Measuring the Degree of Unmatched Patient Records in a Health Information Exchange Using Exact Matching

John Zech[1]; Gregg Husk[2]; Thomas Moore[3]; Jason S. Shapiro[4]

[1]Icahn School of Medicine at Mount Sinai, New York, NY, USA; [2]Lenox Hill Hospital, New York, NY, USA; [3]Healthix, Inc., New York, NY, USA; [4]Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Keywords

Health information exchange, medical record linkage, performance improvement

Summary

Background: Health information exchange (HIE) facilitates the exchange of patient information across different healthcare organizations. To match patient records across sites, HIEs usually rely on a master patient index (MPI), a database responsible for determining which medical records at different healthcare facilities belong to the same patient. A single patient's records may be improperly split across multiple profiles in the MPI.

Objectives: We investigated the how often two individuals shared the same first name, last name, and date of birth in the Social Security Death Master File (SSDMF), a US government database containing over 85 million individuals, to determine the feasibility of using exact matching as a split record detection tool. We demonstrated how a method based on exact record matching could be used to partially measure the degree of probable split patient records in the MPI of an HIE.

Methods: We calculated the percentage of individuals who were uniquely identified in the SSDMF using first name, last name, and date of birth. We defined a measure consisting of the average number of unique identifiers associated with a given first name, last name, and date of birth. We calculated a reference value for this measure on a subsample of SSDMF data. We compared this measure value to data from a functioning HIE.

Results: We found that it was unlikely for two individuals to share the same first name, last name, and date of birth in a large US database including over 85 million individuals. 98.81% of individuals were uniquely identified in this dataset using only these three items. We compared the value of our measure on a subsample of Social Security data (1.00089) to that of HIE data (1.1238) and found a significant difference (t-test p-value < 0.001).

Conclusions: This method may assist HIEs in detecting split patient records.

Correspondence to:

Jason S. Shapiro
Associate Professor, Department of Emergency Medicine
Icahn School of Medicine at Mount Sinai
Box 1620
One Gustav Levy Place
New York, NY 10029
jason.shapiro@mountsinai.org

Appl Clin Inform 2016; 7: 330–340
http://dx.doi.org/10.4338/ACI-2015-11-RA-0158
received: December 1, 2015
accepted: February 26, 2016
published: May 11, 2016
Citation: 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: 330–340
http://dx.doi.org/10.4338/ACI-2015-11-RA-0158

Background

Health information exchange (HIE) facilitates the exchange of patient records between different healthcare organizations in a standardized way [1]. Well-functioning HIE enables healthcare providers to make more informed patient care decisions. In order to connect patient records at multiple sites, a health information exchange utilizes a master patient index (MPI) that connects external, site-level patient medical record numbers (MRNs) to internal MPI identifiers. Each unique patient is intended to have one internal MPI identifier, but may have multiple MRNs at different healthcare sites that map to that single MPI identifier. Through the use of this MPI identifier, a patient's records can be connected across sites.

HIEs are responsible for determining which MRNs should be mapped to which MPI identifiers, which is established by a matching algorithm [2, 3]. Exact matching algorithms match patient records when certain fields in different records are identical. Deterministic matching algorithms link records "based on agreement rules (exact, approximate, and partial) for matching variables, which are often structured hierarchically" [4, 5]. Probabilistic matching algorithms calculate match likelihood scores based on similarity between certain fields in different records and match patient records when the total match likelihood score is above a given threshold [6]. Both deterministic and probabilistic matching algorithms can match records more flexibly than exact matching. These matching algorithms enable the connection of patient records both within and across different sites. Assessing the quality of an MPI's record matching is challenging, as there is no readily available set of validation data that can be used to test the accuracy of record linking.

Records can be incorrectly linked in two ways: false negatives and false positives. A false negative link fails to connect two records that belong to the same patient, and may result in the patient having a split patient record, with portions of the record each associated with different MRNs. A false positive link connects two records that belong to different patients. A failure to correctly match records within and between healthcare sites can compromise the quality of patient care [7–9]. The safety risks of false positive matches, in which clinicians believe incorrect information about the patient to be true, may be greater than the safety risks of false negative matches, in which clinicians lack access to existing information about the patient. As a result, HIEs have typically shown a preference for minimizing false positive matches at the expense of increasing false negative matches. This creates a tendency for HIEs to have significant numbers of split patient records (false negative matches), and we focus on this category of records in this paper.

Past work has noted substantial numbers of split records in clinical databases and examined methods for detecting these split profiles, including the use of exact matching [7, 10–14]. Researchers have previously studied how individuals could be uniquely identified using combinations of different identifiers [3, 15–17]. Organizations are aware of the existence of split records, but approaches to dealing with the issue are not consistent [18]. Much work has focused on the challenges of matching records at the level of an individual organization. HIEs operate on a larger scale than individual provider organizations and expressly include multiple records from the same patients at different organizations. Accordingly, the challenge of matching patient records in the aggregate data of an HIE may be significantly greater.

It would be useful for an HIE to have tools that could assist in the detection of split records. Such tools could alert HIEs of the degree of split records in their database and indicate patient record pairs that are split into multiple MPI profiles, but are likely to belong to the same patient. Organizations could manually examine a number of such patient record pairs that their MPI linking algorithm did not link, and come to understand why their MPI linking algorithm failed to link these records. This could inform tuning and customization of their deterministic or probabilistic matching algorithm and improve MPI performance. This process could be repeated iteratively to remedy common causes of false negative matches. Organizations could specifically analyze subpopulations that are especially likely to have issues with record matching, such as homeless individuals, to ensure that these patients' records are not improperly split [19].

Objectives

In this paper, we investigated how exact matching may be used as a method to detect split patient records in an HIE's MPI. This method may be useful in detecting false negative matches between records that share exactly matching first name (FN), last name (LN), and date of birth (DOB). However, this method cannot detect false negative matches between records lacking exactly matching FN, LN, and DOB, nor can it detect false positive matches, as it relies exclusively on the fact that it is relatively rare for two different individuals to share a FN, LN, and DOB.

We defined a measure to quantify how frequently different individuals share a FN, LN, and DOB, and proposed to use the reference value of this measure from a large national demographic database, the Social Security Death Master File (SSDMF), as a baseline comparison in the US. We demonstrated the application of this proposed method using data from a New York City based HIE, Healthix. We compared the calculated measure from the HIE and the calculated measure from the SSDMF to detect the existence of a substantial number of likely split records in this HIE.

Methods

Data sources included data from the Social Security Death Master File (SSDMF) and MPI data from Healthix, a New York City based HIE.

The Social Security Administration maintains a file containing demographic information for deceased residents of the United States [20, 21]. This file contains first name (FN), last name (LN), date of birth (DOB), and Social Security Number (SSN) for a large number of deceased Americans (85,822,194 individuals as of November 2011). A November 2011 copy of the Social Security Death Master File can be freely downloaded from a privately hosted website, and we used this data in this paper [22]. For comparability with HIE data, we perform no data cleaning on SSDMF data. We were not able to clean Healthix data because patients' names and dates of birth are protected health information, and we relied on other parties to generate the aggregated data that we report in this paper. We give more detailed information on the relative frequency of placeholder values in both datasets in ▶Appendix A.

We chose to use the SSDMF as we did not have access to any alternative manually curated data source covering such a large population and containing first name, last name, date of birth, and a separate unique identifier for individuals (SSN). Although each individual record in the SSDMF was manually confirmed prior to inclusion against a government database including all issued SSNs and corresponding demographic information, and it serves as the best available reference dataset for our study, certain errors in the SSDMF have been identified [23]. These errors include typographical mistakes in recording demographic data, omitting certain individuals who have died from the database, and incorrectly including certain living individuals in the database [23].

Healthix is a New York City based HIE. As of November 2014, their MPI contained 11,604,984 unique patient identifiers linking records across more than 100 participating organizations in New York City and Long Island [24].

Each of these two separate data files (SSDMF and Healthix MPI) contained an assigned identifier intended to designate a unique individual. In the SSDMF data file, a unique SSN identifier was assigned to each individual. In the Healthix dataset, an internal Healthix identifier (Healthix ID) was assigned to each record based on a probabilistic record matching algorithm that utilizes demographic data, including LN, FN, DOB, gender, home address, phone number, and Social Security Number when available. When the Healthix matching algorithm compares two records, each of these fields is compared individually, and a match likelihood score is calculated for each field. These scores can be positive in the case of agreement, or zero or negative in the case of disagreement. Intermediate scores can be given for partial matches by using algorithms to calculate similarity between two records [25]. The individual field match likelihood scores are summed to give a total match likelihood score, and records with a total match likelihood score greater than a pre-selected threshold value are matched to the same Healthix ID. In this way, the same Healthix ID can be assigned to multiple records believed to belong to the same patient across multiple participating organizations or within an institution.

Our proposed measure was to match records exactly on the tuple of FN, LN and DOB. Two records were considered a match only if each of these three fields in the tuple matched exactly. For each unique tuple in a database, we counted the number of unique identifiers that are assigned to this particular tuple (i.e. the number of SSNs for each unique combination of FN, LN and DOB). We calculated the average number of unique identifiers per tuple, and this served as the reference value of our benchmark, indicating how often two individuals shared the same FN, LN, and DOB in the given population. We calculated this measure on the SSDMF database as a baseline for the general US population. We then separately repeated this approach on Healthix data. We matched Healthix records exactly based on the tuple of FN, LN, and DOB. For each unique tuple in the Healthix database, we counted the number of unique identifiers (Healthix IDs) that are assigned to this particular tuple. We calculated the average number of unique identifiers per tuple, and this served as the comparison value of our measure, which we compared against the SSDMF-based reference value to detect the degree of duplicated records in Healthix data. We chose to define our measure in this way because we believed it would be interpretable and straightforward for others to replicate. Likewise, we link records exactly on FN, LN, and DOB to promote straightforward calculation, and we discuss how the approach could be extended in the Discussion section.

We analyzed SSDMF data and Healthix data completely separately, and at no point did we compare any individual records between the two databases. Accordingly, we did not need to match records between these two databases. An example of this calculation on simulated MPI data is demonstrated in ▶Figure 1. Please note that for the Healthix dataset, we could not simply divide total Healthix IDs by the number of unique tuples to find the average Healthix IDs per tuple, as some Healthix IDs were shared among multiple distinct tuples. This would have occurred if the Healthix MPI correctly matched two records belonging to the same patient to the same internal MPI despite differing demographic data, as in the case of typographical errors (e.g., 'JON WILLIAMS' and 'JOHN WILLIAMS') or name changes. It would also have occurred if the Healthix MPI incorrectly matched two records with differing demographic data. The number of Healthix IDs associated with a tuple had to be calculated for each tuple, and an average of those numbers gave the average number of Healthix IDs per tuple. For example, in ▶Figure 1, there are 4 unique tuples and 4 unique MPI IDs, and yet the average RHIO IDs / tuple is 1.25 when properly calculated because Jane Jones born 8/27/1985 was mapped to two separate RHIO IDs.

We calculated this measure on (1) the full SSDMF dataset, (2) a random sample of the SSDMF dataset similar in size to the Healthix dataset, (3) a subset of the SSDMF dataset similar in size to the Healthix dataset that included all overlapping FN, LN, DOB tuples in the SSDMF, (4) the full Healthix dataset, and (5) a reduced Healthix dataset containing records for patients that visited two or more facilities (so-called "crossover" patients) [26, 27]. To determine which patients had visited multiple sites, we matched records exactly on FN, LN, and DOB. Each SSN identifier in the SSDMF dataset was believed to correspond to a unique individual. The SSDMF dataset thus gave a reference value for our measure, indicating how often two individuals might share the same FN, LN, and DOB purely by chance in the general US population. If no two individuals in the SSDMF shared the same FN, LN, and DOB tuple, we would have calculated an average number of SSN identifiers per tuple of exactly 1. The higher the number of individuals who shared a FN, LN, and DOB with other individuals, the higher this SSDMF reference measure would have been.

If individual patients tended to have their records split across multiple identifiers in Healthix because of the difficulty of matching records, we would have observed an inflated value of this measure in Healthix data. The difference between this measure calculated on SSDMF and Healthix data thus gave a measure of the degree of duplicate profiles in Healthix's MPI: the higher the average number of identifiers per tuple was in Healthix data relative to SSDMF data, the greater the degree of likely record splitting and duplicate Healthix IDs that were present in the Healthix MPI. We compared the number of identifiers per tuple in Healthix and SSDMF data using a t-test to confirm the significance of the difference observed. This approach allows us to measure the degree of false negative matches where FN, LN, and DOB are an exact match. It should be noted that it did not measure false negative matches where FN, LN, and DOB are not an exact match, and it does not measure false positive matches.

Results

Our primary results are presented in ▶Table 1. The November 2011 SSDMF data contained 85,822,194 individuals.. There were 85,292,316 unique FN, LN, and DOB tuples, and 1.0062 SSN identifiers per tuple. 1,020,932 individuals shared a FN, LN, and DOB tuple with at least one other individual in the database. 98.81% of individuals did not share their FN, LN, and DOB tuple with any other individual in the database.

In a randomly sampled subset of national SSDMF data in which approximately 13.1% of records were included, there were 11,240,288 individuals. There were 11,230,306 unique FN, LN, and DOB tuples, and 1.00089 SSN identifiers per tuple. 19,779 individuals shared a FN, LN, and DOB tuple with at least one other individual in the sample. 99.82% of individuals did not share their FN, LN, and DOB tuple with any other individual in the sample.

An SSDMF partition of 11,600,000 records containing all of the records with shared FN, LN, DOB tuples had an average of 1.048 SSNs / tuple. The partition including the remaining 74,222,194 records in the SSDMF had an average of exactly 1 SSN / tuple.

The November 2014 Healthix MPI had an overall average of 1.1238 HIE identifiers per tuple for patients overall, an average of 1.2862 HIE identifiers per tuple for those patients who visited two or more sites, and an average of 1.0180 HIE identifiers per tuple for those who visited only one facility. The average number of identifiers per tuple for Healthix patients overall was significantly higher than the number of identifiers per tuple in the comparably sized SSDMF sample (t-test p-value <0.001).

Discussion

We have shown that it was uncommon for two or more legal United States residents to share the same FN, LN, and DOB tuple in a large population of over 85 million individuals. It was even rarer for two individuals to share a FN, LN, and DOB in our sample sized to match the Healthix patient population size: the average individual had only a 0.18% chance of sharing a FN, LN, and DOB tuple with another individual in the sample. The reference value for our measure of identifiers per FN, LN, and DOB tuple was very close to one for the entire 85.8 million individual SSDMF population (1.0062) and the smaller SSDMF sample (1.00089). If an HIE were able to match records perfectly and did not contain more placeholder values for FN, LN, and DOB than SSDMF data, we would expect to see a comparison value of this measure close to one when calculated using their data. Comparison values of the measure significantly larger than one give an indicator that patients' records are being split into multiple profiles. The utility of this measure lies in the fact that we have observed low rates of overlap on FN, LN, and DOB in large demographic databases. This suggests that an exact match on FN, LN, and DOB is strong evidence that two records belong to the same individual. A database in which records that share these three data fields are regularly split into multiple profiles is likely to be incorrectly splitting patient records. A low comparison value of this measure is not sufficient to prove that matching is highly accurate because there could be false negative matches with differing FN, LN, and DOB, but a high comparison value of this measure provides evidence that duplicate records are present. We believe that this measure could be of value as part of a regularly run HIE matching performance analysis to detect the degree of split records, using the approach described in the last paragraph of the Background.

In Healthix data, we saw a comparison value of the measure of 1.2862 for patients who visited two or more sites, significantly greater than the 1.00089 reference value we found in a sample of SSDMF data. It is likely that many patient records were split across multiple profiles. We also note that the comparison value of this measure was substantially higher for patients who visited multiple sites (1.2862) than patients who visited only one site (1.0180), indicating that split profiles were more likely to arise in this HIE when linking records across rather than within sites. We believe that our comparison between Healthix data and a subsample of SSDMF data is limited by the fact that the SSDMF does not directly correspond to the current population of New York City. The distribution of names in a region reflects the ethnic groups present in that area, and popular names change over time. Furthermore, the SSDMF included only legal residents of the United States, but undocu-

mented immigrants receive care at New York City hospitals and would appear in HIE databases. State voter data is an alternative data source that could better reflect the current distribution of names in a given geographic area. While voter data for New York State is available, state law forbids the use of this data for non-election purposes [29].

We compared the total Healthix database against a similarly sized subsample of the SSDMF in as we wished to keep the analysis straightforward and replicable by others on their own datasets. One might instead prefer to compare crossover patients to an appropriate control group, as only crossover patients test the ability of the HIE to connect records across sites. We believe the sample of SSDMF data likely understates the true value of this measure in New York City due to its inability to reflect the particular ethnic groups present in this area. At the same time, we believe that the values calculated on HIE data are substantially in excess of what one might reasonably find on a real population. As an extreme example, we calculated a measure value of 1.048 on a set of 11.6 million SSDMF records containing all records with overlapping FN, LN, and DOB. No 11.6 million record subset of the SSDMF could have a measure value more extreme than this one, and the value we observed on Healthix data is significantly larger than even this result.

In practice, we would not expect to see real-world HIE comparison values of this measure as low as the reference value we might calculate on the true population of an area. We believe it is common for HIE MPI records to share enough data to suggest a match, but not enough data to automatically establish one. In Healthix data, for example, there were over 4 million site-level patient records with a matching score deemed ambiguous by Healthix. In an ambiguous situation, an HIE may reasonably prefer a false negative to a false positive, as false positives carry more problematic clinical and legal consequences. It might be desirable for clinicians requesting HIE data to receive a list of 'possible matches' and have the ability to determine if the returned records correspond to the patient under consideration. However, this is not currently possible in Healthix due to state laws that regulate consent for HIE.

Our exact matching approach is limited in that it missed opportunities to link records when there were typographical errors or variations in name that prevent an exact match ("JONATHON" vs. "JON" vs. "JNOATHON", etc.). Additionally, individuals who change their name (either individually or after marriage) were split into multiple records in our exact matching approach. Exact linking keeps the comparison measure query straightforward and replicable, but an organization could extend our approach to match common variations of names using lookup tables or a phonetic mapping system, such as the New York State Identification and Intelligence System (NYSIIS), if they wished [28]. In that case, an appropriate reference measure would have to be recalculated on SSDMF data using corresponding matching rules.

Our proposed measure is limited in that it reflects only one part of record matching, the inappropriate splitting of records with matching demographic data. It is not intended to give a complete measure of match quality: it does not measure the inappropriate combination of records, and it cannot detect improperly split records where FN, LN, and DOB are not exact matches. One could achieve a comparison value of this measure of exactly one by matching records exactly, but this would not take into account false negative and false positive matches. For example, this would create many false negative matches where clear typographical variants (e.g., 'JONATHON' and 'JNOATHON' and 'JON') are not recognized as identical. It would also create false negative matches among groups with changed last names (i.e., married people who change or hyphenate their last name). It could also create false positive matches among common names (e.g., 'JOHN SMITH') or among records with placeholder values ('UNKNOWN'). In an SSDMF sample sized comparably to the Healthix patient population, we observed that the average individual had a 0.18% chance of sharing a FN, LN, and DOB tuple with another individual in the sample. In the full SSDMF sample, this chance was 1.19%. These give a sense of the frequency of false positives one might expect using strictly exact matching. Given the variety of ways in which names can be reported, misspelled, and changed over time, we would expect false-negative rates in an exact matching algorithm to be substantial.

Conclusions

We believe that this method may serve as a tool to assist HIEs and other healthcare organizations to measure the degree of split patient records in their systems. This could alert organizations to issues with data quality and the process used to match records, and could be used to inform how organizations choose tuning parameters for the algorithms they employ to match records. We would encourage researchers with access to other large demographic datasets to replicate our approach on their data and report the frequency with which they find individuals sharing a FN, LN, and DOB.

Clinical Relevance

Inappropriately split records at the level of an HIE may cause clinical decisions to be made with incomplete information. Detection of split records at the level of an HIE can better ensure that all relevant data is available to practitioners as they provide care to patients. This paper provides a tool that may assist in the detection of split HIE records using exact record matching.

Conflicts of Interests

The authors declare that they have no conflicts of interest in the research.

Protection of Human and Animal Subjects

This study was reviewed by the Mount Sinai IRB and the Healthix Research Committee and deemed not human research.

Acknowledgements

We wish to thank Ben Purkis at Audacious Inquiry, Tina Lowry at Healthix, and Laurence Berg at the New York e-Health Collaborative for explaining how Healthix matches patient records and running queries on the Healthix MPI to provide data for this analysis.

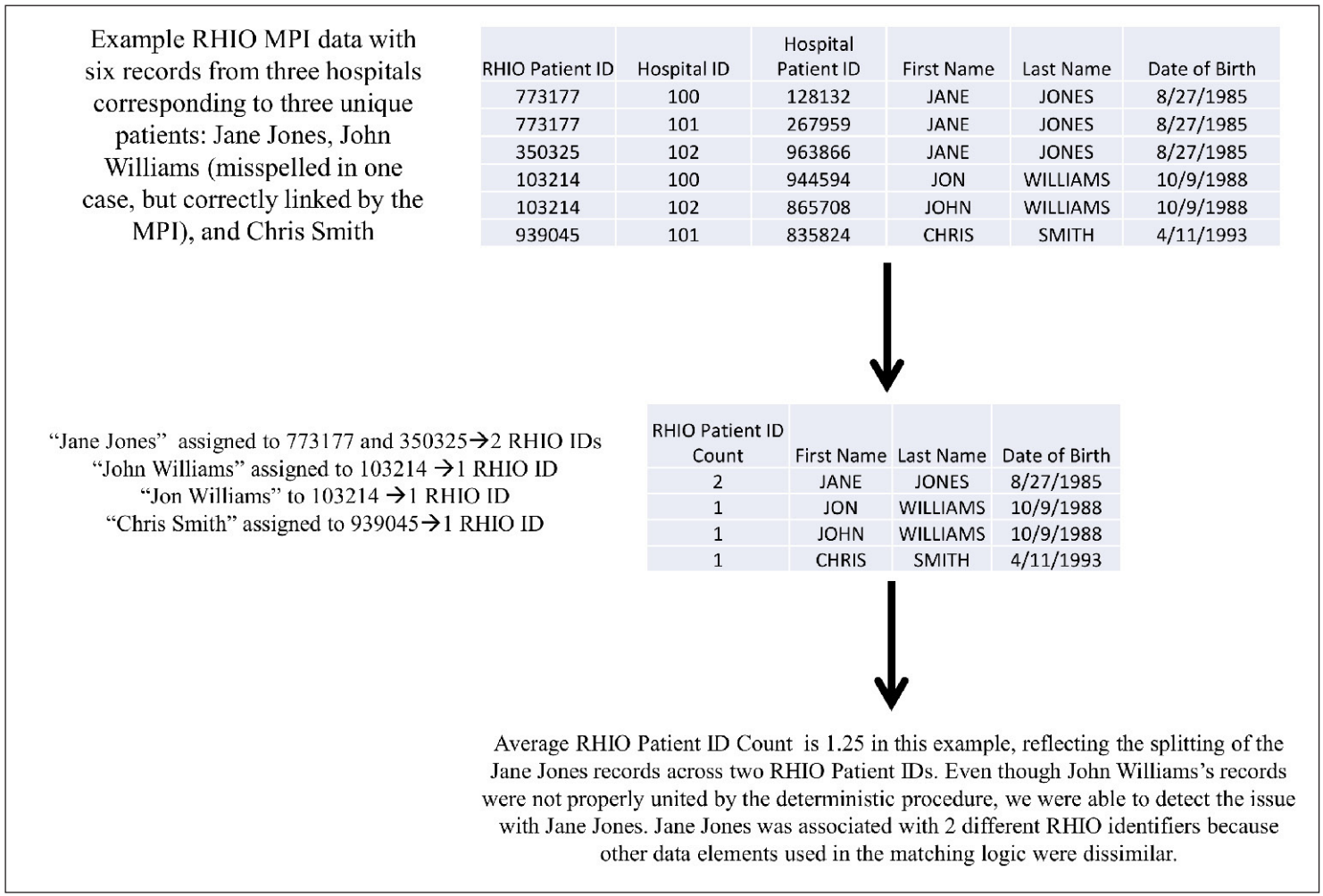

Fig. 1 Example of measure calculation on simulated HIE data.

Table 1 Results for each group analyzed

	Identifiers	FN, LN, DOB tuples	IDs / tuple	Percentage of patients uniquely identified by FN, LN, DOB tuple
Full SSDMF	85,822,194	85,292,316	1.0062	98.81%
SSDMF Sample	11,240,288	11,230,306	1.00089	99.82%
Healthix	11,604,984	11,456,145	1.1238	N/A

Appendix A: Data Quality

We assessed SSDMF data quality by manually examining FN, LN, or DOB that were associated with 100 or more records ('high frequency'). In total, there were 85,822,194 records. 83,075,209 records had a high-frequency FN, 73,312,820 records had a high-frequency LN, and 85,258,198 records had a high-frequency DOB. We determined that a FN or LN was likely a placeholder when it was blank, consisted of one character, had no vowels, or had a space in its second character (often indicating an abbreviation). We excluded the known FN and LN 'NG'. We found there to be 2,619,129 records with a high-frequency placeholder FN and 563 records with a high-frequency placeholder LN. We determined that a DOB was likely a placeholder when it had a month, day, or year value of zero. All other dates with 100 or more occurrences appeared valid. We found there to be 121,707 records with a high-frequency placeholder DOB.

We assessed Healthix data quality by manually examining FN, LN, or DOB that were associated with 100 or more unique site-level MRN records ('high frequency'). In total, there were 19,525,815 site-level MRN records. 16,816,734 records had a high-frequency FN, 12,840,712 records had a high-frequency LN, and 19,277,893 records had a high-frequency DOB. No blank FN, LN, or DOB were observed in these high-frequency data. We determined that a FN or LN was likely a placeholder when it consisted of one character, had no vowels, had a space in its second character (often indicating an abbreviation), or contained a known dummy value used for testing. We excluded the known FN and LN 'NG'. We found there to be 38,432 records with a high-frequency placeholder FN and 8,791 records with a high-frequency placeholder LN. No DOBs with a month, day, or year value of zero were observed. We determined that a DOB was likely a placeholder when it occurred with unusually high frequency or corresponded to a birthdate before 1900. All other dates with 100 or more occurrences appeared valid. We found there to be 113,394 records with a high-frequency placeholder DOB.

References

1. The National Alliance for Health Information Technology Report to the Office of the National Coordinator for Health Information Technology on Defining Key Health Information Technology Terms. 2008.
2. Fellegi IP, Sunter AB. A Theory For Record Linkage. J Am Stat Assoc 1969; 64(328): 1183–1210.
3. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. AMIA Annu Symp Proc 2002; 305–309.
4. Texas A&M Health Science Center Population Informatics Research Group. Record Linkage Basics [Internet]. Available from: http://research.tamhsc.edu/pinformatics/record-linkage-basics/
5. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: Issues, methods, and directions for the future. Health Serv Res 2010; 45(5 PART 2): 1468–1488.
6. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. AMIA Annu Symp Proc 2003; 259–263.
7. McCoy AB, Wright A, Kahn MG, Shapiro JS, Bernstam EV, Sittig DF. Matching identifiers in electronic health records: implications for duplicate records and patient safety. BMJ Qual Saf 2013; 22: 219–224.
8. Joffe E, Bearden CF, Byrne MJ, Bernstam E V. Duplicate Patient Records – Implication for Missed Laboratory Results. In: AMIA Annu Symp Proc 2012. p. 1269–1275.
9. Smith PC, Araya-guerra R, Bublitz C, Parnes B, Dickinson LM, Van Vorst R, Westfall JM, Pace WD. Missing Clinical Information During Primary Care Visits. JAMA 2005; 293(5): 565–571.
10. Joffe E, Byrne MJ, Reeder P, Herskovic JR, Johnson CW, Mccoy AB, Sittig DF, Bernstam E V. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. J Am Med Inform Assoc 2014; (21): 97–104.
11. Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a "basic" deterministic algorithm. Health Informatics J 2008; 14(1): 5–15.
12. Achimugu P, Soriyan A, Oluwagbemi O, Ajayi A. Record Linkage System in a Complex Relational Database – MINPHIS Example. Stud Health Technol Inform 2010; 160(MEDINFO 2010): 1127–1130.
13. Sauleau EA, Paumier J-P, Buemi A. Medical record linkage in health information systems by approximate string matching and clustering. BMC Med Inform Decis Mak. 2005; 5: 32.
14. Arellano MG, Weber GI. Issues in Identification and Linkage of Patient Records Across an Integrated Delivery System. J Healthc Inf Manag 1998; 12(3): 43–52.
15. Hillestad R, Bigelow JH, Chaudhry B, Dreyer P, Greenberg MD, Meili RC, Ridgely MS, Rothenberg J, Taylor R. Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System. 2008.
16. Yancey WE. Expected Number of Random Duplications Within or Between Lists. JSM 2010. 2010; 2938–2946.
17. Grannis SJ, Overhage JM, McDonald C. Real world performance of approximate string comparators for use in patient matching. Stud Health Technol Inform 2004; 107: 43–47.
18. McClellan MA. Duplicate Medical Records: A Survey of Twin Cities Healthcare Organizations. In: AMIA Annu Symp Proc 2009. p. 421–5.
19. Zech J, Husk G, Moore T, Kuperman GJ, Shapiro JS. Identifying homelessness using health information exchange data. J Am Med Inform Assoc 2015; 22(3): 682–687.
20. Social Security Administration. Social Security Death Master File (SSDMF) [Internet]. Available from: https://www.ssdmf.com
21. Social Security Administration. Requesting the Full Death Master File (DMF) [Internet]. Available from: http://www.ssa.gov/dataexchange/request_dmf.html
22. Download the Death Master File Free (SSDMF.info) [Internet]. Available from: http://ssdmf.info/download.html
23. United States Government Accountability Office. Social Security Death Data: Additional Action Needed to Address Data Errors and Federal Agency Access. 2013.
24. Healthix. Healthix: About Us [Internet]. Available from: https://services.lipixportal.org/HealthixPortal/Home/About
25. Levenshtein V. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Sov Phys Dokl 1966; 10(8): 707–710.
26. Finnell JT, Overhage JM, Grannis S. All Health Care is Not Local: An Evaluation of the Distribution of Emergency Department Care Delivered in Indiana. In: AMIA Annu Symp Proc. 2011. p. 409–416.
27. Finnell JT, Overhage JM, Dexter PR, Perkins SM, Lane KA, Mcdonald CJ. Community Clinical Data Exchange for Emergency Medicine Patients. In: AMIA Annu Symp Proc. 2003. p. 235–238.
28. Lynch B, Arends W. Selection of a surname encoding procedure for the Statistical Reporting Service record linkage system. Washington, D.C.: U.S. Department of Agriculture; 1977.
29. New York State Board of Elections. Freedom of Information Requests [Internet]. Available from: http://www.elections.ny.gov/FoilRequests.html