Design and implementation of a privacy preserving electronic health record linkage tool in Chicago

RECEIVED 11 September 2014
REVISED 25 February 2015
ACCEPTED 26 March 2015
PUBLISHED ONLINE FIRST 23 June 2015

Abel N Kho[1], John P Cashy[1,2], Kathryn L Jackson[1], Adam R Pah[1], Satyender Goel[1], Jörn Boehnke[3], John Eric Humphries[3], Scott Duke Kominers[4], Bala N Hota[5], Shannon A Sims[5], Bradley A Malin[6], Dustin D French[7], Theresa L Walunas[1], David O Meltzer[2], Erin O Kaleba[8], Roderick C Jones[9], William L Galanter[10]

ABSTRACT

Objective To design and implement a tool that creates a secure, privacy preserving linkage of electronic health record (EHR) data across multiple sites in a large metropolitan area in the United States (Chicago, IL), for use in clinical research.

Methods The authors developed and distributed a software application that performs standardized data cleaning, preprocessing, and hashing of patient identifiers to remove all protected health information. The application creates seeded hash code combinations of patient identifiers using a Health Insurance Portability and Accountability Act compliant SHA-512 algorithm that minimizes re-identification risk. The authors subsequently linked individual records using a central honest broker with an algorithm that assigns weights to hash combinations in order to generate high specificity matches.

Results The software application successfully linked and de-duplicated 7 million records across 6 institutions, resulting in a cohort of 5 million unique records. Using a manually reconciled set of 11 292 patients as a gold standard, the software achieved a sensitivity of 96% and a specificity of 100%, with a majority of the missed matches accounted for by patients with both a missing social security number and last name change. Using 3 disease examples, it is demonstrated that the software can reduce duplication of patient records across sites by as much as 28%.

Conclusions Software that standardizes the assignment of a unique seeded hash identifier merged through an agreed upon third-party honest broker can enable large-scale secure linkage of EHR data for epidemiologic and public health research. The software algorithm can improve future epidemiologic research by providing more comprehensive data given that patients may make use of multiple healthcare systems.

Keywords: record linkage, privacy protection, health information exchange

BACKGROUND AND SIGNIFICANCE

Because patients may receive care at multiple institutions within a region,[1–3] "single-site" studies may under- or over-represent key clinical features such as the number of affected patients, the severity of disease, or the extent of treatment. Integrating health records across care delivery sites is thus critical to developing a more comprehensive and accurate picture of health and healthcare delivery for the individual, and in aggregate may provide clearer insight into the health of particular populations.

Healthcare laws and federal incentives promoting the use of electronic health records (EHRs)[4,5] have led to a dramatic increase in electronic clinical data. Consequently, researchers and public health officials have expressed increased interest in efficient data linkage for use in cross-site health studies. However, linking EHR data across healthcare institutions requires a balance between data availability and privacy. The Federal Health Insurance Portability and Accountability Act (HIPAA), along with the more recent Omnibus rule, provide clear specifications on what constitutes protected health information (PHI) and outlines required processes and procedures for securing PHI from inappropriate use or access.[6,7]

Within the context of inter-institutional agreements for the sharing of PHI, some regions have implemented health information exchange (HIE) systems that provide healthcare providers with up-to-date clinical data on patients across institutions for safer and more coordinated care.[8] HIEs often use an Enterprise Master Patient Index service to assign a single common identifier for a patient based on the likely match of common patient identifiers (e.g., last name, first name, and date of birth) across institutions.[9–11] Prior work has demonstrated the potential for highly accurate match rates based on these common patient identifiers using linkage algorithms that minimize the need for human review.[11,12] However, operational HIEs sharing patient-level identifiers are still not widely implemented, due to sustainability, privacy, and security concerns, as well as other issues.[13] Outside the United States, some countries assign a single unique patient identification code, greatly simplifying the patient matching process.[14,15] To date, efforts to institute such an identifier in the United States have been unsuccessful and remain unlikely for the foreseeable future.[16]

Prior work suggests that use of secure encryption algorithms may potentially enable individual-level patient record linkage across institutions without the need to expose patient level identifiers,[17–21] but few successful real-world applications exist – particularly in large, urban settings, where multiple healthcare institutions compete for patients. In this paper, we describe the real-world implementation of a software application (Distributed Common Identity for the Integration of Regional Health Data – DCIFIRHD) that performs secure, cross-site aggregation, and linkage of EHR data for research using a standardized and distributed encryption algorithm. We implemented our application in a large metropolitan region (Chicago, IL, USA), aggregating over 5 million patients' clinical records across 6 healthcare institutions as

Correspondence to Dr Abel N Kho, Division of General Internal Medicine, Northwestern University Feinberg School of Medicine, 633 N. St Clair St., 20th floor. Chicago IL, 60611, USA; akho@nm.org

For numbered affiliations see end of article.

part of the HealthLNK research project. Here, we present and describe our methods and results, and explain the steps taken to address the privacy and security concerns arising from patient data aggregation.

METHODS AND MATERIALS

Setting

Chicago is served by 42 hospitals, 19 Federally Qualified Health Centers, a number of large academic medical centers, and myriad small practices. For this study, we partnered with 6 healthcare institutions: 4 large academic medical centers (Loyola University Medical Center, Northwestern Medicine, Rush University Medical Center, and University of Illinois at Chicago Medical Center); 1 large county healthcare system (Cook County Health and Hospital Systems); and a network of community health centers with multiple outpatient care sites (the Alliance of Chicago).

Each site provided pre-existing EHR data, either from local data warehouses or from medical record systems. The data proposed for use in the HealthLNK research project included none of the elements defined as PHI under federal HIPAA regulations other than patient ZIP codes (included to enable later mapping of disease distribution). Nevertheless, inclusion of ZIP code data constituted a limited data set and thus necessitated approval from all Institutional Review Boards (IRBs). We created a master IRB protocol and distributed this to each participating institution for modification and submission. Subsequently, IRB proposals and applications for expedited review were submitted at each participating healthcare institution. Final approval required coordination across all 6 sites and was achieved in 12 months. Data provision at each institution was led by a site lead investigator, with approval from the institution's IRB. Only records of patients aged 18–89 were included. The data were further restricted to only include records of visits to the participating institutions that occurred between January 1, 2006 and December 31, 2012. Specific diagnoses considered sensitive (such as Human Immunodeficiency Virus status) were specifically excluded, in accordance with our approved IRB protocol. We defined a set of acceptable data uses for research purposes and specifically excluded cross-institution comparison unless agreed upon in writing by each institution within the comparison.

For this particular phase of implementation, we limited analysis to structured EHR data elements because of uncertainty as to whether free text data from clinician notes could be reliably deidentified.[17] We identified a set of demographic and clinical variables and provided each participating site with a data dictionary specifying the source and format for each data element.

We reconciled demographic information across linked records to produce only one demographic record per patient, retaining the most frequent response and conservatively deleting responses in fields without a clear majority. For time-sensitive fields (e.g., insurance, ZIP code), data was retained from the site which a patient visited most recently. Clinical data was not reconciled between sites; all clinical data records were retained to represent the overall care of a patient.

Design of DCIFIRHD Data Infrastructure

We preprocessed EHR data at each site as a first step in the DCIFIRHD application (preprocessing steps detailed in Table 1, application workflow detailed in Figure 1). One individual not responsible for managing incoming data gave sites a common hashing seed consisting of a passcode and passphrase. Using the shared seed, the application created up to 17 512-bit hashes for each patient, containing combinations of patient's first and last names, date of birth, Social Security Number (SSN), and gender using SHA-512, a secure HIPAA-compliant cryptographic hash function developed by the National Security Agency (see Supplementary Information S1 for details on the usage of hash functions in medical informatics). Using a large number of hashes makes allowance for common mistyping errors and possible variation in availability of patient data across sites.

Each of the participating sites delivered the seeded HashIDs and demographic and clinical data of interest in a comma-separated file, via a SFTP server to a single host site (Northwestern University) previously agreed up on by all participants to act in the role of an honest broker of the data. A single authorized individual at the host site executed a component of the DCIFIRHD software program to merge records that share identical seeded HashIDs across institutions.

Matching algorithm

Record linkage methods simplistically divide into deterministic and probabilistic algorithms.[22] A deterministic algorithm applies sets of rules to common patient identifiers (e.g., first name, last name, date of birth) to assign matches between records, whereas probabilistic algorithms apply statistical methods that assign a probability of a match between records. Although probabilistic approaches may perform better than deterministic approaches, they generate a spectrum of matched probabilities, with intermediate matched probabilities often requiring human review, which may not always be available or desired due to confidentiality concerns. Moreover, probabilistic methods may not be directly applicable to hashed identifiers. Therefore, we developed a software application which encoded the best performing components of prior deterministic algorithms and assigned weights to variables and variable combinations to create a range of possible matches for future optimization.[23,24]

In this initial implementation, we matched records using the following simple deterministic algorithm. Two records from different sites are considered to "match" (i.e., be associated to the same patient) if they have the same:

1. Seeded HashID of (First Name + Last Name + Date of Birth),
2. Seeded HashID of (Date of Birth + SSN),
3. Seeded HashID of (Last Name + SSN), or
4. Seeded HashID of (Three Letter First Name + Three Letter Last Name + Soundex First Name + Soundex Last Name + Date of Birth + SSN).

These four matching criteria were selected because

- the necessary data are available in almost all medical records,
- it is straightforward to understand what constitutes a match, and
- false positives are very unlikely.

To build in flexibility for future adjustment of matching criteria by different users, we designed the software to match records using a deterministic *threshold crossing* model: for purpose of our initial implementation, we chose weights (1.01) specifically so that only the hashes of the four intuitive matches described above could reach 1, the threshold for declaring a match.

We included an option (combination 1) that generates a match in the absence of an SSN, given that one site was unable to extract SSNs for any of its patients, and another site only had collected SSNs for 28% of its patients. We did not allow social security number agreement alone to constitute a match, in order to avoid matching individuals who may be fraudulent users of others' social security numbers.

Table 1: Methods for data pre-processing and validity checks

Identifier	Preprocessing steps	Acceptable values
First Name	• Convert to lower case • Remove prefix and suffix • Remove all punctuation • Remove all digits • Replace multiple spaces with one space • Trim spaces	Not null
Last Name		
Birth Year		≥1900 and ≤ current year
Birth Month		≥1 and ≤ 12
Birth Day		≥1 and ≤ 31
Social Security Number	Remove all except digits	• Length must be 9 • Numbers with all zero in any of the 3-2-4 digit groups are invalid • Numbers 000, 666, or 900-999 in the 3-digit group are invalid • Numbers from 987-65-4320 to 987-65-4329 are reserved for advertisements and are invalid • 078-05-1120 and 123-45-6789 are invalid
Gender	• Convert to lower case • Female → F • Male → M • All else → X	

All combinations with a score above 1 were clustered using the R *igraph* package, so that matched PatientIDs are as assigned a final StudyID representing only one individual.

Since the purpose of the software was to link records across healthcare sites and enable cross-site research, we intentionally selected matching criteria to be conservative, minimizing the probability of a false match (Type-II error). Erroneously linking patient histories presents more danger in analysis than would inability to link all records of a single individual.

Performance Analysis

To validate the performance of our matching algorithm, we used a subset of patients recruited into the NUgene biobank study at Northwestern.[25] At the time of recruitment, NUgene participants complete an intake form entering key identifiers, including last name, first name, date of birth, social security number, and gender. Sometimes identifiers were misspelled or mistyped, just as in other real-world registration processes. We applied our hashing and matching programs to both the NUgene population as well as to the larger population within the Enterprise Data Warehouse at Northwestern, and manually checked those patients believed to be false positive or false negative matches. The Northwestern group ($n = 2\,336\,466$) was comprised of all patients between the ages of 18–89 seen at either Northwestern Medical Faculty Foundation (NMFF) or Northwestern Memorial Hospital (NMH); the NUgene group was comprised of all patients between the ages of 18–89 registered for the NUgene cohort study ($n = 11\,292$), the majority of whom ($n = 10\,975$) were also seen at either NMFF or NMH, as designated by having a unique patient identifier associated with the Northwestern group. Table 4 describes the demographics of the NUgene cohort. The remaining 317 NUgene patients were not seen at NMFF or NMH, and therefore were not expected to match with a patient from the Northwestern group.

After applying the hashing and matching algorithms, we categorized each match as either true match, a false positive, or a true negative match. NUgene patients who matched to the correct Northwestern patients were internally documented as only true positive matches; NUgene patients who did not match and were not expected to match with a Northwestern patient, as well as all correctly nonmatching pairs, were said to be true negative matches. The remaining pairs were manually checked against identifiers in the database to determine if they were true or false matches.

Data Privacy and Security

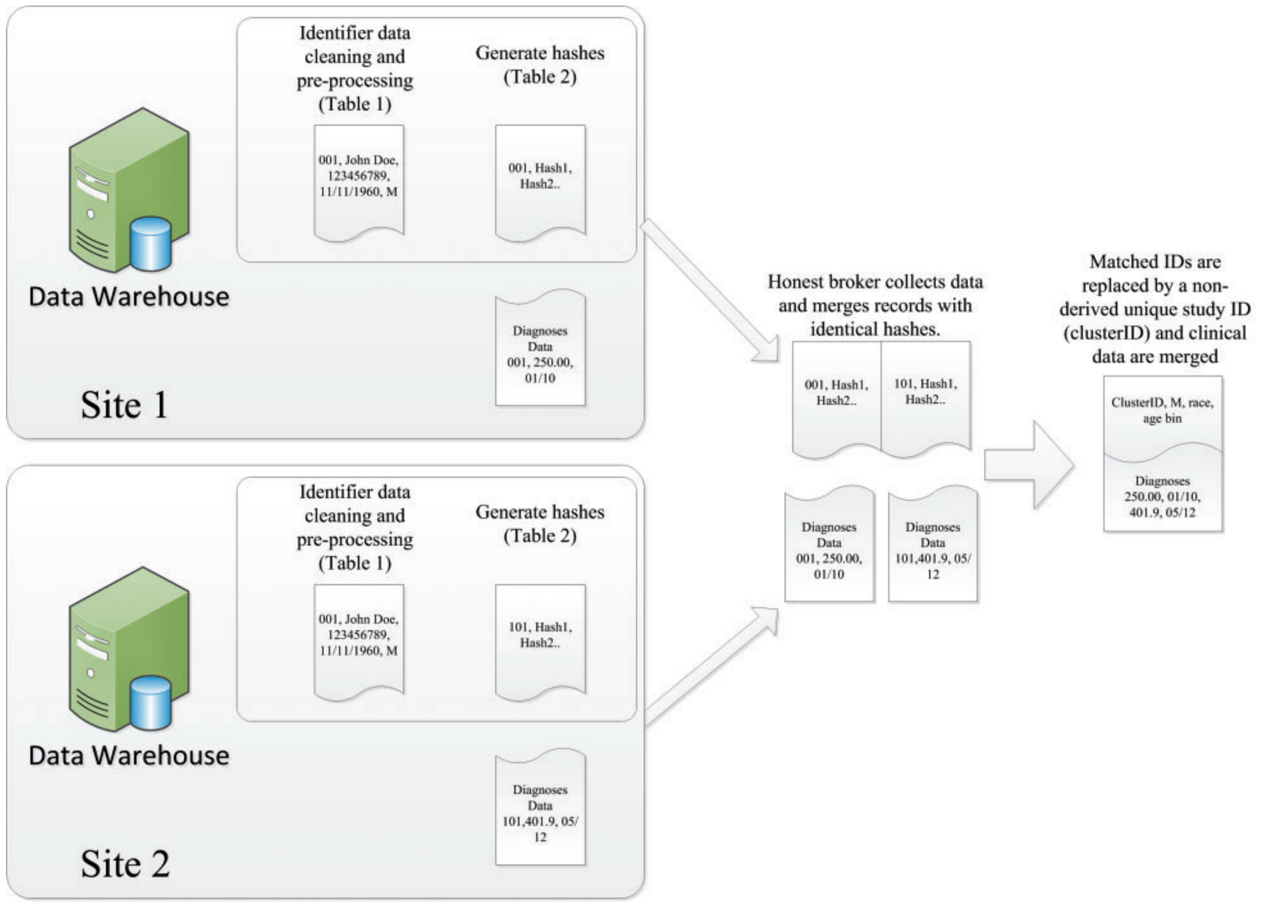
To protect against statistical attacks, any patient information not uniformly distributed in society (e.g., names and birthdays) was hashed in combination with other patient information, in order to create uniformly distributed information groups. (The complete set of available hash structures is listed in Table 2.) To protect against dictionary or “rainbow-table” attacks, we seeded the hash algorithm by requiring users of the hashing application to enter a passphrase and passcode distributed by a team member not involved in managing the inbound hashed files. Seeding the hashing algorithm ensured that only users knowing the seed could use the HashIDs to link individual records.

To ensure full HIPAA-compliance and to create an additional barrier to re-identification, after linkage, each cluster of site-specific PatientIDs was replaced with a nonderived StudyID for use in subsequent analyses.

Statistical methods

We used SAS (version 9.3, SAS Institute, Inc., Cary, NC, USA) to analyze the collected aggregate data. Minimum, maximum, and median percentages of demographics of the HealthLNK population were calculated to show the broad distribution of demographic variables

Figure 1: Workflow of the DCIFIRHD software application. At each hospital site EHR data is cleaned and pre-processed and then the patient identification information is hashed with a site-specific password and passcode. This hashed patient identification is sent along with diagnosis data to an honest broker site, where the hashed output is merged. Matched hashed identifiers are merged and the identifiers are then replaced with a unique study identification number.


Identifier data cleaning and pre-processing (Table 1)
Generate hashes (Table 2)
001, John Doe, 123456789, 11/11/1960, M
001, Hash1, Hash2..
Data Warehouse
Diagnoses Data 001, 250.00, 01/10
Site 1
Identifier data cleaning and pre-processing (Table 1)
Generate hashes (Table 2)
001, John Doe, 123456789, 11/11/1960, M
101, Hash1, Hash2..
Data Warehouse
Diagnoses Data 101,401.9, 05/12
Site 2
Honest broker collects data and merges records with identical hashes.
001, Hash1, Hash2..
101, Hash1, Hash2..
Diagnoses Data 001, 250.00, 01/10
Diagnoses Data 101,401.9, 05/12
Matched IDs are replaced by a non-derived unique study ID (clusterID) and clinical data are merged
ClusterID, M, race, age bin
Diagnoses 250.00, 01/10, 401.9, 05/12

represented by individual single site data compared with aggregated data. Frequencies and proportions were tabulated to describe the composition of the nondeduplicated and deduplicated record sets. We used chi-square tests for categorical variables and *t*-tests for continuous variables to compare demographic features between HealthLNK and US Census data.

Since deduplication of records across sites was a key aim of our application, we computed the numbers of patients diagnosed with three conditions known to be well captured in EHR data – Type II diabetes, myocardial infarction, and asthma[26] – before and after deduplication. We used ICD-9 codes (250.x0 or 250.x2 for Type II diabetes; 410.x1 for myocardial infarction and 493.x for asthma) to define the number of cases for each disease.

RESULTS

Table 4 shows the categories and counts of matches. We identified 8 false positive matches and 447 false negative matches. Sixty percent of the false negative matches came from patients with both a disagreement in their first or last name (either misspelling, hyphenation, pre/suffix, or nickname present) and a missing or invalid SSN. Fifteen percent of the false negative matches were due to patients having different last names in the two databases; 97% of these cases were women, suggesting that our algorithm is failing to match patients who change their last names after marriage. The sensitivity of the matching algorithm was 0.9569 and the specificity was 0.9999.

Participating institutions contributed data on patient demographics, diagnoses, procedures, vital signs, medications, and laboratory tests. Before matching and deduplication, HealthLNK included data associated to over 7 million PatientIDs. The matching and deduplication process reduced this population by 18%, to 5.3 million patients (i.e., clusters of matched PatientIDs). We received additional demographic and clinical data associated with 2.8 million of the 5.3 million patients aged 18–89 in our data who had at least one visit to a participating institution between 2006 and 2012. Further restriction to only those patients residing in zip codes within the city of Chicago provided a data set of 1 492 144 unique patients.

For three conditions known to be well captured in EHRs (Type II Diabetes, Asthma and Myocardial Infarction), the deduplication process reduced the measured number of cases by 24, 28, and 10.9%, respectively (Table 5).

Race and ethnicity demographic characteristics varied widely across HealthLNK sites (Figure 2). The percentage of Caucasian

Table 2: Hashes generated by DCIFIRHD application for an individual patient

Hash Name	Hash Type	Variables Hashed
FNLNDOB*	Normal hash	First name + last name + date of birth
LNFNDOB	Normal hash	Last name + first name + date of birth
FNLNBOD	Normal hash	First name + last name + date of birth (month and day switched)
FNSSN*	Normal hash	First name + SSN
LNSSN*	Normal hash	Last name + SSN
DOBSSN*	Normal hash	Date of birth + SSN
SSN*	Normal hash	SSN
3LFNLNDOB*	Only first three letters of first and last name used in hash	First name + last name + date of birth
3LLNFNDOB*	Only first three letters of first and last name used in hash	Last name + first name + date of birth
3LFNLNBOD	Only first three letters of first and last name used in hash	First name + last name + date of birth (month and day switched)
3LFNSSN	Only first three letters of first and last name used in hash	First name + SSN
3LLNSSN*	Only first three letters of first and last name used in hash	Last name + SSN
SXFNLNDOB*	Soundex	First name + last name + date of birth
SXLNFNDOB	Soundex	Last name + first name + date of birth
SXFNLNBOD	Soundex	First name + last name + date of birth (month and day switched)
SXFNSSN*	Soundex	First name + SSN
SXLNSSN*	Soundex	Last name + SSN

*Hashes used optimized match.
SSN: Social Security Number.

patients at participating sites ranged from 12.0% to 51.2% (median 26.2%), and the percentage of African American patients ranged from 17.5% to 57.5% (median 23.3%). Patients who reported Other, Declined to Answer, or gave no answer (Unknown) for race ranged in values, with 32.7% of patients having one of these categories at a single site. Comparing our combined data (both non-deduplicated and deduplicated) with 2010 US Census within the city of Chicago we made several observations. The software application produced a higher minority rate than that of the Census: 27.4% of our sample was Caucasian compared to 48.1% of the Census population. We had a higher percentage of Other/Declined/Unknown race compared to that of the Census (23.5% vs 14.0%, respectively). In our demographic sample 16.6% of patients reported a Hispanic ethnicity compared to 25.3% from the Census. All differences were significant at $P < .0001$.

Additionally, we compared the percent by age group of our sample, both at the site level and combined, to that of the 2010 Census (Figure 3). Similar to the race/ethnicity percentages, the percentages of patients in 5-year bins varied significantly across sites, especially in the youngest two categories (20–24 year olds and 25–29 year olds). The percentages of patients aged 20–24 and 25–29 ranged from 3.9% to 15.5% (median = 6.8%) and 7.5% to 15.9% (median = 10.5%), respectively. After combining and deduplicating, the proportion of patients by 5-year bins more closely approximated (although still remained statistically different at $P < .0001$) proportions from Census data, especially for age groups aged 55–59 and older. The difference in proportions of patients between EHR data and Census data was in the youngest adult age categories (ages 20–24), which may represent a tendency for young healthy individuals to seek care less often than older individuals.

DISCUSSION

While there has been a significant effort to implement secure HIE across the United States, many regions, including Chicago, still lack central HIEs that meet the needs and concerns of local healthcare systems, researchers, and networks. To overcome this problem, we successfully developed and implemented an IRB-approved approach using a distributed software application that enabled multiple, otherwise unaffiliated (and competing), healthcare institutions to aggregate longitudinal clinical data on approximately 5 million residents of a large United States city. The resulting database of linked, de-identified cross-site patient data (HealthLNK) contains records on a substantial proportion of the population of the city of Chicago. The cross-site database more closely approximated US Census demographics than did any single contributor institution. The Federal Meaningful Use definition creates a common standard for the collection and use of clinical data within EHRs and may improve data quality for (re-)use in research.[27] With the rapid increase in adoption of EHRs as part of routine clinical care, a secure method of aggregating records across care sites may present an efficient complement to prospective data collection for research or public health purposes.

Similar data aggregation projects are in place or in progress nationally. Vanderbilt University created the *synthetic derivative*, a

Table 3: NUgene population demographics compared with the wider HealthLNK population

	Deduplicated HealthLNK population	NUgene
Total Patients	$n = 1\,492\,144$	$n = 11\,292$
White	27.4	59
Black	35	9.9
Asian	3.3	2
American Indian/ Alaska Native	1.1	0.1
Pacific Islander	0.2	0.02
Other/Unknown/ Declined	23.5	25.6
Hispanic (Ethnicity)	16.6	5.5
Median Age (in years)	42	53

Table 4: NUgene to Northwestern Match Results

	Had Northwestern Match	No Northwestern Match
Matched	10 585 (TP)	8 (FP)
Did Not Match	477 (FN)	26 383 363 002 (TN)

Table 5: Numbers of patients identified with type II diabetes, myocardial infarction, and asthma by ICD9 codes and percent reduction after cross-institution deduplication

	Non Deduplicated	Deduplicated
Diabetes (type II only)	$n = 135\,779$	$n = 103\,177$; 24.0% reduction
Asthma	$n = 110\,640$	$n = 79\,563$; 28.0% reduction
Myocardial infarction	$n = 6049$	$n = 5384$; 10.9% reduction

deidentified extract of EMR data linked to an institutional biorepository, *BioVU*.[28] Like our work, the Vanderbilt study used a one-way hash function. However, the Vanderbilt study was single-site, using only a hash of the internal medical record number, and does not include location data. In New York City, the Primary Care Information Project, which formed the foundation for the Office of National Coordinator Regional Extension Center program, installed a particular EHR product in over 400 physician offices. Built atop this, the Hub Population Health System (the Hub) can run deidentified queries against medical practices to identify aggregate counts of at-risk patients, and deliver decision support alerts for specific conditions.[29] However, unlike HealthLNK, the Hub is presently unable to link individual patients across practices. The Informatics for Integrating Biology and the Bedside (I2B2) project has developed a broad and powerful set of open source tools for data aggregation, analysis, and reporting. This system was initially used for research in academic centers, but is now finding more widespread adoption.[30,31] Although the current suite of I2B2 tools does not link subject or patient identity across installations, we are actively exploring the potential for the robust I2B2 tools to complement or enhance our existing approach.[32,33]

Population-level health analysis can be conducted without any transfer of patient data by exporting computations to multiple sites and subsequently aggregating the results at a central server. This distributed query model has been used successfully in various healthcare settings where the participating institutions either are connected to single enterprise data warehouse or have minimal overlap in patient populations.[34–36] However, this method cannot connect records for a given individual who visits multiple healthcare institutions, and thus may lead to double-counting of individuals when aggregating health events across proximate institutions. A combination of distributed query methods with secure and encrypted identity disambiguation of the type we present here may provide an improved framework for accurate measurement of a population's health.

Our results show the benefits of combining and deduplicating patients across multiple sites in order to form a more accurate (although still imperfect) picture of the health of a population. Demographic characteristics of patient populations in our data set varied widely across site; this finding supports research suggesting that single site studies may not accurately represent the population, and can therefore misrepresent the severity of a particular condition. However, by merging and deduplicating patient records across institutions, we were able to mitigate this problem and achieve a more representative sample of the Chicago population. Our percent reduction in duplicated patients across sites ranged from 10.9 to 28%, consistent with our prior findings of patient overlap in a different urban setting.[37] The ability to capture data longitudinally across all 6 sites may provide a more accurate picture of the extent or progression of a disease.

LIMITATIONS

Our intent is to create a platform for cross-institutional cohort discovery, hypothesis generation, and estimation of regional disease burden. Although our data represents a cross-section of the population of the City of Chicago, including a large uninsured and underinsured population, it likely misses residents who rarely seek care (or seek care outside of participating institutions) – a limitation of most studies that rely exclusively on EHR-derived data.[38] Additionally, data captured in EHRs represents only a subset of the important health factors for any given patient, and our initial focus on structured data elements further restricts the analysis. The effectiveness of EHR data for determining the prevalence of a specific condition depends on both the categories of data used in the condition definition, and also on the availability of these data.[39] Careful consideration of and accounting for these sources of bias is critical for drawing accurate conclusions from EHR data and much additional research is needed.

Our current implementation of DCIFIRHD requires effort at each participating institution to extract data, pass it through the hashing application, and then upload it to a server at a central trusted site, which may not always be available in a region. As part of a wider research initiative (PCORnet) we are developing a platform for the distributed query of patient data across sites, minimizing the need for central

Figure 2: Race/ethnic breakdown of deduplicated HealthLNK patients and overall Chicago population.

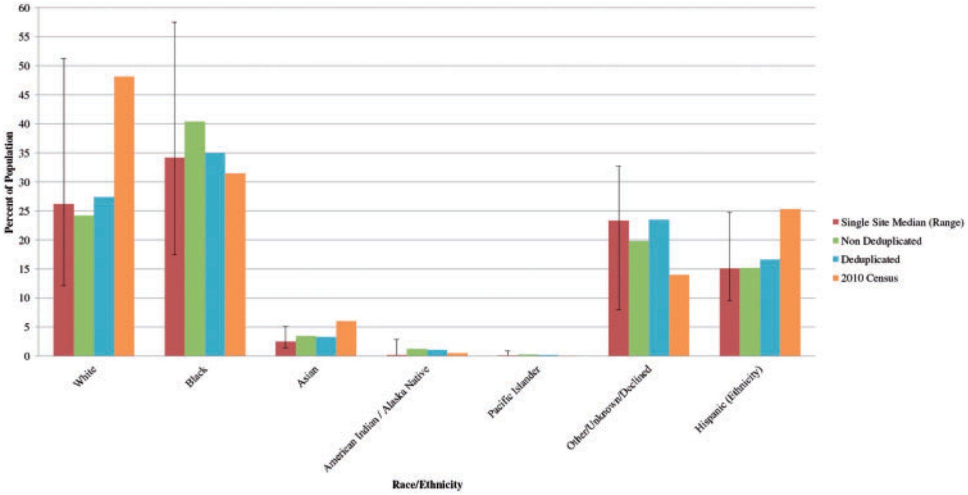
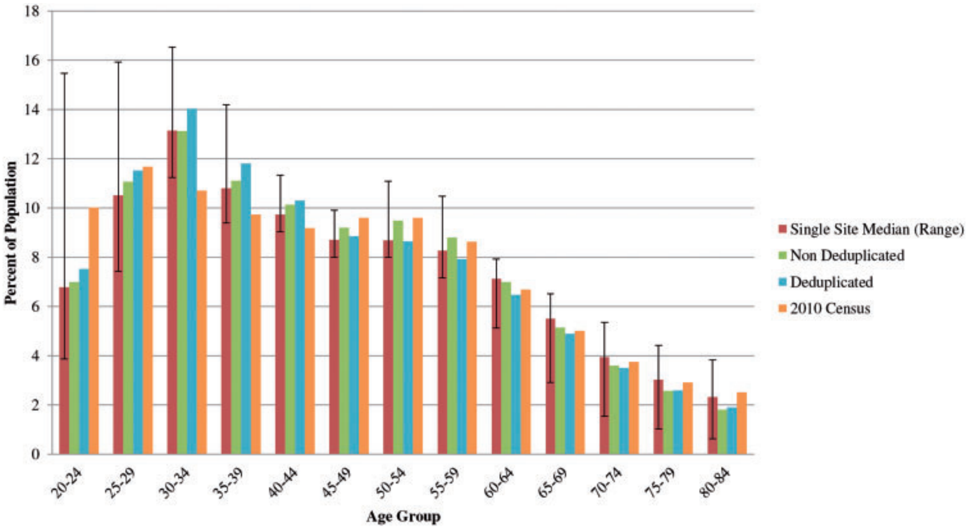


Figure 3: Age group comparison of Chicago HealthLNK and 2010 Census. The proportion of patients by 5-year bins was similar to that of the Census data, with the exception of underrepresentation in the youngest two adult age groups.



aggregation.[40,41] Matching of several million records required significant computational time, although use of blocking techniques and a dedicated computer server reduced match times from 10 days to 2 h.[22] We are currently exploring improvements or alternatives to the DCIFIRHD application, such as use of Bloom filter encodings, which may enhance the performance and security of privacy preserving record linkage.[42]

In this project we used 5 specific patient identifiers (last name, first name, date of birth, gender, and social security number) to generate weighted combinations for matching purposes. Our matching algorithm relied on an *a priori* set of weights derived from literature demonstrating the relatively high performance of specific combinations of these identifiers and was not optimized based on the underlying variability of the available patient features. The addition of other patient features (e.g., phone numbers or home ZIP codes) or use of optimized weights derived from the available variables may further enhance match performance. Matching performance may also be affected by the underlying demographics of the population, for example if there are large proportions of the population with common last names. Our manual validation utilized a population drawn from within one of the systems in our study and differed from other populations within our study catchment. Current work is focused on studying the effects of population demographics and differing weighting strategies on matching performance.

Our intent was to create a practical privacy-protecting means of integrating data for research purposes. Despite deidentification (by removal of HIPAA identifiers from data), the presence of other features (e.g., clinical data) presents a nontrivial risk of reidentification. Careful attention to quantifying reidentification risk, through formal statistical analysis by a qualified expert, should be considered before any data release.

OUTLOOK

Significant work remains to identify the strengths and limitations of EHR-based data to describe the health of populations. Prior studies indicate that EHR data can generate reliable estimates of disease prevalence for a focused geography, with nearly complete capture of the population within a single institution's EHR system or across a wider geography with sampling of diagnoses from primary clinics within a region.[26,43,44] Ongoing work within New York City to compare EHR-derived disease estimates with prospective data collected from a statistically representative sample of the population will likely yield important lessons.[29]

Researchers and public health departments face an environment of constrained resources and shrinking budgets. This stands in contrast to the rapid increase in the adoption of EHRs. Linked EHR data presents an opportunity to efficiently re-purpose existing clinical data to generate new insights and guide regional interventions for researchers and public health officials. Tools such as DCIFIRHD may provide a mechanism for privacy protecting, population level data aggregation for future epidemiologic research.

SOFTWARE ACCESS

Software developed through this project is available for research purposes. To access, please contact the lead author at akho@nm.org.

FUNDING

This work was supported in part by funding from the Otho S.A. Sprague Memorial Institute.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

A.N.K. led the design and development of the DCIFIRHD tool, convened the collaborative, and drafted and revised the paper. J.P.C. developed the initial clustering method for patient matching and performed initial data analysis. K.L.J. conducted data analysis and chart validation and software testing and drafted and revised the paper. A.R.P. conducted software testing. S.G. conducted matching optimization. J.B. wrote software code and performed software testing and revised the paper. J.E.H. provided weighting optimization scheme and revised the paper. S.D.K. provided expertise in matching and optimization and revised the paper. B.N.H. contributed to development of the collaborative and implemented the software. S.A.S. contributed to development of the collaborative and implemented the software. B.A.M. provided expertise on data deidentification and privacy preserving methods for record linkage. D.D.F. revised the paper. T.L.W. revised the paper. D.O.M. contributed to development of the collaborative and implemented the software. E.O.K. contributed to development of the collaborative and implemented the software. R.J. contributed to development of the collaborative. W.L.G. contributed to development of the collaborative and engaged student teams in the initial software development.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the exceptional work of Ariadna Garcia, Pravin Babu, Shazia Sathar, and Ulas Keles in drafting an initial version of the hashing and matching software. Kominers gratefully acknowledges the support of the Harvard Milton Fund.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Kho AN, Lemmon L, Commiskey M, Wilson SJ, McDonald CJ. Use of a regional health information exchange to detect crossover of patients with MRSA between urban hospitals. *JAMIA.* 2008;15:212–216.
2. Finnell JT, Overhage JM, Dexter PR, *et al.* Community clinical data exchange for emergency medicine patients. *Am Med Inform Assoc.* 2003;235.
3. Gichoya J, Gamache RE, Vreeman DJ, *et al.* An evaluation of the rates of repeat notifiable disease reporting and patient crossover using a health information exchange-based automated electronic laboratory reporting system. *Am Med Inform Assoc.* 2012;1229.
4. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *New Engl J Med.* 2010;363:501–504.
5. DesRoches CM, Charles D, Furukawa MF, *et al.* Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012. *Health Aff.* 2013;32:1478–1485.
6. Health Insurance Portability and Accountability Act of 1996. In: Rights OoC, ed, Office of Civil Rights. Rule number 42 U.S.C. Section Number 1320d-9. Enacted August 21, 1996. U.S. Congress. Washington, DC.
7. (2013) Omnibus Rule ("Final Rule"). In: Office of the Secretary Department of Health and Human Services, pp. 5565–5702. Document number 78 FR 5565. Washington, DC.
8. McDonald CJ, Overhage JM, Barnes M, *et al.* The Indiana network for patient care: a working local health information infrastructure. *Health Aff.* 2005;24:1214–1220.
9. Clayton P, Narus S, Huff S, *et al.* Building a comprehensive clinical information system from components. *Methods Inf Med.* 2003;42:1–7.
10. Arellano MG, Weber GI. Issues in identification and linkage of patient records across an integrated delivery system. *J Healthc Inf Manag.* 1998;12:43–52.
11. Weiner M, Stump TE, Callahan CM, Lewis JN, McDonald CJ. A practical method of linking data from Medicare claims and a comprehensive electronic medical records system. *Int J Med Inform.* 2003;71:57–69.
12. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. *Am Med Inform Assoc.* 2003;259.
13. Adler-Milstein J, Bates DW, Jha AK. Operational health information exchanges show substantial growth, but long-term funding remains a concern. *Health Aff.* 2013;32:1486–1492.
14. Morris AD, Boyle DI, MacAlpine R, *et al.* The diabetes audit and research in Tayside Scotland (DARTS) study: electronic record linkage to create a diabetes register. *BMJ.* 1997;315:524–528.
15. Wettermark B, Hammar N, MichaelFored C, *et al.* The new Swedish Prescribed Drug Register—opportunities for pharmacoepidemiological research and experience from the first six months. *Pharmacoepidemiology and drug safety* 2007;16:726–735.
16. Hillestad RJ, Bigelow JH, Chaudhry B, *et al.* Identity crisis: an examination of the costs and benefits of a unique patient identifier for the US health care system: *RAND*, Santa Monica, California. 2008.
17. Durham E, Kantarcioglu M, Xue Y, *et al.* Composite bloom filters for secure record linkage. *IEEE Trans Knowl Data Eng.* 2014;26:2956–2968.
18. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Med Inform Decis Mak.* 2009;9:41.
19. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. *J Biomed Inform.* 2013.
20. Quantin C, Bouzelat H, Allaert FAA, *et al.* How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. *Int J Med Inform.* 1998;49:117–122.
21. Churches T, Christen P. Some methods for blindfolded record linkage. *BMC Med Inform Decis Mak.* 2004;4:9.
22. Winkler WE. Overview of record linkage and current research directions. *Citeseer.* 2006.

23. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Am Med Inform Assoc.* 2002;305.
24. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. *J Am Med Inform Assoc.* 2012;19:e157–e161.
25. Wolf W, Doyle M, Aufox S, *et al.* DNA banking study in an ethnically diverse urban university hospital. *Am J Hum Genet.* 2003;73:423.
26. Violan C, Foguet-Boreu Q, Hermosilla-Perez E, *et al.* Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity. *BMC Public Health.* 2013;13:251.
27. Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med.* 2011;3:79re71.
28. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin Transl Sci.* 2010;3:42–48.
29. Kaye K, Singer J, Newton-Dame R, Shih SC. Health information technology and the primary care information project. *Am J Public Health* 2014;104: e8–e9.
30. Murphy SN, Weber G, Mendis M, *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *JAMIA.* 2010;17:124–130.
31. Natter MD, Quan J, Ortiz DM, *et al.* An i2b2-based, generalizable, open source, self-scaling chronic disease registry. *JAMIA.* 2013;20:172–179.
32. Weber GM, Murphy SN, McMurry AJ, *et al.* The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *JAMIA.* 2009;16:624–630.
33. Holve E, Segal C, Lopez MH, Rein A, Johnson BH. The Electronic Data Methods (EDM) forum for comparative effectiveness research (CER). *Med Care.* 2012;50:S7–S10.
34. Behrman RE, Benner JS, Brown JS, *et al.* Developing the Sentinel System—a national resource for evidence development. *New Engl J Med.* 2011;364: 498–499.
35. Curtis LH, Weiner MG, Boudreau DM, *et al.* Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf.* 2012;21:23–31.
36. Ohno-Machado L, Bafna V, Boxwala AA, *et al.* iDASH: integrating data for analysis, anonymization, and sharing. *JAMIA.* 2012;19:196–201.
37. Kho AN, Doebbeling BN, Cashy JP, *et al.* A regional informatics platform for coordinated antibiotic-resistant infection tracking, alerting, and prevention. *Clin Infect Dis.* 2013;57:254–262.
38. Green LA, Fryer GE, Yawn BP, Lanier D, Dovey SM. The ecology of medical care revisited. *N Eng J Med.* 2001;344:2021–2025.
39. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Chute CG. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform.* 2013;82:239–247.
40. Kho AN, Hynes DM, Goel S, *et al.* CAPriCORN: Chicago Area Patient-Centered Outcomes Research Network. *JAMIA.* 2014;21:607–611.
41. Collins FS, Hudson KL, Briggs JP, Lauer MS PCORnet: turning a dream into reality. *JAMIA.* 2014;21:576–577.
42. Kuzu M, Kantarcioglu M, Durham EA, Toth C, Malin B. A practical approach to achieve private medical record linkage in light of public resources. *JAMIA.* 2013;20:285–292.
43. VanWormer JJ. Methods of using electronic health records for population-level surveillance of coronary heart disease risk in the Heart of New Ulm project. *Diabetes Spectr.* 2010;23:161–165.
44. Tomasallo CD, Hanrahan LP, Tandias A, *et al.* Estimating Wisconsin Asthma Prevalence Using Clinical Electronic Health Records and Public Health Data. *Am J Public Health* 2014;104:e65–e73.

AUTHOR AFFILIATIONS

[1]Department of Medicine, and Center for Health Information Partnerships, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

[2]Department of Veterans Affairs, Pittsburgh PA

[3]Department of Economics, University of Chicago, Chicago, IL, USA

[4]Society of Fellows Department of Economics, Business School, Program For Evolutionary Dynamics, and Center for Research on Computation and Society, Harvard University, Cambridge, MA, USA

[5]Department of Medicine, Rush University Medical Center, Chicago, IL, USA

[6]Department of Biomedical Informatics, School of Medicine, and Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, Nashville, TN, USA

[7]Center for Healthcare Studies and Department of Ophthalmology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[8]Alliance of Chicago Community Health Services, Chicago, IL, USA

[9]Formerly of Chicago Department of Public Health, currently at Ann and Robert H. Lurie Children's Hospital, Chicago, IL, USA

[10]University of Illinois Hospital and Health Sciences System, Chicago, IL, USA