



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Privacy-Preserving Record Linkage for Real-World Evidence

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-10-23 · privacy preserving record linkage · real world evidence · patient matching · patient tokenization · linkage validation · ehr claims linkage · pharmaceutical research · probabilistic linkage · data governance

Original article: <https://intuitionlabs.ai/articles/privacy-preserving-record-linkage-pharma-rwe>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Key Changes: Selecting a Linkage Method
 4. Architecture and Trust Boundaries
 5. Implementation Considerations and Process Changes
 6. Validation for the Intended RWE Analysis
 7. Bias, Monitoring and Procurement
 8. Data Analysis and Evidence
 9. Implications and Future Directions
 10. Conclusion
-

Executive Summary

Privacy-preserving record linkage (PPRL) connects records about the same person while limiting disclosure of identifying information. For pharmaceutical **real-world evidence (RWE)**, the central decision is whether linked electronic health records (EHRs), claims, registries, or trial records can support the intended analysis. Patient matching uses demographic fields, while regulatory assessment also considers the relevance and reliability of the resulting data. ([healthit.gov](https://www.healthit.gov)) (www.fda.gov) A defensible program therefore specifies both its trust boundaries and its acceptable consequences of false and missed matches.

There is no universal match threshold. In a **2026** two-institution evaluation using a proxy reference standard, reported recall changed from **30.6%** to **96.8%** between strict and permissive settings. (pmc.ncbi.nlm.nih.gov) This range should inform validation design rather than become a procurement guarantee. Separately, a systematic review included **five studies**, all using deterministic matching, illustrating the limited scope of that published evidence base. (pmc.ncbi.nlm.nih.gov)

The architecture should assign responsibility for source identifiers, key custody, linkage decisions, and clinical-data access. The Childhood Obesity Data Initiative (CODI) model distinguishes data partners, a key escrow, a linkage agent, and a data coordinating center. (phii.org) **Tokenization is not a legal determination of anonymity:** US Health Insurance Portability and Accountability Act (HIPAA) de-identification has specified methods, while UK guidance treats pseudonymised personal data as within data-protection law. (www.hhs.gov) (ico.org.uk) Applicable permissions, recipient access, and re-identification risk require separate assessment.

The practical recommendation is to validate a frozen configuration against an appropriate **reference sample**, including records that were not linked. Measure precision, recall, uncertainty, subgroup coverage, and sensitivity of the actual study result. The US Food and Drug Administration (FDA) recommends examining linkage errors and robustness to the degree of match; reporting guidance also calls for comparisons of linked and unlinked records. (www.fda.gov) (journals.plos.org) Choose deterministic rules when their identifier requirements fit the population; consider probabilistic or hybrid methods when their additional recovered links justify their error profile and privacy controls. Approval should cover a specific population, data partnership, software configuration, and analytical use, with revalidation triggers for material changes.

Introduction and Background

Real-world data (RWD) can be fragmented across systems that observe different parts of a patient's care. Patient matching identifies records belonging to the same person across those systems; the US Office of the National Coordinator for Health Information Technology describes demographic comparisons using names, dates of birth, telephone numbers, and addresses. ([healthit.gov](https://www.healthit.gov)) PPRL applies privacy-protecting transformations and organizational controls to that identity-resolution problem.

Begin with the evidence question. A proposed study might need claims to establish enrollment and dispensing, EHR data to characterize clinical measurements, and a registry to improve outcome ascertainment. These are proposed analytical designs, not a claim that every available source contains those variables. The European Medicines Agency (EMA) distinguishes linkage, which enriches individual records, from pooling data to increase population breadth. (www.ema.europa.eu) If the question only requires aggregate comparisons, person-level linkage should have an explicit justification.

For IntuitionLabs' life sciences readership, the perspective here is architectural and epidemiological: evaluate how a linkage decision affects the study being planned. A useful specification states the target population, exposure, outcome, follow-up window, and information absent from each source. It also defines what counts as an eligible counterpart record. Without that definition, an unmatched record is difficult to interpret.

Linkage accuracy and data fitness are separate assessments. FDA's final EHR and claims guidance was issued in **July 2024** and addresses relevance and reliability. (www.fda.gov) Correct identity resolution does not, by itself, establish that a variable measures the required endpoint or that follow-up is adequate. EMA's final RWD-specific data quality framework, published **27 March 2026**, likewise provides a broader setting for evaluating source quality. (www.ema.europa.eu)

This report provides a validation blueprint rather than a certification scheme. Its proposed controls and worksheets are recommendations. Published results retain their original populations and reference definitions, and vendor descriptions remain attributed descriptions. The intended output is a documented decision about whether the linked data are sufficiently useful and controlled for a particular RWE analysis.

Key Changes: Selecting a Linkage Method

From matching labels to observable decision rules

Deterministic linkage accepts records when specified identifier combinations agree. **Probabilistic linkage** combines evidence across fields to estimate or score match plausibility. These labels do not describe every implementation detail: a Patient-Centered Outcomes Research Institute (PCORI)-supported evaluation found that probabilistic use of individually hashed fields outperformed the traditional deterministic approaches it evaluated. (www.ncbi.nlm.nih.gov) That is evidence about the tested configurations, not a universal ordering of methods.

A deterministic rule can use several alternative combinations rather than one universal identifier. Conversely, a probabilistic pipeline may begin with exact agreement rules or candidate-generation blocks. The validation specification should record which records can become candidates, how evidence is combined, and how

conflicts are resolved. RECORD reporting guidance explicitly includes blocking variables, completeness, rules, thresholds, and manual review. (journals.plos.org)

Table 1 compares proposed selection criteria and privacy questions. These are design considerations, not product ratings.

Approach	Decision mechanism	Validation priority	Privacy and operational question
Deterministic tokens	Accept specified exact token combinations.	Test missing fields, unstable identifiers, and competing candidates.	What transformations and secrets protect each token? Unkeyed hashing needs particular scrutiny. (ico.org.uk)
Probabilistic matching	Combine field-level agreement evidence and apply a decision threshold.	Examine discrimination, threshold sensitivity, and errors outside candidate blocks.	What information remains inferable from encoded comparisons? Obfuscation alone does not establish privacy. (datasciencecampus.ons.gov.uk)
Hybrid matching	Combine deterministic passes with scored matching and an uncertainty policy.	Validate every pass and the final deduplicated identity graph.	Who can resolve uncertainty, and what identifying information can that role access? Segregation of duties remains relevant. (www.ons.gov.uk)

The table supports a decision sequence: establish available identifiers, specify permissible disclosures, then test candidate configurations. Adding complexity should have a demonstrated benefit for the intended population. An encrypted-linkage evaluation found that results depended heavily on parameter settings, reinforcing the need to test the actual configuration. (link.springer.com)

From field count to usable identifier evidence

More fields do not automatically mean better matching. An Australian evaluation found that including addresses could introduce instability because addresses differed or were missing. (link.springer.com) The relevant question is whether a field supplies reliable distinguishing information in the particular source pair.

Before choosing a method, document:

- **Rule semantics:** Specify exactly what agreement permits a link.
- **Missingness behavior:** State whether absent fields exclude a record, alter its score, or activate another rule.
- **Candidate generation:** Describe which pairs can reach the matching stage.
- **Ambiguity handling:** Define the response to multiple plausible counterparts.
- **Identity constraints:** State whether the study permits one-to-many relationships.
- **Evidence retention:** Preserve the rule or score supporting each accepted link.

These are proposed minimum specifications. Their value is that another analyst can reproduce the selection decision and identify which rule changed an observed result.

From a software choice to a study-specific choice

Method selection should follow the study's objectives and population. The Australian surveillance evaluation discussed below explicitly connects linkage choices to those considerations. (www.jmir.org) A feasibility cohort, an exposure-outcome study, and trial follow-up can assign different consequences to the same error. The sponsor should articulate those consequences before selecting a threshold, then test whether the resulting cohort remains suitable.

Architecture and Trust Boundaries

Separate identity processing from clinical analysis. The Public Health Informatics Institute's CODI documentation assigns hosting and exchange functions to data partners, matching to a linkage agent, key creation to a key escrow, and longitudinal assembly to a coordinating center. (phii.org) The following matrix adapts those roles into a proposed pharmaceutical implementation.

Table 2 describes minimum access for this proposed architecture. Actual access must be checked against deployment contracts and technical controls.

Role	Permitted inputs	Principal output	Boundary to document
Data partner	Its identifiers and clinical records.	Encoded linkage inputs and separately prepared clinical payloads.	Which preprocessing and export operations occur locally?
Key escrow or key custodian	Authorized key requests and key-management metadata.	Controlled cryptographic material or operations.	Can the custodian also access identifiers or clinical payloads? CODI defines a separate key role. (phii.org)
Linkage agent	Authorized encoded identifiers and local record references.	Project mappings, scores, and decision provenance.	Does any workflow expose plaintext identifiers?
Data coordinating center	Authorized mappings and clinical extracts.	Assembled project dataset.	Longitudinal assembly is a distinct CODI function. (phii.org)
Data user	Approved analytical dataset and quality documentation.	Study analyses and controlled outputs.	Can the user reconnect a project identifier to an identity?

The matrix describes intended access, not a guarantee that every service implements the same separation. Review the deployed system, including support access, hosted processing, exception handling, and disaster recovery. The UK Office for National Statistics (ONS) recommends segregation of linkage duties; France's data protection authority, CNIL, calls for keeping the additional identifying information separately. (www.ons.gov.uk) (www.cnil.fr)

A proposed information flow is: the partner validates identifiers locally, creates linkage representations, transmits them to the authorized matching environment, receives project mappings, and releases the approved clinical extract through a separate route. The analyst receives the project identifier and the clinical variables authorized for the study. Each transfer should have an owner, purpose, retention period, and audit record.

Technical protection does not settle legal applicability. HHS describes Expert Determination and Safe Harbor as HIPAA de-identification methods. Expert Determination requires documentation, and HHS discusses conditions for cryptographic codes derived from protected health information, including nondisclosure of associated keys. (www.hhs.gov) De-identifying protected health information is itself a use of that information. (www.hhs.gov)

European deployments require jurisdiction-specific assessment. CNIL states that processing must be lawful and purpose-specific. (www.cnil.fr) The UK Information Commissioner's Office (ICO) pseudonymisation guidance is explicitly under review, so it should not be presented as definitive EU-wide legal advice. (ico.org.uk) Privacy officers and counsel should assess the proposed parties, permissions, recipient environment, and legal route before production.

Implementation Considerations and Process Changes

Profile and normalize before token creation

Identifier quality is an upstream dependency. The PCORI evidence report states that missing fields reduce discriminatory power and matching accuracy. (www.ncbi.nlm.nih.gov) Its evaluation found modest overall improvements from standardization, with benefits varying by population. (www.ncbi.nlm.nih.gov) Normalization should therefore be measured rather than assumed to repair all identifier problems.

The recommended partner-specific completeness profile includes:

- **Availability:** Distinguish an absent column from a missing value.
- **Validity:** Identify impossible dates, placeholder values, and structurally invalid identifiers.
- **Consistency:** Compare telephone, address, name, and date representations.
- **Stability:** Examine how identifiers change across repeated observations.
- **Joint completeness:** Measure availability of the combinations each matching rule requires.
- **Population coverage:** Repeat the profile by site and relevant demographic or enrollment groups.
- **Provenance:** Record the source system, extraction logic, and transformation version.

Freeze a shared preprocessing specification before producing compatible tokens. Define character encoding, whitespace, punctuation, name components, date conventions, and treatment of unknown values. These are recommended engineering controls. NICHD's PPRL report supports standardization with clear definitions of personally identifying information. (www.nichd.nih.gov) Preserve meaningful distinctions and test transformations against approved edge cases rather than broadly deleting unfamiliar characters.

OneFlorida's implementation explains why normalization matters: small differences in inputs prevent consistent hash generation. (pmc.ncbi.nlm.nih.gov) Include test records that exercise apostrophes, compound surnames, transliteration, reordered dates, and changed contact details. Synthetic test fixtures can be exchanged without becoming the empirical accuracy reference. Resource planning also matters: the PCORI report describes computationally intensive preprocessing for optimized deidentified linkage. (www.ncbi.nlm.nih.gov)

Govern tokens and keys as versioned dependencies

A cryptographic hash, encryption operation, and token assignment are different mechanisms. **HMAC**, a keyed-hash message authentication code, combines a cryptographic hash with a secret key in the National Institute of Standards and Technology (NIST) definition. (csrc.nist.gov) Naming a hash algorithm alone does not specify key custody, isolation, authorization, or permitted reuse.

The proposed lifecycle controls are:

- **Creation:** Record the token recipe, algorithm, secret scope, and approved implementation.
- **Custody:** Assign responsibility for generation, storage, access, and recovery.
- **Versioning:** Carry preprocessing, token, and matching versions with each batch.
- **Rotation:** Test a migration plan before changing secrets or token recipes.
- **Retirement:** Define destruction, retention exceptions, and revocation responsibilities.

NIST key-management guidance covers generation through destruction and warns that poor key management can compromise strong algorithms. (nvlpubs.nist.gov) Changing salts or peppers changes resulting hashes, so a rotation plan must address longitudinal compatibility rather than silently mixing incompatible versions. (ico.org.uk) Require a documented migration strategy appropriate to the actual technology.

Validation for the Intended RWE Analysis

Build a reference sample capable of exposing missed matches

A reference standard must be independent enough to test the linkage decision. Brown and colleagues used existing linkages established through exact matching, probabilistic linkage, and intensive clerical review. (link.springer.com) That illustrates a stronger evidentiary basis than treating a second algorithm's output as unquestionable truth.

A proposed validation protocol should specify:

- **Reference construction:** Explain permitted identifiers, source documents, and adjudication evidence.
- **Sampling frame:** Include accepted links, rejected candidates, ambiguous cases, and records outside candidate blocks.
- **Sampling weights:** Correct estimates when difficult cases or score bands are oversampled.
- **Adjudication:** Use documented criteria, assess reviewer agreement, and preserve unresolved classifications.
- **Separation:** Keep tuning records distinct from the final evaluation sample.
- **Uncertainty:** Report intervals and the effect of unresolved reference cases.

Access to identifying information can be necessary for quality evaluation. ONS notes that evaluating pseudonymised matching may require identifiable subsets from the data owner. (www.ons.gov.uk) That access should use an approved, restricted workflow rather than routine disclosure to analytical users. FDA also recommends documenting adjudication procedures for linked-data discrepancies. (www.fda.gov)

Calculate errors with explicit denominators

Define a **true positive (TP)** as an accepted correct link, a **false positive (FP)** as an accepted incorrect link, and a **false negative (FN)** as a true link the process missed. **True negatives (TN)** are correctly rejected nonmatching pairs. Precision concerns the correctness of accepted links; ONS explicitly uses that definition. (www.ons.gov.uk) The F1 score is the harmonic mean of precision and recall. (link.springer.com)

The following worksheet uses original illustrative counts, not published performance.

Hypothetical Example: A fully classified reference contains **990 TP, 10 FP, 110 FN, and 8,890 TN**.

Calculations from these stated inputs give:

- **Precision:** $TP / (TP + FP) = 990 / 1,000 = 99.0\%$.
- **Recall:** $TP / (TP + FN) = 990 / 1,100 = 90.0\%$.
- **F1:** $2TP / (2TP + FP + FN) = 1,980 / 2,100 = 94.29\%$.

For a separate source-record denominator of **2,000 eligible people**, if **1,000** receive an accepted counterpart, the **match rate is 50.0%**, calculated as linked eligible people divided by all eligible people. These are different denominators: candidate-pair counts and source-person counts must not be interchanged. ONS emphasizes that match rate alone does not indicate linkage quality. (www.ons.gov.uk)

Choose a threshold through downstream sensitivity

Predefine which error consequences matter. A false link can attach another person's outcome; a missed link can remove relevant exposure or follow-up information. FDA recommends quantifying false and missed matches and testing robustness to the degree of match. (www.fda.gov) Do not substitute a single F1 target for that assessment.

For each candidate configuration, retain the confusion counts, uncertainty estimates, linked cohort composition, follow-up, outcome counts, and final study estimate. Compare both a fixed cohort and the cohort actually selected by each threshold when feasible. This proposed comparison helps distinguish changed identities from changed cohort membership. The acceptance decision should explain the tradeoff; ONS recognizes that success criteria differ across projects. (www.ons.gov.uk)

Bias, Monitoring and Procurement

Assess coverage separately from correctness

High overall linkage accuracy can coexist with differential error. Linked-data evaluation guidance warns that differential error may produce substantial bias even when overall error rates are low. (ijpds.org) The analysis should therefore ask whose records become linkable and whose clinical history becomes more complete.

A recommended subgroup worksheet contains:

- **Eligibility:** Define whether a counterpart should exist for each group.
- **Identifier coverage:** Report completeness of the combinations used for linkage.
- **Error estimates:** Present weighted precision, recall, and uncertainty where a reference exists.

- **Analytical coverage:** Compare outcome capture and observable follow-up.
- **Selection:** Compare linked and unlinked characteristics within the relevant eligible population.

Do not interpret every unmatched record as a false negative. Some records legitimately have no counterpart; methodological guidance explicitly cautions that not all records are expected to link. (ijpds.org) Consent and source participation can also create selection bias even with perfect matching among participants. (ijpds.org) The US Centers for Disease Control and Prevention (CDC) PPRL evaluation planning includes subgroup assessment, supporting the need to go beyond an overall point estimate. (stacks.cdc.gov)

Person-level linkage also leaves event harmonization work. OneFlorida explicitly notes that its deduplication occurred at the patient level. (pmc.ncbi.nlm.nih.gov) A proposed RWE pipeline should separately define how repeated encounters, overlapping claims and EHR events, and inconsistent timestamps enter the endpoint algorithm. Identity resolution should not silently become event deduplication.

Monitor drift against the validated configuration

A production monitoring plan should distinguish changes in input data, linkage decisions, and analytical coverage. FDA's guidance includes linkage accuracy and completeness over time. (www.fda.gov) The following are recommended monitoring and revalidation triggers, rather than universal schedules:

- **Input drift:** Track missingness, placeholders, format changes, and new source systems.
- **Decision drift:** Track score distributions, ambiguity, rule usage, and multiple-candidate frequency.
- **Population drift:** Examine changes in partner mix, eligibility, and subgroup coverage.
- **System change:** Reassess altered preprocessing, token recipes, keys, models, or thresholds.
- **Analytical impact:** Repeat the relevant endpoint sensitivity analysis after material changes.

Preserve a release manifest with source snapshots, extraction dates, transformations, model settings, and an approved reference evaluation. RECORD recommends showing how individuals progress through linkage stages. (journals.plos.org) A reconciliation report should account for incoming, excluded, deduplicated, candidate, linked, unresolved, and released records without confusing those distinct populations.

Convert build-versus-buy into evidence requests

Published vendor descriptions can establish architectural possibilities. **Datavant** documents local command-line or desktop deployment and customer-cloud deployment, and describes machine learning with referential data. (www.datavant.com) **HealthVerity** describes behind-firewall processing that produces salted hashes and also a hosted modality involving encrypted transfer, processing, and deletion. (healthverity.com) These are vendor descriptions, not independent measurements of study-specific accuracy.

A request for proposal should require the proposed deployment diagram, complete access matrix, key-custody responsibilities, reference-validation plan, error estimates by relevant subgroup, migration procedures, and release provenance. Request permission to inspect supporting evidence rather than accepting a single percentage. A build decision requires the same evidence plus internal operational ownership. EMA recommends clearly defining linkage methods, standardization, and discrepancy resolution practices. (www.ema.europa.eu)

Data Analysis and Evidence

Threshold sensitivity in a contemporary evaluation

Bell and colleagues' **2026 Journal of the American Medical Informatics Association** evaluation, supported by the US National Institutes of Health, compared an XGBoost-based PPRL system with an EHR matching system across nearby institutions, querying **849,157 patients** and manually reviewing a categorized sample of **1,200**. ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/30611111/))

Table 3 reports the study's threshold-specific estimates. The settings are internal scores, not portable acceptance probabilities.

PPRL threshold	Estimated precision	Estimated recall
98	100%	30.6%. (pmc.ncbi.nlm.nih.gov)
95	99.4%	92.2%. (pmc.ncbi.nlm.nih.gov)
90	98.7%	96.8%. (pmc.ncbi.nlm.nih.gov)

These estimates make the tradeoff tangible. Recall used the union of matching systems as a proxy reference; concordant decisions were assumed correct, and uncertain reviewed matches were excluded from the precision denominator. ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/30611111/)) This is a useful published evaluation, with reference and generalizability limits.

Reference design and identifier availability change interpretation

In Mirel and colleagues' hospital-survey-to-death-index evaluation, initial precision and recall were **93.8%** and **98.7%**; refinement produced **98.9%** and **97.8%**. ([journals.sagepub.com](https://journals.sagepub.com/doi/10.1177/0898010119851111)) The reference used clear-text deterministic and probabilistic linkage, and the source identifiers were relatively complete. ([journals.sagepub.com](https://journals.sagepub.com/doi/10.1177/0898010119851111)) After refinement, aggregate death-rate differences from the reference were **less than 1%**. ([journals.sagepub.com](https://journals.sagepub.com/doi/10.1177/0898010119851111)) That outcome comparison illustrates an analytical validation endpoint rather than a universally acceptable error level.

OneFlorida's detailed follow-up reported **seven false positives among 1,000** linked records, or **99.3% precision**, alongside **245 missed links among 1,000** reference records, or **75.5% recall**. ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/30611111/)) Its abstract reports a different upper precision value; the figures here deliberately identify the specific body-text evaluation. ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/30611111/)) A high precision estimate alone therefore cannot establish completeness.

In the Australian surveillance study's larger pathology reference, requiring multiple linkage keys reduced sensitivity to about **66%**, while **21.62%** of records had only one particular key type available. ([www.jmir.org](https://www.jmir.org/2019/12/e12000/)) The researchers could not directly measure specificity there because correctly unmatched records could not be identified. ([www.jmir.org](https://www.jmir.org/2019/12/e12000/)) These findings connect apparent method performance to identifier availability and reference construction.

What the evidence does and does not establish

The systematic review's **five included studies** all used deterministic methods; its evidence should not be treated as comprehensive validation of contemporary probabilistic services. ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) Brown and colleagues reported a highest **F-measure of 0.978** for a specific encrypted-linkage parameter set, not an invariant property of encryption. (link.springer.com)

Reporting quality also limits assessment. A multimorbidity linkage review included **20 reports**; among **eight** studies listing linkage variables, only **one** described their quality. (bmjopen.bmj.com) Collectively, these studies support reporting configurations, denominators, reference limitations, and downstream sensitivity. They do not establish a benchmark that can be transferred unchanged to a new pharmaceutical data partnership.

Implications and Future Directions

The most useful validation artifact is a **linkage evidence package** attached to the study's data release. It should connect the analytical question to source eligibility, identifier availability, the privacy architecture, the chosen matching configuration, and the resulting uncertainty. RECORD's recommendations on linkage rules and comparison of linked and unlinked records provide a reporting foundation. (journals.plos.org)

For a pharmaceutical team, the proposed package should let an epidemiologist explain selection, a data engineer reproduce transformations, a privacy officer inspect access boundaries, and a partnership lead understand permitted reuse. Distinct technical and analytical sign-offs make those responsibilities visible. The combined dataset also warrants its own sensitivity assessment: ONS policy explicitly treats linkage as requiring a new assessment. (www.ons.gov.uk)

Stronger privacy technology still needs empirical validation. ONS's experimental PPRL toolkit reported **99.3% recall and 100% precision** on its stated linkage benchmark. (datasciencecampus.ons.gov.uk) It also advises independent assurance of implementations. (datasciencecampus.ons.gov.uk) Benchmark results should therefore be accompanied by a clear statement of data conditions and followed by testing on the proposed source partnership.

Future evaluations would be more decision-useful if they reported errors outside candidate blocks, uncertainty around subgroup estimates, version changes, and effects on actual epidemiological results. This is a research recommendation rather than a claim that every current service supplies those analyses. Privacy-preserving evaluation arrangements are particularly valuable when analytical users cannot inspect the identifying information needed for reference adjudication.

Governance should remain active after initial release. HHS notes that even correctly applied de-identification methods retain a small, nonzero identification risk. (www.hhs.gov) That observation supports reviewing recipient access and the information available in newly combined datasets. A durable program should preserve the ability to reassess both privacy protection and analytical fitness when its sources, participants, or purpose change.

Conclusion

Privacy-preserving record linkage for real-world evidence is a combined architecture, measurement, and study-design decision. The appropriate method follows from the identifiers available, the permitted information flows, and the consequences of linkage error for the intended analysis. A deterministic, probabilistic, or hybrid label is only the beginning of the specification.

A defensible implementation separates identity processing from clinical-data use, documents responsibility for cryptographic material, freezes preprocessing and matching versions, and preserves the provenance of each release. Its validation includes a suitable reference sample, explicit denominators, uncertainty, and an assessment of records that do not link. The published evidence discussed above makes the importance of those choices visible without supplying a universal performance guarantee.

The proposed acceptance decision has a specific scope: a source partnership, an eligible population, a linkage configuration, and a study purpose. Teams should select thresholds using both matching errors and the sensitivity of the downstream result. Subgroup coverage and changes over time belong in that decision because an attractive overall percentage cannot describe every analytical consequence.

Procurement and internal development should produce the same reviewable evidence package. The operational endpoint is a reproducible linked dataset with documented permissions, known limitations, and defined conditions for revalidation. That gives pharmaceutical RWE teams a concrete basis for deciding whether the additional information justifies using the linked data in the planned analysis.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://healthit.gov/standards-and-technology/patient-identity-and-patient-record-matching>
2. <https://www.fda.gov/media/152503/download?attachment=>
3. <https://pmc.ncbi.nlm.nih.gov/articles/PMC13089510>
4. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11752849>
5. <https://phii.org/module-4/codi-roles-and-scope>
6. <https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html>
7. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/pseudonymisation>
8. <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001885>
9. https://www.ema.europa.eu/en/documents/other/data-quality-framework-eu-medicines-regulation-application-real-world-data_en.pdf
10. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory>
11. <https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/data-quality-framework-medicines-regulation>
12. <https://www.ncbi.nlm.nih.gov/books/NBK618683>
13. <https://datasciencecampus.ons.gov.uk/developing-a-privacy-preserving-record-linkage-toolkit>
14. <https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/datalinkageandmatchingpolicy>

15. <https://link.springer.com/article/10.1186/s12911-017-0478-5>
16. <https://www.jmir.org/2020/6/e16757>
17. <https://www.cnil.fr/en/sheet-ndeg1-identify-personal-data>
18. <https://www.hhs.gov/hipaa/for-professionals/faq/may-a-health-information-organization-de-identify-information/index.html>
19. https://www.nichd.nih.gov/sites/default/files/inline-files/NICHD_ODSS_PPRL_for_Pediatric_COVID-19_Studies_Public_Final_Report_508.pdf
20. <https://pmc.ncbi.nlm.nih.gov/articles/PMC6994009>
21. <https://csrc.nist.gov/pubs/fips/198-1/final>
22. <https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-57pt1r5.pdf>
23. <https://ijpds.org/index.php/ijpds/article/view/2137>
24. https://stacks.cdc.gov/view/cdc/120225/cdc_120225_DS1.pdf
25. <https://www.datavant.com/life-sciences/connect-linkage>
26. <https://healthverity.com/identity-manager>
27. <https://journals.sagepub.com/doi/10.3233/SJI-210891>
28. <https://bmjopen.bmj.com/content/13/5/e069212>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.