Statistical Journal of the IAOS 38 (2022) 413–421
DOI 10.3233/SJI-210891
IOS Press

A methodological assessment of privacy preserving record linkage using survey and administrative data

Lisa B. Mirel[a,*], Dean M. Resnick[b], Jonathan Aram[a] and Christine S. Cox[c]
[a]*Data Linkage Methodology and Analysis Branch, Division of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA*
[b]*Statistics and Methodology Department, NORC at the University of Chicago, Bethesda, MD, USA*
[c]*Health Care Programs Department, NORC at the University of Chicago, Bethesda, MD, USA*

Abstract.
BACKGROUND: The National Center for Health Statistics (NCHS) links data from surveys to administrative data sources, but privacy concerns make accessing new data sources difficult. Privacy-preserving record linkage (PPRL) is an alternative to traditional linkage approaches that may overcome this barrier. However, prior to implementing PPRL techniques it is important to understand their effect on data quality.
METHODS: Results from PPRL were compared to results from an established linkage method, which uses unencrypted (plain text) identifiers and both deterministic and probabilistic techniques. The established method was used as the gold standard. Links performed with PPRL were evaluated for precision and recall. An initial assessment and a refined approach were implemented. The impact of PPRL on secondary data analysis, including match and mortality rates, was assessed.
RESULTS: The match rates for all approaches were similar, 5.1% for the gold standard, 5.4% for the initial PPRL and 5.0% for the refined PPRL approach. Precision ranged from 93.8% to 98.9% and recall ranged from 98.7% to 97.8%, depending on the selection of tokens from PPRL. The impact of PPRL on secondary data analysis was minimal.
DISCUSSION: The findings suggest PPRL works well to link patient records to the National Death Index (NDI) since both sources have a high level of non-missing personally identifiable information, especially among adults 65 and older who may also have a higher likelihood of linking to the NDI.
CONCLUSION: The results from this study are encouraging for first steps for a statistical agency in the implementation of PPRL approaches, however, future research is still needed.

Keywords: National Center for Health Statistics, National Hospital Care Survey, National Death Index

1. Background

The National Center for Health Statistics (NCHS) serves as the nation's principal federal health statistics agency, whose mission is to provide statistical information that can be used to guide actions and policies to improve the health of the American people. NCHS conducts several population-based and establishment health surveys designed to collect important information about the health of the U.S. population. Through the NCHS Data Linkage Program, data from these surveys are linked to mortality data from the National Death Index (NDI), health care utilization data from Medicare and Medicaid administrative records, and federal housing program participation records from the Department of Housing and Urban Development [1]. These data link-

*Corresponding author: Lisa B. Mirel, Data Linkage Methodology and Analysis Branch, Division of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, USA. Tel.: +1 301 458 4087; E-mail: LMirel@cdc.gov.

ages are based on both deterministic and probabilistic linkage algorithms, which rely on the exchange and comparison of personally identifiable information (PII) between data sources. The linked data expand the scientific utility of surveys and enable richer analysis than would be possible with each data source alone. The NCHS linked data resources have supported over 1,000 PubMed-indexed scientific publications [2].

Privacy concerns are one of the greatest barriers to data linkage. Often the personal identifiers used for data linkage are protected by laws and regulations [3,4]. For example, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule prohibits the release of identifiable health information except in certain circumstances [5]. However, the HIPAA Privacy Rule does allow for the release of protected health information that has been de-identified, and the Rule provides specific standards for de-identification [5]. Privacy preserving record linkage or "PPRL" is a method that can be used to link de-identified data [6]. One PPRL technique that meets HIPAA standards is called "hashing" [7]. Hashing converts names, addresses, and dates of birth into unique encrypted codes that protect the original values. Because record linkage often involves different combinations of PII, hashes are often combined to form multiple "tokens" for one individual [8]. One example of a token could include sex, date of birth, Social Security Number (SSN) [9] and another token could include sex, address, name and SSN.

Previous research has shown the feasibility of PPRL and assessed its accuracy [10,11]. A PPRL study that used laboratory and clinical data included multiple records per individual and multiple reporting sites. The records in the multiple sources contained complete name, date-of-birth, ZIP code and a unique health insurance identification number (Medicare ID number). The results from the PPRL produced sensitivity and specificity that were as high as 100% [10]. In a larger dataset, with up to 20% of records missing Medicare ID number, sensitivity estimates reached 95% and specificity estimates were as high as 99% [10]. A study conducted using hospital admission records from Western Australia reported no difference in linkage quality when comparing PPRL to linkage with unencrypted identifiers. The same study reported slightly lower accuracy using PPRL to link records from a different part of Australia which had a higher percentage of missing PII values [11]. These studies demonstrate the potential of PPRL to accurately link healthcare records. They also highlight the importance of data quality and the potential for lower-quality linkages when personal identifiers are missing or incorrect.

PPRL techniques could expand the sources of data that could be used in linkages. However, before implementing these methods it is important to understand the impact of using a new linkage approach on data quality and secondary data analysis. Therefore, this analysis should be treated as a case study where an assessment of data quality is being performed prior to implementing PPRL methods with new sources however, future research is still needed.

1.1. Objective

To assess the potential of using PPRL techniques with NCHS survey data and administrative records, the analysis described here utilized Datavant software to perform data hashing/tokenization followed by utilizing SAS software to link data from the 2016 National Hospital Care Survey (NHCS) to the 2016/17 NDI [12]. Datavant has been used in a variety of settings, including PCORnet which is the National patient centered clinical research network, to de-identify and link patient records (https://pcornet.org/). For this study, the results of the PPRL are compared to a previously conducted linkage of the same data sets using unencrypted identifiers [13]. By comparing the linkage results obtained from encrypting algorithms with the prior gold standard results (the previously linked NHCS-NDI data) the accuracy of the PPRL method is assessed.

This paper provides an overview of the linkage process using unencrypted identifiers and PPRL and compares the results of the PPRL to the standard linkage. It also provides estimates of the impact of PPRL on secondary data analysis. A summary of findings as well as a discussion of future research needs is provided.

2. Materials and methods

2.1. Description of data sources

2.1.1. National Hospital Care Survey (NHCS)

The NHCS is an establishment survey that collects inpatient, emergency department (ED), and outpatient department episode-level data from sampled hospitals. The goal of NHCS is to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, health services utilization, and substance involved ED visits. From participating hospitals, NHCS collects data on all inpatient and ambulatory care visits occurring during the calendar year. The target uni-

verse for NHCS is all inpatient discharges and in-person ambulatory care visits in noninstitutional, nonfederal hospitals in the 50 states and the District of Columbia that have 6 or more staffed inpatient beds. The patient records collected in the NHCS include patient PII (e.g., name, date of birth (DOB), and SSN, which allows for the linkage of episodes of care across hospital units as well as to other data sources, such as the NDI. NHCS is not currently nationally representative due to low response rates, 158/581 = 27%. Still, linking NHCS with the NDI does allow for new analyses, such as studying mortality post hospital discharge, along with specific causes of death [13]. The linkage described here includes only patients with at least one inpatient or ED visit reported by hospitals participating in the 2016 NHCS. Less than one percent of NHCS records that were eligible for linkage are missing values for name, state of residence, sex, or date of birth. The completeness of SSN varied by age, with almost 71% of those 65 and older having an SSN (Table 3).

2.1.2. National Death Index (NDI)

The NDI is a centralized database of United States death record information on file in state vital statistics offices [12]. Working with these state offices, NCHS established the NDI as a resource to aid epidemiologists and other health and medical investigators with their mortality ascertainment activities. The NDI became operational in 1981 and includes death record information for persons who have died in the U.S. or a U.S. territory from 1979 onward. The records, which are compiled annually, include detailed information on the underlying and multiple causes of death. For this analysis the 2016/2017 NDI records were used.

2.2. Submission files

Preprocessing was performed to standardize the identification data in both files. For name values, the standardization steps included using full capitalization, removing punctuation, converting hyphens to spaces, removing name descriptors (e.g., mister, junior), and converting non-English letters (e.g., diacritics) into their English equivalent. Additionally, for names, alternate records were generated that included formal name equivalents of known nicknames and different versions of name parsing [14–16]. Detailed information on the PII standardization practices has been published elsewhere [16]. After data standardization, the data are referred to as "submission files."

2.3. Privacy preserving record linkage methodology

The privacy of identification data on files to be linked is protected through encryption by assigning hashes/tokens to groups of identifier fields before it is shared among organizations that are conducting the linkage. Linkage is conducted by comparing the hashes between files being linked and is similar to a deterministic link where tokens must match exactly to be considered a link.

Token Number	*Token Composition*
1	Last Name + First Name (1st letter) + Sex + DOB
2	Last Name (Soundex) + First Name (Soundex) + Sex + DOB
4	Last Name + First Name + Sex + DOB
5	**SSN + Sex + DOB**
7	Last Name + First Name (1st 3 characters) + Sex + DOB
16	**SSN + First Name**
40	First Name + Last Name + DOB + State

Token 40 (outlined) was custom coded, and HIPAA certified for use in this analysis; High precision tokens shown in **bold**.

Fig. 1. Selected Datavant tokens used in this analysis, all tokens are HIPAA certified.

2.4. Tokens

Rather than separately encrypting the individual identification fields, such as first name, last name, and DOB, these fields are concatenated into tokens to add privacy protections. For example, a single token is used to represent the concatenation of last name, first name, sex, and DOB. An example of a token developed using John Smith, sex = male, born March 27, 1968, could look like "iGy3RRqnKjO7cLUMF1z + er8SuR9F3WAmpqc8vsqCONQ =". Of note, for some tokens, the Soundex function is also used with names. Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English [17]. Datavant has developed a number of tokens for users to select from, including several tokens classified as high precision due to the inclusion of SSN as a unique identification number. Additional variables such as DOB are also included in these high-precision tokens to provide additional verification that in fact the two linked records belong to the same person. Although over 40 tokens were available, for this analysis only tokens for which the full complement of PII in the survey data were available were utilized. Figure 1 above illustrates the tokens used in this analysis which were all HIPAA certified by expert determination and deemed to have minimal risk of patient reidentification [8]. In addition, Datavant's security has been tested and certified by third parties [18].

The actual process of comparing two files, based on the encryption key, is performed outside of the Datavant software package. For this evaluation, the token files were read into SAS statistical software (version

9.4) [19] so that tokens could be compared and all pairs that shared at least one common token between the two files were output to a results file. Note that if any of the identifiers within a given token were missing, that token was not created. File sequence numbers were carried through the process to enable evaluation of link validity and comparisons to a gold standard.

Once all possible token comparisons were made, a de-duplication process was implemented to create a linked file at the patient level. For each NHCS-NDI record pair identified by PPRL analysis as having at least one token in common, the number of linked high-precision tokens was counted. For each NHCS patient record, the paired NDI record having the highest number of high-precision tokens was selected as the link. If there were more than one paired NDI record with the same number of high-precision tokens, then the record was assessed to see if other non-high-precision tokens matched and the record with the highest number of additional tokens was selected as the link for that patient record. The remainder of analysis was conducted on a patient-level based on the links selected in the de-duplication process.

2.5. Identifying the gold standard

A previous linkage of eligible patient records from the 2016 NHCS and 2016/2017 NDI records was used as the gold standard, henceforth referred to as the gold standard linkage. Previous studies have also used the results of a clear-text probabilistic linkage as a gold standard [20]. The gold standard linkage performs the linkage in two passes, the first pass relies on a deterministic approach and the second pass relies on a probabilistic approach [13]. Patient records were considered eligible for linkage if the record had two of the following: valid DOB (month, day, and year), name (first, middle, and last), and/or a valid format 9-digit SSN. The probabilistic approach performed weighting and link adjudication following the Fellegi-Sunter method [21]. The Fellegi-Sunter method is the foundational methodology used for record linkage. The probabilistic approach estimated the likelihood that each pair is a match before linking the most probable matches between a survey record and NDI record. Following this approach, a selection process was implemented with the goal of selecting pairs believed to represent the same individual between the data sources. In sum, the two passes are explained below:

Pass 1. Deterministic linkage which joins on exact SSN, which were validated by comparison of other identifying fields: when validation criteria were met, these records were linked and assigned a probability of being a valid match (match probability) of 1.00 [13].

Pass 2. Probabilistic linkage identified likely matches, or links, between all records, including those already matched in Pass 1. Records were linked and scored as follows (note: SSN is excluded from the analysis for this step):

a. Identified possible matched pairs
b. Scored potential match pairs using probabilistic weights – matches were scored based on the concurrence of these variables: First Name, Middle Initial, Last Name (or Father's Surname), State of Residence, Year of Birth, Month of Birth, Day of Birth, Date of Death (if available on hospital record), Sex
c. Probability modeling – estimate match probability.

Then for each patient record, the linked NDI record with the highest estimated match probability above the match threshold, which is based on minimizing type I (false positive) and type II (false negative) errors, was selected.

To assess the accuracy of the linked records from the two-step process noted above, subsequent record linkage analysis and error analysis was performed. There are two type of errors that were estimated:

- Type I Error: Among pairs that are linked, what percentage of them were not true matches
- Type II Error: Among true matches, how many were not linked

Detailed descriptions of these methods are described elsewhere [13]. The Type I and Type II errors for the gold standard were, 0.2% and 1.1%, respectively [13].

2.6. Evaluating the PPRL links

2.6.1. Initial PPRL

First, we compared the results of the de-duplicated matches to the results of the gold standard linkage. We calculated the number of true positives (TP, correctly linked records), true negatives (TN, correctly unlinked records), false positives (FP, incorrectly linked records), and false negatives (FN, incorrectly unlinked records). Next, we assessed the precision and recall. Precision was calculated as TP/(TP + FP). Recall was calculated as TP/(TP + FN). Cohen's Kappa statistic was used to measure agreement between the two approaches. The standard range of the Kappa statistic is 0 for no agree-

ment and 1 for complete agreement, albeit values from −1 to 0 are possible and would indicate negative correlation. Landis and Koch suggest the following interpretation for the Kappa statistic: < 0.00: Poor; 0.00–0.20: Slight; 0.21–0.40: Fair; 0.41–0.60: Moderate; 0.61–0.80: Substantial; 0.81–1.00: Almost Perfect [22]. The Kappa statistic was used to account for agreement by chance.

2.6.2. *Refined PPRL*

Because certain combinations of tokens generated FP, we assessed if there was a way to minimize type I and type II errors by selecting tokens with low false positive rates. The false positive rate was calculated as the number of FPs (an incorrectly identified link) as a percentage of FPs and TPs. If the false positive rate was greater than 50%, we removed the records generated by the token combination from the returned links. The use of only these links is referred to as the refined approach. The precision and recall were then recalculated once the token combinations with false positive rates greater than 50% were removed.

2.6.3. *Secondary data analysis*

Lastly, to assess the impact of linkage results on secondary data analysis, match rates for the three approaches were assessed (gold standard linkage, initial PPRL assessment, refined PPRL assessment). In addition, 30-, 60- and 90-day post hospital discharge mortality rates were calculated and compared based on the gold standard, the PPRL linkage using all token combinations, and the refined PPRL linkage using only the tokens that had a FP rate < 50%. Death rates were calculated based on the 30, 60, and 90 days from the last known discharge date for each patient.

3. Results

The NHCS linkage submission file included 5,386,469 records of which 1,205,063 were alternate records (i.e., records that were generated to include formal name equivalents of known nicknames and different versions of name parsing). The NDI file included 6,373,038 records, of which 762,101 were alternate records. The gold standard linkage resulted in 212,922 unique links between patient record and the NDI. The de-duplicated (e.g., one record per patient as described above) PPRL resulted in 223,929 unique NHCS-NDI links – a 5.2% increase over the gold standard linkage. A patient that linked to the NDI is described as assumed deceased and those that did not link are assumed to be alive.

Largely the determinations of vital status have high concordance across these two linkages. The overall concordance was 99.6% for linkage-eligible patient records. The kappa statistic was 0.96, suggesting almost perfect agreement. The precision of the initial approach was 93.8% and the recall was 98.7%.

An examination of the links from the gold standard that were not returned by PPRL shows that the majority of these agreed on some part of the name (first or last), date of birth, state of residence and sex in the gold standard (data not shown). Differences between the two approaches may be due to how name fields were compared. In the gold standard, text strings were compared by modification to the Jaro-Winkler string comparator function [23]. By contrast, in the Datavant PPRL, since text strings are encrypted string comparator functions could not be used to compare names and the PPRL algorithm relied on encrypted tokens based on the exact name or indexed strings based on Soundex.

We next looked at the token combinations to assess which ones produced both FPs and TPs. A false positive rate was also calculated. Table 1 illustrates the 29 unique token combinations and the number of TPs and FPs associated with each combination. In the initial PPRL assessment, close to 60% of the TP links (n = 125,182) resulted from two high precision tokens that relied on SSN (i.e., Tokens 5 and 16). Close to 80% (n = 12,566) of the FP links were identified from links found by Token 1 or Token 2 only, which are based on agreement of first or last name, sex, and date of birth (see Fig. 1 for token definitions).

We note that links identified through Token 1, Token 2, the combination of Tokens 1 and 7, and the combination of Tokens 1 and 2 all have false positive rates > 50%. To assess the impact of removing the token combinations with high FP rates, we removed all links identified through these specific token combinations and assessed the new result set referred to as the refined PPRL links.

Table 2 summarizes the results for the initial and refined PPRL approaches.

Largely the determinations of vital status remained having a high concordance across these two linkages. The overall concordance was 99.8% for linkage-eligible patient records. The kappa statistic was slightly higher than the initial assessment at 0.98. The precision for the refined approach was 98.9% and the recall was 97.8%.

There were more FNs in the refined approach and less FPs compared to the initial PPRL assessment. An

Table 1
Token patterns and the number of true positives, false positives and false positive rate associated with each pattern based on the initial PPRL

	High precision tokens		Non-high precision tokens					True positive	False positive	False positive
Row	Token 5	Token 16	Token 1	Token 2	Token 4	Token 7	Token 40	TP count	FP count	FP rate
1	0	0	0	0	0	0	1	57	***	***
2	0	1	0	0	0	0	0	2,705	56	2.0
3	0	1	0	0	0	0	1	77	0	0.0
4	1	0	0	0	0	0	0	2,151	6	0.2
5	1	1	0	0	0	0	0	924	0	0.0
6	0	0	0	0	1	0	0	***	0	***
7	1	1	0	0	1	0	0	***	0	***
8	0	0	0	1	0	0	0	710	3,111	81.4
9	1	0	0	1	0	0	0	10	0	0
10	1	1	0	1	0	0	0	521	0	0
11	0	0	1	0	0	0	0	333	7,473	95.7
12	0	0	1	0	0	1	0	537	626	53.8
13	0	0	1	0	0	1	1	***	0	***
14	1	0	1	0	0	0	0	530	***	***
15	1	1	1	0	0	0	0	***	0	***
16	1	0	1	0	0	1	0	720	0	0.0
17	1	0	1	0	0	1	1	***	0	***
18	1	1	1	0	0	1	0	***	0	***
19	0	0	1	1	0	0	0	158	305	65.8
20	0	0	1	1	0	1	0	991	230	18.8
21	1	0	1	1	0	0	0	128	***	***
22	1	1	1	1	0	0	0	***	***	***
23	1	0	1	1	0	1	0	1,006	0	0.00
24	1	0	1	1	0	1	1	***	0	***
25	1	1	1	1	0	1	0	8	0	0.00
26	0	0	1	1	1	1	0	2,314	1,638	41.4
27	0	0	1	1	1	1	1	79,757	315	0.3
28	1	1	1	1	1	1	0	2,916	***	***
29	1	1	1	1	1	1	1	113,486	114	0.1
								210,059	**13,880**	**6.6**

***Suppressed due to count of 5 or less. Highlighted rows indicate a FP rate greater than 50%; High precision tokens rely on SSN; Token 5 = SSN + Sex + DOB; Token 16 = SSN + First Name; Non-high precision tokens rely on other identifiers and not SSN (see Fig. 1).

Table 2
Results of initial and refined PPRL approaches compared to gold standard

PPRL approach	Assumed deceased in gold standard		% Agreement	Precision	Recall
Initial: Assumed Deceased	No	Yes	*99.6*	*93.8*	*98.7*
No	3,954,604	2,863			
Yes	13,880	210,059*			
Refined: Assumed Deceased	No	Yes	*99.8*	*98.9*	*97.8*
No	3,966,119	4,601			
Yes	2,365	208,321			

*Among the 210,059 true positive links, there are 18 cases where the patient has been determined to have died in both linkages and yet the patient record was linked to a different death certificate in PPRL compared to the standard linkage.

examination of the links from the gold standard that were not returned by refined PPRL shows similar results to the full PPRL result set. Most FNs had agreed on some part of the name (first or last), date of birth, state of residence and sex in the gold standard (data not shown).

3.1. Impact on secondary data analysis

A final assessment looked at the impact of initial and refined PPRL links on secondary data analysis. The match rates by age and sex are presented below for the three approaches: gold standard, initial PPRL and

Table 3
2016 NHCS patients linked to 2016–2017 NDI data by linkage approach and demographics

Characteristic	Total sample	Eligible for linkage[3]		Linked to NDI by gold standard			Linked to NDI by initial PPRL		Linked to NDI by refined PPRL	
		n	% of total	*n*	% of eligible	% of eligible linked with non-missing SSN	*n*	% of eligible	*n*	% of eligible
Total	5,823,165	4,181,406	71.8	212,922	5.1	59.8	223,939	5.4	210,686	5.0
Age[1]										
< 18	1,292,544	1,209,207	93.6	3,537	0.3	10.0	4,100	0.3	3,303	0.3
18–44	1,477,352	1,386,614	93.9	15,086	1.1	29.0	16,347	1.2	14,576	1.1
45–64	920,862	865,729	94.0	51,045	5.9	41.9	54,443	6.3	49,881	5.8
65+	757,631	712,646	94.1	142,174	20.0	70.8	148,254	20.8	142,153	19.9
Missing	1,374,776	7,210	0.5	1,080	15.0	53.5	795	11.0	791	11.0
Sex[2]										
Male	2,600,185	1,853,765	71.3	108,706	5.9	61.9	114,360	6.2	107,544	5.8
Female	3,160,111	2,280,737	72.2	102,216	4.5	56.8	106,497	4.7	100,351	4.4
Missing	62,869	46,904	74.6	2,000	4.3	97.2	3,082	6.6	2,791	6.0

Source: 2016 NHCS – 2016/2017 NDI linked data file. Note: NHCS is the National Hospital Care Survey; NDI is the National Death Index. Data are presented at patient level. [1]Age was calculated by subtracting date of birth from date of first encounter from patient's first encounter record. [2]Sex is based on the reported sex on patient's first encounter record. [3]Eligibility for linkage is based upon having sufficient personally identifiable information in at least two of three data element groups: SSN, name, and date of birth.

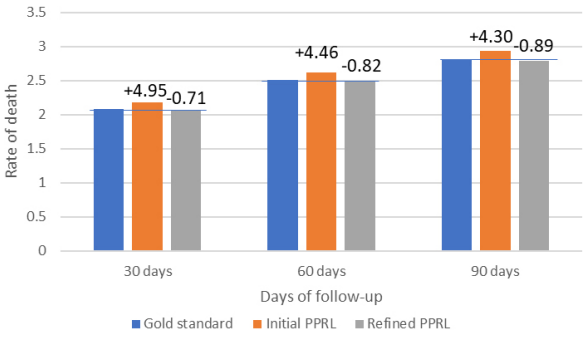

Fig. 2. Patient death rates by linkage approach and follow-up period and percent difference from gold standard.

refined PPRL (Table 3). In addition, the rate of SSN agreement is presented for the gold standard.

The age and sex distribution of eligible linked survey participants appears similar across the three linkage approaches, although a higher number of links occur for those records with missing sex in both PPRL approaches (6.6% and 6.0% compared to 4.3% in the gold standard). On the other hand, there were a lower number of links found for records with missing age in PPRL (Table 3). Linkage rates are highest among adults aged 65 and over, with approximately 20% of adults aged 65 and over matching to the NDI for all three approaches. It is also important to note that this group (aged 65 and over) had the lowest percentage of missing SSN, 70.8% of eligible patients 65 and over had a non-missing SSN.

In addition, we performed a secondary analysis of death rates based on the results of the three linkage approaches: gold standard, initial PPRL, and refined PPRL links. Figure 2 shows a comparison of overall death rates during several follow-up periods: 30, 60, and 90 days from the last known discharge date for each patient.

Figure 2 shows that the initial PPRL approach leads to death rates being overestimated by 4.3% in the 90 day follow up period (and somewhat higher for shorter follow-up periods). By contrast, the refined PPRL links lead to a slight under-estimation of mortality (by less than 1%) for all follow up periods.

4. Discussion

PPRL is a technique that allows organizations to conduct linkages without sharing direct PII. As a federal statistical agency, it was important to assess data quality prior to implementing PPRL techniques to create data for use in official statistics. It should be noted that while we performed an assessment of data quality, we did not conduct a separate privacy analysis of the encoded databases. Prior to a full implementation of PPRL at NCHS, a privacy risk-reidentification analysis would need to be conducted as well [24–26]. This would enable us to identify any potential known risks and assess if there are any mitigation strategies that could be implemented to maintain the protection of privacy of the NCHS survey participants as required under Title III of Foundations for Evidence-Based Policymaking Act of 2018, Public Law 115–435 (the Confidential In-

formation Protection and Statistical Efficiency Act or CIPSEA).

Nevertheless, the data quality results of this case study are encouraging. We examined the efficacy of PPRL compared to a gold standard by re-linking previously linked NHCS and NDI data using privacy protected, encrypted PII values (hashes/tokens) created in Datavant software rather than the actual data values. The standard NCHS linkage methodologies are based on deterministic and probabilistic techniques, using clear text matching. This research demonstrates that PPRL approaches, particularly the refined PPRL approach, produce results similar to a gold standard linkage (kappa statistics suggesting almost perfect concordance). The refined approach highlights the importance of the selection of tokens. When we removed the tokens that relied solely on first or last name, sex, and date of birth the concordance with the gold standard increased.

In terms of secondary data analysis of the match rates by age and sex, all three approaches produced similar results. Having a unique identifier such as SSN increases the likelihood that the records will match in all three approaches as noted in Table 3. In addition, in this analysis the percent with non-missing SSN is highest in the 65 and over age group which is also the age group most likely to link to the NDI.

This analysis quantifies how data linkage quality may impact inferences in secondary data analysis, which underscores the role of continued methodologic work to improve data linkage quality. The aggregated death rates computed from PPRL are less than 5% different than those from the gold standard and after refinement based on false positives, this falls to less than 1%. Next steps could include comparing the three approaches to assess death rates for different subpopulations and distributions by cause of death.

In addition, while this initial work laid the much-needed groundwork for assessing the accuracy of PPRL-created linked data when compared to traditional linkage methods, additional evaluation is needed to assess the quality of PPRL- based linkages when using data with lower quality PII. This initial analysis used two sources with high quality PII. Given the varying quality of PII within the wide range of data needed to conduct health related research it is important to expand the assessment of PPRL tools to include data sources with varying PII quality and completeness. While work to evaluate PPRL techniques is being conducted in the extramural research community, NCHS is in a unique position as the Department of Health and Human Services (HHS) federal statistical agency to assist with evaluating different PPRL strategies to potentially broaden linkage capabilities within HHS. It is of utmost importance to assess the quality of linkage processes and results when integrating data sources, using PPRL, so that researchers can determine whether the resulting data sets are suitable for inference and generalization to other populations.

In real-life scenarios researchers would not necessarily have a gold standard to be used for comparison or another way to determine linkage quality, therefore, these results could be used as a guide for the selection of tokens to reduce Type I errors. The generalizability of these findings for survey-based self-reported identification data compared to hospital patient records is still unclear and additional evaluations assessing PPRL techniques, including other PPRL methods, that rely on tokenization or hashing with lower quality PII is forthcoming. Future work will build on this analysis by testing a variety of scenarios, including PII that is non-standardized, incomplete (e.g., missing unique identification numbers such as Social Security Number (SSN)), and of varying levels of quality (e.g., informal names provided in survey collections) using previously linked NCHS data files as the gold standard to benchmark against PPRL-based results.

5. Conclusion

This research demonstrates that PPRL can be an effective record linkage technique that produces results similar to a gold standard. However, it is important to note that these results were obtained using sources that had a high rate of complete identifiers, including SSN.

6. Significance

This novel evaluation highlights the first time NCHS, a federal statistical agency, has utilized PPRL and compared it to an already released linkage that relied on clear text matching. The assessment builds on many of the guiding principles outlined in the Federal Committee on Statistical Methodology report “A Framework for Data Quality” [27]. The approach focused on assessments of threats, like accuracy and coherence, when utilizing new software to perform linkage and illustrates a way (e.g., the selection of refined tokens) to maintain scientific integrity and credibility while adhering to the protection of privacy. This is an important first step toward advancing the potential use of PPRL to integrate data without sharing direct identifiers.

Funding

This research was funded in part by the Department of Health and Human Services' Office of the Secretary Patient Centered Outcomes Research Trust Fund (OSPCORTF).

Conflict of interest

The authors have no competing interests to declare.

References

[1] National Center for Health Statistics. NCHS Data Linkage Activities 2021 [updated 7/8/2021]. Available from: https://www.cdc.gov/nchs/data-linkage/index.htm.
[2] National Center for Health Statistics. Linked Mortality Files Citation List 2021 [updated 9/1/2021]. Available from: https://www.cdc.gov/nchs/data/datalinkage/Linked-Mortality-Files-Citation-List-20210901-508.pdf.
[3] Confidential Information Protection and Statistical Efficiency Act of 2002, 107th US Congress. 2002.
[4] Health Insurance Portability and Accountability Act of 1996, 104th US Congress. 1996.
[5] HIPAA Privacy Rule. 2003.
[6] Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Information Systems. 2013; 38(6): 946–69.
[7] Office for Civil Rights – US Department of Health and Human Services. Guidance Regarding Methods for De-identifying of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule 2012 [updated 11/26/2021]. Available from: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.
[8] Datavant Inc. Overview of Datavant's De-Identification and Linking Technology for Structured Data [Available from: https://datavant.com/wp-content/uploads/dlm_uploads/2018/09/WhitePaper_-De-Identifying-and-Linking-Structured-Data.pdf].
[9] Social Security Administration, Office of Retirement and Disability Policy. The Story of the Social Security Number. Social Security Bulletin, Vol 69, No 2, 2009.
[10] Nguyen L, Stoové M, Boyle D, Callander D, McManus H, Asselin J, et al. Privacy-preserving record linkage of deidentified records within a public health surveillance system: Evaluation study. J Med Internet Res. 2020; 22(6): e16757-e.
[11] Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. Journal of Biomedical Informatics. 2014; 50: 205–12.
[12] National Center for Health Statistics. About the National Death Index 2021 [updated February 5, 2021]. Available from: https://www.cdc.gov/nchs/ndi/about.htm.
[13] National Center for Health Statistics. The Linkage of the 2016 National Hospital Care Survey to the 2016/2017 National Death Index: Methodology Overview and Analytic Considerations. Hyattsville, MD; 2019.
[14] Lloyd PC, Helms VE, Simon AE, Golden C, Brittain J, Call E, et al. Linkage of 1999–2012 National Health Interview Survey and National Health and Nutrition Examination Survey Data to U.S. Department of Housing and Urban Development Administrative Records. Vital and Health Statistics Ser 1, Programs and Collection Procedures. 2017(60): 1–40.
[15] National Center for Health Statistics. The Linkage of National Center for Health Statistics Survey Data to the National Death Index – 2015 Linked Mortality File (LMF): Methodology Overview and Analytic Considerations, March 2019. In: Centers for Disease Control and Prevention, editor. 2019.
[16] Sayers J, Campbell S, Thompson C, Jackson G, eds. Data Linkage with an Establishment Survey. Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference; 2018; Washington DC.
[17] The U.S. National Archives and Records Administration. Soundex System 2020 [May 30, 2007]. Available from: https://www.archives.gov/research/census/soundex.
[18] Datavant Security Overview [Available from: https://datavant.com/wp-content/uploads/2021/09/Datavant-Security-Overview.pdf].
[19] SAS Institute Inc. 2016. SAS/CONNECT® 9.4 User's Guide, Fourth Edition. Cary, NC: SAS Institute Inc.
[20] 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(1): 83.
[21] Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. 1969; 64(328): 1183–210.
[22] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1): 159–74.
[23] Resnick D, Mirel L, Roemer M, eds. Adjusting Match Weights to Partial Levels of String Agreement in Data Linkage. Joint Statistical Meetings, American Statistical Association; 2020; Alexandria, VA.
[24] Gkoulalas-Divanis A, Vatsalan D, Karapiperis D, Kantarcioglu M. Modern privacy-preserving record linkage techniques: An overview. IEEE Transactions on Information Forensics and Security. 2021; 16: 4966–87.
[25] Randall S, Brown A, Ferrante A, Boyd J. Privacy preserving linkage using multiple dynamic match keys. International Journal of Population Data Science. 2019; 4.
[26] Vidanage A, Ranbaduge T, Christen P, Randall S. A Privacy Attack on Multiple Dynamic Match-key based Privacy-Preserving Record Linkage. International Journal of Population Data Science. 2020; 5.
[27] Federal Committee on Statistical Methodology. A Framework for Data Quality. In: National Center for Educational Standards, editor. 2020.