CDC Foundation
Together our impact is greater

Data Linkage and Identity Management - Privacy Protecting Record Linkage (PPRL)

Meeting Summary Prepared by HLN Consulting | March 2023

Introduction

In October 2022 the CDC Foundation convened a set of stakeholders to discuss Privacy Protecting Record Linkage (PPRL). Nearly twenty representatives of leading public health organizations and their industry partners joined nearly twenty of their Centers for Disease Control and Prevention (CDC) colleagues in discussing the potential benefits, barriers and sustainable business models for PPRL implementation.

PPRL is a strategy that allows records to be linked together without revealing identifying information. With PPRL, records from two different sources are linked by encrypting ("hash-ing") each person's identifying information within each record. A third party can then compare the hashed values to see if a pair of records are from the same person, without revealing that person's identity. One key to securing PPRL is a trusted third party (not the entity sending the data, and not the entity ultimately wanting to match the data) who performs the match of the hashed values to produce the linked, de-identified data set.

There are several powerful uses of this technology that may improve public health surveillance and prevention. Using PPRL, public health can connect data sets that were generated independently, enabling new analysis opportunities. Similar data sets can be combined reliably to identify and remove duplicate events, reducing inflated case counts or immunization rates. PPRL can allow public health data sets to be linked to external data sets, which may provide new, richer analytical potential. This is achieved while preserving the privacy of the original records in a HIPAA-compliant manner. PPRL was used by CDC to link COVID-19 case and vaccination data and to improve COVID-19 case counts, with CDC receiving de-identified data (data with hashed identifiers) from various sources, then linking the hashed identifiers together to more complete and de-duplicated yet de-identified records, better tracking the spread and prevention of COVID-19. In addition to its value at the federal level, PPRL also has great potential to help state, tribal, local, and territorial (STLT) agencies. Examples and useful references are included in the bibliography, below.

This content is supported by the Centers for Disease Control and Prevention (CDC) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award totaling $1,915,915 with 100 percent funded by CDC/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by CDC/HHS, or the U.S. Government.

Focus Group Outcomes

A. Benefits

- Enables sharing of anonymized data
- Simplifies policy compliance
- Opens data to wider audiences
- May improve matching at the originating agency
- Facilitates record completion
- Leverages investment

B. Barriers to adoption

I. Resources

- Cost

II. Quality

- Confidence
- Quality compromised by missing data

III. Practicality

- Newness
- Modification of existing systems

B. Barriers to adoption (continued)

III. Practicality (continued)

- Most benefits accrue to the data recipient
- Challenging matches
- Often inappropriate for clinical care
- Risk of data re-identification
- Multiple PPRL vendors
- Reliance on vendor

IV. Policy

- Politically sensitive
- Policy barriers

C. Sustainability Strategies

- Evaluate risks constantly
- Standardize data at the source
- Provide PPRL to the source
- Create useful documentation
- Promote trust
- Shared funding
- Leverage Health Information Exchanges (HIEs)
- Consider “tokenization bridges”

A. Benefits

After a brief presentation of PPRL features, participants addressed three key questions in an open discussion about this technology. The focus group identified the following key **perceived and realized benefits of PPRL:**

Enables sharing of anonymized data: Patient data is anonymized at the source by software provided by a trusted third party before transmission, so the risk of inappropriate exposure is far lower than it would be with matching methods that require sharing identified data.

Simplifies policy compliance: The data associated with a PPRL identifier is anonymized. For some participants, there may be less need for institutional review board (IRB) processes or complicated data use agreements with data exchange partners.

Opens data to wider audiences: Within the constraints of relevant data use agreements, PPRL may enable agencies to share data reliably not only across programs, but also across jurisdictions when sharing identified data is not allowed..

May improve record matching at the originating agency: Where there are restrictions on sharing identified data within the organization, PPRL techniques can be used to match records within a single organization for deduplication of records or linking data sources.

Facilitates record completion: Combining records with PPRL can help “fill in the gaps” in a data set by combining data from different sources that would not otherwise be available. For example, clinical data sources that often do not capture ethnicity data might be linked to data sets with more complete ethnicity data, allowing identification of ethnicity-related gaps in clinical care.

Leverages investment: Though the initial investment in PPRL may be sizable, once the infrastructure is in place, it can be leveraged across data sets and projects.

B. Barriers to adoption

Next, the focus group identified the following key **barriers to PPRL adoption:**

I. Resources

Cost: PPRL may involve expensive, specialized and often proprietary vendor services; gaining access to existing third-party PPRL-enabled data can be even more expensive.

II. Quality

Confidence: No matching algorithm is perfect. While PPRL can achieve low false positive matching rates (*i.e.,* the risk that two records are determined to be for the same person when they are not), most testing has not been done specifically on public health surveillance data. Potential public health users may wonder about the accuracy of record matching on their data. In addition, since the source data used to create a PPRL hash is shielded from participants, users of PPRL cannot easily verify or validate the accuracy of the algorithm for themselves when the data originates outside of their own agency.

Quality compromised by missing data: Certain public health data sets commonly are missing key data typically used by PPRL algorithms, potentially compromising the results or limiting the usefulness of PPRL processing.

III. Practicality

Newness: PPRL is a relatively new matching technique and promulgating its use will require some education and promotion.

Modification of existing systems: Because it is new, most existing public health systems cannot incorporate their record linkage processes to PPRL without at least some modification or augmentation.

Most benefits currently accrue to the data recipient: While the *source* of the data does the work of enabling PPRL, most of the benefits accrue to the *recipient* of the data. This imbalance of who does the work versus who benefits might be addressed through sharing of PPRL results, especially if the data recipient, like the CDC, is combining data from multiple sources into one super-set that can be shared back to all sources.

Challenging matches: Some types of matches are inherently challenging. For instance, multiple births (e.g., twins) can be very challenging to match since key demographic data used to discriminate between records is often very similar or nearly identical. Because examination of the source data cannot be used for corroboration, the matches may be suspect.

Often inappropriate for clinical care: Like any matching method, even good PPRL matching may produce mismatches; the resulting data may be good enough for accurate population-level statistics, but may not be reliable at the individual person level; i.e., if matching is not highly reliable, it may be unwise to use the matched data in individual clinical care.

Risk of data re-identification: Whenever disparate records are linked together, the richness of the resulting data sets may expose them to a greater risk (or perceived risk) of re-identification.

Multiple PPRL vendors: The real strength in using PPRL is the ability to relate disparate data sets together. Data sets processed by different vendors cannot be readily related together.

Reliance on vendor: At its core PPRL rests on the trust its users instill in the service providers who manage it and the users who interact with it. Because of its newness and the closed, “black box” nature of some of its processes, some skeptics may still feel this to be too great a risk.

IV. Policy

Politically sensitive: Record matching and consolidation can be a politically charged issue in some circles, especially among consumer privacy advocates who fear inappropriate data disclosure. New and different technologies like PPRL may be especially suspect and resisted. On the other hand, when data from different sources are to be linked, privacy advocates may prefer PPRL to methods requiring direct identifiers.

Policy barriers: Regardless of its privacy preserving attributes, PPRL may nonetheless face agency policy barriers or perceived policy barriers that inhibit or delay implementation.

C. Sustainability Strategies

Finally, the focus group identified the following **sustainability strategies might facilitate PPRL adoption:**

Evaluate risks constantly: Risk assessment is not a “once and done” activity, especially in what can be for some a very politically charged topic.

Standardize data at the source: All matching algorithms, including PPRL, rise and fall with the quality of data at the source. When comparing across data sets this becomes especially important. The more compatible (*i.e.,* standardized) data is at the source the better likelihood that high-quality matches can be established.

Deliver value to data providers: Getting buy-in and participation from data sources requires figuring out how to provide value to them for their participation. Possible benefits to them might include improved matching or de-duplication capabilities or receiving valuable information back from the matched dataset.

Create useful documentation: Creating documentation that is general enough for reuse will help further PPRL adoption in the field.

Promote trust: Promotion and education around the confidentiality advantages of PPRL may generate wider acceptance and participation. Validating PPRL results by testing with multiple service providers may also build trust among participants.

Shared funding: PPRL implementations can be expensive. Federal-STLT funding partnerships could go a long way to ease the cost of implementation and support.

Leverage HIEs: Health Information Exchanges (HIEs) manage health data from many sources and already have expertise in matching records. They may provide a centralized forum for implementing PPRL, relieving that task from their customers.

Consider "tokenization bridges": Different organizations or programs may use different PPRL implementations. Tokenization bridges are a strategy to allow linking data across different PPRL implementations. Tokenization bridges essentially provide an extra PPRL cycle, providing a new PPRL identifier (a token), created to link PPRL identifiers that were generated through different PPRL implementations, i.e., through two initial PPRL processes that use different hashing algorithms.

Bibliography

STLT Experience

Drawz, Paul E, Malini DeSilva, Peter Bodurtha, Gabriela Vazquez Benitez, Anne Murray, Alanna M Chamberlain, R Adams Dudley, et al. "Effectiveness of BNT162b2 and MRNA-1273 Second Doses and Boosters for SARS-CoV-2 Infection and SARS-CoV-2 Related Hospitalizations: A Statewide Report from the Minnesota Electronic Health Record Consortium." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, February 7, 2022, ciac110. https://doi.org/10.1093/cid/ciac110.

Kho, Abel N, John P Cashy, Kathryn L Jackson, Adam R Pah, Satyender Goel, Jörn Boehnke, John Eric Humphries, et al. "Design and Implementation of a Privacy Preserving Electronic Health Record Linkage Tool in Chicago." *Journal of the American Medical Informatics Association* 22, no. 5 (September 1, 2015): 1072–80. https://doi.org/10.1093/jamia/ocv038.

Federal Experience

"COVID-19 Vaccine IT Overview: Vaccination Reporting | CDC," May 18, 2022. https://www.cdc.gov/vaccines/covid-19/reporting/overview/IT-systems.html.

Kompaniyets, Lyudmyla, Ryan E Wiegand, Adewole C Oyalowo, Lara Bull-Otterson, Heartley Egwuogu, Trevor Thompson, Kaʻimi Kahihikolo, 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, January 1, 2022–March 31, 2022." Clinical Infectious Diseases, February 8, 2023, ciad063. https://doi.org/10.1093/cid/ciad063.

Mirel, Lisa B. "Privacy Preserving Techniques : Case Studies from the Data Linkage Program." Edited by National Center for Health Statistics (U.S.), May 19, 2021. https://stacks.cdc.gov/view/cdc/114623.

Mirel, Lisa B., Dean M. Resnick, Jonathan Aram, and Christine S. Cox. “A Methodological Assessment of Privacy Preserving Record Linkage Using Survey and Administrative Data.” Statistical Journal of the IAOS 38, no. 2 (January 1, 2022): 413–21. https://doi.org/10.3233/SJI-210891.

National Center for Advancing Translational Sciences. “N3C Data Overview,” August 31, 2020. https://ncats.nih.gov/n3c/about/data-overview.

Raad, Jason H, Elizabeth Tarlov, Abel N Kho, and Dustin D French. “Health Care Utilization Among Homeless Veterans in Chicago.” *Military Medicine* 185, no. 3–4 (2020): e335–39. https://doi.org/10.1093/milmed/usz264.

Technology

Baker, Dixie B., Bartha M. Knoppers, Mark Phillips, David van Enckevort, Petra Kaufmann, Hanns Lochmuller, and Domenica Taruscio. “Privacy-Preserving Linkage of Genomic and Clinical Data Sets.” IEEE/ACM Transactions on Computational Biology and Bioinformatics 16, no. 4 (2019): 1342–48. https://doi.org/10.1109/TCBB.2018.2855125.

Dixie Baker, Mark Phillips, David van Enckevort, Peter Christen, Ken Gersing, Maximilian Haeussler, Cenk Sahinalp, and Adrian Thorogood. “Technology Primer: Overview of Technological Solutions to Support Privacy-Preserving Record Linkage.” International Rare Diseases Research Consortium (IRDiRC) & Global Alliance for Genomics and Health (GA4GH), December 7, 2017. https://irdirc.org/wp-content/uploads/2018/03/PPRL-Technical-Primer-V4-2.pdf.

Assessments

Brown, A. P., S. M. Randall, J. H. Boyd, and A. M. Ferrante. “Evaluation of Approximate Comparison Methods on Bloom Filters for Probabilistic Linkage.” International Journal of Population Data Science 4, no. 1 (May 23, 2019): 1095. https://doi.org/10.23889/ijpds.v4i1.1095.

Garfnkel, Simson. “De-Identifying Government Data Sets.” Gaithersburg, MD: National Institute of Standards and Technology, 2022. https://doi.org/10.6028/NIST.SP.800-188.3pd.

Lee, Joyce. “Why Privacy-Preserving Record Linkage Is Having a Moment: Interview with Abel Kho.” Datavant (blog), November 21, 2020. https://medium.com/datavant/why-privacy-preserving-record-linkage-is-having-a-moment-interview-with-abel-kho-837808a03a07.

Lim, David, Sean Randall, Suzanne Robinson, Elizabeth Thomas, James Williamson, Aron Chakera, Kathryn Napier, et al. “Unlocking Potential within Health Systems Using Privacy-Preserving Record Linkage: Exploring Chronic Kidney Disease Outcomes through Linked Data Modelling.” Applied Clinical Informatics 13, no. 4 (August 2022): 901–9. https://doi.org/10.1055/s-0042-1757174.

Nguyen, Long, Mark Stoové, Douglas Boyle, Denton Callander, Hamish McManus, Jason Asselin, Rebecca Guy, Basil Donovan, Margaret Hellard, and Carol El-Hayek. “Privacy-Preserving Record Linkage of Deidentified Records Within a Public Health Surveillance System: Evaluation Study.” Journal of Medical Internet Research 22, no. 6 (June 24, 2020): e16757. https://doi.org/10.2196/16757.

Rainer Schnell and Christian Borgs. “Implementing Privacy-Preserving National Health Registries.” Proceedings of Statistics Canada Symposium 2018, 2018. https://www.statcan.gc.ca/en/conferences/symposium2018/program/09a3_schnell-eng.pdf.

Randall, Sean, Helen Wichmann, Adrian Brown, James Boyd, Tom Eitelhuber, Alexandra Merchant, and Anna Ferrante. “A Blinded Evaluation of Privacy Preserving Record Linkage with Bloom Filters.” BMC Medical Research Methodology 22, no. 1 (January 16, 2022): 22. https://doi.org/10.1186/s12874-022-01510-2.

Zimmerman, Lindsay P., Satyender Goel, Shazia Sathar, Charon E. Gladfelter, Alejandra Onate, Lindsey L. Kane, Shelly Sital, et al. “A Novel Patient Recruitment Strategy: Patient Selection Directly from the Community through Linkage to Clinical Data.” *Applied Clinical Informatics* 9, no. 1 (January 2018): 114–21. https://doi.org/10.1055/s-0038-1625964.