determinants of health. The COVID-19 pandemic demonstrated how a lack of accurate, timely data can cripple efforts to proactively respond to a public health crisis.[1,2] The fragmentation of the US health care and public health ecosystem continues to pose challenges to connecting and linking data from disparate sources. Federal initiatives, such as the Centers for Disease Control and Prevention (CDC) Data Modernization Initiative and the Food and Drug Administration Enterprise Modernization Action Plan, have identified expanding data linkages as a key priority for a connected, response-ready public health system.[3,4] In addition, the Evidence Act of 2018 and the Creating Helpful Incentives to Produce Semiconductors (CHIPS) and Science Act of 2022 support the use of data for evidence-based policymaking while protecting privacy. The National Science and Technology Council and the National Institute of Standards and Technology have also focused on challenges to and mechanisms for privacy protection when linking fragmented data.[5]

Traditionally, linking data about individuals across disparate sources requires sharing personally identifiable information (PII). Using such identifiers raises privacy concerns due to the risk of identifying individuals and breaching protected health information (PHI). Federal, state, and tribal privacy laws, such as the Health Insurance Portability and Accountability Act (HIPAA), may also prohibit sharing PII and PHI.

Received: February 8, 2024; **Revised:** June 27, 2024; **Editorial Decision:** July 11, 2024; **Accepted:** July 12, 2024
Published by Oxford University Press on behalf of the American Medical Informatics Association 2024.

Privacy preserving record linkage (PPRL), a process for linking de-identified data, eliminates the need to share direct PII for record linkage. PPRL substantially reduces privacy and security risks and addresses a key barrier to data sharing and linkage.

While PPRL techniques have evolved substantially over the last 2 decades, bloom filter based techniques are considered as a reference standard for PPRL.[6] Applications of PPRL are emerging across domains including health care, public health, crime and fraud detection, and national security because of its potential benefit by connecting disparate data sets and improving accuracy by de-duplicating records.[7–9] Large-scale public health initiatives have adopted PPRL to enhance collaboration across health care networks, data repositories, and research networks. For example, the National Institutes of Health (NIH) National COVID Cohort Collaborative (N3C) is piloting PPRL technology to link real-world COVID-19 data to help clinicians, researchers, and patients better understand the disease.[10] CDC also used PPRL to link COVID-19 case and vaccination data and then track and analyze vaccination patterns and case counts to improve the understanding of COVID-19 vaccination effectiveness and inform public health recommendations.[11] The implications of data gaps that the COVID-19 pandemic revealed highlight the need to urgently explore the value of PPRL and increase knowledge about its applications, benefits, and implementation challenges. This study aims to understand the current landscape and value of PPRL in the clinical and public health domains, provide a high-level assessment of estimates of PPRL accuracy and privacy, and evaluate the opportunities and challenges in adopting and implementing PPRL for public health action.

Methods

Literature scan

The aim of the literature scan was to understand the current landscape of PPRL applications in public health from published peer-reviewed literature that was publicly available as of January 2023. The primary research question examined was whether PPRL is effective for data linkage in terms of accuracy, data privacy, and scalability. Although the review focused primarily on public health applications, articles applying PPRL to clinical care data were included because of overlapping data sources across the 2 fields. A search strategy was developed for the National Library of Medicine (NLM) National Center for Biotechnology Information (NCBI) PubMed literature database and Google Scholar. The review included articles published between January 1, 2010, and January 1, 2023. While the literature focusing on PPRL is extensive, literature about PPRL in public health is sparse. To limit the number of irrelevant articles for review, we used “OR” and “AND” operators and exact phrases, see Table 1. For example, several review articles about privacy enhancing technologies contain references to privacy preserving record linkage but are not focused on PPRL use cases or implementation. A combination of “privacy preserving record linkage” with “PPRL” or “hash” helped retrieve articles with multiple references to PPRL which improved identification of relevant articles that capture real-world use cases of PPRL. To identify state-level applications of PPRL, we also combined the search terms with individual state names.

Table 1. Search strategy for literature scan

Database	Search strategy	Retrieved	Retained
PubMed	((((PPRL) OR (hash)) AND (privacy preserving record linkage)) NOT (plasma)) NOT (prolactin)	29	19
GoogleScholar	PPRL OR hash “privacy preserving record linkage” -plasma -prolactin OR PPRL OR hash AND “privacy preserving record linkage” AND *statename* -plasma -prolactin	768	11

The search resulted in an initial retrieval of 797 unique articles. Because PPRL is relatively new in the public health domain, we also searched websites of federal agencies to identify PPRL applications that may not have been published. Using the terms “PPRL” and “hash” separately on these websites did not retrieve additional unique articles or documents. Titles and abstracts were scanned to narrow the initial retrieval to a subset of publications that focused on application of PPRL methods or evaluation of PPRL effectiveness; articles that aimed to develop or refine PPRL techniques were excluded. This approach resulted in an initial subset ($n = 124$) of articles. Further, of these, 84 articles were subsequently excluded that did not focus on real-world applications or use cases of PPRL in public health or clinical care. An additional 13 articles were excluded that focused on evaluating the performance metrics of specific PPRL software tools. The resulting review included the remaining 30 articles. Information about public health questions answered by the study, the funding agency, datasets and number of records, type of PPRL techniques used, and key findings categorized into accuracy, privacy, and scalability were abstracted.

Key informant interviews

The literature scan focused on understanding PPRL use cases, effectiveness of PPRL to answer public health questions, and limitations of PPRL. To understand how these aspects of PPRL translate into opportunities, strategies, and challenges to PPRL adoption, implementation, and scale-up in public health, 12 interviews with subject matter experts (SMEs) from CDC were conducted. Several of the SMEs had led PPRL implementation, including for the Clinical and Community Data Initiative (CODI),[12] the COVID-19 vaccination surveillance project,[10] and an evaluation of PPRL performance against clear text linkage methods.[13] SMEs also had expertise with IT and data governance, technology, informatics, and legal services related to PPRL. We developed a semi-structured interview protocol that focused on 3 key topic areas: *the value of PPRL* for public health action, *opportunities and strategies for PPRL adoption and scale-up, and gaps to be addressed* for successful PPRL implementation in public health. The interview protocol contained questions tailored to each interviewee’s role at CDC. The interview protocol developed for members of the PPRL strategy leaders is shown in Supplementary Appendix A.

Interviews were held virtually via Microsoft Teams and generally lasted 45 min. All interviews were recorded with verbal consent from the interviewees. Two qualitative researchers transcribed the interviews for thematic analysis

(LS, MB). Three researchers (LS, MB, and AP) reviewed transcripts and mapped text from the transcripts to organize the data into the 3 predefined key topic areas. For interview analysis, we followed an inductive approach to coding through iterative reading of transcripts and emergent coding. With iterative reading of the transcripts, we were able to delineate multiple interpretations of the raw data and identify cross-mapping of text to multiple topic areas. Emergent coding helped data-driven identification of key themes. Themes agreed upon by all 3 researchers to be representative of interviewee opinions were included in the final results. NVivo software was used to support data coding and analysis.

This activity was reviewed by CDC and deemed not to be research, and IRB review was not required. Activity was conducted consistent with applicable federal law and CDC policy.

Results

Literature scan

Accuracy

In real-world clinical data, PPRL has demonstrated linkage quality and accuracy comparable to clear-text linkage methods using unencrypted PII.[14–18] For example, using multiple data sources, including laboratory data, electronic health records, and payer claims in the Chicago region, data linkage using PPRL methods resulted in sensitivity and specificity as high as 100%.[19] Several studies comparing PPRL linkages to a gold-standard data set—where health records belonging to the same patient were already known based on clear text linkage—have also shown that PPRL produced high accuracy, with precision and recall exceeding 90%.[20,21]

Accuracy depends on several components, such as PPRL technique, and the proportion of missingness and errors in underlying data.[6,22] For example, although there was no difference in linkage accuracy when comparing PPRL to linkage with unencrypted identifiers in hospital admission records from Western Australia, the same study reported slightly lower accuracy using PPRL to link records from a different part of Australia that had a higher percentage of missing PII values.[23] Results from linking National Center for Health Statistics survey data with the National Death Index confirmed PPRL provides linkage accuracy comparable with clear text record linkage (precision ranged from 93.8% to 98.9% and recall ranged from 97.8% to 98.7%), specifically for data sets with a low proportion of missing identifiers.[12]

Data privacy

Compared with clear text record linkage, PPRL improves data privacy by design because no direct PII must be shared for linkage. Various components of PPRL implementation—including choice of PPRL method, type of PII used to generate tokens, protocols to access and store data, and participation of a linkage agent—can affect the level of privacy protection. The security of various PPRL methods has been rigorously evaluated for vulnerability to a variety of attacks and resulting modifications have led to stronger and more secure PPRL methods.[24–27] Safeguards regarding access to and the use, storage, and destruction of encrypted data are as important as the PPRL method chosen to prevent privacy breaches. For example, the HIPAA expert determination method where a privacy expert must certify that the dataset produced from the proposed linkage involves a low level of risk based on the data set, purpose of data usage, and ability of an anticipated recipient to identify an individual is a standard practice to ensure privacy and prevent re-identification.[28–30]

Scalability

Recent large-scale record-linkage efforts have demonstrated the scalability of PPRL for public health. The National Patient-Centered Clinical Research Network (PCORnet) project successfully linked more than 170 million records to accurately measure clinical characteristics and disease prevalence.[31] Similarly, NIH N3C and the CDC COVID-19 project linked millions of records from various sources and demonstrated the scalability of PPRL for public health surveillance.[9,10]

Supplementary Appendix B provides a glossary of relevant terms and describes a typical workflow for PPRL implementation involving 2 data owners. Table 2 lists large-scale applications of PPRL identified from the literature scan that demonstrate the value of PPRL for public health.

Limitations and opportunities of PPRL. PPRL is intended to protect the privacy of the individuals who are being linked. However, the literature scan also revealed weaknesses of PPRL that can impact the security and privacy of PII. The linkage quality of PPRL relies on high quality PII to construct the hash tokens, and poor data quality and linkage errors can degrade linkage quality.[16,12,23] In addition, PPRL is computationally intensive which can pose a cost and resource burden and limit the scale-up of PPRL in public health. While PPRL does not require sharing of PII, different types of adversary attacks could make PII vulnerable despite the PPRL process.[39] But recent developments such as BLIP (BLoom and flIP) method for bloom filter hardening (a differentially private method for flipping bits in the Bloom filter with certain probabilities) provide provable privacy protection against re-identification risks and can help mitigate these risks.[40,41]

Key informant interviews

The key informant interviews revealed 4 common themes regarding gaps and strategies in PPRL adoption and scale-up:

1) ***Awareness and understanding of PPRL.*** All 12 interviewees noted that knowledge of PPRL is limited to SMEs and public health professionals who have previously implemented PPRL, and the success of PPRL depends on increasing understanding about PPRL methods, feasibility, and appropriateness for public health. All twelve interviewees noted that input from stakeholders on barriers and facilitators to PPRL implementation is important. One interviewee explained, “it took a lot [of effort] to just get to the point where there is buy-in…for funding.” Much of that effort was for education about PPRL, because there are only “a few…programs and handfuls of individuals within those programs who understand [PPRL].” Interviewees also indicated the importance of increasing awareness among public health experts and leadership in federal agencies about the benefits and limitations of PPRL, the cost and resource burden, and the return on investment for PPRL applications.

   To improve understanding of PPRL in the public health community, interviewees identified strategies such as establishing communities of practice and creating graphics and dissemination material clarifying PPRL processes and privacy preservation mechanisms. Sharing successful use cases

Table 2. Selected applications of PPRL in public health.

Years	Agency/organization	Description
2010*	NIH	NIH developed the Global Unique Identifier Tool, so researchers can share study participant–specific data without exposing PII, and match participants across labs and research data sets.[32]
2013*	Patient-Centered Outcomes Research Institute	PCORnet has de-identified and integrated the electronic health record data of over 66 million patients across the USA, which can be accessed by researchers conducting observational studies, clinical trials, population health studies, and more.[31] Several research studies have utilized and expanded used of PPRL with data within the PCORnet network to answer a variety of research questions relevant to clinical care and public health.[33,34]
2018*	CDC	CODI linked individual-level data across 3 health systems and 3 community-based partners in Denver, Colorado, to assess the prevalence of pediatric obesity across the Metro Denver region and evaluate the effectiveness of pediatric weight management interventions and other community programs.[11]
2019*	CDC	The National Center for Immunization and Respiratory Diseases' surveillance of COVID-19 vaccine administration linked vaccine administration data from 9 Immunization Information Systems data partners, 21 national and regional retail pharmacy networks, 2 dialysis partners, and 1 federal agency to track individuals' vaccination events reported by multiple different entities (eg, vaccination clinics, pharmacies, and health care providers).[10]
2020*	NIH	The PPRL pilot within NIH N3C includes data from 75 institutions for over 6 million patients across the USA, for the purpose of studying the evolving coronavirus and its treatments.[9]
2020	US Department of Veterans Affairs (VA)	The Chicago HealthLNK Data Repository (HDR) used PPRL to link electronic health records across health care systems in the Chicago area. Data from VA were merged with data from the HDR to identify veterans eligible for VA services who were homeless or at risk of becoming homeless.[16,35]
2020	NIH	As part of the objective to recruit 1 million Americans in the All of Us Research Program Electronic Health Record data, the analysis aimed to determine the extent of fragmentation in care across 7 health provider organizations in 3 states (Wisconsin, Illinois, and Indiana).[36]
2022	National Cancer Institute	The Frederick National Laboratory for Cancer Research and Evaluation used source data from 6 US cancer registries, CVS, and Walgreens to evaluate PPRL solutions on criteria including ease of use, pre- and post-processing requirements, match quality, performance, and scalability.[37]
2022	CDC	Using data from Minnesota Electronic Health Record Consortium linked to data from Minnesota's immunization information system, Homeless Management Information System, and Department of Corrections, the research team conducted a retrospective, observational cohort study evaluating COVID-19 vaccine VE against SARS-CoV-2 related hospitalization among patients who had experienced homelessness or incarceration.[38]
2022	CDC	The National Center for Health Statistics conducted a case study comparing initial and refined linkages of a PPRL solution to those of an established, standard method of data linkage.[12]

* Indicates ongoing PPRL implementation as of January 1, 2023.

in public health through presentations and short publications was frequently cited as a strategy for increasing trust in the benefits of PPRL across partners.

2) ***Data owner buy-in.*** Most respondents stated that buy-in and data owner participation are critical to PPRL's success in public health. Further, all SMEs identified the engagement of state, tribal, local, and territorial public health departments (STLTs)—primary sources of public health data—was the most challenging aspect of PPRL implementation. For example, one interviewee described the difficulty engaging STLTs with the COVID-19 PPRL project was "because of the response situation and how resources are thinly stretched [during the COVID-19 pandemic], it was hard to get data owners to engage." Multiple interviewees pointed out legal, technical, and logistical challenges that could hamper data owner consent and fidelity to PPRL protocol. All 12 interviewees highlighted the need to demonstrate the value of PPRL to data owners (eg, deduplicating records and building community infrastructure, as demonstrated by CODI) as key to improving participation in PPRL.[11] Furthermore, providing education to STLT data owners about risks and benefits of PPRL and level of effort for PPRL implementation was deemed crucial to help STLTs make informed decisions regarding participating in PPRL.

3) ***Governance.*** All 12 interviewees emphasized the importance of an appropriate agency-level governance structure for PPRL, including adapting existing governance frameworks, regulations, and data-use agreements for information technology and data investments to include PPRL. Interviewees identified aspects of the PPRL process that would require developing guidance for PPRL projects, such as defining privacy and security thresholds; identifying trusted linkage agents; and creating standardized guidelines for data retention, storage, destruction, and further linkages. Ten interviewees noted that establishing processes for coordination across PPRL projects is important. Interviewees highlighted the importance of developing process checklists and guides as a way to enhance governance by outlining "at which stage do you need to know what…

that road map, or an implementation path is needed [for PPRL]" and describing existing approaches and resources "a checklist for folks to go through…what data are you trying to link, usage agreements already in place, this is how our agency is approaching it, this is how we know external partners (CMS, NIH, FDA, etc.) are approaching it—we need to provide that".

Nine interviewees indicated that selecting a single PPRL technology or vendor within an organization or agency was paramount for establishing PPRL governance. One key informant described this as one of "the biggest [obstacles] that lay before PPRL." Various open-source and vendor-provided PPRL solutions are available, but the use of different PPRL encryption methods and keys may preclude linking across these tools. Using a single authorized vendor or solution in an agency could overcome this issue. However, choosing a single vendor would entail time and expense, particularly if vendors or vendor ownership should change. One interviewee also noted that linking data collected by different federal agencies would require each to use the same PPRL tool, necessitating considerable coordination across agencies. Potential strategies identified to address this issue included evaluating PPRL solutions on criteria such as ease of use, pre- and post-processing requirements, match quality, license costs, performance, and scalability.

4) *Collaboration within and across public health agencies.* Several interviewees noted that lack of coordination and collaboration between programs implementing PPRL within public health agencies was an important barrier to its adoption and scale-up. One key informant highlighted the issues with a project-based on domain-based approach to PPRL within an agency as "if you're just doing it in your domain, sure it'll serve one purpose. You can come out with a research paper, but has it served the bigger purpose of bringing a cohesive ecosystem, more responsive and data-driven and science-driven CDC?" Other potential issues identified included missed opportunities for data linkage, duplication of cost and effort, and the significant burden on data owners participating in multiple PPRL programs. Several interviewees indicated the need for greater collaboration across federal, state, and local agencies to share lessons learned and inform the use of PPRL for public health planning and surveillance. For example, one interviewee noted "we want to be able to find [PPRL efforts across agencies] though and learn from their experience, but there's no way to find [information about PPRL applications from federal or state agencies] unless we dig deep-…maybe a snowball type of search. We have not addressed it in a systematic way, only when we hear about it and someone's willing to share." Another interviewee stressed the importance of aligning with external partners to "save people time in the learning process, …. ccwe don't want them to go down rabbit holes with vendors or inappropriate techniques." Eight interviewees noted that providing legal, contractual, and technical guidance is important to improving PPRL adoption and scale-up.

Discussion

Although PPRL has been used for years to link disparate data across domains, knowledge of its utility in public health remains limited. Our study examined published literature and publicly available information to understand the landscape of PPRL applications and assess PPRL's effectiveness in terms of linkage accuracy, privacy, and scalability in clinical care and public health.

Public health agencies and organizations have revealed several powerful, practical, and actionable use cases for PPRL in public health. For example, NIH N3C has used PPRL to improve research regarding the spread of and treatments for COVID-19, and the CDC COVID-19 project used PPRL to track and analyze COVID case counts and vaccination records to assess vaccine effectiveness. PPRL linkage quality was comparable to that of clear text linkage methods (requiring direct PII) for linkage across various settings and research questions.[9,10] PPRL links data in a HIPAA-compliant manner by eliminating the need to share direct PII. PPRL improves data privacy, although the level of privacy protection depends on the policies and protocols governing the storage, retention, and destruction of PPRL tokens and keys. Large-scale implementations of PPRL linking millions of observations—including PCORnet, NIH N3C, and the CDC COVID-19 project have demonstrated the scalability of PPRL for public health applications.[31,9,10]

CDC subject matter experts described opportunities and challenges for adoption and scale-up of PPRL. Interviewees reported 4 main gaps and strategies in PPRL adoption and scale-up: increasing awareness and understanding of PPRL, improving data owner buy-in, establishing appropriate governance structure and policies and implementation oversight, and developing a coordinated strategy for PPRL implementation within and across public health agencies. Various agencies and organizations have recently started to develop governance frameworks for PPRL. For example, the Eunice Kennedy Shriver National Institute of Child Health and Human Development assessed potential governance and technical approaches for implementing PPRL across pediatric COVID-19 studies.[42] PCORnet has also described the governance considerations and process to establish a standardized and scalable infrastructure for a national clinical research network.[43] In addition, several PPRL tools follow the secure hash standards issued by the National Institute of Standards and Technology, ensuring a robust and secure methodology for PPRL implementation.[44] Interviewees highlighted the urgent need for organizations to examine the costs, benefits, and risks including vendor lock of establishing a single approved PPRL technology solution, and the trade-offs between open-source and commercial solutions. Similar themes emerged from a CDC Foundation convening of representatives from public health organizations, industry partners, and CDC experts to discuss the potential benefits of, barriers to, and sustainable business models for PPRL implementation.[45] Interviewees suggested strategies to facilitate PPRL adoption, including demonstrating value to data owners participating in PPRL, sharing lessons learned from prior PPRL projects, and developing PPRL process checklists to help staff new to PPRL. Some of these strategies have already been applied in projects, providing insights into future PPRL applications. For example, CODI has published lessons learned from planning, designing, and implementing a clinical-community infrastructure enhancement that used PPRL technology to link data.[11] Similarly, the Frederick National Laboratory for Cancer Research and Evaluation used source data from 6 US cancer registries, CVS, and Walgreens to evaluate PPRL technology solutions on such criteria as ease of use, pre- and post-processing requirements, match quality, performance, and scalability.[37]

Our study has a few limitations. The literature scan did not include articles published after January 1, 2023, and does not reflect findings from recent PPRL applications in public health. All SMEs interviewed for the study were affiliated with 1 federal agency (CDC), and therefore may not be representative of the public health community as a whole. The opportunities, successes, challenges, and limitations identified through the literature scan and SME interviews may not be fully representative of the field; as availability and adoption of PPRL evolves, it is likely that the opportunities and challenges will also evolve. Further research is needed to incorporate views of experts from other federal, state, and local agencies, health systems, researchers, and community organizations to understand barriers and facilitators to PPRL implementation across the public health community.

Conclusions

By linking patient-level clinical and public health data without sharing unencrypted PII, PPRL could bridge crucial data gaps such as lack of timely and relevant data, thus creating new opportunities for public health research and surveillance. Recent applications have demonstrated the feasibility, scalability, and value of PPRL for public health. Although gaps must be addressed before PPRL can be implemented widely across use cases, this process offers a promising solution to some of the public health ecosystem's current data linkage challenges. Tangible next steps for improving PPRL adoption include (1) sharing lessons learned from PPRL applications, (2) checklists for those involved in PPRL selection and adoption, (3) evaluating costs, benefits, and risks of single approved PPRL technology solution, and (4) developing governance and oversight guidelines for PPRL applications.

Author contributions

The study was conceived by Adi V. Gundlapalli, and Corrine Ferdon, with Steven Luxenberg. Aditi Pathak led the design and analysis with collaboration from Daniela Zapata, Laina Serrer, and Meera Bhalla. Subject matter experts including Arunkumar Srinivasan, Raymond King, Lisa B. Mirel, Thom Sukalac, and Patrick Baier provided valuable contributions regarding real-world PPRL use cases, opportunities, challenges, and resource materials to inform the research and writing.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was funded by the Centers for Disease Control and Prevention [Contract Number HHSD2002013M53937B].

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data underlying this article cannot be shared publicly for the privacy of the individuals who participated in the key informant interviews.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the US Centers for Disease Control and Prevention and/or National Science Foundation and/or American Institutes for Research. This research was conducted while Lisa B. Mirel was employed at the National Center for Health Statistics, Centers for Disease Control and Prevention prior to becoming an employee at the National Center for Science and Engineering Statistics within the National Science Foundation.

References

1. Galaitsi SE, Cegan JC, Volk K, et al. The challenges of data usage for the United States' COVID-19 response. *Int J Inf Manage*. 2021;59:102352. https://doi.org/10.1016/j.ijinfomgt.2021.102352
2. Bekemeier B, Heitkemper E, Backonja U, et al. Rural public health data challenges during the COVID-19 pandemic: The case for building better systems ahead of a public health crisis. *J Public Health Manag Pract*. 2023;29(4):496-502. https://doi.org/10.1097/PHH.0000000000001726
3. U.S. Food and Drug Administration. *Enterprise Modernization Action Plan (EMAP)*. 2022. Accessed April 20, 2023. https://www.fda.gov/media/158206/download
4. Centers for Disease Control and Prevention. *Data Modernization Initiative Strategic Implementation Plan*. 2021. Accessed April 20, 2023. https://www.cdc.gov/surveillance/pdfs/FINAL-DMI-Implementation-Strategic-Plan-12-22-21.pdf
5. Fast-Track Action Committee on Advancing Privacy–Preserving Data Sharing and Analytics, Networking and Information Technology Research and Development Subcommittee of the National Science and Technology Council. *National Strategy to Advance Privacy–Preserving Data Sharing and Analytics*. Executive Office of the President of the United States; 2023. Accessed May 9, 2023. https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf
6. Vatsalan D, Christen P, Verykios V. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*. 2013;38(6):946-969.
7. Hamp AD, Doshi RK, Lum GR, Allston A. Cross-jurisdictional data exchange impact on the estimation of the HIV population living in the District of Columbia: evaluation study. *JMIR Public Health Surveill*. 2018;4(3):e62. Accessed April 20, 2023. https://doi.org/10.2196/publichealth.9800
8. Jonas J, Harper JC. Effective counterterrorism and the limited role of predictive data mining. *Policy Anal*. 2006;584:1-12. Accessed April 20, 2023. https://www.cato.org/sites/cato.org/files/pubs/pdf/pa584.pdf
9. Phua C, Smith-Miles K, Lee VC, Gayler R. Resilient identity crime detection. *IEEE Trans Knowl Data Eng*. 2012;24(3):533-546. Accessed April 20, 2023. https://ieeexplore.ieee.org/abstract/document/5677523
10. National COVID Cohort Collaborative. *N3C Privacy-Preserving Record Linkage: Enabling Data Connectivity, Ensuring Data Security*. National Institutes of Health. Accessed April 20, 2023. https://covid.cd2h.org/PPRL
11. Kompaniyets L, Wiegand RE, Oyalowo AC, et al. Relative effectiveness of COVID-19 vaccination and booster dose combinations among 18.9 million vaccinated adults during the early SARS-CoV-2 Omicron period—United States. *Clin Infect Dis*. 2023;76(10):

ciad063. https://doi.org/10.1093/cid/ciad063. Accessed April 19, 2023. https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciad063/7030940

12. King RJ, Heisey-Grove DM, Garrett N, et al. The childhood obesity data initiative: a case study in implementing clinical-community infrastructure enhancements to support health services research and public health. *J Public Health Manag Pract*. 2022;28(2):E430-E440. https://doi.org/10.1097/PHH.0000000000001419
13. Mirel LB, Resnick DM, Aram J, Cox CS. A methodological assessment of privacy preserving record linkage using survey and administrative data. *Stat J IAOS*. 2022;38(2):413-421. Accessed April 19, 2023. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210891
14. Centers for Disease Control and Prevention. Evaluation report for COVID-19 vaccination administration data privacy preserving record Linkage (PPRL). MITRE. Task Order No. 75D30119F05691. August 2022.
15. Bernstam EV, Applegate RJ, Yu A, et al. Real-world matching performance of deidentified record-linking tokens. *Appl Clin Inform*. 2022;13(4):865-873. Accessed May 18, 2023. https://www.thieme-connect.com/products/ejournals/html/10.1055/a-1910-4154#N117CC
16. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. *J Biomed Inform*. 2014;50:205-212. https://doi.org/10.1016/j.jbi.2013.12.003. Accessed April 20, 2023. https://www.sciencedirect.com/science/article/pii/S1532046413001949
17. Nguyen L, Stoové M, Boyle D, et al. Privacy-preserving record linkage of deidentified records within a public health surveillance system: evaluation study. *J Med Internet Res*. 2020;22(6):e16757. Accessed April 19, 2023. https://www.jmir.org/2020/6/e16757/y
18. Jarrett M, Hills B, Zhao Y, et al. Evaluating PPRL vs clear text linkage with real-world data. *IJPDS*. 2020;5(5). https://doi.org/10.23889/ijpds.v5i5.1542
19. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *J Am Med Inform Assoc*. 2015;22(5):1072-1080. Accessed April 19, 2023. https://academic.oup.com/jamia/article/22/5/1072/930113
20. Irvine K, Smith M, de Vos R, et al. Real world performance of privacy preserving record linkage. *IJPDS*. 2018;3(4). https://doi.org/10.23889/ijpds.v3i4.990
21. Brown A, Borgs C, Randall S, et al. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. *BMC Med Inform Decis Mak*. 2017;17(1):83. https://doi.org/10.1186/s12911-017-0478-5
22. Bian J, Loiacono A, Sura A, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. *JAMIA Open*. 2019;2(4):562-569. Accessed May 18, 2023. https://doi.org/10.1093/jamiaopen/ooz050
23. Randall S, Wichmann H, Brown A, et al. A blinded evaluation of privacy preserving record linkage with Bloom filters. *BMC Med Res Methodol*. 2022;22(1):22. Accessed May 18, 2023. https://doi.org/10.1186/s12874-022-01510-2
24. Christen P, Schnell R, Vatsalan D, Ranbaduge T, et al. Efficient cryptanalysis of bloom filters for privacy-preserving record linkage. In: Kim J, eds. *Advances in Knowledge Discovery and Data Mining: Lecture Notes in Computer Science*. Springer International; 2017:628-640.
25. Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-preserving record linkage for big data: current approaches and research challenges. In: Zomaya AY, Sakr S, eds. *Handbook of Big Data Technologies*. Springer International; 2017:851-895.
26. Schnell R, Borgs C, et al. Protecting record linkage identifiers using a language model for patient names. In: Hübner U, eds. *German Medical Data Sciences: A Learning Healthcare System*. IOS Press; 2018:91-95.
27. Stammler S, Kussel T, Schoppmann P, et al. Mainzelliste SecureEpiLinker (MainSEL): privacy-preserving record linkage using secure multi-party computation. *Bioinformatics*. 2022;38(6):1657-1668. Accessed April 19, 2023. https://doi.org/10.1093/bioinformatics/btaa764
28. Code of Federal Regulations. *Security and Privacy, Other Requirements Relating to Uses and Disclosures of Protected Health Information*. 45 CFR §164.514(c). 2019. Accessed April 20, 2023. https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164
29. Marsolo K, Kiernan D, Toh S, et al. Assessing the impact of privacy-preserving record linkage on record overlap and patient demographic and clinical characteristics in PCORnet®, the National Patient-Centered Clinical Research Network. *J Am Med Inform Assoc*. 2023;30(3):447-455. https://doi.org/10.1093/jamia/ocac229. Accessed April 19, 2023. https://academic.oup.com/jamia/article/30/3/447/6855148?login=true
30. U.S. Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. 2022. Accessed May 9, 2023. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
31. Patient-Centered Outcomes Research Institute. About. PCORnet: the National Patient-Centered Clinical Research Network. Accessed April 20, 2023. https://pcornet.org/about/
32. Center for Information Technology. Introducing BRICS. Biomedical Research Informatics Computing System. Accessed April 20, 2023. https://brics.cit.nih.gov/intro
33. Canterberry M, Kaul AF, Goel S, et al. The patient-centered outcomes research network antibiotics and childhood growth study: implementing patient data linkage. *Popul Health Manag*. 2020;23(6):438-444. http://doi.org/10.1089/pop.2019.0089
34. Agiro A, Chen X, Eshete B, et al. Data linkages between patient-powered research networks and health plans: a foundation for collaborative research. *J Am Med Inform Assoc*. 2019;26(7):594-602. https://doi.org/10.1093/jamia/ocz012
35. Trick WE, Hill JC, Toepfer P, Rachman F, Horwitz B, Kho A. Joining health care and homeless data systems using privacy-preserving record-linkage software. *Am J Public Health*. 2021;111(8):1400-1403. https://doi.org/10.2105/AJPH.2021.306304
36. Kho AN, Yu J, Bryan MS, et al. Privacy-preserving record linkage to identify fragmented electronic medical records in the All of Us Research Program. In: Cellier, P., Driessens, K. eds, *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science*. Vol 1168. Springer, Cham; 2020. https://doi.org/10.1007/978-3-030-43887-6_7
37. Frederick National Laboratory for Cancer Research. *Evaluating the Performance of Privacy Preserving Record Linkage Systems (PPRLS)—Part One*. National Cancer Institute; 2022. Accessed May 9, 2023. https://surveillance.cancer.gov/reports/TO-P2-PPRLS-Evaluation-Report-Part1.pdf
38. DeSilva MB, Knowlton G, Rai NK, et al. Vaccine effectiveness against SARS-CoV-2 related hospitalizations in people who had experienced homelessness or incarceration—findings from the Minnesota EHR Consortium. *J Community Health*. 2023;49(3):448-457. https://doi.org/10.1007/s10900-023-01308-3
39. Vidanage A, Ranbaduge T, Christen P, Schnell R. A taxonomy of attacks on privacy-preserving record linkage. *JPC*. 2022;12(1). https://doi.org/10.29012/jpc.764
40. Alaggan M, Gambs S, Kermarrec A-M. BLIP: non-interactive differentially-private similarity computation on Bloom filters. In: *Symposium on Self- Stabilizing Systems*; 2012: 202-216.
41. Schnell R, Borgs C. Randomized response and balanced Bloom filters for privacy preserving record linkage. In: *ICDMW DINA*, Barcelona; 2016.

42. Eunice Kennedy Shriver National Institute of Child Health and Human Development, Office of Data Science and Sharing. *Privacy preserving record linkage (PPRL) for pediatric COVID-19 studies. Final report*. National Institutes of Health; 2022. Accessed May 9, 2023. https://www.nichd.nih.gov/sites/default/files/inline-files/NICHD_ODSS_PPRL_for_Pediatric_COVID-19_Studies_Public_Final_Report_508.pdf
43. Kiernan D, Carton T, Toh S, et al. Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet®, the National Patient-Centered Clinical Research Network. *BMC Res Notes*. 2022;15(1):337. https://doi.org/10.1186/s13104-022-06243-5
44. National Institute of Standards and Technology. Secure Hash Standard. (Information Technology Laboratory, Gaithersburg, MD.), Federal Information Processing Standards Publications (FIPS PUBS); 2015:180-4. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf
45. CDC Foundation. *Data Linkage and Identity Management—Privacy Protecting Record Linkage (PPRL)*. HLN Consulting; 2023.

Published by Oxford University Press on behalf of the American Medical Informatics Association 2024.

Journal of the American Medical Informatics Association, 2024, 31, 2605–2612
https://doi.org/10.1093/jamia/ocae196
Research and Applications