Research Article

Privacy-Preserving Duplicate Record Detection Across Decentralized Healthcare Institutions

Kotha Raja Dinesh[1], Vijaya Kumar Reddy Radha[2]

[1, 2] Department of CSE, Koneru Lakshmaiah Foundation, Vaddeswaram, AndhraPradesh, India
[1] rajadinesh2314@gmail.com, [2]rvijayakumarreddy@kluniversity.in

Abstract

Patient information can be duplicated across providers, which may lead to bias in patient counts and to loss of longitudinal analysis; however, patient demographics may not be allowed to be shared directly. Some of the most important issues that arise from HHN data sources are similar to those found in other typical hospital data sources, such as the MITRE Synthetic Denver FHIR dataset, which is used in this paper to implement and test a privacy-preserving record linkage (PPRL) benchmark. A total of 6,354 Patient resources were extracted from 5 simulated provider partitions using a synthetic identity label (ID) separate from the provider-partitioned matcher input elements. It normalizes seven demographic fields, creates Cross-Site Candidates via a Union of 4 HMAC-SHA-256 blocking keys, and compares records by Exact or Character 3-gram Weighted similarity per field. A Bloom filter exists for every position generated using a set of 3 HMAC-positions, and for the encoded method, the Bloom filter is encoded into a 128–1,024 bit size Bloom filter. It uses a held-out 65% identity-group split to test, and a deterministic 35% development split for thresholds. In the test set 604 true cross provider pairs were retrieved as blocking counts 554 of which (91.72%). At 1,024 bits, the Bloom-filter matcher achieved 0.941 precision, 0.891 recall, and 0.915 F1; a plaintext q-gram reference achieved 0.950, 0.889, and 0.919, respectively. Not actual results from real hospitals - these are synthetic results measured in a "realistic setting. The existing HMAC and Bloom tokens are pseudonymous codes, and can reveal known information about membership in the blocks and about similarities; there was no evaluation of any secure multiparty computation, differential privacy or end-to-end security protocol.

Keywords: Privacy-preserving record linkage, patient matching, Bloom filter, HMAC, healthcare interoperability, synthetic data, FHIR.

1. Introduction

Most health systems have multiple patient indexes. Inconsistent spelling of names, out-of-date contact information and missing fields can lead to multiple identities if the same person receives care from more than one organization. Sensitive identifiers like name, birth date, phone number and address can be used for linking, e.g., to help with deduplication, building research cohorts and as a common link to enable cross-system continuity. Patient-matching is a fundamental interoperability concern as recognized by the nationally available patient-matching guidance, and existing FHIR guidance provides for a cross-organizational Patient $match workflow [1, 2].

The goal of privacy-preserving record linkage (PPRL) is to match records in such a way that very little identifying values are disclosed. The measures, as demonstrated in operational deployments, illustrate PPRL's ability to enable analysis of patient overlap that can be performed at the network level or across multiple networks [3], [4], [5]. But it's not a single security property—"privacy-preserving" is, as used here, a category of security. Secure multiparty computation (MPC) can be used to create cryptographic protocols with more rigorous formal guarantees in certain threat models, while token based and Bloom-filter methods can reveal equivalence, frequency, and/or similarity details [6]–[10]. A linkage system has to be assessed in terms of linkage utility and its explicit linkage model.

This work does a small reproducible comparison of deterministic exact matching, plaintext q-gram matching and HMAC-keyed Bloom-filter matching. It leverages the Mitre Synthetic Denver subset [11], which includes simulated FHIR Patient resources broken down by provider site, and provides synthetic cross-site identity labels for these resources. We present our contributions: (1) a public-data extraction pipeline that extracts matcher's features and preserves the identity truth; (2) multi-pass keyed blocking, field-wise similarity, and threshold selection using an identity-group holdout; (3) measured precision, recall, F1, and blocking recall for 4 Bloom-filter sizes; and (4) a threat model that describes whether it is a pseudonymous token comparison or a full-plot private computation.

The purpose and scope is intentionally limited. The source data is simulated and the five site labels are the partition for the dataset in the database and not the hospitals that participate in the benchmarking. The benchmark is executed with a single process and the same demonstration key. For this reason it is not a realistic way to estimate real clinical matching performance, real regulatory compliance or resistance to a determined attacker.

Related Work

A. Healthcare PPRL in practice

PCORnet reported the deployment of PPRLs as distributed to monitor overlap in the patient populations within a national research network and provide deduplicated population summaries [3],[4]. Across-institution linkage within the national COVID cohort collaborative, as part of the N3C experience, places a focus on linkage as a means to infrastructure for a learning health system [5]. Bloom-filter linkage, compared with clear-text linkage, was concordant with the groupings reported for hospital and death-registration records from Western Australia [6]. The following studies illustrate their use but their governance and control, data quality and actions/controls differ from this implementation.

Recent papers have stressed that linkage accuracy is dependent upon the completeness, stability, and blocking of the field. Found a small body of evidence in U.S. real world PPRL evaluation in a 2025 systematic review, with some combinations having greater precision and lower recall, and alternative identifier pattern combinations improving recall (and PPRL coverage) [7]. Similarly, public-health adoption models designate governance, architected compatibility, purpose of use, and assessment of privacy as life cycle factors rather than ones offered by a matching algorithm [8] [9].

B. Formatting the hard drive and encryption methods

Bloom filters are another viable PPRL representation and they provide approximate comparison of the sets of encoded q-grams. Bit-vector encoding does not provide them with any level of privacy, as published attacks succeed in reconstructing sensitive input patterns with frequency, from background knowledge, or from leakage on the similarity score [10, 12]. In a sound key management, some offline guessing can be prevented by using a keyed pseudorandom function, but doesn't prevent hiding equality within the key domain, the block size or the output of the matcher. In the present benchmark, this distinction is brought to life by HMAC with its hard-coded key – for it isn't an actual patient's key that it should protect, it's for reproducibility purposes only.

Other approaches include secure MPC linkage frameworks which do not rely on a trusted third party as guaranteed by its cryptographic assumptions [13]; locality-sensitive hashing for fuzzy linkage [14]; and multi-party computation for medical record linkage [15]. In the realm of security, output semantics, and communication cost, PSI, threshold PSI as well multiparty protocols remain evolving [16]–[19]. We implement and do not compare their protocol level guarantees. We are not using one of the baseline techniques from the q-gram community as a deployable privacy method—they are only used as an accuracy check.

C. Standards and synthetic evaluation

This context of interoperability is relevant to the fields used, such as FHIR patient matching guidance and patient demographic standards [1, 2, 20]. Synthetic records can be used to provide repeatable tests without risking the lives of patients, and synthetic records can also not guarantee an accurately re-created reporting period (in terms of error distributions, missingness, population diversity, and workflow context) of every real institution [21, 22]. The Synthetic Denver resource is particularly helpful since records are linked across its simulated provider partitions via identity labels [11].

Research Gap

Previous PPRL deployments and evaluations in the healthcare sector demonstrate the potential of linkage across institutions to enable research and public health workflows, but limited usefulness is found across factors of population, identifiers available, data quality, and governance models [3]–[9]. This makes it possible to compare a problem and get good pair-classification results when the true pairs are not seen in any of the step of candidate generation, and having structural information when this is not directly shown in the original values when using a token based approach[10,12]. Their accuracy alone is insufficient to characterize the behavior of a linkage pipeline.

While synthetic records can enable repeatable experiments, they can also lack the missingness, entry errors, population diversity, and workflow variances that can be present between real-world providers [21], [22]. The research gap being filled here is a compact and reproducible evaluation that provides a differentiated report of candidate-generation recall

performance from held-out pair classification when used for the same split to represent whether the representation is an exact, plaintext q-gram or keyed Bloom-filter and to indicate the leakage and deployment boundaries, in addition to the outcomes reported. While this work fills a methodological gap on a public synthetic FHIR benchmark, testing in real working organizations and security/user end-to-end analysis are still open research challenges.

Objective

The main objective is to implement and evaluate a reproducible method for detecting duplicate patient records across simulated healthcare-provider partitions while limiting direct exposure of identifier values during comparison. The study has five specific objectives. First, extract the available Patient resources and create separate feature and ground-truth files so synthetic identity labels are not used as matcher inputs. Second, normalize selected demographic fields and generate cross-provider candidate pairs using multiple keyed blocking rules. Third, implement and compare three pair-scoring approaches: exact equality, plaintext character q-gram similarity, and HMAC-derived Bloom-filter similarity at four vector lengths. Fourth, tune decision thresholds on development identities and measure candidate recall, precision, recall, and F1 on held-out identity groups, including pairs excluded by blocking in end-to-end recall. Fifth, document runtime, reproducibility steps, assumptions, and information exposed by the single-coordinator design. These objectives define a software and benchmark evaluation; they do not imply clinical validation, use of real hospital records, or a formal end-to-end privacy guarantee.

Methodologies

A. Source and extraction

The benchmark uses the MITRE Synthetic Denver FHIR archive, a generated dataset designed for identity-matching experiments. The downloaded archive was 295 MB; SHA-256: c87f6f29d9c5dad42b0bb6319d032ed73ab39b035c7acb3d5df735fd04dfac31. From the five provider partitions, the extractor retained 6,354 Patient resources with a synthetic link identifier. This is three fewer than the published nominal total of 6,357 because three archive items did not contain an extractable Patient resource with the required label. Extracted counts are CH=2,039; DH=2,039; GotR=119; HFC=119; and KP=2,038. These names denote dataset partitions and do not represent real institutions.

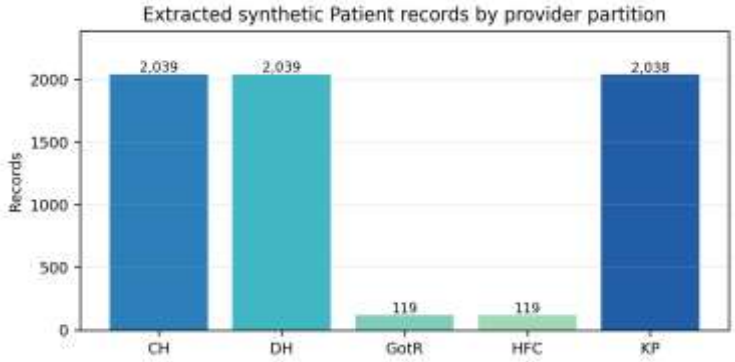


Fig. 1. Extracted Patient-resource counts by simulated provider partition.

Feature file includes: record ID, Partition, Given Name, Family Name, Birth Date, Gender, Postal Code, Phone and Email. Record ID and Synthetic Link ID are the only information provided in the ground-truth file. This is only read to split assignments and also for scoring – never as a matcher value. All information fields that designate direct identity are not deleted from the local benchmark files, so that the experiment is not a denial of the existence of raw source records in a central file.

C. Count down and divide.F. Divide and count up.

The records are said to be a pair if they have the same synthetic link ID but different provider partitions. Duplicate detection is not part of the task. Pair truth comes from the identity label, rather than a clinical identity. Deterministic split based on link_id: SHA-256(link_id) % 100, meaning that 35% of identity_groups belong to development, and 65% to test. This avoids any records from spilling over between the threshold-tuning and evaluation sets for one synthetic person. There are 604 pairs of true cross-providers in the realized test and 360 pairs in the development.

Only pairs of records that are in the same partition are considered the candidate universe. I hope you can see in that table that the possible space of all of the crosssites can reach into the millions of different pairs. That space alone shrinks down to 1,133,724 pairs when you block it because there are a few held-out test groups included. There are 476,229 of those held-out test site pairs. Blocking (no similarity score) is the percentage of all positive test pairs that remain after blocking that have passed pre-processing. There are 5,478 different synthetic identities identified by the labels: 4,687 are singly labeled, and 791 are more than one-partition labeled. There were 964 cross-site positive record pairs throughout the benchmark. Pair multiplicity should be considered because each identity with 3 or 4 records adds several pair labels that cannot be counted and disregarded as person counts.

D. Document and define why it is important.

5,478 synthetic link IDs were matched to the extracted records. Of these, 4,687 have IDs that only appear once in the extracted data; 791 have IDs that span multiple site partitions. There are 709 identities with 2 records, 79 with 3 records and 3 with 4 records, in this multi-site set. The total of 964 cross site + pairs are all possible inter-partition combinations for each label. A matching ID with three sites may be able to add three pairings; with four sites can add up to six pairings. The same approach to the evaluation was used, treating pairs as the unit of classification while maintaining identity-group separation for the development/test split.

Table I. Extracted Record And Label Profile

Profile item	Count
Patient resources retained	6,354
Distinct synthetic link IDs	5,478
Link IDs seen once	4,687
Identities spanning multiple sites	791
Cross-site positive pairs	964
Positive pairs: development / test	360 / 604
Empty email field	275 (4.33%)
Empty family-name field	7 (0.11%)
Empty postal-code field	15 (0.24%)

The extraction manifest records the archive digest, site counts, and exclusions, and allows a reviewer to compare the local archive with the intended public source. Fields with very low missingness in this extract should not be mistaken for representative production completeness. Email is the only attribute with a nontrivial blank rate in the parsed fields. Other quality dimensions, including stale values, inconsistent formatting that remains after normalization, family relationships, shared contact details, and culturally specific naming variation, are not measured by blank-field frequency.

D. Evaluation questions and design

Feature files include record ID, Partition, Given Name, Family Name, Birth Date, Gender, Postal Code, Phone and Email. Record ID and Synthetic Link ID are the only information provided in the ground-truth file. This is only read to split assignments and also for scoring – never as a matcher value. All information fields that designate direct identity are not deleted from the local benchmark files, so that the experiment is not a denial of the existence of raw source records in a central file.

E. Divide and count.

The records are said to be a pair if they have the same synthetic link ID but different provider partitions. Duplicate detection is not part of the task. Pair truth comes from the identity label, rather than a clinical identity. Deterministic split based on link_id: SHA-256(link_id) % 100, meaning that 35% of identity_groups belong to development, and 65% to test. This avoids any records from spilling over between the threshold-tuning and evaluation sets for one synthetic person. There are 604 pairs of true cross-providers in the realized test and 360 pairs in the development.

Only pairs of records that are in the same partition are considered the candidate universe. I hope you can see in that table that the possible space of all of the crosssites can reach into the millions of different pairs. That space alone shrinks down to 1,133,724 pairs when you block it because there are a few held-out test groups included. There are 476,229 of those held-out test site pairs. Blocking (no similarity score) is the percentage of all positive test pairs that remain after blocking that have passed pre-processing. There are 5,478 different synthetic identities identified by the labels: 4,687 are singly labeled, and 791 are more than one-partition labeled. There were 964 cross-site positive record pairs throughout the benchmark. Pair multiplicity should be considered because each identity with 3 or 4 records adds several pair labels that cannot be counted and disregarded as person counts.

F. Determine and name the distribution.

5,478 synthetic link IDs were matched to the extracted records. Of these, 4,687 have IDs that only appear once in the extracted data; 791 have IDs that span multiple site partitions. There are 709 identities with 2 records, 79 with 3 records and 3 with 4 records, in this multi-site set. The total of 964 cross site + pairs are all possible inter-partition combinations for each label. A matching ID with three sites may be able to add three pairings; with four sites can add up to six pairings. The same approach to the evaluation was used, treating pairs as the unit of classification while maintaining identity-group separation for the development/test split.

System Architecture

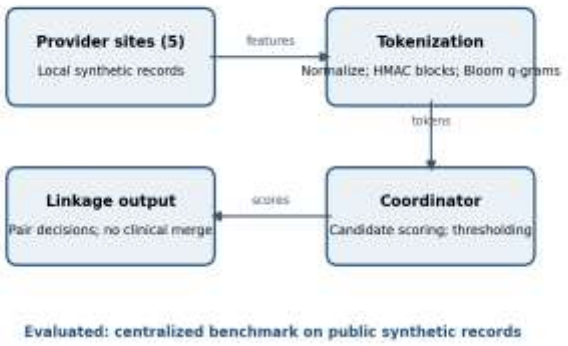


Fig. 2. Conceptual workflow and implementation boundary. Token comparison is benchmarked; institution-isolated deployment and protocol protections are not.

The implementation is Python 3 standard library for extraction, HMAC encoding, candidate generation, scoring, and metrics, with Matplotlib for figures and python-docx for the manuscript. It runs as a single coordinator over extracted records. Figure 2 describes the intended institutional flow but labels which pieces are conceptual. No secure transport, key service, access-control model, MPC, PSI, confidential execution, or distributed protocol was deployed. Runtime was 48.390 seconds for the reported benchmark in this workspace; it is an environment-specific observation, not a scalability result.

Algorithm And Implementation

The extraction program accesses the zipped archive and accesses the patient resources in the 5 different directories. It includes just demographic fields used by the experiment and the synthetic identity label. The identity label is written into ground_truth.csv with a record key field, the matcher feature file does not have a record key field. Good for repeated experiments, but does not give a privacy preserving preprocess environment, as patient demographic information is recorded as clear text in records.csv. In a production system, this CSV would not be exported by a production process that performs feature normalization and token generation locally, in an access-controlled manner.
The benchmark itself relies on python core modules to normalize unicode, to read/writes in and out csv, to hash with hmac and sha-256, to score and to write and read json. Result files give reproducible plots with Matplotlib. This is for DOCX generation, using python-docx. The project package contains all outputs necessary for an understanding of the

test split reported, including the feature records, identity labels, benchmark summary, method metrics, site-pair blocking metrics and scripts. The raw source archive has not been copied into the package as it is fairly large; it can be downloaded through the official route indicated in the README and is exactly 593cdbd1d352ab2ee3a3946530fc38d38f0e75b713292ae00a837dd39454c7b.

This is not checksummed with an external signing authority – the digest sense is not to be a means to know who has created or where the archive has come from or to establish authorship or chain of custody – but it is a means to notice an accidental mismatch with this run. An independent user should confirm that the official source page has been correct, that the "digest" has been archived and is consistent with the number of times it has been extracted, and that the results of the benchmarks are correct. Depending on the upstream change in the archive, the result might slightly vary even if changes are not made on scripts. The acquisition package includes the acquisition date and the package's checksum in the package's provenance manifest.

The first is setting up decision and measurement boundaries.The first one is decision and measurement boundaries.

A false positive links two records with entirely separate synthetic identities, and a false negative prevents any attempt to link a record with one synthetic identity with another related record for the same identity. In the real workflows, the effects of these error classes are asymmetric. A false merge allows another person's clinical history to be exposed or linked with another person and a missing link might fragment care or under count a cohort. F1 is equally weighty when it comes to precision and recall, so it is an easy case study to summarize, but not a clinical utility function. The level of system alarm or trigger is not included in this dataset and it must be related to clearly specified harm estimates and human alarm decisions.

Field weights (nonstatic) are used in the record score. There is no probability calibration - a score of 0.80 does not mean that there is an 80% probability that records are the same person. Similarly, threshold is both data and dataset-dependent, and may change from one data-entry practice to another within the data. In a real deployment evaluations would be required for calibration, confidence intervals, site-stratified errors, drift, and adjudicator consistency. It also should examine if the rates of exclusion differ for demographic subgroups as a result of candidate generation, when this benchmark does not have enough information to make a responsible assessment.

There was no set made for manual review. This typically includes labels derived from the link ID and it is this label that is used to score all predicted pairs. The manufactured construction provides us the reproducible truth when it comes to comparing algorithms but cannot replace human-checked identity resolution. Separate measures of the number of uncertain pairs, adjudication reversals, and of the frequency of modification of a previously linked cluster made by an entity-resolution policy depending on new data should be measured in an operational study.

B. Scaling considerations

There are currently 1,130,000 pairs in the full cross-site candidate set - across 6,354 records. In the python implementation, candidates are stored in memory, as well as per-pair score maps, which works fine for the small reward budget, but may be sub-optimal or inefficient for use with large record collections. Duplicate godown events may produce a large cross-product: if organizations are large, the largest cross-product is unlikely to be block skew. Possible engineering are: Frequency aware block suppression, Pair caps, Sorted stream processing, Site-pair partitioning, Parallel scoring. Each alters in either recall or operation/setting and must be measured independently. Since blocks are not necessarily evenly distributed, and accessing them also turns out to be expensive with memory pressure, extrapolated runtime is not reported.

Even the Bloom filter implementation creates one bit vector for each record and for each field itself and it uses the same bit vector for multiple records and fields. So, it is more efficient than rebuilding q-grams for each candidate; and does perform a field-wise similarity against each candidate. To decrease the memory overhead, a compact representation in bits might be used. An implementation in cryptography would add key exchange, secure protocol rounds which would bring in networks timing, overhead in message size, availability of participants, and error handling. These costs would have to be identified in a local Python scoring system in comparison with the MPC. These costs are not reported in this paper.

C. Responsible interpretation

A synthetic identity and/or provider partitions may be willful safety and repeatability decisions. They enable an inspection and sharing of the benchmark without hospital records being implied. Crediting the data as “real world performance” is inappropriate as it is not de-identified clinical information. When there is a high F1 for that data, it only means that the implementation can create a defined matching task on the specified synthetic data sets. It does not

show that a health network is able to implement the method, that identifiers would not be able to leave institutions, or that a private approach would be approved by a privacy regulator.

D. What's the difference between Candidate-Result and Record-Result?

The experiment doesn't establish an “unified record” at the person level; instead it establishes a pairwise identity match. This distinction avoids a common misinterpretation—that is, that 604 test positive pairs equates to 604 different patients. A link ID can be used by more than one set of positive pairs. The precision and recall are reported on the pair level and the split is done at the link-ID level. If an alternative record level metric were used, it would depend on an explicitly specified cluster construction as well as a policy for combining a cluster's transitive edges, and using an alternative object to measure the edges on a pair would require carefully defined cluster metrics that compare the objects to the nodes, such as pairwise completeness, pairwise quality, and component-level purity.

Similarly, a positive pair does not assure proper reconciliation of the complete clinical histories. There is no clinical condition data, medication or encounter data in the model features of the project. Identifies if two synthetic Patient resources refer to one generated identity, based on their demographics. There is no monitoring or documentation of downstream chart review/care continuity or functional outcomes. The article does not imply automatic deletion or merging of records, but instead refers to “duplicate record detection,” which deterprises it.Duplicate record detection is not synonymous with automatic deletion or with merging records, it detertipers that.

E. Why this evaluation does not prove the performance in real world

The real-world performance needs to be based on data collected in an organisation that has been approved by the protocol and an agreed-on validation target representative of the intended use. The MITRE dataset was created specifically for system testing, and includes carefully-designed, user-readable cross-site matching labels. This is suitable to be a methods demonstration and reproducible software test. It doesn't contain natural variation in hospital registration flows as it would in a live network, nor case-mix, vendor interfaces, nor staff corrections nor patient-driven changes in case-mix, patient driven or staff errors. Even within this extract the field-completeness rates are a characteristic of this source release and cannot be presented as prevalence rates for healthcare systems.

Those who would wish to validate a real-world external data source would require at least two institutions authorized to provide the data, a reference window (population and timeframe), data-use approvals, a trusted source of reference linkage, an adjudication sample and a prespecified method for choosing the threshold. It should not include leakage, report confidence intervals considering clustered pairs, and should test the performance by site and strata of relevant populations. A blinded study would also make it so that the creators of the linkage labels don't even know who the matching team is until after the evaluation stopped being modified. None of these in the real world can be mimicked by relabelling the existing data.

The current metrics are useful as they are a direct on a well-documented public test split. They do not represent the scientific value of their clinical validation; their value is to be scientific and reproducible, and to give an indication of blocking loss. Any statement of hospital recall, precision, privacy compliance or impact on patient safety would need new evidence that goes beyond this level.

F. Pseudocode and Complexity

The steps that are implemented using the candidate and score pipeline are summarized in Algorithm 1 below. Ythe ground-truth file is used after prediction is made. Records are grouped by each block key and are cross-sited to create Cartesian products within each keyed token which produce candidate pairs. Duplicated candidates are removed from the output of multiple passes by using set-union. Cost of feature encoding is O(n*Q), where n is the number of normalized q-grams (and Q is the number of fields). The cost for generating candidates is the sum of the product of the block populations, over the tokens and site pairs. Once encodings are generated, similarity evaluation is O(CF) with C the number of unique pairs of candidates and F the 7 evaluated fields. The set of observed C is {1,133,724} However in pathological common block cases, the space of the candidates could approach the entire cross-site pair space.

Input: records R, shared demo key K, field weights W, Bloom length m, score threshold τ

1. For each record r ∈ R: normalize the seven identity fields.
2. Construct four canonical block strings from birth year, gender, name initials, and birth month-day.
3. Replace each block string with HMAC-SHA-256(K, block); index records by token.
4. For each token and each pair of distinct provider sites: add Cartesian-product record pairs to candidate set C.
5. For each candidate (a,b) ∈ C: compute exact, plaintext q-gram, or field-wise Bloom Dice similarities.
6. Renormalize weights over fields present in at least one record; aggregate a weighted score.

7. Predict match iff score ≥ τ.
Output: pair decisions; evaluate only afterward against identity labels.

G. Security Property Boundaries

Table III separates properties implemented in the experiment from properties that would require additional protocol components. HMAC is a keyed message authentication function used here as a pseudorandom token generator. It is not a homomorphic encryption system, and the Bloom filter is not encryption. The coordinator sees token equality classes and can observe approximate similarity derived from Bloom bits. Since the key is embedded in the demo code, anyone who receives the code can reproduce the tokens and perform dictionary tests against candidate values. These choices are acceptable for a controlled synthetic benchmark only.

Table V. Implemented Properties And Unmet Security Requirements

Property	This benchmark	Deployment requirement
Raw data removal	No; extracted demographic features are scored centrally	Keep source features within each institution
Token generation	HMAC, fixed demo key	Protected key service, rotation, separation
Fuzzy comparison	Bloom q-gram Dice	Leakage analysis and reviewed protocol
Coordinator trust	Fully trusted for experiment	MPC/PSI or explicit trusted-party controls
Transport	Local process only	Authenticated encryption and access control
Formal privacy	No DP or formal end-to-end claim	Threat model, proof/review, legal governance

A formal end-to-end claim must specify which parties learn the candidate list, match scores, pair identifiers, and final decisions; whether parties may collude; how keys are generated; which inputs can be malicious; and what is disclosed by the output. If match outcomes are returned to source sites, the outcome itself can reveal whether a patient is known to another site. Reducing token leakage does not eliminate inference from the linkage result. For population-level releases, small cell suppression, query governance, and possibly differential privacy may also be required, but those are distinct from pair matching.

H. Reproduction Checklist

It comes with companion package containing the following files: extraction/output code, benchmark/output code, input feature CSV, the groundtruth CSV is separate, check sum in archive and result CSV/JSON files, README instructions. To be able to reproduce, download the archive from its official download page, verify the checksum, run prepare_dataset.py, and then run run_benchmark.py with the provided input/output paths and data. As noted in the README, the 295 MB source archive refers to the source code that's purposefully not included in the benchmark package. Before looking at the metrics of classifiers, consider site counts, 6,354 extracted records, deterministic split counts, and the test candidate-recall value. The actual threshold chosen for each method is included in the scripted output.

The follow-up measures are meaningful if a preregistered protocol is used, if a path of data governances is configured at the organization level, if authorization is secured for all actual datasets and before the actual identifiers are transformed, an independent security assessment is carried out. A real world validation should present information on the construction of linkage labels, the adjudication procedure, definition of cohort, detail on missingness level by site,

false merge and missed match costs, confidence intervals, and cross-site external validation. There are no synthetic public benchmarks for those items, and a prerequisite necessity for an interpretation of an operational claim.

I. Reproducibility Parameters

This appendix consolidates the fixed scoring and blocking parameters for readers reproducing the experiment. Field weights were specified before test evaluation and were not learned from the held-out pairs. The blocking rules are combined with set union; duplicate candidates produced by multiple rules are counted once. Each row is part of the benchmark configuration and does not imply that the feature is appropriate for a real clinical deployment.

Table VII. Scoring Field Weights

Field	Weight	Similarity
Given name	0.18	Exact / 3-gram Dice
Family name	0.22	Exact / 3-gram Dice
Birth date	0.25	Exact / 3-gram Dice
Gender	0.05	Exact / 3-gram Dice
Postal code	0.12	Exact / 3-gram Dice
Phone	0.10	Exact / 3-gram Dice
Email	0.08	Exact / 3-gram Dice

Table VIII. Keyed Blocking Rules

Pass	Canonical values before HMAC
1	Birth year + gender
2	Birth year + given-name initial
3	Birth year + family-name initial
4	Birth month-day + gender

Normalization is case-folding, Unicode accent removal, and removal of non-alphanumeric characters. Missing fields are omitted from the pair-score denominator when both records lack the field. HMAC-SHA-256 tokens are generated with a fixed public demonstration key in the source code. Bloom-filter lengths are 128, 256, 512, and 1,024 bits, with three positions per q-gram. The development thresholds are chosen independently for each method from the grid in run_benchmark.py; held-out test pairs are not used to select a threshold.

VIII. Results And Discussion

A. Candidate generation

The four-pass candidate generator retained 554 of 604 positive test pairs, yielding 91.72% candidate recall. The earlier single-pass birth-year-plus-gender block retained 516/604 (85.43%). The union adds 38 additional pairs, or 6.29 percentage points. Candidate pairs increased from 994,866 to 1,133,724 across the complete dataset (13.96%). In the test split, 476,229 candidate pairs were scored. Therefore, 50 true test pairs were unrecoverable by the selected blocking rules regardless of the downstream threshold.

B. Pair classification

Table IV reports test results. The plaintext q-gram reference reached the highest F1 (0.919), narrowly above the 1,024-bit Bloom filter (0.915). The latter's precision was 0.941 and recall 0.891, compared with 0.950 and 0.889 for plaintext q-grams. The 512-bit Bloom variant had the highest recall (0.897) but lower precision (0.922). The exact-equality baseline had 0.955 precision and 0.825 recall. These results show a precision-recall tradeoff rather than a single uniformly superior method.

Table IV. Held-Out Pair Classification Results

Method	τ	TP	FP	FN	P	R	F1
Exact equality	0.35	498	23	106	0.956	0.825	0.885
Plain q-gram	0.55	537	28	67	0.950	0.889	0.919
Bloom 128 bit	0.65	527	32	77	0.943	0.873	0.906
Bloom 256 bit	0.6	534	31	70	0.945	0.884	0.914
Bloom 512 bit	0.55	542	46	62	0.922	0.897	0.909
Bloom 1024 bit	0.55	538	34	66	0.941	0.891	0.915

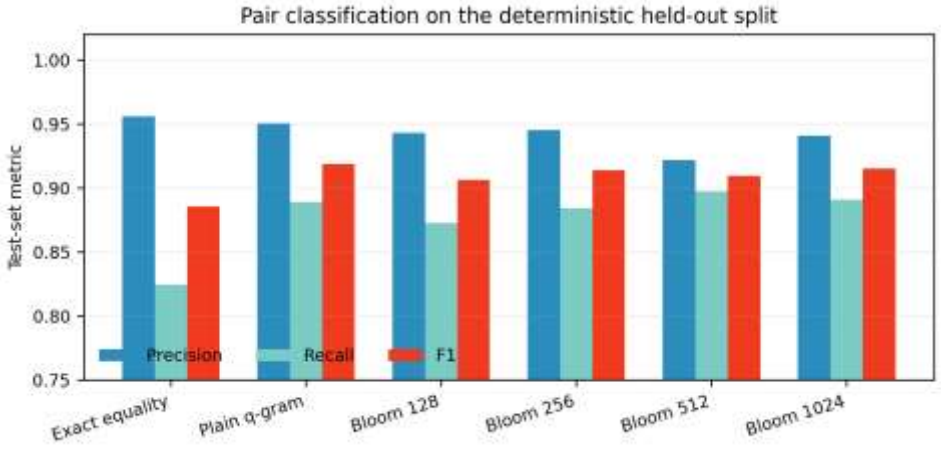


Fig. 3. Precision, recall, and F1 for the held-out synthetic test pairs.

C. Blocking ablation and site-pair variation

To isolate the role of each pass, we measured the candidate set produced by each key individually and by cumulative union. The birth-year-plus-gender key is the widest pass and yields the largest candidate pool. The other keys add candidates at lower marginal cost. The union of all four passes captures 38 additional positive test pairs over the first pass. This is a measurable recall gain, while still leaving 50 positives outside the candidate set.

Table II. Blocking Pass Ablation

Block pass	All pairs	Test positives found	Test recall
Birth year + gender	994,866	516 / 604	85.43%
Birth year + given initial	135,761	409 / 604	67.72%
Birth year + family initial	121,171	436 / 604	72.19%
Month-day + gender	22,304	389 / 604	64.40%
Union of all four	1,133,724	554 / 604	91.72%

The pair-level behavior is not uniform across site combinations. CH–KP produced the most test positives (148) among the larger partitions and blocked 131; HFC–KP blocked 21 of 24. Small partitions have fewer total candidates, but one missing positive can materially change their percentage. TABLE III reports the held-out pair coverage, not full matcher accuracy by pair because the current output contains aggregate matcher metrics only.

Table III. Test Blocking Recall By Site Pair

Site pair	Test positives	Blocked	Recall
CH–DH	154	141	91.56%
CH–GotR	24	23	95.83%
CH–HFC	28	26	92.86%
CH–KP	148	131	88.51%
DH–GotR	24	22	91.67%
DH–HFC	25	23	92.00%
DH–KP	150	140	93.33%
GotR–HFC	1	1	100.00%
GotR–KP	26	26	100.00%
HFC–KP	24	21	87.50%

The individual-pass rows overlap and must not be summed. Any pair that is seen is added cumulatively to the union pairs. This is important because the cost of computing resource usage follows the volume of the unique union of the two passes (not the total of the volumes for the two passes). In this dataset there are 476,229 test pairs that remain after the union. This is still a considerable portion of the cross site space, and reflects the preference for recall over widespread candidate reduction currently afforded by simple block keys.

D. Reproducibility and Runtime, and decision operating points

It is deterministic, depending on the downloaded archive and extraction rules, on the version of the implementation, and on the hard-coded demonstration key in the system. Train/Test membership is determined by the same synthetic identity groups and is rerun with SHA-256, and is not based on random seeds. The threshold grid is explicitly stated in the source; and test metrics for each method are based on the selected development threshold. In this environment, the time for the 48.390 run results comprised feature normalization, block construction, scoring and result serialization. There is no download of the original archive, no generation of the documents, it is not a throughput measurement; it is only an observation of a local reproducible result.

To visualize the conservative operating point of a workflow with a high positive predictive value, the exact-match baseline is shown. For a use case where recall is the priority, the 512-bit Bloom version will have the highest false-positive rates but the highest recall when compared against tested other Bloom use cases. Adjudication capacity, harm from false merges, reversibility and downstream use would have to be part of the operational threshold as it cannot be decided based solely on F1. A clinical safety threshold is not suggested here.

E. Field scoring and assumptions

The current weights are not estimated from a representative clinical cohort, not varying and do not sum to one before missing value renormalization. The most important weight is given to birth date, which is a discriminative demographic factor, but an error in the date might reduce the weight of what would otherwise be correct pairs. The weight of gender is small as well as there are 2 block keys. This implies either a gender imbalance in the generation of the pair or there may be an imbalance in the scoring of the pair. The postal codes or contact information may be incorrect, outdated, or not available, as well as names may be written in differing characters or there may be typos. The intention of the rule-based normalization in the benchmark is to be transparent, yet there are phenomena that can't be modelled with such normalization.

The code employed are 3 grams of characters and Dice coefficient. The implementation pads with markers for start/end of string for short strings; implement values of less than 3 characters after left-padding with the markers. Each normalized attribute is considered as a set of distinct character grams, thus multiple character grams of a specific character do not contribute to the frequency of their character gram. Collisions in the bloom could result in multiple occurrences of the same two (or more) q-grams being seen in the same location, which would create seemingly high similarity between the same two. The number of Bloom bits directly decreases the probability of collisions but the F1

response measured (accuracy) is dependent on threshold retuning, feature distribution, etc., and cannot be relied upon as security or accuracy to guarantee.

F. Compile and review process (s).

The benchmark returns all pair-wise match decisions. Doesn't cluster-transitively, doesn't choose a surviving master record, doesn't reconcile conflicting attributes nor change a patient record. Decision by pairs can result in inconsistent components (A-B, B-C, but not A-C). A production deduplication workflow would need a policy for defining a policy for entity-resolution; consistency rules for the cluster; confidence calibration; imperfect resolution operation review via humans; audit trails; corrective actions for incorrect resolutions; and safe merge/undo operations. These functions are not included in the measurable implementation.

G. Bloom-filter size

All vectors of vector length increase from 128 to 1,024 bits measured F1 from 0.906 to 0.915, but it is not linearly monotonic, which is 0.909 at length of 512 bits. The larger bit vector can help to minimize the collisions but does not necessarily give the best pair classification as its performance is also influenced by the threshold used, overlaps between the q-grams used in the fields, and errors in the data. Development data were used to independently tune each size.

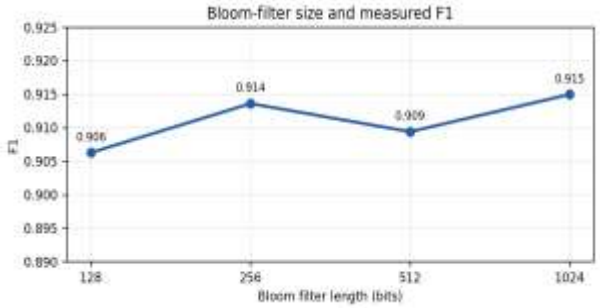


Fig. 4. F1 against Bloom vector length; points are measured on the same held-out split.

H. Dataset and experiment summary

Item	Measured value
Extracted records	6,354
Simulated provider partitions	5
Positive cross-site pairs (development/test)	360 / 604
All candidate pairs after blocking	1,133,724
Test candidate pairs	476,229
Test candidate recall	91.72%
Runtime	48.390 s; environment-specific

The benchmark shows that by augmenting the down stream matcher with multi-pass keyed blocking, synthetic candidate recall can be improved without altering the down stream matcher. This is in keeping with the overall lesson that combinations of identifiers should take into account variations in the field and absence of certain identifiers [7]. However, the rate of candidate recall is still 91.72% with rest of the errors limiting the end to end recall. If only match scoring were optimized, the system might seem correct, but it could still be missing tough pairs at the blocking phase. Future assessments should be reported for both the candidate generation and pair classification process with information on subgroup analysis and missingness.

The 1,024-bit Bloom matcher gets close to the plaintext q-gram reference on this generated set but that doesn't make privacy-utility equivalence. Normalized identifiers are directly connected to plain text baseline. On the other hand, there are cases of structural leakage in Bloom tokens and the matcher is given with a candidate-block and score information.

Results are restricted to a known synthetic data generator and fixed normalization pipeline, hand selected field weights, and an identity label development / test partition. They must not be projected in external populations or use.
Development identity groups of the same data set used for threshold calibration and disjoint identity groups of the data set used for test. This is more than a random split of pairs as no person's records can be on both halves. Nevertheless, the split does not represent scope shift (geographic, language, age, system vendor, time). Results could vary when there is unevenness in field distributions, variation in error mechanisms, nonrandom sampling of subjects, and varying overlap of sites.

a. Privacy*

The modeled participants include source partitions and a central matching process. If the key is kept secret and never reused in different contexts, the direct disclosure of the clear block string to a party without the key is prevented with HMAC. Bloom vectors minimize direct access to each q-gram but can offer good pattern hints that repeated and similar q-grams exist. Though the benchmark coordinator knows all the extracted feature records and the demonstration key, the implementation does not prevent the coordinator from being compromised or a key from being disclosed, collusively obtained, or attacked with a chosen input, or from inferring from the outputs. In this evaluation, the fields are synthetic and there is no real threat of PHI; the design should not be construed to be secure for PHI.
For a production version, an adversarial model and security proof; separate key provisioning to the project, rotation, and revocation; strong rate limits and output controls; leakage assessment; hardening the implementation; access control governance, and institutional, legal, and privacy review are necessary. MPC/PSI can increase the distrust of the coordinator, and impose protocol and operational costs [13]–[19] on account of this technique, depending on the requirements. Differential privacy deals with aggregate disclosure and, it is not a substitute for secure record-level linkage per se. No formal privacy budgets or differential-privacy mechanism were put in place.

J. Limitations

The first, the data source is synthetic, the partitions represent providers. Second, although 6354 of the potential 6357 of available Patient resources were obtained through extraction, the number of the omitted resources was recorded in the manifest. Third, a synthetic label, benchmark ground truth, is the ground truth, NOT an independently adjudicated clinical one. Fourth, It is a classifier that follows a fixed feature set, each with one threshold, weighted similarity rule and not a learned and calibrated clinical model. Fifth: the percent of test positive pairs (Tp) missed by blocking equality clusters and equality misses is 8.28%. Sixth, no adversarial attack evaluation, formal proof, end-to-end deployment or real world audit were performed on all claims regarding privacy. Third and last, an individual execution time in a local machine cannot back a general throughput or scale claims.

Conclusion

This paper presents a readily reproducible implementation and evaluate using held-out data of exact, plaintext q-gram, and HMAC-keyed Bloom-filter matching techniques applied to the five-partition Synthetic Denver FHIR data set developed by MITRE. For 604 positive cross provider test pairs, four pass blocking stage lost 6.28% of the Candidates (false rejection rate). The Bloom matcher improved the F1 score to 0.941, compared with 0.919 by the plaintext q-gram reference, with both having their own 1,024-bit message size. These are results on simulated records only. In this implementation neither HMAC nor Bloom filters provide privacy nor encryption. As part of the future step in the research, a separately managed evaluation under an explicit threat model with valid real multi-organization data and the use of a protocol like PSI/MPC will be carried out before application in clinical practice, which will also be reviewed from a privacy and security perspective.

Future Work

Going forward, the study should be replicated with officially issued records from at least two health care organizations given institutional, legal and privacy approval. The evaluation needs to rely on independently adjudicated labels, threat-model prespecification, and independently reviewed linkage protocol (e.g., PSI, MPC, etc.) where labels of the coordinator should not contain raw identifiers. It should be able to give missingness at site level, recall of candidates, entity and pair-level errors and confidence intervals, review burden for false-merge and variation or differences across subgroups, and review burden due to operation and costs. Protected key management, authenticated transport, output controls, auditability, correction and rollback procedures, and a security assessment would also be required prior to any live use. These steps are suggested but not done in this benchmark.

The proposed Real-World Validation Plan is as follows:

An upcoming study for a real world claim should be conducted step-by-step. The following sequence is recommended for the design and is not an experiment that has been conducted in this paper. The acceptance criteria is independent for each stage, allowing the team to break off before getting onto processing live identifiers if governance, security or data-quality requirements are not satisfied.

Table VI. Proposed Validation Stages

Stage	Evidence to collect	Decision gate
Governance and purpose	Institution approvals, permitted use, retention and deletion plan	Written authorization and data steward sign-off
Data characterization	Field availability, missingness, formats, and site-specific error patterns	Approve features and normalization per site
Security design	Adversary model, key custody, protocol leakage, transport and output controls	Independent review before identifier processing
Calibration	Separate development sample with identity-group split	Freeze thresholds and blocking rules
Blinded evaluation	Independent reference labels on held-out inter-site sample	Report pair and entity metrics with uncertainty
Operational review	False-merge adjudication, workload, correction and rollback drills	Clinical and privacy leadership acceptance

The study should publish a protocol that states the unit of analysis, permitted identifiers, reference-label generation procedure, adjudicator training, inclusion criteria, sample-size rationale, and evaluation period. The development sample should reflect the intended sites and population but remain distinct from the blinded test period. Thresholds and block passes should be frozen before unblinding. If repeated threshold tuning is required after seeing test errors, that sample becomes development data and a new held-out sample is needed.

The primary endpoint should include pairwise precision and recall with confidence intervals that account for repeated records from the same person and repeated site pairs. Candidate-generation recall should be reported separately, along with end-to-end recall, false-merge rate, and the number of pairs sent for review. If the system emits entity clusters, cluster-level purity and completeness should be specified in advance. A result should be stratified by site pair and by data-quality variables that can be safely analyzed under the study approvals. Small groups should be protected from disclosure in published tables.

A parallel security evaluation should verify the production implementation rather than infer security from the benchmark. Reviewers should test key compromise, dictionary attacks against likely demographic values, frequency and intersection leakage, malicious submissions, replay, unauthorized queries, output inference, transport security, and incident response. If MPC or PSI is selected, the protocol assumptions and outputs should be documented, and independent cryptographic review should precede a live test. Governance should include key rotation, access revocation, breach response, retention limits, audit trails, and a method to correct mistaken links without silently erasing provenance.

Before clinical use, a controlled pilot should compare the linkage output with current patient-identity workflows and measure how staff handle uncertain matches. False-positive and false-negative consequences should be reviewed with patient-safety and privacy representatives. Any claim of real-world performance should identify the exact participating data sources, period, cohort, label method, thresholds, and error rates. The public synthetic benchmark reported here can serve as an initial regression test for code changes, but cannot satisfy these gates or substitute for authorized clinical evidence.

References

1. HL7 International, “Interoperable Digital Identity and Patient Matching Implementation Guide, Version 2.0.0,” FHIR R4, 2024. [Online]. Available: https://hl7.org/fhir/us/identity-matching/
2. Office of the National Coordinator for Health Information Technology, “Patient Identity and Patient Record Matching,” updated Dec. 8, 2025. [Online]. Available: https://healthit.gov/standards-and-technology/patient-identity-and-patient-record-matching/
3. D. Kiernan et al., “Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet,” BMC Res. Notes, vol. 15, no. 1, Art. no. 337, 2022, doi: 10.1186/s13104-022-06243-5.
4. K. Marsolo et al., “Assessing the impact of privacy-preserving record linkage on record overlap and patient demographic and clinical characteristics in PCORnet,” J. Amer. Med. Inform. Assoc., vol. 30, no. 3, pp. 447–455, 2023, doi: 10.1093/jamia/ocac229.
5. U. Tachinardi et al., “Privacy-preserving record linkage across disparate institutions and datasets to enable a learning health system: The N3C experience,” Learn. Health Syst., vol. 8, no. 1, e10404, 2024, doi: 10.1002/lrh2.10404.
6. S. Randall et al., “A blinded evaluation of privacy preserving record linkage with Bloom filters,” BMC Med. Res. Methodol., vol. 22, no. 1, p. 22, 2022, doi: 10.1186/s12874-022-01510-2.
7. K. Tyagi et al., “Accuracy of privacy preserving record linkage for real world data in the United States: A systematic review,” JAMIA Open, vol. 8, no. 1, ooaf002, 2025, doi: 10.1093/jamiaopen/ooaf002.
8. A. Pathak et al., “Privacy preserving record linkage for public health action: Opportunities and challenges,” J. Amer. Med. Inform. Assoc., vol. 31, no. 11, pp. 2605–2612, 2024, doi: 10.1093/jamia/ocae196.
9. A. Pathak et al., “Proposed framework for adopting privacy-preserving record linkage for public health action,” J. Public Health Manag. Pract., vol. 31, no. 1, pp. E26–E33, 2025, doi: 10.1097/PHH.0000000000002027.
10. F. Armknecht et al., “Strengthening privacy-preserving record linkage using secure Bloom filters,” Proc. Privacy Enhancing Technologies, 2023, pp. 1–20, doi: 10.56553/popets-2023-0054.
11. MITRE, “Synthetic Denver FHIR dataset,” Synthea downloads, accessed Oct. 24, 2026. [Online]. Available: https://synthetichealth.github.io/downloads.html.
12. W. Yin et al., “Differential cryptanalysis of Bloom filters for privacy-preserving record linkage,” IEEE Trans. Inf. Forensics Security, 2024, doi: 10.1109/TIFS.2024.3421292.
13. S. Stammler et al., “Mainzelliste SecureEpiLinker (MainSEL): Privacy-preserving record linkage using secure multi-party computation,” Bioinformatics, vol. 38, no. 6, pp. 1657–1668, 2022, doi: 10.1093/bioinformatics/btaa764.
14. A. Adir et al., “Privacy-preserving record linkage using local sensitive hash and private set intersection,” arXiv:2203.14284, 2022.
15. Ş. S. Mağara et al., “Accelerating privacy-preserving medical record linkage: A three-party MPC approach,” arXiv:2410.21605, 2024.
16. X. Yu et al., “Multiparty threshold private set intersection protocol with low communication complexity,” Secur. Commun. Netw., vol. 2022, Art. no. 9245516, 2022, doi: 10.1155/2022/9245516.
17. S. D. Gordon, C. Hazay, and P. H. Le, “Fully secure PSI via MPC-in-the-head,” Proc. Privacy Enhancing Technologies, vol. 2022, no. 3, pp. 291–313, 2022, doi: 10.56553/popets-2022-0073.
18. A. Bay, Z. Erkin, J. H. Hoepman, S. Samardjiska, and J. Vos, “Practical multi-party private set intersection protocols,” IEEE Trans. Inf. Forensics Security, vol. 17, pp. 1–15, 2022, doi: 10.1109/TIFS.2021.3118879.
19. Z. Niu, H. Wang, Z. Li, and X. Song, “Privacy-preserving statistical computing protocols for private set intersection,” Int. J. Intell. Syst., vol. 37, pp. 10118–10139, 2022, doi: 10.1002/int.22420.
20. Office of the National Coordinator for Health Information Technology, “United States Core Data for Interoperability (USCDI), Version 4,” 2023. [Online]. Available: https://www.healthit.gov/isp/united-states-core-data-interoperability-uscdi.
21. K. El Emam et al., “An evaluation of the replicability of analyses using synthetic health data,” Sci. Rep., vol. 14, Art. no. 6978, 2024, doi: 10.1038/s41598-024-57207-7.
22. MITRE, “Synthea: Synthetic patient population simulator,” project documentation, updated 2024. [Online]. Available: https://github.com/synthetichealth/synthea.
23. M. Baumgartner et al., “Health data space nodes for privacy-preserving linkage of medical data to support collaborative secondary analyses,” Front. Med., vol. 11, Art. no. 1301660, 2024, doi: 10.3389/fmed.2024.1301660.

24. [24] A. B. Su, Y. Wang, D. Schiavazzi, and F. Liu, “Privacy-preserving data synthesis via differentially private normalizing flows with application to electronic health records data,” Proc. AAAI Symp. Ser., vol. 1, no. 1, pp. 161–167, 2023, doi: 10.1609/aaaiss.v1i1.27495.
25. [25] K. Han et al., “Private computation on common fuzzy records,” Proc. Privacy Enhancing Technologies, vol. 2025, no. 1, 2025, doi: 10.56553/popets-2025-0031.
26. [26] HHS Office for Civil Rights, “Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule,” reviewed 2022. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.
27. [27] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.
28. [28] National Institute of Standards and Technology, “The NIST Cybersecurity Framework (CSF) 2.0,” NIST CSWP 29, 2024, doi: 10.6028/NIST.CSWP.29.
29. [29] D. Meeker, C. Kallem, Y. Heras, S. Garcia, and C. Thompson, “Case report: Evaluation of an open-source synthetic data platform for simulation studies,” JAMIA Open, vol. 5, no. 3, ooac067, 2022, doi: 10.1093/jamiaopen/ooac067.