The Impact of Name Transformation on Match Rates Within a Large Consumer Database

Jonah Leshin, PhD[1], Arjun Sanghvi, MS[1], Kavi Ravuri, BS[1], Matthew Owen, BS[1], Abel Kho, MD, MS, FACMI[2]

[1]Datavant, San Francisco, CA; [2]Northwestern University, Chicago, IL

Abstract

Accurate record linkage depends on the availability and quality of features such as first name and last name. Privacy preserving record linkage methods using tokenization is sensitive to perturbations in the patient features used as inputs. In this study we evaluated the impact of name transformations on the accuracy of patient matching using a large commercial dataset. We used a set of 68 million records representing 59 million unique individuals, and implemented and evaluated eight name transformation strategies, and generated precision, recall and F1 scores. Transforming names to include the most common nicknames resulted in a significant gain in recall while maintaining precision, and generated the highest F1 score compared with no name transformation (0.905 vs 0.807). Strategies tailored to transforming patient features can improve the precision and recall of patient matching, and make it possible to create high quality, linked datasets for research purposes.

Introduction

Patient data relevant for clinical care, research, and public health is commonly fragmented across multiple sources.[1,2,3] Accurate linking of data on the same patient is critical to capture the most complete picture of the health and health care for any given individual. The most common way to connect fragmented data has been to use patient demographic information, such as first name, last name, date of birth, and other identifiers. Accordingly, accurate patient matching across these diverse settings is dependent on the availability, completeness, and quality of these identifiers.[4] In prior studies, standardization of specific patient features, such as address and last name, improved match rates significantly.[5,6]

Using healthcare data for research purposes presents additional challenges. Absent consent or the use of the data for QA/QI purposes requires that the data is de-identified before it is used for research. Within an institution, this can be done with an honest broker who can first link, then de-identify the data. However, across multiple institutions, this approach can be difficult. To be able to link data across different organizations without the use of an honest broker requires different techniques. In some settings, fragmented data are being linked using tokens, or alphanumeric identifiers generated cryptographically from patient demographic information.[1,7] Record linkage with tokenization preserves patient privacy and has gained favor as an accurate and effective means to link fragmented records and has been used to enable research across a wide variety of settings and health conditions.[8,9,10,11,12,13] However, given the exquisite sensitivity of tokenization to perturbations in input data, attention to data preparation, standardization and transformation may be as or more critical as with linking of un-encrypted (clear text) patient data.

Because first and last names are prominent identity features affecting match quality we set out to evaluate the impact of name transformations on the accuracy of patient matching using a large commercial dataset. Here we characterize the relative effect of name transformation for matching across a variety of use cases, and show the resulting accuracy tradeoffs with the goal of enabling practitioners to make evidence-based decisions when transforming names for the purpose of patient matching.

Objectives

The objective of this study is to evaluate the impact of a series of name transformations on patient matching accuracy. The transformations are applied to plain text fields, but can subsequently be anonymized to support the exchange of HIPAA de-identified data. In our analysis, we applied Datavant's tokenization software to the transformed raw text, which uses a combination of hashing and encryption to convert plaintext to a Base64 encoded character string of length 44 (a "token"). In practice, healthcare data sources use this tokenization process to link disparate sources of data.

We used a large dataset of 680 million person-level records, last updated in 2020, with birth years spanning from 1924 to 2002. Data was sourced from a large consumer data company. Each record contains raw demographic information and a gold standard ID (hereon denoted by "ID") that is derived from both transaction level data and household level information, and is consistent across distinct records belonging to the same individual. For computational feasibility, we worked with a random 10% sample of records– 68 million records, which consisted of 59 million unique individuals.

From these 68 million records, we generated an evaluation dataset consisting of all record pairs that could plausibly be considered matches based on shared demographic information.

We evaluated a series of name transformations, all aimed at consolidating different variants of the same first or last name; for example, one transformation type consisted of normalizing nickname variants to a single canonical form.

Each name transformation T results in new versions T(first name), T(last name) of a record's first and last name, respectively. For each record R in the dataset, we can then form the record-level identifier associated with this transformation given by the concatenation

$$T(R) := T(\text{first name}) + T(\text{last name}) + \text{gender} + \text{date of birth}$$

To obtain a final form of the identifier for analysis, we tokenized the transformed raw text T(R) using Datavant's tokenization software. These tokens are HIPAA certified, enabling us to work in a de-identified data environment for analysis, and more broadly allowing them to be shared if generated from data owned by a covered entity. From hereon, we use T(R) to refer to this tokenized, transformed identifier.

We incorporated date of birth and gender into the definition of T(R) to reduce name collisions. Date of birth is permanently static, and is therefore a valuable field to include in a person-level identifier across records over time. While gender has historically been considered a binary construct (in our dataset, all non-null gender values were Male or Female), recent societal shifts have created greater diversity in how this field may be captured, although data standards lag behind in reflecting this.

For each pair of records (R1, R2) with associated IDs (ID(R1), ID(R2)) coming from our dataset we can then compare whether there is agreement on (ID(R1), ID(R2)) and on (T(R1), T(R2)). From here, we can quantify the transformation's impact on matching accuracy.

Methods

Data Preparation

All name transformations and subsequent analyses were processed on Amazon Web Services (AWS) cloud computing machines, and Snowflake's (Snowflake Inc., Bozeman MT) data warehouse. Our data pipeline consisted of the following steps:

1) Form a dataset with one row for each record R in the original record level dataset by creating one column for the gold standard ID corresponding to R, and one column for each transformation T that contains the value T(R).

   Table 1a provides an illustrative example with three transformations for each record, in addition to the Record ID column that increments with each row and the gold standard person-level ID for that record.

2) Generate a dataset with one row for each record pair that agrees on either the gold standard ID or on at least one qualifying combination of demographic information. Qualifying combinations consisted of various concatenations of (possibly transformed) PII fields, including first and last name, gender, date of birth, address, zip code, city, state, email address, and cell phone number. The qualifying combinations included all instances of the transformations we evaluated.

   Table 1b provides an illustrative example of a single row of this dataset. The table is comparing Record ID's 1 and 2 from Table 1a, and indicating that the two records have the same values for T2 and T3, and different values for ID and T1.

   These record pairs serve as candidate matches to be evaluated, with the idea being that any matching pair of records should agree on at least one qualifying combination of PII fields, and that evaluating all possible pairs in a dataset with 68 million records is not computationally feasible.

Table 1a. Illustrative example of record level dataset.

Record ID	ID	T1	T2	T3
1	A	x	y	z
2	B	w	y	z

Table 1b. Illustrative example of pairwise level dataset.

Record ID 1	Record ID 2	ID agreement	T1 agreement	T2 agreement	T3 agreement
1	2	False	False	True	True

Description of Name Transformations

We analyzed a variety of transformations in order to consolidate distinct first and last names in our dataset that belong to records corresponding to the same individual. Prior to any of these transformations (including the case of no transformation at all), we performed data cleaning on first and last name consisting of removing leading and trailing white space, uppercasing all characters, and removing any non-ascii characters.

Nicknames

In this section only, we write "name" for first name.

We used our full 680 million record dataset to generate nickname resolutions at varying levels of confidence, which could then be applied to our evaluation subset. Here, a resolution consists of groupings of names so that all names within a group are treated as equivalent.

Each resolution stemmed from an initial clustering of names obtained by grouping together names with the same ID. Confidence was then quantified within each cluster by evaluating *Relative Frequencies* and *Pair Counts* for each pair of names in the cluster. We then implemented minimum thresholds for the Relative Frequency and Pair Count values in order to remove names from a cluster that were less likely to be valid name variants of the other names within the cluster.

To provide a concrete example of the Relative Frequency and Pair Count criteria, consider the names *Bob* and *Robert*. Across all clusters containing a record with the name Robert, if 60% of the records in these clusters that do not have the name "Robert" have the name "Bob," then we would say the pair (Robert, Bob) has a Relative Frequency of .6. With regard to Pair Count, if our dataset has 10 individuals for which there are records with the first name "Bob" and for which there are also records with the first name "Robert", then the Pair Count for the pair (Robert, Bob) would be 10.

Conceptually, the pair (Robert, Bob) has a high Relative Frequency if Bob is frequently a nickname for Robert *relative* to other nicknames, and (Robert, Bob) has a high Pair Count if Bob frequently appears as a nickname for Robert by *total count*.

We can obtain final clusterings at various minimum levels of Relative Frequency and Pair Count values by filtering out all pairs below the minimum input levels provided. Any name that is filtered out is treated as its own cluster of size 1, and so this process associates each name in our dataset with a cluster. In this way, we can refer to these cluster assignments via a pair *(<Relative Frequency>,<Pair Count>);* for example, the pair *(.5, 10)* refers to the clustering obtained by taking the original clustering, and removing all pairs of names with a Relative Frequency below .5 and a Pair Count below 10.

Such a clustering corresponds to a name transformation T for which T(X) is the same for any first name X within the same cluster. Smaller values of Relative Frequency and Pair Count correspond to larger name clusters, and the corresponding transformation is more likely to consider distinct records with loosely associated names as belonging to the same individual.

Truncation

We evaluated two methods of truncation: We truncated first name down to the first letter, and down to the first three letters. The value of truncation is the ability to associate name variants belonging to the same individual that are due to nicknames, name changes from name extensions (e.g. “Jean” versus “Jean-Pierre”), or typographical errors.

Phonetic Algorithms

We applied two phonetic algorithms to first and last names: Metaphone[14] and Soundex.[15] Each of these aims to reduce a name to its phonetic components; for example, eliminating the distinction between “Chris” and “Kris”.

One-Character Corrections based on names available in Census

We used US Census data as a set of reference first and last names that we could use to identify first and last names in our dataset that may have a typographical error. For first names, we used Census data from 1930-2010, and for last names we used 2010 only. The process worked as follows:

Given a first or last name, if the name was already in the Census, it was left unchanged. If the name was not in the Census, then we looked for all names in the Census with a Levenshtein edit distance of 1 from the name in question. If there were no such names in the Census, the original name was left unchanged. If there was at least one such name in the Census, then among all such names, we chose the one that most frequently occurred in the Census, and replaced the original name with that name.

For example, if the first name “Joshhua” appears in our data, we first check if this name is in the Census data. If so, it is left unchanged. If not, then we search the Census data for all first names that are an edit distance of 1 away from “Joshhua”. If we do find at least one, we select the one with the most frequent occurrence in the Census, in this case “Joshua”. If we do not find any, we leave the name “Joshhua” unchanged.

The primary goal of this transformation is to reduce the quantity of records that are incorrectly unmatched due to typographical errors.

Combined Transformations

We also evaluated several of these transformations in combination, for example, applying a nickname mapping, followed by metaphone.

Evaluation

For each transformation type, across a set of record pairs, we compared whether the transformation T agreed and whether the gold standard (ID) agreed, and used this data to compute true positives (TP), false positives (FP), and false negatives (FN), and from there precision, recall, and F1, as defined below for a given pair of records (R1, R2):

TP : T(R1) = T(R2) and ID(R1) = ID(R2)

FP: T(R1) = T(R2) and ID(R1) != ID(R2)

FN: T(R1) != T(R2) and ID(R1) = ID(R2)

Precision: TP/(TP+FP)

Recall: TP/(TP+FN)

F1: 2*Precision*Recall / (Precision + Recall)

For all metrics, we only considered record pairs for which both records had a valid ID and contained enough information for the transformation T to be computed.

Results

In total we conducted over 300 record transformations involving names. Here we provide results for a selection that encompasses a diversity of type and performance (for a full set of transformation results please contact the primary author). Table 2a defines the transformations we evaluated, and Table 2b provides the metrics associated with each Transformation ID. In Table 2a, transformations are given an ID and defined as combinations of the operations in the remaining columns. Nickname values refer to *(Relative Frequency, Pair Count)* used for the nickname resolution; Truncation values are number of characters after which name was truncated; Phonetic values are MPH

for Metaphone and SDX for Soundex; One Character Correction is the name type (first or last) for which a one-character correction was made using Census data.

The data from Table 2b is plotted in Figure 1.

Table 2a. Name transformation IDs and definitions.

Transformation ID	Nickname	Truncation	Phonetic	One Character Correction
1	-	-	-	-
2	(.5, 10)	-	-	-
3	(.8, 10)	-	-	-
4	(.3, 20)	-	MPH	-
5	(.5, 10)	-	MPH	-
6	(.5, 10)	-	SDX	-
7	-	-	-	First name
8	(.5, 10)	3	-	-
9	(.5, 10)	1	-	-

Table 2b. Accuracy metrics by transformation ID.

Transformation ID	TP	FN	FP	Precision	Recall	F1
1	4,955,176	2,080,346	282,186	0.946	0.704	0.807
2	6,065,263	970,259	423,692	0.935	0.862	0.897
3	5,635,459	1,400,063	326,535	0.945	0.801	0.867
4	6,264,829	770,657	549,711	0.919	0.890	0.905
5	6,176,965	858,521	501,395	0.925	0.878	0.901
6	6,185,077	850,445	521,319	0.922	0.879	0.900
7	5,020,367	2,244,933	305,086	0.943	0.691	0.797
8	6,270,264	765,258	590,272	0.914	0.891	0.902
9	6,329,932	705,590	1,321,415	0.827	0.900	0.862

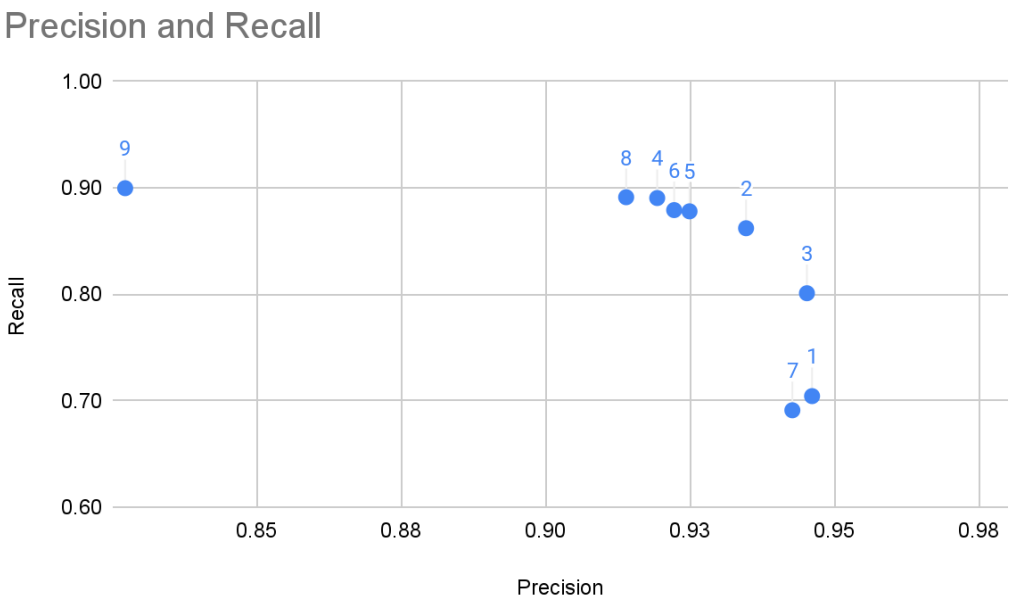

Figure 1. Plot of precision and recall values for the transformations in Table 2b. Data points are labeled by Transformation ID.

Discussion

As expected, the highest level of precision is achieved without any name consolidation (Tr 1); however, one gains significant improvement in recall with minimal cost to precision by employing the most conservative nickname mapping (Tr 3) in the sense that the mapping only includes name associations that are strongest relative to other similar associations. The best overall performing transformations, as measured by F1 score, all involved the (.3, 20) nickname mappings, of which we have shown one example with Tr 4. This illustrates the increased value of relatively aggressive nickname consolidation, as these nickname mappings have a lower bar for the relative strength of a name association.

The negative impact of first name truncation on recall stands out, with minimal performance gain in recall to compensate, suggesting that truncating to 3 characters is more effective at balancing precision and recall than taking only the first initial of first name.

The single character corrections referencing known census names had minimal impact, and in fact when applied to first name (Tr 7), the resulting performance was weaker in both precision and recall. The minimal impact is not entirely surprising since we found only 3.6% of all records were changed with this first name correction mechanism. The negative impact on recall comes from the fact that the number of true positives gained was significantly fewer than the number of false negatives incurred.

Study Limitations and Future Work - Due to variation in naming conventions across ethnicities, conducting separate name transformation studies for different ethnic groups would shed light on the bias that particular transformations may have on certain name types. In particular, it would be interesting to consider groups where it is common to have two first names or two last names. This type of analysis would provide guidance for practitioners who are able to tailor their matching algorithms to their specific population.

In this study, we used gender and date of birth as the demographic fields to append to first and last name to form the person-level identifiers used to evaluate name transformations. Particularly with respect to gender, future research will need to consider the shifting nature of demographic data capture and standards. Additionally, future work may involve experimenting with other choices of demographic fields, or using transformed names as input into a predictive algorithm that determines whether a pair of records match.

Conclusion

The problem of data fragmentation and the growing opportunity presented by novel data sources relevant to health will require mechanisms to accurately and securely link data. Privacy preserving linkage techniques can help with linking fragmented data across different data sets while maintaining the privacy of the patient records, and obviating the need for identifiers to be shared. These techniques however, require careful name transformation to optimize the precision and recall of matching using PPRL techniques. Our paper evaluates different name transformations and identifies those that are the most effective in improving precision and recall.

Nickname consolidation proved to be an impactful transformation for improving recall with minimal cost to precision; however, such consolidation requires the curation of a mapping at the desired confidence level. Self-contained transformations involving phonetic algorithms or truncation may be better suited when resources are limited. The transformations in this study can be used as HIPAA certified identifiers for patient matching, and can also form the basis of more complex matching algorithms that use these identifiers as inputs to machine learning models.[16]

For optimal matching it is critical to have comprehensive, longitudinal patient data and to understand the population it represents, and the quality and completeness of these data. This is particularly important for the use of real world data for clinical investigation which can be of varying quality but may be enhanced by linking different data sets together (such as claims data and EHR data) to provide a more comprehensive view of a patient and their outcomes. We believe that improvements in name transformations will improve the precision and recall of patient linking, and make it possible to create high quality, linked datasets for research purposes.

Growth in the volume and diversity of healthcare data has enabled new research opportunities requiring record linkage. Optimizing match performance is both art and science and requires attention to shifting demography and cultural trends and successful future strategies will need to continuously improve and adapt to both the available data and shifts in how patients self-identify.

References

1. Kho AN, Yu J, Bryan MS, et al. Privacy-preserving record linkage to identify fragmented electronic medical records in the all of us research program. In: Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing; 2020. p. 79–87.
2. Mays JA, Jackson KL, Derby TA, et al. An evaluation of recurrent diabetic ketoacidosis, fragmentation of care, and mortality across Chicago, Illinois. Diabetes Care. 2016;39(10):1671–6.
3. Walunas TL, Jackson KL, Chung AH, et al. Disease outcomes and care fragmentation among patients with systemic lupus erythematosus. Arthritis Care Res (Hoboken). 2017;69(9):1369–76.
4. Culbertson A, Goel S, Madden M, et al. The building blocks of inter-operability: A multisite analysis of patient demographic attributes available for matching. Appl Clin Inform. 2017;08(02):322–36.
5. Grannis SJ, Xu H, Vest JR, et al. Evaluating the effect of data standardization and validation on patient matching accuracy. J Am Med Inform Assoc. 2019;26(5):447–56.
6. Churches T, Christen P, Lim K, Zhu JX. Preparation of name and address data for record linkage using hidden Markov models. BMC Med Inform Decis Mak. 2002;2(1):9.
7. N3C linkage honest broker [Internet]. Regenstrief Institute. 2021 [cited 2022 Mar 8]. Available from: https://www.regenstrief.org/n3c-lhb/
8. 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(4):562–9.
9. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc. 2015;22(5):1072–80.
10. Trick WE, Hill JC, Toepfer P, Rachman F, Horwitz B, Kho A. Joining health care and homeless data systems using privacy-preserving record-linkage software. Am J Public Health. 2021;111(8):1400–3.
11. Agiro A, Chen X, Eshete B, et al. Data linkages between patient-powered research networks and health plans: a foundation for collaborative research. J Am Med Inform Assoc. 2019;26(7):594–602.
12. Raad JH, Tarlov E, Kho AN, French DD. Health care utilization among homeless Veterans in Chicago. Mil Med. 2020;185(3–4):e335–9.
13. Ahmad FS, Chan C, Rosenman MB, et al. Validity of cardiovascular data from electronic sources: The Multi-Ethnic Study of Atherosclerosis and HealthLNK. 2017;136(13):1207–16.
14. Philips L. Hanging on the Metaphone. Computer Language, 1990;7(12).

15. Soundex system [Internet]. National Archives. 2016 [cited 2022 Mar 8]. Available from: https://www.archives.gov/research/census/soundex
16. Grannis SJ, Kho A, Phua J, Kasthuriranthne S. Evaluation of Token Collections and Matching Models to Support Privacy- Preserving Record Linkage (PPRL). In: AMIA 2021 Annual Symposium; 2021.