Developing Methods for Standardized Patient Cohort Deduplication Across the Carolinas Collaborative, A Regional Data Research Network.

Robert L Bradford, BS, CCIS[1], Jordan Brittingham, MSPH[2], Steven Evans, BS[3], Brian Ostasiewski, BS[4], Wonhee Yang, MSC[1], Kira C Bradford, PhD[1], Ashok Krishnamurthy, PhD[1]

[1]University of North Carolina at Chapel Hill, Chapel Hill, NC; [2]Health Sciences South Carolina, Charleston, NC; [3]Duke University, Durham, NC; [4]Wake Forest Baptist Medical Center, Winston-Salem, NC

Description of the Problem

Throughout a patient's medical history it is not uncommon for patients to change providers or hospitals; however, this can lead to patient overlaps and duplication within data sharing networks composed of multiple healthcare organizations. Within clinical data research networks (CDRN), patient overlaps could cause patients to be double counted, missed during analysis, or provide incomplete data affecting study outcomes. For example, the state of Massachusetts has evaluated a 31% overlap of their patient population across at least two hospitals[1]. When determining study feasibility, knowing that approximately 30% of the patients returned in a CDRN may be patient overlap could prevent the study from achieving statistical power. Establishing methods to determine baseline and study specific patient overlaps will empower investigators to improve study design and analysis.

In a regional CDRN (Carolinas Collaborative) composed of Health Sciences South Carolina, UNC Chapel Hill, Duke University, and Wake Forest Baptist Medical Center, there is a hypothesized patient overlap due to site proximity. All four academic healthcare sites reside within 220 miles of each other, with two sites being within 10 miles (UNC, Duke) leading to significant acute care overlap. As a service to researchers utilizing the Collaborative, the network sought to investigate methods and develop procedures to allow for deduplication of patients across institutions.

Methods

The onset of the project established a workgroup incorporating technical and regulatory leaders from the network to collaborate and develop a procedure for deduplication. For regulatory simplicity, the procedure would require no sharing of identified patient data, and rely solely on computed hashes where patient values are combined and encrypted. For initial testing, the workgroup chose four patient demographic values (See **Figure 1**) to evaluate the procedure's ability to deduplicate without sharing direct patient identifiers.

In order to evaluate the efficacy of the deduplication process, we identified a gold standard for testing by utilizing internal data from Electronic Medical Record (EMR) crossovers. Each site in the network underwent a crossover from a legacy EMR to a new EMR, causing a patient's record to exist in both the legacy data set and active data set from the current EMR. Each site maintains an existing, curated linkage(s) between legacy and active data such as simple matches on medical record numbers (MRN), shared database keys, or imputed master patient indexes. These curated linkages allow for direct comparison of collision results (matching hash values) and perform a direct evaluation of the number of successful and failed collisions (misses). In other words, this methodology provides a denominator of expected results from the curated linkage to be compared with the number of successful collisions.

Results

After the initial application of our deduplication procedure on internal data, sites determined that additional steps were necessary to standardize data and remove patient records that would consistently fail (e.g. missing values, erroneous SSN) to ensure maximum linkage. Sites were able to obtain a matching

rate of ~90% on expected matches between legacy and active data after standardization. The most prevalent cause for misses occurred on differences between first names (See **Table 1**). The evaluation of the internal collisions demonstrated the necessity for very specific standardization of each value in the hash to achieve accurate collision rates. Due to the nature of applying a hash algorithm, single character differences can result in vastly different hash values.

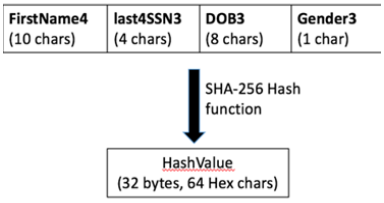

Figure 1: Carolinas Collaborative Hash Algorithm Definition

Data Issue	Example Value
First Name Includes Middle Initial	John C
First Name Includes Alphanumeric Characters	Mary-Joe Anne-Marie
Erroneous Last 4 SSN	-9999 , -0000 , -1234
Null Values	*Null*
Control Characters	Carriage Return
Date Format Differences	DD/MM/YYYY vs. MM/DD/YY

Table 1: Data Discrepancies Between EMR Databases

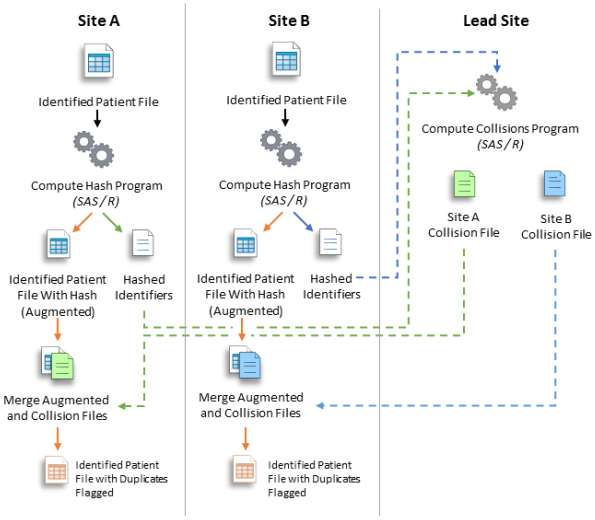

Figure 2: Overview of Deduplication Process for the Carolinas Collaborative

Establishing baseline success rates of the chosen hashing mechanism using local data provided a basis for broad testing across the network. Low collision results (under 1%) with UNC and Duke elucidated specific differences between sites that were not discovered during internal testing. Initial full population collisions were performed utilizing a standardized process (**Figure 2**) to identify base population overlap across all sites. The largest overlap occurred between UNC Health Care and Duke University (21% of UNC population at Duke, 31% of Duke population at UNC), and 0.1% network overlap of patients across all four institutions.

Discussion

The ability to efficiently and accurately deduplicate patients across a CDRN can be invaluable and will have an impact in designing studies for rare disease cohorts. Initial endeavors within the Carolinas Collaborative demonstrate the ability of this procedure to identify patient collisions with high accuracy. However, this method is limited to an all or nothing approach. Additional work is required to determine if additional identifiers are needed or an additional step for probabilistic matching based off of partial matches is needed.

Conclusion

The results presented demonstrate the Collaborative's ability to identify patient population overlaps within a network which improved with the implementation of standardized data formatting and processes. The methodology and programs developed will ideally be applicable to other institutions and data networks. Connecting patients across healthcare systems in combination with data harmonization within CDRN's will enable researchers to examine a patient's complete disease case history.

Attendee's Take-away Tool

Understanding of tools and procedures being developed to perform deduplication of patients within a CDRN (See **Figure 2**).

References

1. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010;170:1989–95.