Real-World Matching Performance of Deidentified Record-Linking Tokens

Elmer V. Bernstam[1,2] Reuben Joseph Applegate[1] Alvin Yu[3] Deepa Chaudhari[1] Tian Liu[3] Alex Coda[3] Jonah Leshin[3]

[1] School of Biomedical Informatics, The University of Texas Health Science Center, Houston, Texas, United States
[2] Division of General Internal Medicine, Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, United States
[3] Datavant, Inc., San Francisco, California, United States

Address for correspondence Elmer V. Bernstam, MD, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Suite 600, 7000 Fannin Street, Houston, TX 77030, United States (e-mail: Elmer.V.Bernstam@uth.tmc.edu).

Appl Clin Inform 2022;13:865–873.

Abstract

Objective Our objective was to evaluate tokens commonly used by clinical research consortia to aggregate clinical data across institutions.

Methods This study compares tokens alone and token-based matching algorithms against manual annotation for 20,002 record pairs extracted from the University of Texas Houston's clinical data warehouse (CDW) in terms of entity resolution.

Results The highest precision achieved was 99.9% with a token derived from the first name, last name, gender, and date-of-birth. The highest recall achieved was 95.5% with an algorithm involving tokens that reflected combinations of first name, last name, gender, date-of-birth, and social security number.

Discussion To protect the privacy of patient data, information must be removed from a health care dataset to obscure the identity of individuals from which that data were derived. However, once identifying information is removed, records can no longer be linked to the same entity to enable analyses. Tokens are a mechanism to convert patient identifying information into Health Insurance Portability and Accountability Act-compliant deidentified elements that can be used to link clinical records, while preserving patient privacy.

Conclusion Depending on the availability and accuracy of the underlying data, tokens are able to resolve and link entities at a high level of precision and recall for real-world data derived from a CDW.

Keywords
- patient records
- electronic health records
- privacy
- research
- dataset

Background and Significance

Health care data are fragmented across numerous collection points (electronic health records, insurance claims, pharmacy prescriptions, etc.) depending on where the patient has interacted with the health care system. Exchanging identified health care data is problematic due to ethical and regulatory requirements to protect patient privacy.

Record linkage is an entity resolution problem where information about the same individual is integrated into a single cluster, despite the individual being referenced differently by different data sources. Traditional record linkage requires personally identifying information (PII), such as name, date of birth (DOB), and address to be available in two or more datasets.[1] In contrast, privacy-preserving record linkage (PPRL) allows two or more datasets to be linked (e.g.,

received
July 12, 2022
accepted after revision
July 22, 2022
accepted manuscript online
July 27, 2022

DOI https://doi.org/10.1055/a-1910-4154.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

to recognize the same individual within separate datasets) without sharing sensitive identifiers. Therefore, PPRL solutions are attractive, particularly for research networks involving multiple independent institutions.[2]

PPRL methods can be divided into deterministic PPRL and probabilistic PPRL.[3] Both approaches start with demographic data about an individual and involve one-way hashing of identifying data such that these identifying data can no longer be connected to the originating patient. Most often, the input to the one-way hash is a string (e.g., first name), and the output is a deterministically determined string that cannot be tied back to the input string on its own.

In a deterministic PPRL system, patient demographic data items are concatenated and hashed, and the resultant random string is used directly as a unique identifier for that patient. This leads to high precision but limits recall if data are inaccurate or missing. Moreover, deterministic methods cannot account for frequently changing data such as addresses, zip codes, etc., or variants such as nicknames, alternative spellings, or misspellings.

In a probabilistic PPRL system, multiple demographic identifiers are first separately encrypted, taking care to preserve enough variability in each encrypted output string to prevent dictionary attacks (i.e., brute force approach that attempts to break the encryption by matching an encrypted string against every possible encrypted string generated from some universe of inputs). The resultant collection of random strings is used as the feature set to establish a probabilistic linkage between two records. This probabilistic linkage preserves patient privacy by only admitting the minimum set of hashed elements into the feature space, but at the same time preserves as much information as possible to allow for increased recall, especially in cases of missing, changing, or inaccurate data within certain demographic elements.[4]

Multiple PPRL systems exist in both academic and commercial settings. In general, these systems allow record linking while obfuscating PII. One such system was created by Datavant (Datavant, Inc., San Francisco, CA). Several hundred health care entities across the United States exchange datasets that have been deidentified by means of generating Datavant tokens from raw PII. These entities span the health care continuum, including, for example, laboratories, academic research institutions, and the U.S. National Institutes of Health. They represent a diverse set of use cases, and there are a variety of medical data fields across the datasets exchanged, ranging from physician National Provider Identifier numbers to laboratory test results to insurance claim charges. In addition to medical data fields, patient PII fields may vary across datasets. Moreover, the underlying populations across these sources of medical data vary considerably with respect to age, gender, and ethnicity.

Objectives

In previous work,[5] we tested a variety of record-linking algorithms and compared their performance. In this paper, we describe the results of a matching study to evaluate the matching performance of commonly-used deidentified tokens, using a large, real-world, human-annotated identified EHR dataset as a gold standard. By analyzing multiple deterministic token-based matching algorithms on real-world clinical data, this study provides a benchmark of real-world performance. In addition, we offer specific guidance regarding the utilization of these algorithms and the required data based on empirical evaluation against a large, real-world, manually reviewed dataset.

Methods

Data were derived from the University of Texas Health Science Center at Houston's clinical data warehouse (CDW). At the time that this dataset was generated, the CDW contained 2.61 million distinct medical record numbers. Some of these 2.61 million medical record numbers represented duplicate records (i.e., patient John Smith has two or more records in the database). The eight fields that were most often present in patient records, first name, middle name, last name, DOB, social security number (SSN), gender, primary address, and primary phone number, were extracted for each record.

Datavant's patient matching software requires that the underlying raw data contain the PII fields necessary to generate constructs derived from PII, but not containing PII. These constructs are referred to as "tokens." To match patient records using Datavant tokens, one needs to employ a deterministic approach that relies on token comparisons. In a given individual use case, one may build on top of these deterministic algorithms by making use of any additional data elements that are available.

Datavant's token-based matching uses heuristics built on top of approximate deterministic PPRL and consists of software installed on-premises by each identified data source that obfuscates PII to create an output file containing unique encrypted tokens (also called Patient Keys) for each patient. These tokens are coupled to Health Insurance Portability and Accountability Act-compliant deidentified clinical records, which can then be exchanged with data partners and linked to other matching records without revealing a patient's identity.

Blocking Strategy for the Manually Reviewed Dataset

To decrease the computational cost of identifying duplicates and to increase the yield of the manual review, we used a common blocking strategy to exclude record pairs that were not likely to be duplicates.[6] Specifically, we identified records as potential duplicates if they matched on: first and last names; first name and DOB; last name and DOB; or SSN (to increase recall of the blocking search we encoded names using Soundex[7]). This generated approximately10 million distinct potential duplicates.[5] In total, 20,002 record pairs were then randomly sampled from this set for annotation. This study has been approved by the Committee for the Protection of Human Subjects (the UTHSC-H IRB) under protocol HSC-SBMI-13-0549.

Manual Review

Two reviewers independently reviewed each of 20,002 randomly-selected record pairs as described.[5] Reviewers assigned a match score between 1 and 5 representing their

Table 1 Match algorithms used in this evaluation

A. Token descriptions			
Name	**Token description**		
Token 1	Last name + 1st initial of first name + gender + DOB		
Token 2	Last name (soundex) + first name (soundex) + gender + DOB		
Token 3	Last name + first name + DOB + Zip 3 (three digit zip code)		
Token 4	Last name + first name + gender + DOB		
Token 5	SSN + gender + DOB		
Token 7	Last name + 1st three characters of first name + gender + DOB		
Token 9	First name + address		
Token 16	SSN + first name		
Token 22	Cell phone number (United States)		
B. Token combinations			
Name	**Tokens used**	**Description**	**Evaluation requirement**
Single token match	1 or 2, or 3 or 4, OR 5 or 16	Two records match if they share at least a single token in common.	At least one of tokens 1,2,3,4,5, and16 is present
Demographic	1 and 2	Two records match on both of these tokens to indicate the records have the same name, age, and gender.	Tokens 1 and 2 are present
Net tokens	Any subset of 1, 2, 4, 5, 7, 9, 16	Two records match if more tokens match than do not. Note, tokens based on email, phone, or address are excluded from this list because they are often most prone to error on input.	At least 3 of tokens 1,2,4,5,7,9, and16 are present
SSN	5 or 16	Tokens 5 and 16 use SSN (United States). Two records match if either token 5 or token 16 match.	Token 5 or 16 is present

Abbreviations: DOB, date of birth; SSN, social security number.

subjective confidence in the classification: (1) definite mismatch; (2) probable mismatch; (3) uncertain; (4) probable match; and (5) definite match.[8] Reviewers were asked to designate a record pair as a match (4 or 5) or nonmatch (1 or 2) "only if they would have been comfortable with a computer making the same assertion automatically based on the available data." In case of disagreement between reviewers, meaning one reviewer thought the records matched (4 or 5) while the other did not (1 or 2), or if one of the reviewers thought it was impossible to assert match status (3) with the available data, "the records were forwarded to an evaluation by four independent reviewers." Record pairs "that were not assigned a match/nonmatch status unanimously (or by three reviewers when the fourth reviewer was uncertain [3]) went to further review by open discussion of the entire review panel (six reviewers). Only 48 record pairs could not be adjudicated by four reviewers. These were assigned by consensus (10 matched and 38 nonmatched). In all but 48 cases (0.24%) reviewers felt that the eight demographic data fields were sufficient to assign match status without requiring additional data.

Datavant software was used to create eight different encrypted tokens for each of the 40,004 records (20,002 pairs). Tokens rely on demographic factors such as first name, last name, gender, DOB, etc., to generate tokens for matching purposes. Generally, the patient's zip code would be included in the token methodology; however, the zip code was unavailable in this dataset and was therefore excluded.

►**Table 1** describes the eight tokens used in this evaluation. With single-token comparison, two records match if the relevant tokens match (i.e., Token 1 derived from record A matches Token 1 derived from record B). ►**Table 1** describes multitoken approaches, which were selected by considering common matching strategies across sites using the Datavant token. Tokens are generated in a two-step process—one-way master token generation and then site-specific token encryption (►**Fig. 1**).

One-Way Master Token Generation

The first step in the linkage process is to create a set of encrypted hashed tokens based on the input PII of each patient. The underlying PII is validated, concatenated, and

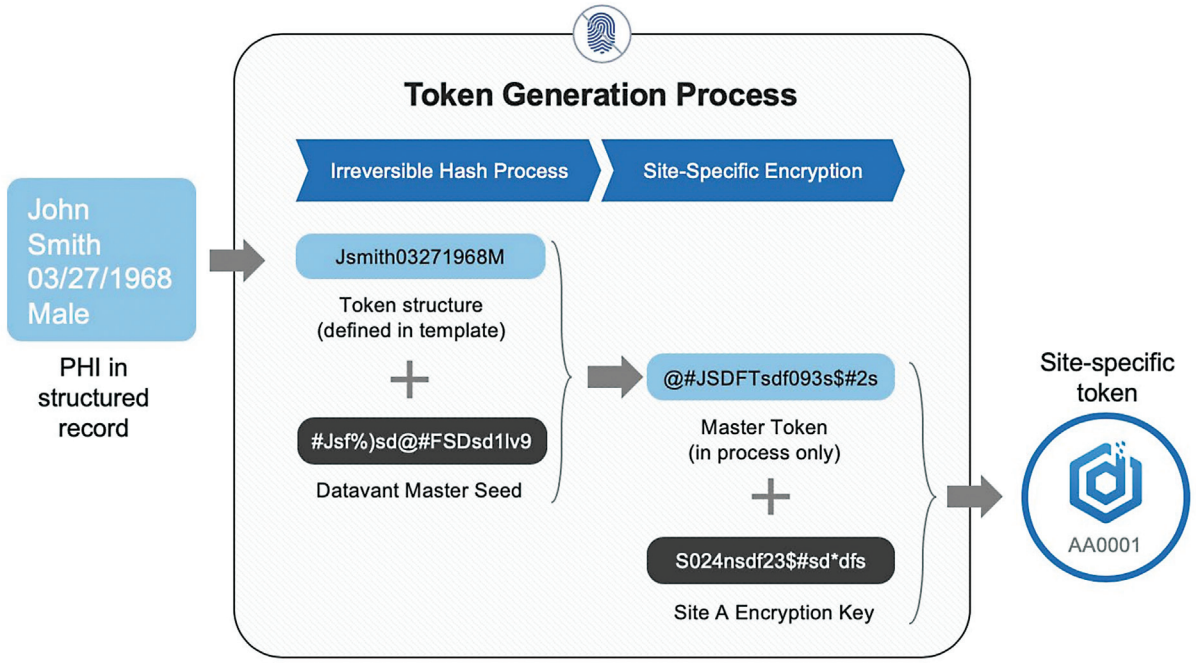

Fig. 1 Token generation.

irreversibly hashed using the SHA-256 algorithm[9] into a series of master tokens using a secure, fixed random string that is added to the concatenated string before creating the final token. The irreversible hashing mechanism ensures that the patient's PII used to create the tokens cannot be recovered from the output value.

Encrypted Site-Specific Token Generation

The master tokens are then encrypted using a site-specific AES-128[10] key. The same PII will always generate the same set of master tokens, but PII is never present in any output or log stream. Only the site-specific encryption tokens are written to the output file. Since tokens are site specific, a breach at one site will not propagate across the Datavant ecosystem, which prevents the reidentification of patients across datasets at different sites and allows for a governance mechanism that prevents linking of patient records across datasets without the permission of both parties.

After tokenization, records were matched and the results were compared with manual annotation, which was considered to be the ground truth. We calculated precision, recall, and F1 using standard definitions (►Table 2.

It is important to note that the dataset had inconsistent fill rates of PII (fill rate = 1- missing rate) and therefore generation rates for individual tokens varied (►Table 3). To avoid bias related to the fill rates for our dataset, we reported recall based only on record pairs in which each record contained the data required to compute the specific token or combination of tokens (see "Evaluation requirement" field in ►Table 1). The "Pair fill rate" in ►Tables 2A and 2B is the proportion of all record pairs for which the required data were available.

Results

►Table 3 shows the demographics and rates of missing values for the study population. The data reflect inconsistent coding practices. For example, the race was sometimes listed as "Hispanic" in addition to, or instead of, ethnicity. Similarly, age was calculated based on DOB and an index date of May 05, 2011 (the date the manual review set was created) which may include errors that are reflected in the table (e.g., DOB 1/1/1900 = unknown). Since our goal was to evaluate real-world performance of PPRL, we did not harmonize the data (e.g., remove Hispanic race).

Token Matching Evaluation

Token 5 based on SSNs had very high precision and good recall, but relatively low fill rate (Table 2). Compared with Token 5, Token 16 had similar precision but lower recall. This may be due to the additional PII elements required by Token 5 (gender, DOB) versus Token 16 (first name); the results imply that true matches are more likely to share gender and DOB than first name.

Tokens 1, 2, and 4 use a combination of name, DOB, and gender. While each element is not uniquely identifying when used separately (e.g., there are many people named John), combinations of these elements can precisely distinguish unique individuals. Tokens 1 and 2 optimize recall, whereas Token 4 optimizes precision. Token 4 had high precision as it used exact matches of first and last name but had lower recall likely due to different spellings of those names (e.g., Stephen vs. Steven, or Nick vs. Nicholas). Tokens 1 and 2 have higher recall because they allow more flexibility with names but lower precision because, for example, they would generate

Table 2 Precision, recall, F1, and fill rates for the eight token types and algorithms tested in this evaluation

Token or algorithm	True positives (TP)	False negatives (FN)	False positives (FP)	Precision[b]	Recall[a]	F1[c]	Valid pairs	Pair fill rate
Token 1	1,098	118	24	97.9%	90.3%	94%	20,002	100.00%
Token 2	955	259	14	98.6%	78.7%	88%	20,000	99.99%
Token 4	787	427	1	99.9%	64.8%	79%	20,000	99.99%
Token 5	355	50	1	99.7%	87.7%	93%	779	3.89%
Token 7	1,076	138	16	98.5%	88.6%	93%	20,000	99.99%
Token 9	271	888	2	99.3%	23.4%	38%	18,163	90.81%
Token 16	247	157	1	99.6%	61.1%	76%	778	3.89%
Token 22	476	437	22	95.6%	52.1%	67%	13,603	68.01%
Single Token Match	1,161	55	36	97.0%	95.5%	96%	20,002	100.00%
Demographic	925	289	4	99.6%	76.2%	86%	20,000	99.99%
Net Tokens	910	304	1	99.9%	75.0%	86%	20,000	99.99%
SSN	368	37	2	99.5%	90.9%	95%	779	3.89%

Abbreviations: SSN, social security number.
[a]Recall = TP/(TP + FN).
[b]Precision = TP/(TP + FP).
[c]F1 = 2* [precision*recall]/[precision + recall].
Note: Token 3 is not listed because zip code was not included in the manual review data; therefore, the fill rate was 0%.

the same token value for distinct names such as "Maria" and "Marie."

Match Approach Evaluation

We tested four matching approaches to see how they performed relative to matching using identified data (►Table 2).

Single Token Match

A matching strategy that leveraged multiple token types (Tokens 1, 2, 4, 5, and 16) to handle inconsistent fill rates yielded a balance of precision (97.0%) and recall (95.5%).

Demographic

Both Tokens 1 and 2 must match. While both Token 1 and Token 2 increase recall through fuzzy matching (just first initial is used in Token 1 and the soundex[7] value is used for names in Token 2), when used together these tokens allow precision of 99.6% without sacrificing much recall compared with the individual tokens. A comparison of precision and recall using Token 4, which required exact match on first and last names, implies that soundex to first name was improved F1.

Net Tokens

The number of matching tokens must exceed the number of nonmatches when comparing the rest of the tokens available (essentially, majority rules). The advantage is that this approach considers all of the tokens available and is robust to varying fill rates. This approach performs well on precision (approaching 99.9%) though the recall was somewhat lower than other approaches at 75%.

Social Security Number

If the underlying SSN for each record is reliable, this algorithm yields high precision (99.5%) and good recall (90.9%).

Hispanic Ethnicity

Hispanic ethnicity is common in our cohort. People who identify as Hispanic are the second fastest-growing racial or ethnic group in the United States 2000 to 2019.[11] Further, previous studies have compared algorithm performance on Hispanic versus non-Hispanic populations.[12,13] Therefore, we divided the population into two distinct groups: at least one record in the pair was of Hispanic ethnicity versus neither record was of Hispanic ethnicity (or missing ethnicity data). Performance was generally similar across the two groups (►Table 4), apart from lower recall for token types and algorithms that rely on first name match (exact or soundex): Token 2, demographic (which uses Token 2), and net tokens (which uses Token 2 and Token 4). From this, one may infer that there are more variants of the same patient's first name in the dataset and that for this dataset, matching on Token 1, or using more permissive matching criteria such as single token match, yielded higher F1 scores. We have omitted precision and recall in cases with fewer than 50 true positive pairs as these results are not likely to be generalizable.

Optimizing Matching using Different Tokens

Using different tokens, either individually or in combination, changes the precision/recall tradeoff (►Fig. 2).

Discussion

We found that a token-based matching system based on commonly available PII performed well. For use cases that require high precision, Token 5 (derived from SSN, gender, and DOB) had a precision of 99.7% and recall of 87.7%. For high recall, Token 1 (utilizing last name, first name, gender, and DOB) yielded a recall of 90.3% while maintaining

Table 3 Study population and dataset ($n = 40{,}004$; categories as listed in the dataset)

Field	Value/Range	%	Fill rate (%)
Age			99.5
	0–10	11.05	
	11–20	10.33	
	21–30	16.15	
	31–40	21.32	
	41–50	16.35	
	51–60	12.09	
	61–70	7.00	
	71–80	3.32	
	81–90	1.59	
	91–100	0.29	
	101–110	0.03	
Gender			100
	M	44.5	
	F	55.5	
	Other	0.1	
Race			58.4
	African American	5.09	
	All other	12.9	
	American Indian, Esk [i]mo, or Aleut	0.08	
	Asian or Pacific Islander	0.25	
	Caucasian	7.43	
	Hispanic or Latino	1.29	
	Latin American	23.29	
	Other	7.72	
	Other race	0.39	
Ethnicity			59.8
	Hispanic	17.4	
	Non-Hispanic	41.6	
First name			100
Middle initial			19.9
Last name			100
Date of birth			100
Phone number (United States)			94.6
Address first line (United States)			97.5
Zip (three digit)			0
Social security number			37.2

precision at 97.9%. Combinations of tokens can perform better than individual tokens. For example, single token match (at least one pair of Tokens 1, 2, 3, 4, 5, or 16 matches) yielded a precision of 97.0% and recall of 95.4%; performance remained high for pairs that included Hispanic ethnicity.

When missing PII fields are inconsistent across records, a multiple-token strategy is necessary. Based on matching results for individual tokens, one may also devise custom strategies, for example, in use cases where SSN is not present, one may rely on tokens derived from name, gender, and DOB.

Strengths of the study included a large, real-world, manually reviewed dataset based on 20,000 manually reviewed record pairs (i.e., 40,000 individual records). The manual review process is described in detail[5] but includes multiple independent reviews for questionable cases, possibly decreasing errors. Previous real-world PPRL evaluations such as[14,15] compared PPRL against "gold standard" matching that used unencrypted records (i.e., PPRL vs. non-PPRL). In contrast, our gold standard consisted of human-reviewed record pairs (i.e., absolute performance of PPRL). The large dataset, as well as the relatively high prevalence of Hispanic ethnicity (►Table 3), allowed us to evaluate the effect of Hispanic ethnicity on match accuracy.

Our work has several limitations. First, our data were selected from a single academic health system and thus our results may not generalize to other settings. However, the Houston metropolitan area is arguably the most diverse in the country.[16] Second, the manual review was limited by the available data. Thus, some errors may be undetected. As an example, infant twins are difficult to distinguish because they share many demographics including DOB, address, phone number, last name, etc., and may lack distinguishing data such as SSN. Third, we used a blocking strategy to create the dataset used for evaluation. We did this to ensure that the set contained matching records. However, it is possible that performance was altered by removing record pairs that were very unlikely to match. Since blocking eliminated "obvious" mismatches, including these cases would likely have improved performance. Finally, we did not exhaustively test all possible identifier combinations and relied upon Datavant software.

Previous studies found (or theorized) that Hispanic ethnicity was associated with lower match accuracy.[12,13] In contrast, we found that Hispanic ethnicity was not consistently associated with lower recall or precision. Notably, Hispanic ethnicity is variably recorded in real-world EHR data. Ethnicity may be underreported[12] and the ethnicity field is used inconsistently. We may have underrecognized Hispanic ethnicity. If so, then this would be expected to decrease the match accuracy of non-Hispanic record pairs. However, match accuracy remained high for both Hispanic and non-Hispanic record pairs.

Unlike matching systems that create a single patient ID for all datasets, the different precision and recall values of each token, or token combination, allow users to choose the best approach for their use case. Below we discuss different use cases.

Cohort Identification (Recall > Precision)

Examples include looking for patients with rare diseases or identifying locations with the most patients eligible for a clinical trial. In these cases, a user may decide to optimize recall to avoid missing any eligible patients, at the cost of

Table 4 Precision, recall, and fill rates for the token types and algorithms by ethnicity

Token or algorithm	Ethnicity	TP	FN	FP	Valid pairs	Pair fill rate	Precision	Recall	F1
Token 1	Not Hispanic	1,029	110	23	13,890	69.44%	97.81%	90.34%	94%
	Hispanic	69	8	1	6,112	30.56%	98.57%	89.61%	94%
Token 2	Not Hispanic	901	236	13	13,888	69.43%	98.58%	79.24%	88%
	Hispanic	54	23	1	6,112	30.56%	98.18%	70.13%	82%
Token 4	Not Hispanic	744	393	1	13,888	69.43%	99.87%	65.44%	79%
	Hispanic		34	0	6,112	30.56%			
Token 5	Not Hispanic	334	48	1	673	3.36%	99.70%	87.43%	93%
	Hispanic		2	0	106	0.53%			
Token 7	Not Hispanic	1,007	130	15	13,888	69.43%	98.53%	88.57%	93%
	Hispanic	69	8	1	6,112	30.56%	98.57%	89.61%	94%
Token 9	Not Hispanic	259	827	0	12,428	62.13%	100.00%	23.85%	39%
	Hispanic		61	2	5,735	28.67%			
Token 16	Not Hispanic	233	148	1	672	3.36%	99.57%	61.15%	94%
	Hispanic		9	0	106	0.53%			
Token 22	Not Hispanic	449	411	18	9,334	46.67%	96.15%	52.21%	68%
	Hispanic		26	4	4,269	21.34%			
Single token match	Not Hispanic	1,086	53	34	13,888	69.43%	96.96%	95.35%	96%
	Hispanic	75	2	2	6,112	30.56%	97.40%	97.40%	97%
Demographic	Not Hispanic	874	263	4	13,888	69.43%	99.54%	76.87%	87%
	Hispanic	51	26	0	6,112	30.56%	100.00%	66.23%	80%
Net tokens	Not Hispanic	859	278	1	673	3.36%	99.88%	75.55%	86%
	Hispanic	51	26	0	106	0.53%	100.00%	66.23%	80%
SSN	Not Hispanic	345	37	2	13,890	69.44%	99.42%	90.31%	95%
	Hispanic				6,112	30.56%			

Abbreviations: FN, false negative; FP, false positive; SSN, social security number; TP, true positive.
Note: Token 3 is not listed because zip code was not included in the manual review data; therefore, the fill rate was 0%.

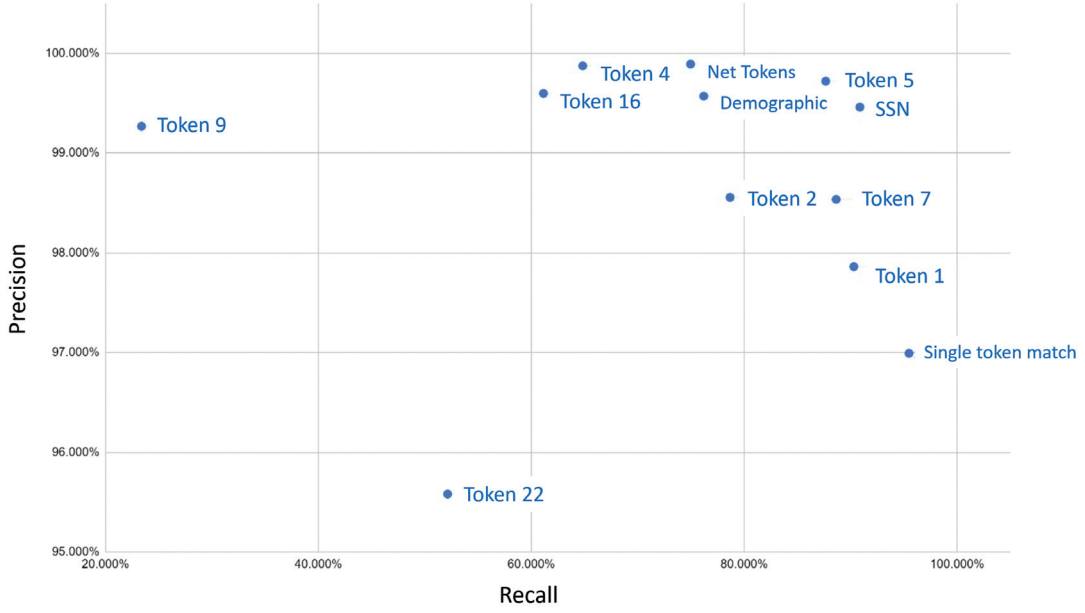

Fig. 2 Precision and recall of different matching strategies.

lower precision. The user might match on Token 1, or Token 5 or 16 if SSN is present.

Cohort Analytics (Balanced Recall and Precision)

For most analytics such as outcomes research, cost analysis, patient segmentation, drug adoption patterns, etc., it is important to have both a large sample and accurate matching. In such cases, the user might match on Token 4 alone or Token 1 and 2 together, either of the SSN-based tokens or single token match.

Clinical Decision Support, Drug Safety, and Intervention (Recall < Precision)

As real-world evidence is increasingly used to support drug approval decisions, risk stratification, and to recommend treatment, the underlying data must be accurate. In these cases, there is little tolerance for false-positive matches, and users may choose to optimize precision at the cost of recall. The user might match using Token 4, 5, or 16 alone, or using the net tokens match, or require all available tokens to match exactly. In contrast, drug safety monitoring may benefit from higher recall at the cost of precision to capture rare events.

Conclusion

Token-based matching systems can link deidentified patient records accurately. Using different token designs or combinations of tokens, users can adjust precision and recall to match their use cases.

Clinical Relevance Statement

Privacy-preserving record linkage (PPRL) is most commonly used in clinical research. Datavant tokens are used for National Institute of Health-sponsored multiinstitutional clinical trials and data-enabled research networks such as the Patient-Centered Outcomes Research Institute Clinical Data Research Networks. More direct clinical applications are possible such as those focusing on transitions of care across institutions and interinstitutional quality improvement projects. Health care consumers can use tokens to log into applications without revealing their identity.

Multiple Choice Questions

1. The token-based matching system used in this study:
   a. Requires all personally identifying information (PII) to be shared between institutions that wish to share data.
   b. Requires some PII to be shared between institutions that wish to share data.
   c. Requires PII to be shared with a trusted third party.
   d. Requires no PII to be shared and thus can be considered a form of privacy-preserving record linkage.

   Correct Answer: The correct answer is option d. Software used to create tokens is installed on premises; therefore, no PII needs to leave the institution.

2. The performance of token-based matching system used in this study:
   a. Is independent of the dataset.
   b. Depends on the distribution of clinical data such as vital signs, laboratory results and clinical notes.
   c. Depends on the distribution of demographic information.
   d. Depends on the speed of the processor used to calculate the token hashes.

   Correct Answer: The correct answer is option c. The tokens are created using one-way hash functions of demographic information. The distribution of the demographic information, therefore, determines the resulting output.

Protection of Human and Animal Subjects

This study has been approved by the Committee for the Protection of Human Subjects (the UTHSC-H IRB) under protocol HSC-SBMI-13–0549.

Author Contributions

R.J.A. and A.Y. wrote the initial manuscript. R.J.A., D.C., A.Y., A.C., and T.L. performed the data analysis. J.L. and J.L. revised the manuscript. E.V.B. provided the data. All authors reviewed and approved the manuscript prior to submission.

Data Availability Statement

The data underlying this article cannot be shared publicly due to the fact that these data are individually identifiable and represent real-world patients.

Funding

This work was supported in part by the National Center for Advancing Translational Sciences (NCATS) under awards UL1TR003167 and U01TR002393; the Cancer Prevention and Research Institute of Texas (CPRIT), under award RP170668, Datavant, Inc., and the Reynolds and Reynolds Professorship in Clinical Informatics.

Conflict of Interest

T.L., J.L., A.C., and A.Y. made contributions to this study while being employees of Datavant, Inc.

References

1 Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc 1969;64(328):1183–1210
2 Bian J, Loiacono A, Sura A, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. JAMIA Open 2019;2(04):562–569
3 Stausberg J, Waldenburger A, Borgs C, Schnell R. Combining different privacy-preserving record linkage methods for hospital admission data. Stud Health Technol Inform 2017; 235:161–165
4 Nguyen L, Stoové M, Boyle D, et al. Privacy-preserving record linkage of deidentified records within a public health surveillance system: evaluation study. J Med Internet Res 2020;22(06):e16757
5 Joffe E, Byrne MJ, Reeder P, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual

review datasets in duplicate records reconciliation. J Am Med Inform Assoc 2014;21(01):97–104
6 Baxter R, Christen P, Churches T. A Comparison of Fast Blocking Methods for Record Linkage. In: Kdd 2003 Workshops.; 2003:25–27
7 Soundex System|The Soundex Indexing System. Published online 2007. Accessed July 16, 2021 at: https://www.archives.gov/research/census/soundex
8 Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. Health Informatics J 2008;14(01):5–15
9 Penard W, van Werkhoven T. On the secure hash algorithm family. In: Cryptography in Context.; 2008:1–18. Accessed July 16, 2021 at: https://web.archive.org/web/20160330153520/http://www.staff.science.uu.nl/~werkh108/docs/study/Y5_07_08/infocry/-project/Cryp08.pdf
10 Announcing the ADVANCED ENCRYPTION STANDARD (AES) Published online November 26, 2001. Accessed July 16, 2021 at: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf
11 Budiman A, Ruiz NG. Asian Americans are the fastest-growing racial or ethnic group in the U.S. Pew Research Center. Accessed June 28, 2022 at: https://www.pewresearch.org/fact-tank/2021/04/09/asian-americans-are-the-fastest-growing-racial-or-ethnic-group-in-the-u-s/
12 Palloni A, Arias E. Paradox lost: explaining the Hispanic adult mortality advantage. Demography 2004;41(03):385–415
13 Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. J Aging Health 2011;23(08):1263–1284
14 Irvine K, Smith M, de Vos R, et al. Real world performance of privacy preserving record linkage. Int J Popul Data Sci 2018;3(04): . Doi: 10.23889/ijpds.v3i4.990
15 Brown AP, Borgs C, Randall SM, Schnell R. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. BMC Med Inform Decis Mak 2017;17(01):83
16 Houston Still Most Diverse City in the Nation. Report Finds. Accessed July 16, 2021 at: https://www.houston.org/news/houston-still-most-diverse-city-nation-report-finds

