# Record Linkage Implementation Checklist

The Record Linkage Implementation Checklist is a resource for guiding decisions that must be made prior to designing and implementing a strategy for linking data from multiple sources and sharing and using the linked datasets for research. The checklist incorporates both data governance and technology decisions and was developed based on findings from an assessment of existing record linkage implementations.

## Overview

The *Eunice Kennedy Shriver* National Institute for Child Health and Human Development (NICHD) Office of Data Science and Sharing (ODSS) is leading an effort to develop frameworks and tools to support responsible individual-level record linkage for research with NICHD populations. In response to significant research and public interest in maximizing the value of data contributed by pediatric COVID-19 patients, by applying Privacy Preserving Record Linkage (PPRL) to link multiple datasets, NICHD ODSS assessed 13 existing record linkage implementations and derived 8 considerations for developing any new record linkage implementation, using PPRL, or another linkage method. These 8 considerations provide the foundation of the Record Linkage Implementation Checklist.

The full assessment report is available at: [Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies].

Prior to linking data from multiple datasets at the individual-level, researchers, data stewards, and other stakeholders must ensure that linkage is appropriate, based on governance factors that apply to the data, such as legal, policy, and/or consent-based requirements. They must also consider technical aspects that impact which PPRL tool should be used, if using PPRL.

***Figure 1: The PPRL Process***

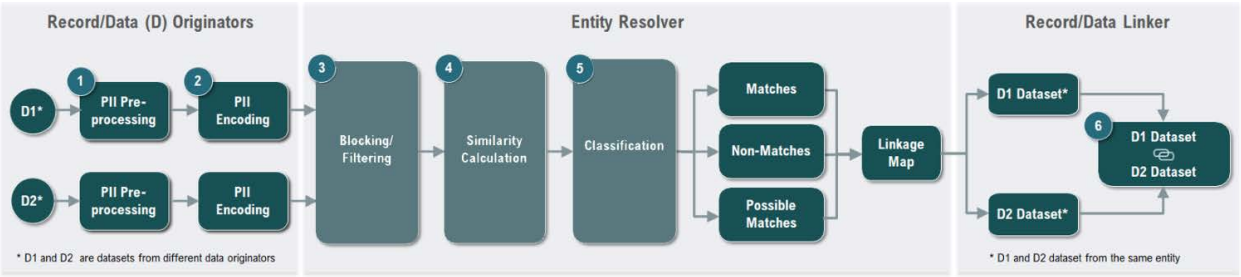

Figure 1 illustrates the PPRL process. This approach uses secure software to enable users to link data from multiple sources to the same individual, without revealing personally identifiable information (PII). PPRL requires that PII is entered into the software to create cryptographically encoded (hashed) codes or "tokens," but the PII does not leave the data originator.

Primary contact: NICHDecosystem@nih.gov

*Details:*

- **Created by:** NICHD ODSS
- **NIH contacts:**
    - Rebecca Rosen, Director, NICHD ODSS
    - Valerie Cotton, Deputy Director, NICHD ODSS
    - Elizabeth Clerkin, Data Science and Policy Specialist, NICHD ODSS

# Record Linkage Implementation Checklist: Governance & Technical Considerations

Prior to designing and implementing a record linkage strategy, funders, researchers, data repositories, and other stakeholders should collaborate to make a series of governance and technical decisions.

## Data Governance Considerations

Data governance is the collective set of rules and controls that define and enforce appropriate collection, sharing, linking, access, and use of data. Good governance is critical to protect research participant privacy, manage risks, address ethical considerations, and respect participant trust. Data governance considerations for record linkage include the following.

### 1. Determine the scope of linkage (which datasets to link).

All record linkage implementations should make up-front determinations regarding which datasets would be linked, and whether the linkage would apply to one specific study (study-specific) or to multiple datasets from one or multiple repositories (linked database model), thereby supporting many studies.

The linked database model is the most sustainable and reasonable approach for fostering reproducible and innovative research in a federated data ecosystem, as it could encompass multiple current and future datasets across multiple repositories, as well as a variety of secondary analyses. The linked database model is often achieved by having a central party maintain either a common database of participant-level globally unique (GUIDs) and/or linkage maps that document how different participant-level IDs from different sources map to each other.

### 2. Obtain approval or authorization to link.

Approval to include a given dataset in a linkage implementation could come from a combination of parties or authorities such as: **research participants** (in the form of explicit informed consent), **data generators/contributors** (such as a study investigator and their institution sometimes in the form of a data submission agreement), **Institutional Review Boards (IRBs)** (waiver of consent or other determination), **federal authority** (e.g., for statistical agencies), and/or **governance bodies** (such as a network steering committee), depending on the nature of the data sources.

- If feasible, studies should consent research participants for the linkage of their data across sources (and data repositories, if applicable). The consent language should address the scope of linkage—that is, which datasets will be linked (see Consideration 1)—and how the linked data will be shared, without overly restricting the scope in a way that could prohibit future valuable record linkage opportunities (e.g., adding new datasets or repositories that may not yet exist).
    - The following provides example consent language that addresses consent as well as assent for when a research study involves children. This example is based on existing consent language but has not necessarily been approved by an IRB in its entirety. Revisions should be made to fit the circumstances of a given record linkage implementation, as appropriate. Note: for the legal guardian, the language refers to "your child"; for the child, the language refers to "you."

---

*If [you/your child] join this study, we will gather data about [you/your child]. What we learn in this study will be put in a secure NIH-designated storage location, called a data repository, where these data would be shared for future research. Information about [you/your child] will be "de-identified," which means it will not include anything that identifies [you/your child]. [NIH] will approve researchers from all over the world to access information from the repository. Researchers will agree not to attempt to identify [you/your child]. It is possible that if [you/your child] participate[s] in more than one study, researchers may be able to combine de-identified data from multiple studies to ease the burden on researchers and participants alike. The purpose of sharing this information is to make more research possible that may improve children's and everyone's health. This sharing of information will be done without obtaining additional permission from [you/your child]. If [you/your child] no longer want [your/your child's] de-identified data to be shared with researchers and combined with other data about [you/your child], you can request [your/your child's] data to be withdrawn from the data repository and destroyed. Please note that any data that has already been shared with researchers cannot be withdrawn. If [you/your child] turns 18 years old while taking part in this study, [you/your child] will be asked to review and sign an informed consent form as an adult if [you/your child] wants to continue to be in the study.*

---

    - Explicit consent for record linkage in the research setting may be particularly prudent when linking with data originally collected for purposes other than research and therefore not originally consented for research use (e.g., data from health care systems or public health surveillance and other administrative sources). This type of consent may warrant additional data source-specific language or at minimum specific examples of the type of external administrative data that will be linked (e.g., pharmacy records, health insurance records, or cancer registries), so participants understand that the scope extends beyond data generated in the research context.

- It may also be appropriate to document approval from the investigator and associated institution that generated and is contributing the dataset, perhaps with input from an IRB and/or equivalent body, especially when re-consent is not feasible. Data repositories that facilitate record linkage of submitted data often include relevant terms and conditions in the associated data submission agreement. This can be done in a manner that requires data submitters to participate in record linkage or gives them to option to "opt in." Whether and the extent to which IRBs are engaged may vary, especially for PPRL-based approaches that link and share only de-identified data, which may not qualify as "human subjects research" that is subject to the requirements of the Federal Policy for the Protection of Human Subjects (also known as the Common Rule).

**3. Identify policies that apply to each dataset including rules specific to certain data types or participant populations.**

Linkage implementers must understand how linked datasets inherit rules and controls from the original datasets contributed to the linkage implementation. This may require reviewing a variety of data governance documentation (consent forms, data submission/use agreements, repository policies, federal or state laws) and analyzing whether they conflict (thus possibly prohibiting the linkage) or how they intersect (thus impacting how the linked data can be shared and used). For example, if one dataset's IRB requires it be used for COVID research only and the consent form of another allows the data to be used for any type of health research, the linked dataset must be used for COVID research only. Certain policies may prohibit the data from being linked. For example, the NIH Genomic Data Sharing Policy requires consent for data sharing and sharing data that follows the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor and Common Rule definitions of de-identified, and therefore cannot be linked with data that has not

been consented for sharing or includes identifiers that violate these de-identification requirements (e.g., HIPAA Limited Datasets that include dates and certain geographical information).

*NICHD is actively developing a metadata schema that will facilitate collecting, structuring, and standardizing dataset-level data governance information such that it can be exchanged in a machine-readable format to facilitate understanding of how these rules intersect.*

**4. Establish which party should link the data.**

One option is to have a central party share already merged and de-duplicated datasets with users. Alternatively, providing researchers individual datasets alongside linkage maps or GUIDs allows them to link the data themselves while maintaining dataset-specific provenance. In the latter approach, researchers may need to navigate different rules depending on the combinations of datasets they choose to link.

**5. Use a variety of controls for mitigating re-identifiability risk to account for unknown re-identification risk introduced by linkage.**

Common risk mitigation strategies include:

- Applying a common definition of de-identified to all linked datasets (e.g., HIPAA Safe Harbor and/or Common Rule)
- Using controlled access approaches to share linked data and/or linkage information (e.g., GUIDs or linkage maps). Controlled access processes are used to verify appropriate use of shared data prior to access and often include requiring verification of requestor identity, committee approval of proposed research use, and signing a data use agreement
- Prohibiting re-identification of participants in the linked and shared data

More sensitive data may warrant more stringent controls such as:

- Systematic transformations (modifications) or aggregations of certain data elements
- Formal disclosure review, expert determination, or other re-identification risk assessments prior to and/or after linkage

- Access via a physical or virtual enclave (i.e., where data cannot leave the access environment)

## Technical Considerations

If using PPRL, the same tool must be used for all datasets that are part of the implementation. Technical factors to consider when selecting a PPRL tool include:

**6. Collect and standardize a broad set of PII elements.**

A broad set of PII elements are required to generate high-quality linkage, regardless of the PPRL technology used. These PII elements should be collected at the outset and in a standardized manner even if they are typically not collected in the course of a research project.

- Common elements include first name, last name, date of birth, sex or gender, and some form of "location" information. Standard definitions are important especially for sex and/or gender. Location of birth (e.g., city/municipality of birth) is more stable than household address, which tends to change over time, and is therefore a better choice especially for facilitating longitudinal linkage. Social Security Number (SSN), phone, and email are also sometimes used; however, the use of SSN is subject to complex security requirements and regulations and phone numbers and email addresses often do not apply for children.

- New PII combinations for a given tool require rigorous statistical assessments using relevant gold standard datasets to inform the best configuration (e.g., weighting).

**7. Select PPRL software that meets basic requirements.**

Many PPRL tools (commercial, open source, and government owned) can accommodate a broad and flexible set of PII, support large scale implementations, prohibit vendor rights to the data, and appropriately protect PII. The PPRL tools diverge on certain desirable features associated with usability (e.g., graphical user interfaces, preprocess/data cleaning), functionality (matching algorithm tuning), and security certifications (e.g., FedRAMP), which may factor into deciding which software is best for a given implementation.

**8. Consider PPRL software sustainability for long-term implementations.**

Assuming basic requirements are met, software sustainability could be a primary driver in PPRL tool selection. Long-term implementations require that the hashed codes persist over time and may benefit from the use of government-owned and maintained software to avoid continual commercial vendor contracts, recurring or use-based costs, and risk associated with business model modifications (e.g., mergers/acquisitions, bankruptcy) that may lead to tool deprecation.