Published online: 2019-08-12

Development and Evaluation of Record Linkage Rules in a Safety-Net Health System Serving Disadvantaged Communities

William E. Trick[1] Kruti Doshi[1] Michael J. Ray[1] Francisco Angulo[1]

[1] Cook County Health, Chicago, Illinois, United States

ACI Open 2019;3:e63–e70.

Address for correspondence William E. Trick, MD, Cook County Health and Hospitals System, 1950 W. Polk, Chicago, IL 60612, United States (e-mail: wtrick@cookcountyhhs.org).

Abstract

Background There is a need for flexible, accurate record-linkage systems with transparent rules that work across diverse populations.

Objectives We developed rules responsive to challenges in linking records for an urban safety-net health system; we calculated performance characteristics for our algorithm.

Methods We evaluated encounters during January 1, 2012 through September 30, 2018. We compared our algorithm, using name (first-last), date-of-birth (DOB), and last four of social security number to our electronic health record (EHR) system's reconciliation process. We applied our algorithm to unreconciled real-time Admission-Discharge-Transfer registration data, and compared match results to reconciled identities from our enterprise data warehouse. We manually validated matches for randomly sampled discordant pairs; we calculated sensitivity/specificity. We evaluated predictors of discordance, including census tract information.

Results Of 771,477 unique medical record numbers, most (95%) were concordant between systems; a substantial minority (5%) was discordant. Of 38,993 discordant pairs, most ($n = 36{,}539$; 94%) were detected by our local algorithm. The *sensitivity* of our algorithm was higher than the EHR process (99% vs. 81%), but with lower *specificity* (98.6% vs. 99.9%). Our highest-yield rules, beyond full first and last name plus complete DOB match, were first three initials of first name, transposed first-last names, and DOB offsets (+1 and +365 days). Factors associated with discordance were homelessness (adjusted odds ratio [aOR] = 2.4; 95% confidence interval [CI], 2.2–2.6) and living in a census tract with high levels of poverty (aOR = 1.4; 95% CI, 1.3–1.4).

Conclusion Our algorithm had superior sensitivity compared to our EHR process. Homelessness and poverty were associated with unmatched records. Improved sensitivity was attributable to several critical input-variable processing steps useful for similar difficult-to-link populations.

Keywords

- record-linkage
- clinical informatics
- disadvantaged
- poverty
- Latino

Background and Significance

Opportunities to monitor and improve the health of populations through record linkage across clinical, community, government agencies, and public health domains are abundant, but unfortunately, largely unrealized. In our research unit, linkage requests are common and increasing in complexity (e.g., linkage across domains, systems, and agencies). Although there are increasing concerns for appropriate privacy protection, there is conditional acceptance of record linkage projects.[1]

received
December 21, 2018
accepted after revision
May 22, 2019

DOI https://doi.org/10.1055/s-0039-1693129.
ISSN 2566-9346.

Linking data sets poses several of the following well-documented challenges: (1) privacy concerns, particularly with large-volume aggregated data sets[2]; (2) accuracy in the context of a limited number of input variables, dynamic variables (e.g., name changes), and frequent data entry errors[3]; and (3) technical challenges in linking records while retaining protected health information at local sites, particularly among sites with limited technical proficiency. An approach to addressing legal barriers and privacy concerns is to link records securely using a privacy-preserving record linkage (PPRL) system.[4] There are promising options for such systems; however, often the rules around identifying a match are not transparent,[5] limitations in input data quality are not recognized or explicitly addressed, or the cost for use of these systems may be prohibitive for many public health or community-to-clinical collaborations. We developed data-driven rules and code to process data for a diverse population for which, for various reasons, record linkage is more complicated[6,7]; we generate hashed identifiers and built matching rules to accommodate flexibility in input variables. We report on the performance of these rules for our large, urban, safety net population.

Objectives

Methods

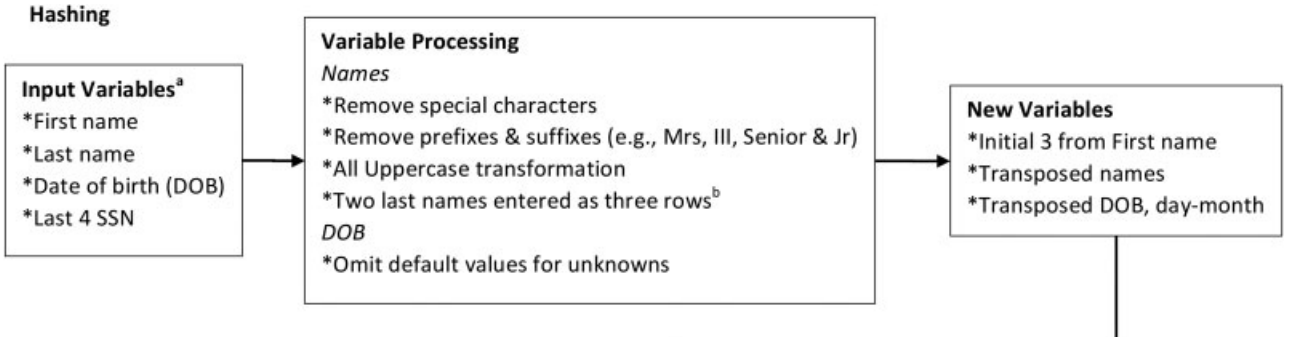
Setting

We extracted electronic health record (EHR) data for patients of Cook County Health, during January 1, 2012 through September 30, 2018. These visits represented the totality of health system visits; thus, not restricted to encounters that have the most comprehensive data collection. We included visits to the following settings: ambulatory clinics, emergency departments, and hospitalizations. We excluded patients > 105 years of age. Cook County Health is the primary safety net health system for Cook County, providing care for uninsured and historically underserved populations. A disproportionate number of patients are foreign born, homeless, victims of trauma, and do not have, or do not report, social security numbers (SSNs). Institutional review board review was exempted as this was a quality improvement project to improve record linkage for our health system. The authors declare that they have no conflicts of interest in the project.

Input Variables for Match

We identified potential input variables through literature review, prior clinical experience, and numerous record linkage projects both within Cook County (intragovernmental clinical sites without unified medical record numbers [MRNs]), and with community-based organizations. The following principles guided variable selection for matching: temporal stability, discriminatory value, prior validation, and availability in the context of restrictions to data sharing due to privacy concerns.

Given the aforementioned considerations, we included name (first and last), date-of-birth (DOB), and SSN restricted to last four integers (►Fig. 1). Due to concerns about performance and interest in preserving specificity, we decided against including phonetic representations of names (e.g., Soundex)[8] and gender/sex, which is increasingly recognized



Composite identifiers	Explanation
FN+LN+DOB	Name, DOB match
LN+FN+DOB	Transposed name, DOB match
FN+LN+T_DOB	Name, transposed DOB month-year match
FN(3)+LN+DOB	3 initials first name, last name, DOB match
FN+LN+(DOB+1 day)+Last 4 SSN	Name, DOB + 1 day, SSN match
FN+LN+(DOB+1 year)+Last 4 SSN	Name, DOB + 1 year, SSN match

Fig. 1 Diagram illustrating local processes to prepare input variables for matching. [a]Unknown date-of-births (DOB) are defaulted to 1/1/1900. Default values are institution specific, discoverable through frequency distributions. [b]Example: Smith-Jones or Smith Jones becomes three observations: Smith-Jones; Smith; and Jones. [c]Code can accommodate additional variables of interest (e.g., address components, gender); the matching rules are customizable.

as fluid, nonbinary, and provides relatively little discriminatory value.[9,10]

Through iterations of validating prototypes through manual chart review, we recognized that considerable attention was needed for processing input variables, especially names. We omitted special characters, spaces, and suffixes (e.g., Junior, II, Jr.) from names; thus, the surname <O'Malley Jr.> would match <Omalley Junior> as "OMALLEY." We developed a library of values that represented defaults for unknowns (likely institution specific), and blocked these values from matching. Examples of names blocked from matching included "unknown"; "unktrauma"; "trauma"; "Male"; "Female"; "JohnDoe"; "JaneDoe"; "Baby boy/girl"; and "Twin."

For patients who could not provide a legitimate DOB, our data set's default value was "1/1/1900," we did not create hash values. We identified this default DOB through a frequency distribution of DOB values. Finally, we processed SSNs by converting repeat integers (e.g., 9999) to null values. We identified commonly used default SSNs by evaluating frequency distribution graphs.

Date of Birth Evaluation

From prior record-linkage projects, a common reason for an individual to have > 1 MRN was DOB data entry error. We reasoned that common names in our system (e.g., Maria Garcia) were more likely to be unique individuals, and therefore have randomly distributed differences in the disparate dates of birth. In contrast, rare concatenated first and last names would be more likely to be the same person, which would reveal spikes in the frequency of DOB differences, indicating common data entry errors. We evaluated the distribution for DOB differences between individuals who shared the same name but with unique MRNs and DOB. We stratified our evaluation by how common the occurrence of the concatenated name was in our database, as follows: 2–4, 5–10, 11–20, 21–50, and > 50.

Matching

We constructed matching logic in the following hierarchical order: (1) complete name + DOB; (2) transposed complete name + DOB; (3) complete name + transposed DOB (month-day); (4) first three initials of first name + last name + DOB; (5) complete name + DOB (with 1 day offset) + (last four SSN); (6) complete name + DOB (1 year offset), + (last four SSN) (►Fig. 1). We implemented a deterministic matching process and created a single "master" patient-level identity (ID) across input values.

We explored factors associated with missing SSN, or for matching discordance between the EHR and our local matches. We focused on the following social factors: homelessness (identified by patient address [emergency shelter or clinical site address]), and a social vulnerability index and the four domains that comprise the index.[11,12] The social vulnerability index was obtained by geocoding patient addresses, and linking their address to census tract data. We used indicators present in the census data for census tracts in which ≥ 90% of residents had limited English, no high school diploma, were a minority, lived in poverty, or were unemployed. After bivariable analysis, we constructed multivariable logistic regression models using backwards stepwise procedures and report the adjusted odds ratios (aORs) and associated 95% confidence intervals (CIs).

Validation

We validated the performance of our matching algorithm using the EHR's reconciliation process; that is, postregistration, the EHR process creates a person-level identifier meant to consolidate disparate MRNs for the same person. Our evaluation was uniquely possible because our research data warehouse captures two sources of Admission-Discharge-Transfer events: (1) real-time Health Level Seven (HL7) messages that we parse and store before identity reconciliation has occurred, and (2) our enterprise data warehouse after the system-level person identifier has been generated.

After comparing matches from the two systems, we performed an unblinded manual review of a sample of discordant results, that is, matches only identified through the EHR or our local algorithm (match = yes/no or match = no/yes). To calculate the sensitivity and specificity for the two systems with a precision of ± 3% for a 95% CI assuming a value of 95%, we needed 200 discordant records for each discordant pair (Match-No match and No match-Match); unintentionally, we reviewed one additional case for the local rule match-EHR unmatched discordant pair. We defined true matches through manual chart review by clinicians. Clear evidence of a match included agreement beyond the respective system for at least one of the following variables: patient address, name or telephone number of preferred contact, or an uncommon clinical event documented in visits for disparate MRNs (e.g., gunshot wound to a specific body part). For adjudication of episodes in which there were more than two records in the bundle of discordant results, we required a match across all records.

Results

We evaluated 771,477 unique MRNs. A high-proportion of patients were from minority populations, were missing an SSN, and lived in socially vulnerable neighborhoods (►Table 1). For most unique MRNs (78%), neither our local rules matching system nor the EHR reconciliation process identified individuals who had been assigned more than one MRN; a substantial minority had multiple MRNs reconciled to a single person ID by both processes, that is, concordant matches (►Fig. 2). Among discordant matches, our local algorithm identified over 15-fold more MRN matches compared to the EHR reconciliation process (►Fig. 2). The highest yield rules for discordant matches identified by our local algorithm were truncating the first name to three characters, transposition of first-last name, transposition of day-month in DOB, then DOB offset (1 day and 1 year) (►Table 2). After applying our original rules, we performed a post hoc processing of hyphenated names, which yielded a large number of additional matches (►Table 2). Since we implemented this rule after the validation sample selection, the hyphenated name rule is not included in other results.

Table 1 Patient characteristics for unique medical records, Cook County Health, January 1, 2013 through August 31, 2018

Characteristics	*N* = 771,477	
	Mean	**SD**
Age, years	44	18
	N	**%**
Sex		
Female	392,853	51
Male	378,011	49
Transgender	195	< 1
Unspecified	418	< 1
Race-ethnicity		
Non-Hispanic African-American or black	412,255	53
Hispanic	203,120	26
Non-Hispanic white	89,088	12
Asian	26,714	3
American Indian or Alaska Native	4,726	1
Social characteristics		
Homeless, by address	9,304	1
Missing SSN	297,543	39
	Median	**IQR**
Social vulnerability index, percentile	74	63–90

Abbreviations: IQR, interquartile range; SD, standard deviation; SSN, social security number.

Table 2 Rules for matching hashed IDs, and the number of matches identified for each rule, but missed by the EHR process

Rule	No. of events[a]	Examples of matches for each rule	
First 3 initials, first name	7,623	Priscilla	Priscila
Transposed name[b]	1,622	Michael Ray	Ray Michael
Transposed DOB	454	3/10/1950	10/3/1950
DOB offset, 1 day[b]	371	8/10/1950	8/11/1950
DOB offset, 365 days[b]	313	1/1/1951	1/1/1952
Two last names[c]	1,988	Maria Vega-Reyes	Maria VegaReyes Maria Vega Maria Reyes

Abbreviations: DOB, date of birth; EHR, electronic health record; ID, identity; SSN, social security number.

[a]Represents the number of matches beyond a complete name + DOB match.

[b]For these matches, we required a match on the last four of the SSN.

[c]The processing step for two last names was added post hoc in response to validation findings. In the raw data, last names were separated by hyphen or space; we created two additional observations for matching. The number of matches identified for this rule was in addition to all other processing rules.

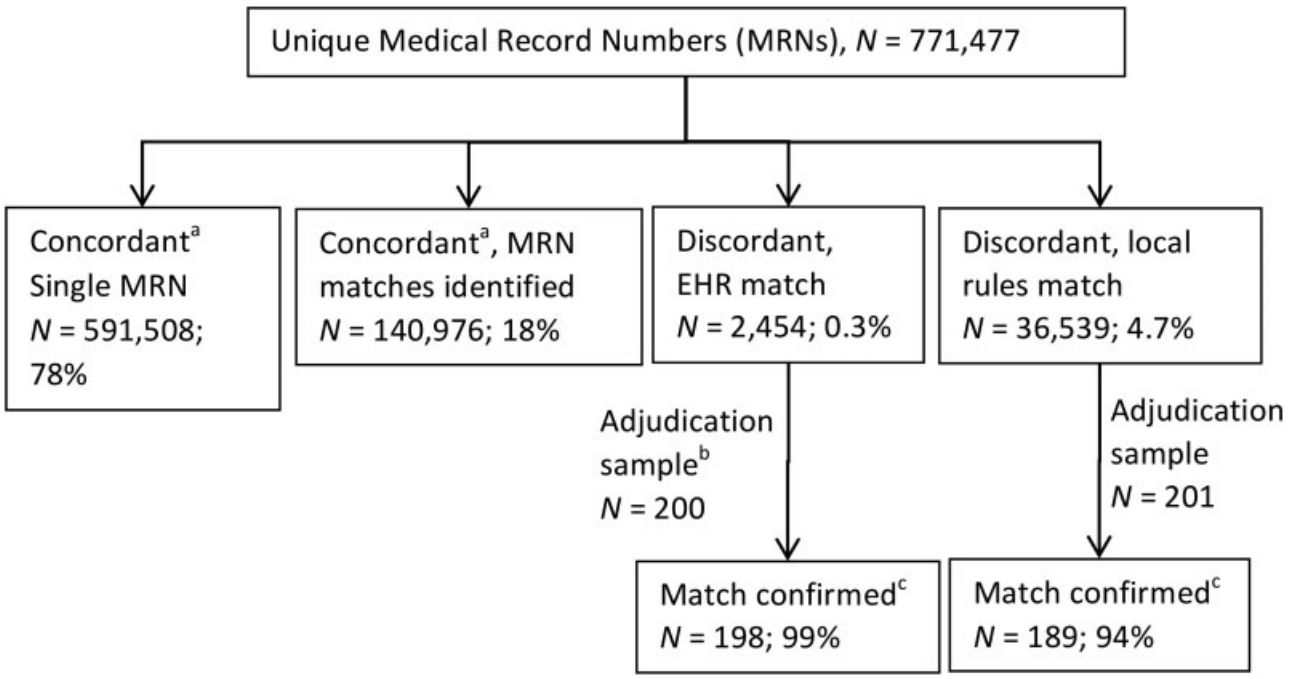


Fig. 2 Frequency of which our electronic health record system agreed with local rules for matching records, Cook County Health. Estimated sensitivity and specificity: Local rules (99.6%: 98.6%), EHR process (80.7%; 99.99%). Calculated with the following assumptions: (1) Concordant matches between systems were true; (2) Matches not definitively confirmed by manual chart review were not true matches. [a]Concordant between the EHR's reconciliation process and our local algorithm. [b]Sampled at the level of the person identity (ID) rather than medical record number (MRN). [c]For most episodes for which a match could not be confirmed, social security number (SSN) and other corroborating data to confirm a match were missing.

By multivariable analysis, the following factors were most strongly associated with missing SSN: census tract with a high proportion (> 90%) of limited English (aOR = 2.9; 95% CI, 2.8–2.9), homelessness (aOR = 1.6; 95% CI, 1.5–1.6), or no high school diploma (aOR = 1.3; 95% CI, 1.3–1.4). The strongest socioeconomic factors associated with discordant matches were homelessness (aOR = 2.4; 95% CI, 2.2–2.6) and living in a census tract with high levels of poverty (aOR = 1.4; 95% CI, 1.3–1.4) or minorities (aOR = 1.4; 95% CI, 1.3–1.4).

Date of Birth Evaluation

We found several distinct spikes in the frequency distribution of DOB differences among patients with rare names assigned separate MRNs. As expected, these spikes diminished incrementally as names became more common (►Fig. 3). The most pronounced spikes for DOB differences in descending order were 1 day, 365 days, 10 days, 2 years, and 30 days. The magnitude of these spikes was less pronounced as names became common, consistent with a random distribution that would be expected with true differences in patient identity. Because of these findings, we added 1- and 365-day DOB differences to our hashing algorithm and matching rules accommodated these differences; however, we required presence of a match to the last 4 of the SSN for DOB date-offset matches (►Fig. 1).

Validation

Among sampled discordant episodes in which our local algorithm identified the match, but the EHR process did not ($N = 201$), 189 were confirmed as true matches (►Fig. 2). Among the 12 episodes unconfirmed as a match, for 11 adjudication was not possible (i.e., absence of data to confirm or refute the match), and one patient was mismatched. The mismatch resulted from the first name rule in which two distinct first names (same first three initials) with a common last name (Smith) and shared DOB was misinterpreted as a match. For the 11 unconfirmed episodes, it is likely that most were "true" matches as the first and last names, and DOB were exact matches. The most common reason the EHR reconciliation missed a match was missing or erroneous SSN, first name spelling errors (encountered past the initial three characters), transposed first-last names, a middle initial concatenated with the first name, and DOB errors (►Table 3).

Among sampled discordant records in which the EHR process found a match but our local algorithm did not ($N = 200$), we confirmed 198 as true matches (►Fig. 2); a

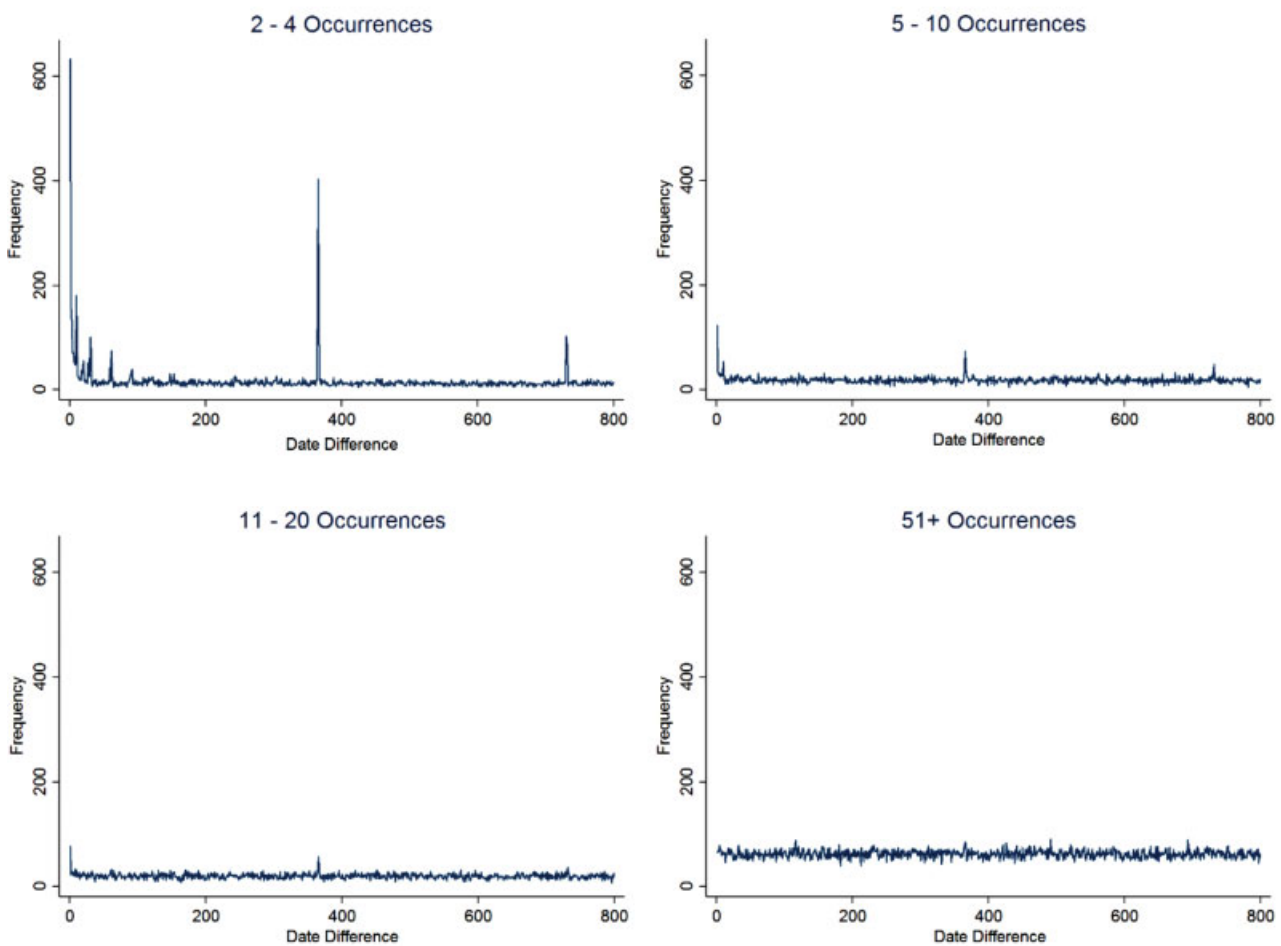


Fig. 3 Justification for including a date-of-birth (DOB) offset for record linkage from the distribution of DOB differences among rare (2–4 occurrences) through common (> 50) concatenated first-last names. Spikes in DOB differences at 1 and 365 days for rare names suggest keystroke errors; such differences dramatically diminished as names became common. Less dramatic increases were observed at 10, 30, and 730 (i.e., 2-year) date differences.

Table 3 Reasons our local rules missed matches and why matches were missed, by system

Reason local rules missed EHR match ($N = 200$)	n[a]	%
Hyphenated last name[b]	50	24
Last name changed	48	23
Date-of-birth (DOB) difference	47	23
Last name spelling error	39	19
First name spelling error	16	8
Two first names	7	3
Parsing error: Last name "Right eye pain"	1	1
Reason EHR process missed local rules match ($N = 201$)	n[a]	%
SSN missing or default value	94	47
First name error	24	12
SSN mismatched[c]	23	11
Middle initial concatenated to first name	11	5
Transposed first-last names	9	4
DOB error	7	3
No reason for mismatch identified	48	24

Abbreviations: DOB, date of birth; EHR, electronic health record; SSN, social security number.
[a]Sums to > 200 samples due to multiple error types for some records.
[b]We modified name processing after project completion to correct these mismatches.
[c]One digit error for 13 mismatched SSNs.

definitive determination could not be made for two episodes. The most common reason our local algorithm missed matches identified by the EHR process was hyphenated names followed by unrecognizable last name changes, DOB errors, last and first name spelling errors, and two first names (►Table 3). It was not always clear how records had been reconciled by the EHR (e.g., a complete last name change); however, for many episodes, there was a complete and legitimate SSN; for some records, we suspect that a manual reconciliation process had been activated by the patient or clinician.

Extrapolating results from our manual chart review to the entire population, we estimated a sensitivity and specificity for our local algorithm or 99.6 and 98.6%, respectively, and for the EHR, 80.7 and 99.99%, respectively.

Name Processing

Given the consistency in availability of names in linking records across disparate systems, we performed several iterations for processing and hashing names. Our process was informed by a relatively large proportion of Hispanic names, which had the following two unique features: Less variability—nearly all of the 20 most common first and last name combinations are associated with Hispanic ethnicity; and hyphenated names (two last names) are relatively common. Our post hoc decision to devise a method for processing hyphenated names was informed through our validation process, which yielded a substantial number of additional matches undetected by the EHR (►Table 2).

Discussion

Compared to patient reconciliation performed by our EHR process, our local algorithm identified over 15-fold more matches of individuals with discrepant MRNs. Despite the much higher rate of patient disambiguation, representing approximately 5% of all MRNs, false positive matches were rare—the single definitively confirmed false positive match resulted from a truncated first name. Our rules likely were highly successful as they were derived for, and applied to, our highly diverse population in a busy safety net health system inclusive of emergency department visits for which data often are missing or default values recorded. Critical to the success of record linkages in diverse populations with high levels of poverty and homelessness, is detailed name and DOB processing to accommodate data entry errors, missing data, and default values.[13]

We developed our matching algorithm with the guiding principle of developing rules and processes applicable to low socioeconomic status (SES) populations, which are notable in that data completeness and reliability often is compromised. We believe that future record linkage projects will emphasize data joins across health sectors, including health systems, community-based organizations, and governmental agencies.[14,15] Linkages across these disparate health sectors, which transcend traditional medical encounters, raise concerns about capture of reliable linkage factors, such as full SSN and stable phone numbers and addresses. While optimizing computational methods for record linkage is important, we believe that such efforts will result in marginal improvements relative to the critical process of standardizing and cleaning input variables and values. In particular, when such processes are performed at local sites to preserve patient privacy.

Our algorithm identified a dramatically higher number of matched patients with disparate MRNs than our EHR process. We expected higher sensitivity for our algorithm as EHRs must be highly specific to avoid inappropriate merging of records. The consequence of a vendor's need to emphasize specificity was manifest in our population, which has a high frequency of erroneous or missing SSNs. Unfortunately, for current and future data linkage efforts, whereas capture of several data fields has been relatively constant over time, SSN documentation has been decreasing over time.[16] A unique feature of our population was the relatively high proportion of individuals with Hispanic names, which can complicate matching due to more frequent use of hyphenated last names and common first-last name combinations. For similar situations, it is possible that executing separate processes with human review could be an opportunity for better fidelity to the record linkages.

Efforts to improve matching individuals both within medical systems and between health entities for longitudinal assessments and provision of services have been ongoing for many years.[17,18] Despite considerable success in prior efforts, there is an ongoing need to evaluate systems through

detailed manual review across diverse populations.[19] The accuracy of record linkage is dependent on availability of input variables, reliability of data entry, and population characteristics (race-ethnicity, SES, homeless status); we found that homelessness and poverty were associated with mismatched records. Given these challenges, opportunities to optimize record linkage have been described for the following domains: incorporation of biometrics, standardized demographic inputs, expanded number of inputs, and use of referential data sources external to the health system.[20] Rather than expand the number of fields for data input, we evaluated a relatively parsimonious linkage system that can work across disparate data sources; we focused on processing reliably present key fields, which also are unlikely to identify false matches.[21] Also, we are keenly aware of institutional, investigator, and individual concerns about processing sensitive information. We designed our system with the intent of linking data between health systems, public health agencies, and community-based organizations each of which will have unique concerns about how and which data are shared. Thus, we evaluated our system as if data had been shared using PPRL by concatenating multiple fields, and hashing with a seed, which mitigates the risk of reidentification through frequency attack algorithms.[4] We provided clear and transparent rules for processing and matching data, which often are obscure.[5]

We had a unique source of data for evaluating our matching and data processing rules; two discrete sources of reconciled EHR data. One source was a real-time registration (HL7) feed that we extracted before our vendor's reconciliation process and the other source was our enterprise data warehouse after EHR reconciliation. We focused our validation on records with discordant matching results. The most common reason the EHR process failed to identify true matches was missing or erroneous SSN—but, there were no episodes for which a default SSN (e.g., nine repeat integers) was the basis for a match. The most common reasons for our algorithm to miss true matches were the presence of hyphenated last names, last name change, errors in DOB, and name spelling errors. Of note, a substantial number of records matched by our rules due to month-day transposition, possibly because many of our patients emigrated from regions with a dd/mm/yyyy date convention rather than mm/dd/yyyy (e.g., Mexico, Central and South America, and parts of Africa). The sensitivity of our local rules could have improved through use of full SSN, but as mentioned, we intentionally restricted SSN to the last four integers.[22]

Complicating our assessment was that the medical record reconciliation process permits manual (patient or clinician) identification of mismatched records, who can communicate the need for reconciliation through the medical records department. Another limitation of our project was that we used observations from a single health system; however, a strength was that we evaluated a unique population that has disproportionate representation of individuals born in other countries, and busy emergency and trauma departments for which data are more likely to be incomplete or unknowable (e.g., overdose and unaccompanied trauma victims). To optimize linkage success, we employed a data-driven approach; for example, we evaluated frequency distributions for DOB differences by the rarity of concatenated first-last names, and we evaluated DOB and SSNs to identify default values.

We developed an open source system for distribution across multiple platforms and with transparent rules.[23] We enhanced our data processing to match hyphenated names, and we intend to expand our library of processing steps to work across institutions (e.g., expansion to other data set-defined default names, such as our patients named "unknowntrauma"). Having such open source code allows health systems to evaluate their process for patient disambiguation by running two parallel processes, and labeling discordant pairs of records for manual review.

Conclusion

The value of linking records across health sectors particularly for populations that experience health disparities are increasingly being recognized. Integration of clinical data with public health and community-based organizations is the ultimate goal. Such linkages will be advanced through low cost, transparent systems with flexible rules for processing commonly available input variables that can be included in PPRL software systems. Our processing and hashing rules successfully identified mismatched records for the diverse, urban, safety-net population our health system serves.

Clinical Relevance Statement

Health care professionals benefit from patient information to facilitate smooth transitions of care. These professionals (care coordinators, nurses, and physicians) will benefit from secure methods to link individuals' care records not only among traditional clinical care entities but also across health sectors, such as housing status and substance use treatment history. Integration of data systems will help clinicians and care coordinators address the social determinants of health often responsible for poor health outcomes.

Protection of Human and Animal Subjects

This study was performed as a quality improvement project to improve record linkage within separate domains of a large integrated health system. We conferred with the Institutional Review Board, and it was determined that review was not necessary.

Conflict of Interest

None declared.

Acknowledgments

This work was performed through support of the investigators' home institution; there were no external funds obtained for the project. We acknowledge Keiki Hinami for guidance during the conception and design of the project.

References

1 Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016;17(01):73

2 Kshetri N. Big data's impact on privacy, security and consumer welfare. Telecomm Policy 2014;38(11):1134–1145

3 Bohensky MA, Jolley D, Sundararajan V, et al. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res 2010;10(01):346

4 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(05):1072–1080

5 Harron K, Wade A, Muller-Pebody B, Goldstein H, Gilbert R. Opening the black box of record linkage. J Epidemiol Community Health 2012;66(12):1198

6 Zingmond DS, Ye Z, Ettner SL, Liu H. Linking hospital discharge and death records–accuracy and sources of bias. J Clin Epidemiol 2004;57(01):21–29

7 Maizlish NA, Herrera L. A record linkage protocol for a diabetes registry at ethnically diverse community health centers. J Am Med Inform Assoc 2005;12(03):331–337

8 Patman F, Shaefer L. Is Soundex Good Enough for You? On the Hidden Risks of Soundex-Based Name Searching. Herndon, VA: Language Analysis Systems, Inc.; 2001–2003

9 Meadow T. "A Rose is a Rose" on producing legal gender classifications. Gend Soc 2010;24(06):814–837

10 Quantin C, Binquet C, Bourquard K, et al. Which are the best identifiers for record linkage? Med Inform Internet Med 2004;29 (3-4):221–227

11 Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry/Geospatial Research, Analysis, and Services Program. Social Vulnerability Index 2014 Database, Illinois. Available at: https://svi.cdc.gov/data-and-tools-download.html. Accessed October 30, 2018

12 Flanagan BE, Gregory EW, Hallisey EJ, Heitgerd JL, Lewis B. A social vulnerability index for disaster management. J Homel Secur Emerg 2011;8(01):3

13 Ansolabehere S, Hersh ED. ADGN: an algorithm for record linkage using address, date of birth, gender, and name. Stat Public Policy (Phila) 2017;4(01):1–10

14 Holman CDAJ, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records in Western Australia: development of a health services research linked database. Aust N Z J Public Health 1999;23(05):453–459

15 Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Annu Rev Public Health 2011; 32:91–108

16 Culbertson A, Goel S, Madden MB, et al. The building blocks of interoperability. A multisite analysis of patient demographic attributes available for matching. Appl Clin Inform 2017;8(02): 322–336

17 Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995;14(5-7):491–498

18 Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. J Am Med Inform Assoc 1997;4(03):233–237

19 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

20 Pew Charitable Trusts. Enhanced patient matching is critical to achieving the full promise of digital health records. Secondary Pew Charitable Trusts. Enhanced patient matching is critical to achieving the full promise of digital health records. Available at: https://www.pewtrusts.org/-/media/assets/2018/09/healthit_en hancedpatientmatching_report_final.pdf. Accessed: November 1, 2018

21 Zech J, Husk G, Moore T, Shapiro JS. Measuring the degree of unmatched patient records in a health information exchange using exact matching. Appl Clin Inform 2016;7(02):330–340

22 Naessens JM, Visscher SL, Peterson SM, et al. Incorporating the last four digits of social security numbers substantially improves linking patient data from de-identified hospital claims databases. Health Serv Res 2015;50(Suppl 1):1339–1350

23 Linkja: Open Source Privacy Preserving Record Linkage. GitHub Repository initiated January 31, 2019. Available at: https://linkja.github.io/. Accessed July 1, 2019