Westat

Improving Lives Through Research®

Privacy-Preserving Record Linkage

Balancing Privacy, Linkage Quality, and Statistical Confidence in Cross-Organization Data Collaboration

November 16, 2026

Authors
Tom Krenzke
Marcelo Simas
Michael Mitchell
John Riddles
Gizem Korkmaz
Liz Mannshardt
Jay Clark

This capability statement includes data that shall not be disclosed outside the government and shall not be duplicated, used, or disclosed—in whole or in part—for any purpose other than to evaluate this capability statement. If, however, a contract is awarded to this offeror as a result of—or in connection with—the submission of data, the government shall have the right to duplicate, use, or disclose the data to the extent provided in the resulting contract. This restriction does not limit the government's right to use information contained in data if it is obtained from another source without restriction. The data subject to this restriction are contained on all sheets.

Table of Contents

1. Introduction **1**

2. The Data Collaboration Challenge **1**

Data Is More Valuable When Connected 2
Why Traditional Data Sharing Is Increasingly Difficult 2

3. What Is Privacy-Preserving Record Linkage? **3**

Understanding Record Linkage 3
What Makes PPRL Different? 4
Core Technology Components 4

4. The Evolution of PPRL Technology **6**

First-Generation Approaches 6
Bloom Filters and Cryptographic Long-Term Keys 7
Modern PPRL Ecosystems 7

5. The Accuracy Challenge: Why Match Quality Still Matters **7**

Privacy Alone Is Not Enough 7

The Tradeoff 8

6. The Missing Dimension: Statistical Quality of Linked Data **8**

Why Analysts Care About More Than Match Rates 9
Statistical Risks Introduced by Linkage Errors 9

7. A Practical Framework for Evaluating PPRL Solutions **11**

8. Illustrative Evaluation of a Modern PPRL Solution **13**

Data and Methodology 13
Privacy Assessment 13
Results and Lessons Learned 14

9. Real-World Applications **15**

Public Health 15
Clinical Research 15
Healthcare Quality 15
Education and Social Policy 16

Table of Contents (continued)

10. The Future of Privacy-Safe Analytics **16**

Conclusion **17**

References **18**

Exhibits

Exhibit 1. Potential solutions to data sharing challenges 2
Exhibit 2. The PPRL process 6
Exhibit 3. Three pillars of PPRL evaluation 11
Exhibit 4. Solutions to data sharing rated for the three pillars of PPRL evaluation and scalability 12

1. Introduction

Organizations increasingly need to connect information across sites, clinics, healthcare systems, research institutions, commercial enterprises, and other organizations. Valuable insights often emerge not from a single dataset, but from integrating disparate datasets and analyzing how patterns and trends emerge and evolve over time. However, growing privacy concerns make traditional approaches to data sharing increasingly difficult.

Privacy-Preserving Record Linkage (PPRL) has emerged as a critical capability for enabling real-world data collaboration without exposing personally identifiable information (PII). Rather than exchanging names, addresses, dates of birth, or other sensitive identifiers, organizations transform identifying information into privacy-preserving representations that support linkage while minimizing disclosure risk. PPRL allows participants to determine whether records correspond to the same individual without directly sharing underlying identifiers (Vatsalan et al., 2022; Schnell et al., 2009).

In practice, organizations invest in record linkage to support program evaluation, healthcare research, fraud detection, customer analytics, public health surveillance, education and social policy, and AI applications. Over the past 2 decades, PPRL has evolved from relatively simple hash-based methods into sophisticated ecosystems that combine tokenization and cryptographic protections, probabilistic matching, distributed governance models, and scalable computational architectures. Many discussions of PPRL focus almost exclusively on cryptography and security.

Yet privacy protection alone is not enough: If a linkage solution produces excessive false matches, misses substantial portions of the population, or introduces systematic bias, the resulting analyses may be misleading regardless of how well privacy is protected. For this reason, organizations evaluating PPRL solutions should consider three pillars simultaneously:

1. **Privacy protection.** How effectively does the solution minimize disclosure risk and protect sensitive information?
2. **Linkage performance.** How accurately and efficiently does the solution identify matching records?
3. **Statistical fitness for use.** Does the resulting linked dataset support valid downstream analytics and decision-making?

Modern approaches can achieve linkage quality comparable to traditional methods while significantly reducing privacy risks (Brown et al. 2017). The most successful PPRL implementations recognize that privacy, linkage quality, and analytical utility are inseparable. The goal is not to maximize any single dimension in isolation, but to achieve an appropriate balance among all three.

2. The Data Collaboration Challenge

This paper describes the evolution and strengths of PPRL in supporting program evaluation, healthcare research, fraud detection, customer analytics, public health surveillance, education and social policy, and AI applications. It describes the integration of statistical fitness for use (i.e., the statistical risks and considerations for linking human subjects' data across entities and systems) into PPRL evaluation. It details emerging solutions and presents a practical framework for evaluating PPRL applications. The framework describes both linkage considerations and fitness for use for population representativeness and downstream analytics. Finally, it describes the future of privacy-safe analytics. We begin with the data collaboration challenge.

Data Is More Valuable When Connected

Modern organizations operate in increasingly fragmented data environments. A patient may appear in hospital systems, insurance claims databases, disease registries, public health systems, and clinical research programs. A person may interact with multiple government programs administered by separate agencies (e.g., Medicaid, Veterans Health Administration, Supplemental Nutrition Assistance Program). A consumer may generate information through surveys, loyalty programs, digital interactions, and transaction histories. Each dataset provides a partial view; greater value emerges when information can be linked across multiple datasets.

The ability to integrate information across organizational boundaries can improve and expand service delivery, policy evaluation, population health management, fraud and anomaly detection, and customer understanding.

Examples of greater value through linkage

- State agencies combining healthcare, social services, and public health data
- Health systems linking clinical and claims records
- Educational organizations studying long-term student outcomes
- Research organizations enriching survey data with administrative records
- Financial institutions building more complete customer risk profiles

Why Traditional Data Sharing Is Increasingly Difficult

Historically, linkage often relied on exchanging PII and performing linkage using raw names, addresses, dates of birth, and other identifiers. This approach has become increasingly challenging because of

- Expanding privacy concerns
- HIPAA and other regulatory obligations
- Cybersecurity risks
- Data ownership considerations
- Public trust concerns
- Organizational governance requirements

Many organizations are understandably reluctant to transfer sensitive identifiers beyond their own trust boundaries. The result is a growing tension between two objectives: maximizing the value of data through collaboration and minimizing privacy and security risk. Today, several approaches exist to address this challenge, as shown in Exhibit 1. This paper focuses on PPRL approaches.

Exhibit 1. Potential solutions to data sharing challenges

Method	Description
Clear-text probabilistic linkage	Accuracy benchmark; not privacy-preserving.
Trusted third-party tokenization PPRL	Operationally mature; requires trust in central token provider. See Petersen et al. (2023).
Deterministic hashing PPRL	Exact match only; brittle to data quality issues. See Vatsalan et al. (2022).
Bloom filter PPRL	Strong balance between privacy and approximate matching performance. See Petersen et al. (2023) and Vatsalan et al. (2022).
Zero-trust tokenization PPRL	Decentralized key ownership, no trusted intermediary, recipient-specific transcription, designed for interoperability and large-scale Spark processing.*

Method	Description
Secure multiparty computation (SMPC)	Protects data by splitting the computation across parties so no participant can see anyone else's data. Strong cryptographic guarantees but higher computing costs. See Vatsalan et al. (2022) and Gkoulalas-Divanis et al. (2021).
Trusted execution environments (TEE)	Secure hardware-based enclaves that allow sensitive record linkage computations to be performed in an isolated, attested environment, protecting data from the host operating system and other unauthorized processes. Protects data by performing near-clear-text matching and computation inside a secure hardware enclave. See Nie et al. (2025).
Homomorphic encryption	Enables computations to be performed directly on encrypted data without decrypting it first, allowing parties to conduct record linkage or analytics while keeping sensitive information hidden throughout the processing pipeline. Currently the most computationally expensive option. See Vatsalan et al. (2022) and Gkoulalas-Divanis et al. (2021).

Note: PPRL = Privacy-Preserving Record Linkage.

* For more information, see https://token.spindlehealth.com/.

3. What Is Privacy-Preserving Record Linkage?

Understanding Record Linkage

Record linkage seeks to determine which records across distinct datasets correspond to the same real-world entity. Traditional linkage typically uses one of three approaches:

- **Deterministic matching** relies on fixed rules with exact agreement among selected fields to link records.
- **Probabilistic matching** evaluates evidence across multiple fields, including the frequency of values and partial agreement among selected fields, and links records depending on whether likelihood ratios meet specified thresholds.
- **Entity resolution** encompasses broader techniques for identifying duplicates and relationships across datasets.

For decades, probabilistic linkage methods derived from the Fellegi-Sunter framework (Fellegi-Sunter, 1969) have served as the statistical foundation for large-scale linkage systems. The key metrics are $\log\left(\frac{m}{u}\right)$ for matching fields that agree and $\log\left(\frac{1-m}{1-u}\right)$ for matching fields that disagree, where

- m = probability that a field agrees given a true match
- u = probability that a field agrees given a non-match

Probabilistic linkage methods derived from the Fellegi-Sunter framework quantify matching evidence by summing these likelihood ratios over all matching fields for each given pair of records. Fields that are highly predictive of true matches receive larger positive scores, while common agreements contribute less evidence. Recent record linkage systems may also employ machine learning techniques that learn matching patterns directly from data, extending or augmenting traditional probabilistic linkage approaches.

Westat has extensive experience developing and applying record-linkage methodologies across large-scale federal research, survey, and administrative data systems. For the U.S. Department of Agriculture's Child Nutrition Analysis Modeling Blank Purchase Agreement, Westat developed a novel methodology to link Common Core of Data district-level files with Food and Nutrition Service administrative records in the absence of a unique identifier. After evaluating multiple alternatives, the team implemented a hybrid deterministic and probabilistic matching approach that addressed data harmonization challenges such as deduplication, name standardization, and match quality, and deployed the solution in SAS, R, and Python.

Under multiple contracts with the National Cancer Institute, Westat conducted longitudinal linkages for the National Institutes of Health AARP Diet and Health Study, encompassing more than 500,000 participants and over 1.35 million records. Westat designed customized matching systems to efficiently link cohort records to state cancer registries, the National Death Index, Centers for Medicare & Medicaid Services data, and contextual geographic datasets while balancing sensitivity and specificity and minimizing manual review. Westat has also applied probabilistic matching methods to assess disclosure risk in public-use data products. For the Office of Personnel Management's Federal Employee Viewpoint Survey, Westat estimated the probability that records in a partially synthetic public-use file could be matched back to original respondents and used these probabilities to evaluate release risk. Similarly, for the U.S. Department of Agriculture's National Household Food Acquisition and Purchase Study, Westat evaluated the re-identification risks associated with geographic clustering and used probabilistic record-linkage methods to identify counties with elevated disclosure risk.

Across these efforts, the Fellegi-Sunter record-linkage framework, embedded in Westat's *WesLink* software, has consistently produced high-quality matches; however, the resulting linked datasets can be susceptible to disclosure risk and therefore require careful confidentiality and data-protection assessments.

What Makes PPRL Different?

PPRL extends record linkage into privacy-sensitive environments. Instead of sharing raw identifiers, participating organizations generate privacy-preserving representations, often called tokens, from identifying information. These transformed values can then be used to support matching without directly exposing underlying identifiers. PPRL allows organizations to

- Match records across systems or studies
- Identify overlapping populations
- Build longitudinal datasets
- Support federated collaboration and interoperability
- Substantially reduce risk of exposure of sensitive information

Core Technology Components

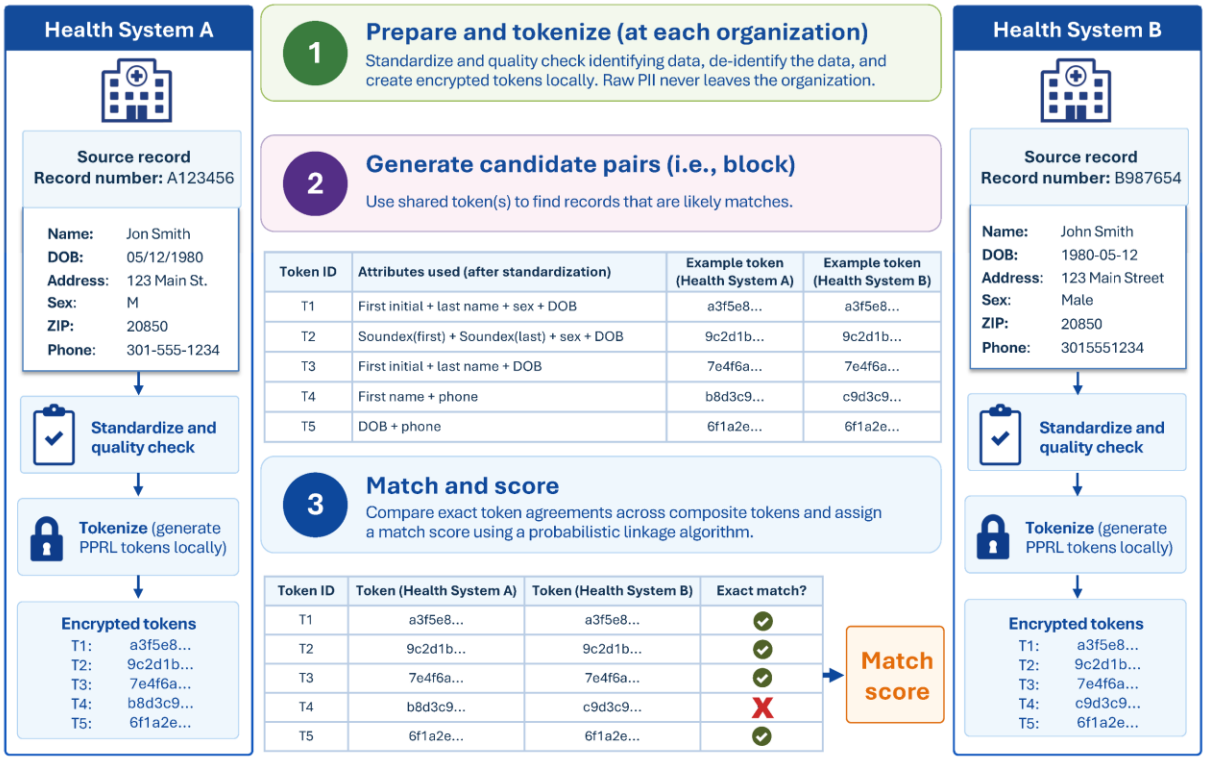
PPRL begins long before identifiers are transformed into privacy-preserving tokens. As shown in Exhibit 2, consider two records that refer to the same patient in separate health systems. One source may record "John Smith" living at "123 Main St." with a date of birth entered as "05/12/1980," whereas another source may contain a record with "John Smith," "123 Main Street," and "1980-05-12." Small differences in formatting, abbreviations, missing values, or data-entry practices can prevent records from linking correctly if they are not addressed before tokenization.

As a result, successful PPRL implementations depend on a sequence of interconnected processes, including data standardization, quality assessment, tokenization, candidate generation, matching, quality evaluation, and governance. Together, these components determine both the accuracy and scalability of PPRL systems, often having a greater impact on linkage quality than the tokenization method itself. Therefore, most modern PPRL architectures include several common elements:

- **Data standardization and quality assessment.** Names, addresses, dates, and other identifiers are cleaned, normalized, and assessed for completeness and consistency before tokenization.
- **Tokenization and cryptographic transformation.** Identifiers are converted into privacy-preserving representations (e.g., tokens as shown in Exhibit 2). This includes the practice of defining multiple tokens from combinations of identifiers to support blocking in PPRL and to allow some degree of "fuzzy" matching (i.e., one token may use "full first name, last name, date of birth," whereas another uses "first name initial, last name, date of birth").
- **Blocking and candidate generation.** Potential matches are narrowed to a computationally manageable subset. Subject matter expertise can be important in reducing dimensionality.
- **Probabilistic matching.** Multiple pieces of evidence are evaluated jointly.
- **Linkage quality assessment.** Matching results are reviewed to detect differential match rates across key characteristics.
- **Linkage governance and auditing.** Controls help ensure transparency and accountability.

A common misconception is that token generation is the most difficult aspect of PPRL. In reality, transforming identifiers is usually computationally straightforward. The more challenging problem is efficiently identifying likely matches within very large populations. Linkage can quickly become computationally expensive because candidate comparisons scale rapidly with dataset size. Effective blocking, indexing, and probabilistic matching strategies are therefore essential components of scalable PPRL implementations.

Exhibit 2. The PPRL process



Note: DOB = date of birth; PII = personally identifiable information; PPRL = Privacy-Preserving Record Linkage.

4. The Evolution of PPRL Technology

First-Generation Approaches

Early PPRL efforts focused on enabling linkage without sharing raw identifiers. One of the most influential contributions came from Schnell et al. (2009), who demonstrated that Bloom filter encodings could support approximate record linkage using encrypted identifiers. Their work established the foundation for much of modern PPRL research. These early approaches demonstrated feasibility but also revealed limitations, including susceptibility to certain attack models mentioned in the following.

Among the early PPRL efforts, researchers were also advancing broader privacy-preserving data publishing techniques. For example, Terrovitis et al. (2011) developed local and global recoding approaches for anonymizing set-valued data, demonstrating how information can be transformed to reduce disclosure risk while preserving analytical utility. Although not a record linkage method itself, this work helped establish many of the conceptual tradeoffs that continue to shape PPRL today, particularly the need to balance privacy protection with information loss and downstream analytical performance.

Bloom Filters and Cryptographic Long-Term Keys

Bloom filters were originally selected because they allow for computing distances between them, which can be used to indicate how much agreement exists between the input data used to generate different tokens. This characteristic allowed for the direct application of them in the context of probabilistic matching and supported the production of high-quality matches. However, they can be vulnerable to frequency, cryptanalysis, and pattern-mining attacks. In such attacks, adversaries may exploit the statistical properties of encoded identifiers to reconstruct sensitive personal information or infer underlying values using publicly available data, such as a list of common names or addresses. These findings prompted the development of more secure encoding, salting, and cryptographic techniques within the PPRL literature.

Subsequent research, such as by Brown et al. (2017), introduced composite Bloom filters and cryptographic long-term keys, allowing multiple identifying attributes to be combined into multiple privacy-preserving linkage representations. Large-scale evaluations demonstrated that these techniques could achieve high precision and recall while preserving privacy. However, they often required careful parameter tuning and greater computational resources than conventional linkage approaches. Researchers, such as Stetsikas et al. (2026), also identified vulnerabilities associated with frequency attacks, dictionary attacks, and repeated use of encoded identifiers across multiple linkage exercises. These concerns encouraged development of next-generation approaches that strengthen privacy while reducing dependence on centralized trust structures.

Modern PPRL Ecosystems

As tokenization in PPRL frameworks advance, organizations are increasingly separating token generation from linkage operations. This distinction is important because tokenization solutions generate privacy-preserving identifiers, whereas linkage solutions evaluate evidence across records and determine matching relationships. A token generator alone does not necessarily solve the record linkage problem. For example, some modern tokenization frameworks provide highly scalable privacy-preserving transformations but rely primarily on deterministic token agreement. These approaches may perform well when data are highly standardized but can struggle with typographical errors, nickname variations, missing values, incomplete fields, and other common real-world data quality problems.

Probabilistic linkage code packages such as SPLINK in Python address these challenges through statistical matching methodologies, candidate generation strategies, and linkage quality evaluation frameworks. Organizations increasingly combine privacy-preserving tokenization technologies with probabilistic linkage engines to achieve both strong privacy protections and high linkage performance.

5. The Accuracy Challenge: Why Match Quality Still Matters

Privacy Alone Is Not Enough

Privacy protections are essential, but they are not the ultimate objective. Organizations pursue linkage because they want reliable answers. A privacy-preserving system that links the wrong records or misses substantial portions of the target population may undermine the very purpose of the analysis. Consequently, decision-makers should ask: Are records being matched correctly? How many true matches are missed? How are linkage rates measured? How are results validated? In an

evaluation environment, where the ground truth (set of true matches) is known, key measures of linkage performance include

- **Precision** measures the proportion of identified matches that are correct. High precision minimizes false positives.
- **Recall** measures the proportion of true matches successfully identified. High recall minimizes false negatives (missed matches).
- **False positives (FPs)** occur when records from different individuals are incorrectly linked. These errors can contaminate analyses and create misleading relationships.
- **False negatives (FNs)** occur when records that belong to the same individual fail to link. These errors reduce completeness and population coverage.

Precision is measured as $\frac{TP}{TP + FP}$, where TP is the number of true matches correctly identified. Recall is measured by $\frac{TP}{TP + FN}$. A precision of 99 percent indicates that 99 percent of declared matches are correct, and a recall of 95 percent indicates that 95 percent of true matches are successfully identified. Effective PPRL solutions seek to optimize both measures simultaneously. Some measures integrate precision and recall into a single measure, such as the F1 score, which is defined as $2 \times \frac{Precision \times Recall}{Precision + Recall}$.

When ground truth is unavailable, *m*- and *u*-probabilities can be estimated directly from the observed comparison patterns using latent class or Expectation-Maximization methods. The EM algorithm iteratively estimates (1) how many pairs are likely matches, (2) the probability of agreement among those likely matches (the *m*-probabilities), and (3) the probability of agreement among likely nonmatches (the *u*-probabilities). It treats true match status as an unobserved variable and iteratively estimates the probability of attribute agreement among likely matches (*m*-probabilities) and likely nonmatches (*u*-probabilities). As a result, linkage quality can be assessed without requiring a fully labeled set of matched and unmatched record pairs.

The Tradeoff

Research, such as by Mirel et al. (2022), has demonstrated that stronger privacy protections sometimes introduce modest reductions in precision, recall, or computational efficiency compared with linkage using clear-text identifiers. The objective is therefore not absolute privacy at any cost, nor maximum linkage rates at any cost. The objective is optimization. Organizations need sufficient privacy protection while retaining enough linkage performance to support valid analytical outcomes.

6. The Missing Dimension: Statistical Quality of Linked Data

Beyond linkage accuracy, an important consideration is the statistical quality of the resulting linked dataset. Even when true match status is unavailable, evidence of linkage bias can be evaluated by assessing whether linkage success differs across population subgroups and whether the linked sample remains representative of the target population. These factors ultimately determine the validity and generalizability of analyses using linked data.

Why Analysts Care About More Than Match Rates

Valuable data come from linkages across systems, such as linking traffic accident data to hospital records, student loan data to STEM innovation, or Internal Revenue Service data to housing and real estate. Evaluating the data becomes essential for judging its statistical utility. Many PPRL evaluations stop at linkage accuracy metrics, but precision and recall represent only intermediate outcomes. The ultimate question is whether linked data support valid analysis and decision-making. Good linkage improves

- Sampling frame creation
- Family and generational matching (i.e., mother–child records)
- Population estimation
- Program evaluation
- Patient outcome research
- Customer analytics
- Predictive modeling
- AI applications

Statistical Risks Introduced by Linkage Errors

Coverage bias. Certain populations may be systematically harder to match because of data quality differences, mobility, language variation, or incomplete identifiers. If these groups are underrepresented among linked records, resulting analyses may become biased. This is the missing dimension, that is, it measures the statistical quality of linked data by computing

$$\text{Coverage rate} = \frac{\text{Linked records}}{\text{Eligible records}}, \text{ or better yet,}$$

$$\text{Coverage bias} = \text{Coverage}_{Group\ A} - \text{Coverage}_{Group\ B}.$$

Even when overall linkage accuracy appears high, differences in coverage rates across demographic groups can indicate potential bias. For example, a 95 percent linkage rate for one population and an 85 percent rate for another may lead to systematic underrepresentation in downstream analyses.

Certain populations may be systematically more difficult to link because of differences in identifier quality, residential mobility, naming conventions, language variation, or missing data. When linkage rates vary across subgroups, the resulting linked dataset may underrepresent portions of the target population.

Linkage bias. Matched records may not accurately represent the broader target population. Differential linkage success can produce a linked sample that differs from the population that investigators intend to study. As a result, analytical findings drawn from linked records may not generalize to the full target population. For example, suppose younger individuals move more often and are harder to link, such that the overall linkage rate is 94 percent, the age 18–29 linkage rate is 82 percent, and the age 65-plus linkage rate is 98 percent. The resulting linked sample would overrepresent older adults. Similarly, women are more likely to change their last names and may have lower match rates. Subsequently, the analyses of healthcare utilization may be biased because the linked population is no longer representative of the target population.

Another approach is to compare characteristics in the source files to those in the linked sample. Substantial differences suggest selective linkage. In addition, approaches may include a sensitivity analysis to gauge the stability of results.

In general, differential linkage rates (or differences from benchmark percentages) across demographic groups can introduce linkage bias, which in turn produces selection bias in the linked analytic sample. As a result, estimates derived from linked records may not accurately reflect characteristics or outcomes in the full target population.

Measurement error. False matches create relationships that do not exist. Clerical or consistency checks can be conducted to ensure the results make sense. Missed matches conceal relationships that do exist. Both conditions distort analytical results.

Equity considerations. Linkage quality is rarely uniform across demographic groups. Disparities in data quality, naming conventions, housing stability, or identifier completeness can lead to different linkage rates across populations. These differences can have important implications for policy, healthcare, and program evaluation.

Evaluating fitness for purpose. A linkage solution should not be judged exclusively by privacy metrics or match rates.

Research such as by Tyagi and Willis (2024) on real-world data has shown that carefully implemented PPRL approaches can achieve high linkage accuracy and, at the same time, produce analytical results comparable to traditional linkage methods. However, these outcomes should be demonstrated rather than assumed.

Organizations should also evaluate the following:

- Population coverage
- Representativeness
- Demographic consistency
- Bias assessments
- Downstream analytical impacts

As mentioned previously, without ground truth, we can often assess representativeness and robustness better than accuracy. In other words, we may be able to detect linkage bias reasonably well, but estimating the true FP and FN rates is much harder.

7. A Practical Framework for Evaluating PPRL Solutions

Organizations evaluating PPRL solutions should consider the three pillars shown in Exhibit 3. Effective PPRL requires more than protecting identifiers. Solutions must balance privacy protection, linkage performance, and statistical fitness for use. A system that performs well in only one area may still fail to meet program objectives. The three pillars provide a practical framework for assessing whether a PPRL solution adequately protects sensitive information, produces accurate and scalable linkages, and supports valid downstream analysis.

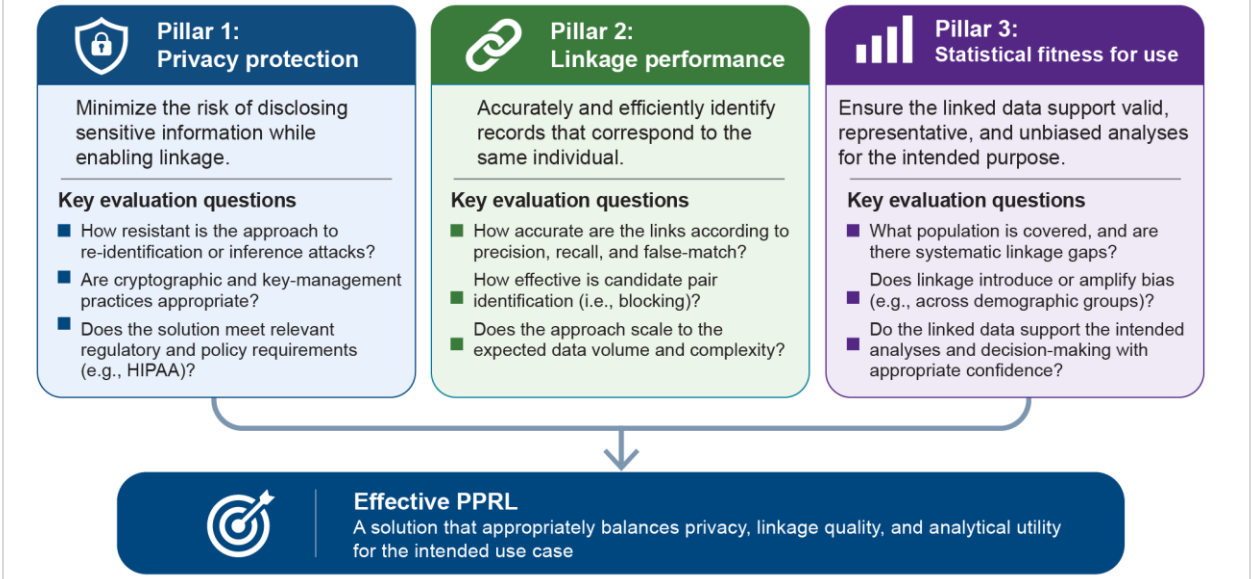
Pillar 1: Privacy protection. Evaluates disclosure risk, attack resistance, key management, governance, and compliance controls.

Pillar 2: Linkage performance. Evaluates match quality, blocking effectiveness, scalability, and computational efficiency.

Pillar 3: Statistical fitness for use. Evaluates coverage, representativeness, linkage bias, transparency, and suitability for the intended analytical purpose.

These pillars assess whether a PPRL solution achieves an appropriate balance between privacy, linkage quality, and analytical utility.

Exhibit 3. Three pillars of PPRL evaluation



As mentioned previously, there are several approaches to the problem. Exhibit 4 provides a high-level assessment of the privacy, linkage performance, statistical fitness for use, and scalability of the variable approaches.

Exhibit 4. Solutions to data sharing rated for the three pillars of PPRL evaluation and scalability

Method	Privacy*	Linking performance**	Statistical fitness for Use	Scalability
Clear-text probabilistic linkage	Very low	Very high	Very high	Very high
Trusted tokenization PPRL	Moderate	High to very high	Very high	Very high
Deterministic hashing PPRL	High	Low	Low to moderate	Very high
Bloom filter PPRL	Moderate to high	High	High	High to very high
Zero-trust PPRL	High to very high	High to very high***	High to very high	Very high
TEE-based linkage	High to very high	Very high	High to very high	High
SMPC	Very high	High	Moderate	Low to moderate
Homomorphic encryption	Very high	High	Low to moderate	Very low to low

Note: PPRL = Privacy-Preserving Record Linkage; SMPC = secure multiparty computation; TEE = trusted execution environments.

* The privacy ratings reflect a combination of factors, including exposure of plaintext identifiers, resistance to record-reconstruction and re-identification attacks, trust assumptions, cryptographic protections, and the consequences of key compromise.

** Linkage performance reflects the accuracy of the complete matching end-to-end workflow, including token generation, blocking, comparison, and scoring. Data cleaning and standardization are not linkage outcomes themselves, but they substantially influence performance and should therefore be considered part of the overall linkage solution when evaluating real-world deployments.

*** Although this approach uses deterministic cryptographic tokens, linkage accuracy is achieved through probabilistic comparison of multiple independent attribute-level tokens rather than exact matching on a single composite hash. This allows the system to tolerate missing values, typographical errors, phonetic name variations, and other data quality issues that would cause traditional deterministic hashing approaches to fail.

8. Illustrative Evaluation of a Modern PPRL Solution

To illustrate the practical performance of a contemporary zero-trust PPRL approach, we evaluated tokens generated using Spindle Health's[1] Open-Source Privacy-Preserving Record Linkage (OPPRL) framework using Spindle Token version 2.4.2. The probabilistic record linking was performed using SPLINK version 5. The evaluation was designed to simulate a decentralized token-generation process followed by centralized linkage. The analysis focused on three questions: (1) whether the approach accurately identifies true matches while minimizing FPs, (2) whether the linkage process is computationally feasible at scale, and (3) what privacy, data quality, and performance tradeoffs emerge when using token-based linkage methods.

Data and Methodology

The evaluation was conducted in a controlled two-source environment in which the true match status of all records was known. Although this provides a useful benchmark for assessing linkage accuracy, real-world PPRL deployments are typically more complex. Production environments often involve multiple data providers, duplicate records, missing information, heterogeneous data quality, and the absence of ground truth against which linkage performance can be directly measured. The evaluation used approximately one million records, including roughly 500,000 true overlapping records between source datasets. A perturbed comparison dataset was created to simulate common data quality issues encountered during operational record linkage. Modifications included birth-date inconsistencies, formatting variations, and other attribute-level errors intended to mimic realistic reporting and data-entry conditions.

Using the OPPRL framework, linkage tokens were generated using five standard token specifications composed of combinations of birth date, first and last name encodings, gender, and Social Security Number. Prior to token generation, limited preprocessing was performed to correct selected birth-date formatting issues and other invalid values that would otherwise prevent token creation.

Candidate record pairs were generated through token-based blocking. Rather than relying on demographic fields such as state or birth year, records were considered candidate pairs only when at least one OPPRL token matched exactly. Because OPPRL tokens are designed for exact agreement rather than fuzzy comparison, this approach substantially reduced the number of comparisons required while maintaining computational efficiency. Linkage scoring and classification were then performed using SPLINK's probabilistic matching framework.

Privacy Assessment

The linkage process replaced direct identifiers with cryptographically generated tokens. The resulting linkage dataset contained tokenized representations of PII rather than the underlying values themselves. The generated tokens cannot be reversed to recover the original identifiers without access to the corresponding cryptographic key material.

Nevertheless, tokenized records should be regarded as de-identified or pseudonymized rather than fully anonymous. As with any privacy-preserving record linkage implementation, organizations should apply appropriate governance, security controls, access restrictions, and data-use agreements when conducting linkage activities.

[1] For more information, see https://github.com/spindle-health/spindle-token.

Results and Lessons Learned

As mentioned above, the model was evaluated on 1,000,000 records, including approximately 500,000 true overlapping records between datasets. The linkage process identified 372,412 true matches while producing only 43 FP matches and 127,253 FN matches. This corresponded to a precision of 99.99 percent, recall of 74.5 percent, and an F1 score of 85.4 percent.

These results demonstrate that token-based PPRL approaches can achieve exceptionally high match confidence while maintaining strong computational performance at scale. The near-zero FP rate indicates that links identified by the model are highly reliable. At the same time, the evaluation highlights an important tradeoff between precision and recall. Approximately one-quarter of true matches were not recovered, reflecting the conservative nature of the linkage strategy.

The evaluation further demonstrated strong scalability. By requiring exact agreement on at least one OPPRL token during candidate generation, the linkage process reduced the number of record comparisons dramatically and processed approximately one million records in roughly five minutes within the test environment. These findings suggest that modern PPRL methods can support large-scale linkage applications without requiring exhaustive comparison of all possible record pairs.

Several lessons emerged from the evaluation. First, token design has a substantial influence on linkage performance. Four of the five OPPRL tokens used in this evaluation incorporated date-of-birth information. Consequently, perturbations affecting birth date could invalidate multiple tokens simultaneously and reduce the probability that a true match would survive the blocking stage. Note that birthdate is typically not incorrect in health systems, perturbing birthdate was a way to help illustrate the competing objectives. It also shows that it is important to have good data quality checks after linkage occurs. Second, highly restrictive blocking strategies can improve computational efficiency and precision but may also reduce recall by preventing otherwise matching records from being compared. Finally, the evaluation illustrates that achieving strong privacy protection, operational scalability, and linkage quality requires balancing competing objectives rather than optimizing any single performance metric.

From a privacy perspective, the evaluation also illustrates the value of multi-attribute token construction. Single-attribute tokens can be vulnerable to frequency-based inference attacks because common and rare values often produce distinguishable token distributions. Multi-attribute tokens reduce this risk by combining information from several attributes, increasing the number of possible token values and making frequency patterns less informative. This consideration is particularly important for attributes with fewer categories such as sex, state, race, or other variables with a limited number of possible values. By combining multiple attributes within a token, modern OPPRL approaches can strengthen privacy protections while still supporting accurate record linkage.

9. Real-World Applications

Public Health

Public health agencies frequently need to integrate immunization registries, disease surveillance systems, claims data, and program participation records to support population health monitoring and intervention programs. PPRL enables these data assets to be linked without broad sharing of PII, helping organizations balance data utility with privacy protection. Benefits include

- Deduplication
- More complete population-level views
- Improved disease surveillance and outbreak response
- Enhanced program evaluation and performance measurement
- Reduced disclosure and re-identification risk

Clinical Research

Clinical research programs increasingly depend on data integration across clinical trials, electronic health records, registries, laboratory systems, and administrative data sources. PPRL supports

- Longitudinal participant follow-up
- Real-world evidence generation
- Clinical trial enrichment and external control arms
- Linkage of research and clinical data while protecting participant privacy
- Use of services for individual health journeys
- Respecting local healthcare system trust boundaries

As multisite and decentralized research initiatives expand, privacy-preserving approaches allow organizations to collaborate more effectively without exposing direct identifiers.

Healthcare Quality

Healthcare organizations seek comprehensive patient views spanning providers, health systems, payers, and care settings to measure quality, outcomes, and utilization. PPRL can support

- Quality measurement and reporting
- Value-based care initiatives
- Care coordination and care management
- Population health and outcomes analytics

Emerging decentralized architectures are particularly attractive because they preserve local stewardship of identifiers while enabling secure cross-organizational linkage for quality improvement activities.

Education and Social Policy

Researchers and policymakers frequently seek to combine survey, administrative, behavioral, workforce, and program data to better understand outcomes and evaluate interventions. PPRL supports

- Linking education, workforce, and social services data
- Enriching survey data with administrative records
- Improving program evaluation and policy research
- Reducing data acquisition and respondent burden
- Strengthening privacy protections for vulnerable populations

These applications highlight the importance of assessing not only linkage quality, but also the downstream impact of linkage error on analytic validity and policy conclusions.

10. The Future of Privacy-Safe Analytics

The future of data collaboration increasingly points toward environments where privacy and analytics are designed together rather than treated as competing objectives. Several trends are shaping this evolution:

- Privacy-enhancing technologies
- Federated analytics
- Distributed governance models
- Secure multiparty computation
- Trusted research environments
- Responsible AI

PPRL is becoming the foundational infrastructure for these ecosystems. At the same time, future systems will likely place greater emphasis on decentralized architectures that reduce reliance on centralized brokers, preserve local control of identifiers, and support flexible interoperability among participating organizations.

The most promising approaches increasingly separate privacy-preserving token generation from statistical linkage and quality assessment. This modular architecture enables organizations to evaluate privacy, linkage performance, governance, and analytical outcomes independently while continuously improving each component.

Conclusion

PPRL has evolved from an experimental research concept into a practical capability supporting large-scale data collaboration across healthcare, government, research, and commercial environments. Several lessons have emerged from the field's evolution.

First, organizations need privacy-safe approaches to data collaboration. Traditional identifier sharing is increasingly difficult to justify operationally, legally, and ethically.

Second, modern PPRL technologies have matured substantially. Advances in tokenization, cryptography, distributed governance, and scalable computation have made privacy-preserving linkage feasible at production scale.

Third, successful linkage requires balancing privacy, accuracy, and analytical utility. Strong privacy protections alone do not guarantee useful outcomes.

Fourth, organizations should evaluate PPRL solutions on three dimensions simultaneously: privacy protection, linkage performance, and statistical fitness for downstream use.

The most effective PPRL implementations recognize that privacy protection is not the destination. It is the foundation that enables trusted data collaboration, high-quality linkage, and statistically sound decision-making. When those elements are combined successfully, organizations can unlock the full value of cross-organizational data while maintaining the trust and stewardship responsibilities that modern data ecosystems demand.

References

Brown, A. P., Borgs, C., Randall, S. M., & Schnell, R. (2017). Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. *BMC Medical Informatics and Decision Making, 17*(83). https://doi.org/10.1186/s12911-017-0478-5

Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., & Kantarcioglu, M. (2021). Modern privacy-preserving record linkage techniques: An overview. *IEEE Transactions on Information Forensics and Security, 16*, 4966–4987. https://doi.org/10.1109/TIFS.2021.3114026

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Mirel, L., Resnick, D., Aram, J., & Cox, C. (2022). A methodological assessment of privacy preserving record linkage using survey and administrative data. *Statistical Journal of the IAOS*. 2022;38(2), 413–421. https://doi.org/10.3233/SJI-210891

Nie, Z., Tyndall, B., Brannock, D., Gentles, E., Parish, E., & Banger, A. (2025). A framework for the design of privacy-preserving record linkage systems. *Journal of Cybersecurity and Privacy*, *5*(3), 44. https://doi.org/10.3390/jcp5030044

Petersen, S. J., Lieberthal, R. D., Miller, K. J., & Vakil, N. H. (2023). Privacy preserving record linkage (PPRL) strategy and recommendations. National Institute on Aging. https://www.nia.nih.gov/sites/default/files/2023-08/pprl-linkage-strategies-preliminary-report.pdf

Schnell, R., Bachteler, T., & Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making, 9*(41). https://doi.org/10.1186/1472-6947-9-41

Stetsikas, L., Karapiperis, D., Papadakis, G., & Koubarakis, M. (2026). Privacy-preserving record linkage: Past, present and yet-to-come. Open Proceedings. https://openproceedings.org/2026/conf/edbt/paper-T2.pdf

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2011). Local and global recoding methods for anonymizing set-valued data. *The VLDB Journal, 20*(1), 83–106. https://doi.org/10.1007/s00778-010-0192-8

Tyagi, K., & Willis, S. J. (2024). Accuracy of privacy preserving record linkage for real world data in the United States: A systemic review. *JAMIA Open, 8*. https://doi.org/10.1093/jamiaopen/ooaf002

Vatsalan, D., Karapiperis, D., & Verykios, V. S. (2022). Privacy-preserving record linkage. *Encyclopedia of Big Data Technologies*. https://doi.org/10.48550/arXiv.2212.05682