Privacy-Preserving Record Linkage to Identify Fragmented Electronic Medical Records in the All of Us Research Program

Abel N. Kho[1](✉), Jingzhi Yu[1], Molly Scannell Bryan[2,3], Charon Gladfelter[1], Howard S. Gordon[2,3], Shaun Grannis[4], Margaret Madden[1], Eneida Mendonca[4], Vesna Mitrovic[1], Raj Shah[5], Umberto Tachinardi[4], and Bradley Taylor[6]

[1] Northwestern University, Evanston, IL 60611, USA
akho@nm.org
[2] University of Illinois at Chicago, Chicago, IL 60612, USA
[3] Veterans Affairs Medical Center, Chicago, IL 60612, USA
[4] Regenstrief Institute, Indianapolis, IN 46202, USA
[5] Rush University, Chicago, IL 60612, USA
[6] Medical College of Wisconsin, Milwaukee, WI 53226, USA

Abstract. As part of a national study in the United States to recruit one million Americans (All of Us Research Program) and their Electronic Health Record data, we set out to determine the degree to which care is fragmented across a sample of participating health provider organizations (HPOs). We distributed a previously validated Privacy-Preserving Record Linkage (PPRL) tool to participating sites to generate a unique set of keyed encrypted hashes for seven participating institutions across three States in the Upper Midwest of the U.S. An honest broker received the resulting encrypted hashes to identify patients with the same encrypted hashes shared across any combination of more than one institution as a proxy for patients receiving care across institutions. Out of 5,831,238 individuals, we identified 458,680 patients with data at more than one institution. Care fragmentation varied significantly by State and by Institution ranging from 6.1% up to 32.7%. Patients with fragmented care were more likely to be black (11.8% vs 10.8%), and slightly older (Median birth year 1968 vs 1969) compared with patients receiving care at only one participating institution. In contrast, patients who maintained an address in a warmer state ("snowbirds") were the least likely to be black (7.5%) of all study groups. We identified conflicting or inconsistent demographic information in 49.1% of patients with care fragmentation compared with 5.6% of patients without care fragmentation. Privacy-preserving record linkage can be an effective means to identify populations with care fragmentation and poor data quality for focused clinical and data improvement efforts.

Keywords: Record linkage · Privacy preservation · Ecology of care

P. Cellier and K. Driessens (Eds.): ECML PKDD 2019 Workshops, CCIS 1168, pp. 79–87, 2020.
https://doi.org/10.1007/978-3-030-43887-6_7

1 Introduction

1.1 The All of Us Research Program

In 2016, the United States Congress launched the Precision Medicine Initiative (PMI) with $200M in funding in order to advance the development and application of individualized care based on a person's unique lifestyle, environment, and biology. A core foundation of the Precision Medicine Initiative, the All of Us Research Program (AoURP) was initially allocated $130M to create a national cohort of over one million Americans broadly representing the rich diversity of the U.S. population. Widespread adoption of Electronic Health Records (EHRs) across the U.S. was identified early in the design of the AoURP as a potentially rich source of data on patient health conditions and treatments.

The AoURP designated and funded over 40 Health Care Provider Organizations (HPOs) nationally to serve as recruitment centers. As part of the enrollment process, HPOs are required to send EHR data for consented participants to the AoURP Data and Research Center after verifying the identity of the participant and standardizing the EHR data into the Observational Medical Outcomes Partnership (OMOP) data model [1].

1.2 Data Fragmentation Across Institutions

However, healthcare in the United States is delivered across a wide variety of care settings and lacks the availability of a universal patient identifier. As a result, patient records may be fragmented across each location where a patient receives care, and unavailable both for patient care, but also for aggregation for research purposes such as those envisioned by the AoURP. Health Information Exchanges (HIEs) emerged as a means to address data and care fragmentation, and use a master patient index to consistently track the same patient across different care settings but are not available in many regions in the United States, or have struggled to remain financially viable [2]. Some EHR systems can link health records across institutions which use the same EHR system for routine clinical care, but do not currently integrate these data together for research purposes [3]. Because the AoURP aims to aggregate as much information about a participant as possible, investigators at participating HPOs questioned how often participants might receive care at a different care site than the HPO at which they might be enrolled. But without cross-institutional data sharing agreements in place to allow for patient identifiers to be shared across sites, and with many HPOs not part of HIEs, an alternate mechanism to link the same patient record across sites was needed.

1.3 Prior Use of Privacy-Preserving Record Linkage

We previously developed software to generate keyed hashes of patient identifiers that is fully compliant with HIPAA de-identification methods and could enable privacy preserving record linkage across AoURP HPOs [4]. A key finding of the initial linkage across seven healthcare institutions was the significant degree of data fragmentation across care sites ranging from 11 to 28% over a several year span. We subsequently demonstrated similar care fragmentation for specific populations including patients with diabetic ketoacidosis [5] and systemic lupus erythematosus [6]. Notably, we identified worse clinical outcomes for patients with fragmented care vs those without care fragmentation, a finding consistent across each condition we studied. Relevant to a cohort study such as the AoURP, we linked individual data between a longitudinal cohort study (the Multi-Ethnic Study of Atherosclerosis or MESA) and EHR data in our region, and identified gaps in data coverage in both sources of data even for conditions as seemingly obvious as a myocardial infarction [7]. The combination of both multi-institutional EHR data and prospectively collected data for a cohort study created a more complete set of data for a given research study participant than any one source alone.

With this background and with the endorsement of the AoURP Steering Committee, we set out to use our previously validated privacy preserving record linkage method to determine how often patients receive care across participating AoURP institutions within a geographically proximate region of three adjoining States in the Upper Midwest of the United States. Our goal was to identify the degree of data fragmentation across AoURP sites in order to determine whether to pursue additional data sources to fully characterize research cohort participants.

2 Methods

We submitted and received approval for this study of de-identified patient level data from the Northwestern Institutional Review Board. We defined the study population as patients seen at participating institutions from January 1, 2011 through May 1, 2018. We excluded patients aged 90 or over as of April 30, 2018 to comply with HIPAA Safe Harbor restrictions on age. Seven institutions participated in the study, three based in the State of Wisconsin, three in Illinois, and one in Indiana which had access to data from the statewide Health Information Exchange.

At a kickoff meeting hosted in Wisconsin and through subsequent discussion, all participating institutions agreed upon a common data dictionary to define key demographic and clinical fields to extract along with keyed hashes to uniquely identify a patient (Table 1).

Table 1. Key data fields extracted by institutions to characterize the demographics and diagnoses of the study population.

Demographics	Diagnoses
Birth year	Year
Gender	Encounter type (e.g. Inpatient, Emergency Department)
Race	Terminology (ICD9, ICD10, SNOMED)
Ethnicity	Primary diagnosis (yes or no)
Insurance status (most recent)	
3 digit ZIP code	

We distributed an executable software program with known matching performance characteristics as described in our prior publication. Participating institutions installed the software locally, and collectively identified a key to be used to hash the patient identifiers that was kept separate from the group aggregating the data on behalf of the study. Using a combination of last name, first name, date of birth, and social security number (where available), sites encrypted multiple concatenated combinations of these features in order to generate up to 17 secret key encrypted hashes. The central site (Northwestern University) team, acting as an honest broker, received the keyed hashes, along with attached demographic and clinical data as defined by the study data dictionary.

We matched the data across the participating institutions to evaluate the degree of care fragmentation within each State, across States, and across all institutions. Because we included three digit ZIP codes in our data set (which is a broad enough level of geography to still be considered de-identified by HIPAA), we could identify the subpopulation of patients who also have a home address in a considerably warmer region of the United States (the States of Alabama, Arizona, Arkansas, California, Florida, Georgia, Louisiana, Mississippi, New Mexico, and Texas) during the winter months (colloquially referred to as "snowbirds"). We analyzed the differences in demographics between those patients who have fragmented and non-fragmented care, as well as between "snowbirds" and those less capable of escaping the cold winter weather in the Upper Midwest.

Several data fields required additional translation between data terminologies in order to be consistent for further analyses. Diagnoses in EHRs arrived as ICD9, ICD10, and SNOMED codes and required significant re-mapping to a consistent and common terminology, in this case MS-DRG-CM. We identified data quality issues including missing data and data which conflicted across sites.

Due to of the large size of the total number of records, we conducted analyses using Python 3.7 with *pandas* and *numpy* packages.

3 Results

In total, we received records on 5,831,238 individuals across the three states. We identified 458,680 patients with data at more than one institution. Table 2 describes the demographics for our total study population, and the populations of patients with non-fragmented care, fragmented care, and "snowbirds". Demographics information that was declined or missing at the point of recording, as well as patients that had conflicting demographics information from multiple patient records were given the same category. Considerable patient race information were found to be conflicted or missing, and as high as 44.8% in fragmented patients.

Table 2. Demographics of the total study population, patients with non-fragmented care, fragmented care, and "snowbirds".

		Total n = 5,831,238	Non-fragmented n = 5,372,558	Fragmented n = 458,680	Snowbirds n = 79,701
Age	Median birth year	1969	1969	1968	1964
Gender	Female	46.8%	46.2%	54.3%	48.3%
	Male	43.6%	44.0%	45.7%	44.2%
	Other	8.9%	9.7%	0.0%	7.4%
	Conflicted or missing	0.7%	0.0%	8.3%	0.1%
Race	White	58.1%	60.0%	35.4%	62.1%
	Other	15.5%	16.2%	6.9%	14.6%
	Black or African American	10.9%	10.8%	11.8%	7.5%
	Declined or missing or conflicted	12.1%	9.4%	44.8%	11.9%
	Asian or other Pacific Islander	2.5%	2.6%	1.0%	3.4%
	Hispanic or Latino	0.5%	0.6%	0.0%	0.2%
	American Indian/ Alaskan Native	0.4%	0.4%	0.1%	0.3%
Ethnicity	Not Hispanic or Latino	89.5%	89.9%	84.3%	91.7%
	Hispanic or Latino	6.6%	6.8%	4.4%	4.9%
	Conflicted or Missing	3.9%	3.3%	11.4%	3.4%

3.1 Patient with Care Fragmentation

The distribution of patients with care fragmentation was unevenly distributed by State and Institutions. The percent of patients with care fragmentation differed by state ranging from 4.9% to 11.7% (Table 3).

Table 3. Care fragmentation by State.

State	Counts	Total	% of fragmented patients within state
Illinois	328,544	2,811,941	11.7%
Wisconsin	108,996	2,240,339	4.9%
Indiana	88,423	846,241	10.4%

The percent of patients with care fragmentation varied by site ranging from 6.1% to 32.7% (Table 4).

Table 4. Fragmentation by care site.

Site	Counts	Total	% of fragmented patients within site
Northwestern University	253,543	1,931,853	13.1%
Rush University Medical Center	213,946	653,358	32.7%
University of Illinois at Chicago	150,918	516,593	29.2%
University of Wisconsin Madison	72,561	636,585	11.4%
Medical College of Wisconsin	63,252	1,031,119	6.1%
Marshfield Clinic	46,952	646,404	7.3%
Regenstrief Institute	88,423	846,241	10.4%

3.2 Data Quality Issues

We identified a significant percentage of records with conflicting demographic information, with the majority of discrepancies for race (Table 5 and Fig. 1).

Table 5. Number of records with conflicting demographic information by feature.

Race	Ethnicity	Gender	Birth year
466,302	59,888	39,547	3,373

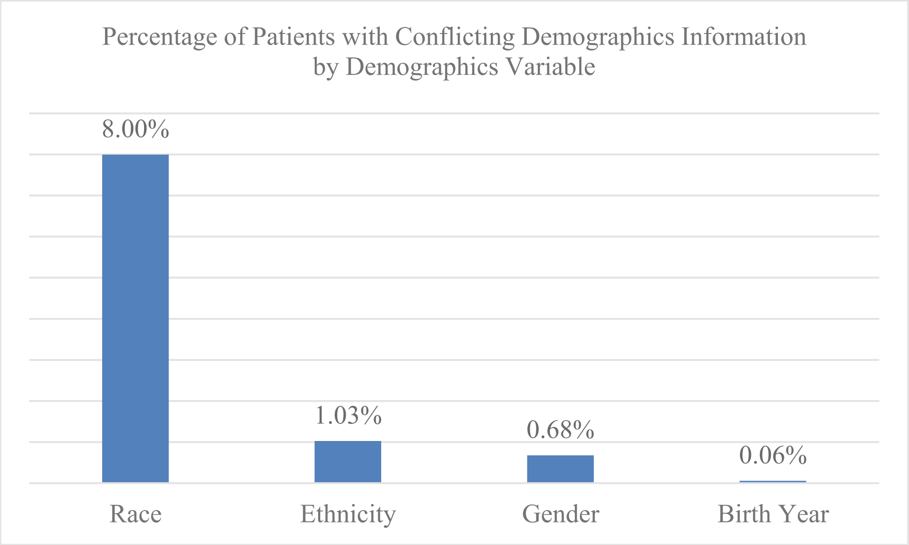

Fig. 1. Most common demographic features with conflicting information.

Patients with care fragmentation had conflicting information at a much higher rate than those without care fragmentation (49.1% vs 5.6%, Table 6)

Table 6. Counts and percentage of patients with conflicting information by fragmentation status.

	# of patients w/conflicted information	# of patients w/o conflicted information	Percentage of patients with conflicted information
Patients that are fragmented within state	225,313	233,367	49.1%
Patients that are not fragmented within state	301,700	5,070,858	5.6%

3.3 Geographic Analysis to Characterize “Snowbirds”

Patients with home addresses (by 3 digit ZIP code) varied by State (Table 7) and by Institution (Table 8).

Table 7. Snowbirds by State

State	Counts	Total	%
Illinois	49,996	2,811,941	1.78%
Wisconsin	26,882	2,240,339	1.20%
Indiana	3,025	846,241	0.36%

Table 8. Snowbirds by Institution

Site	Counts	Total	% of snowbirds out of total patient population
Northwestern University	38,846	1,931,853	2.01%
Medical College of Wisconsin	10,825	1,031,119	1.05%
University of Wisconsin Madison	8,748	636,585	1.37%
Rush University Medical Center	7,763	653,358	1.19%
Marshfield Clinic	7,449	646,404	1.15%
University of Illinois at Chicago	4,333	516,593	0.84%
Regenstrief Institute	3,025	846,241	0.36%

4 Discussion

We used a previously validated privacy preserving record linkage method based on generating keyed hashes of patient identifiers to identify the degree of data fragmentation across a sample of HPOs within the AoURP. Data fragmentation varied from 3.6% to 32.7% with the greatest percentage at sites within IL and the more population-dense Chicago-based institutions. Consistent with prior studies, patients with care fragmentation were more likely to be black and younger. In contrast, patients with the ability to “snowbird” to warmer climes were least likely to be black.

A common problem with linking data across sites is the issue of conflicting data, e.g. one site lists race as “Caucasian” and another site may list race as “unknown”. We identified conflicting demographic information for 49.1% of those patients receiving care at more than one institution. Even in patients who receive care at the same institution, demographic information captured over time had conflicting information 5.6% of the time. Race was the most common demographic feature with conflicting information.

There are several limitations to our study. Our study only included a small number of institutions within each State (those that participate in the AoURP), e.g. in the Chicagoland area alone there are over 40 distinct healthcare institutions. Thus our estimates of data fragmentation are likely significant underestimates. Because we focused on sharing only demographic features compliant with HIPAA de-identification criteria, we could not evaluate more specific geographic features beyond 3 digit ZIP code. Geographic features such as home address are likely to change over time for patients as they move, or to be collected in non-standardized fashions, and could be a common feature at risk of conflicting across care sites. We defined “snowbirds” as having a listed address in the EMR from one of several warm winter month states. However, many “snowbirds” may only list their local address so our estimates likely significantly underestimate the population size.

Our study demonstrated the utility of a privacy-preserving record linkage tool to characterize care fragmentation across institutions spanning three contiguous States. Our findings are consistent with prior findings that care fragmentation is associated with at-risk populations but also demonstrates a novel association with significantly higher proportion of conflicting data. We have ongoing work to analyze the differences in insurance status and diagnoses across the study population and to use study results to guide strategies to capture more comprehensive clinical data for patients enrolled in the All of Us Research Program.

Acknowledgements. This study was funded under a supplement to NIH award 1 U2C OD023196-01 (All of Us Research Program Data and Research Center). During the study period, authors EM and UT were faculty at the University of Wisconsin, Madison.

Statement of Conflicts. ANK is an advisor to Datavant, Inc., which supports Privacy-Preserving Record Linkage software. Datavant acquired Health Data Link, Inc. which ANK co-founded based on this earlier version of the software.

References

1. The OMOP data model. https://www.ohdsi.org/data-standardization/the-common-data-model/. Accessed 14 June 2019
2. Holmgren, A.J., Adler-Milstein, J.: Health information exchange in US hospitals: the current landscape and a path to improved information sharing. J. Hosp. Med. **12**(3), 193–198 (2017)
3. Epic Care Everywhere. https://www.epic.com/careeverywhere/. Accessed 14 June 2019
4. Kho, A.N., Cashy, J.P., Jackson, K.L., et al.: Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J. Am. Med. Inform. Assoc. **22**(5), 1072–1080 (2015)
5. Mays, J.A., Jackson, K.L., Derby, T.A., et al.: An evaluation of recurrent diabetic ketoacidosis, fragmentation of care, and mortality across Chicago, Illinois. Diabetes Care **39**(10), 1671–1676 (2016)
6. Walunas, T.L., Jackson, K.L., Chung, A.H., et al.: Disease outcomes and care fragmentation among patients with systemic lupus erythematosus. Arthritis Care Res (Hoboken) **69**(9), 1369–1376 (2017)
7. Ahmad, F.S., Chan, C., Rosenman, M.B., et al.: Validity of cardiovascular data from electronic sources: The Multi-Ethnic Study of Atherosclerosis and HealthLNK. Circulation **136**(13), 1207–1216 (2017)