Building Longitudinal Records for Research while Preserving Patient Privacy

Thomas W. Carton, PhD, MS\textsuperscript{1}, Andrew J. Gregorowicz, MS\textsuperscript{2}, Toan Ong, PhD\textsuperscript{3}  
\textsuperscript{1}Louisiana Public Health Institute, New Orleans, LA;  \textsuperscript{2}The MITRE Corporation, Bedford, MA;  \textsuperscript{3}University of Colorado, Anschutz Medical Campus, CO

Abstract

Conducting patient centered outcomes research typically involves a process of assembling information on the individuals being studied. When answering research questions, there is often the need to build a longitudinal record, which includes a comprehensive clinical summary of the patient. Generating a longitudinal record requires gathering all relevant information across multiple information systems and organizational boundaries. To successfully gather this information, researchers must be able to correctly identify patients across organizations and maintain their privacy in the process. This panel will discuss the approaches taken by the Childhood Obesity Data Initiative (CODI) to address this issue. First, the panel will discuss the use of Privacy Preserving Record Linkage (PPRL) techniques. This allows the linkage of patient records across organizational boundaries without disclosing personally identifiable information (PII) between the parties. This will include a brief primer on fuzzy matching with PPRL. Then the panel will review how these concepts have been applied to queries on PCORnet, the infrastructure leveraged by CODI. In closing will be a discussion and question and answer session on how CODI is building longitudinal records across organizations to answer research questions related to childhood obesity interventions.

Learning Objectives:

- What PPRL is, how it works, and how it has been adapted to support a pediatric population
- How PPRL has been applied to support research at regional and national scale

Introduction

Performing research in the area of childhood obesity benefits from having a complete picture of a child’s care, from the healthcare they receive in a clinical setting to interventions in the community setting. However, building a dataset of longitudinal records for a pediatric population presents a number of challenges. Regardless of patient age, linkage of health records often involves exposing personally identifiable information (PII) outside of an organization to a third party that can attempt to match individuals. At a regional scale, a small collection of organizations may work with a third party to provide record linkage services. Unfortunately, scaling this approach beyond a small number of organizations is difficult as all the organizations need to place their trust in a single third-party. Further, this third party becomes a larger target for identity thieves as it stores a greater amount of PII. An alternative approach to this problem is the use of Privacy Preserving Record Linkage (PPRL) techniques. This technology allows for record linkage by obfuscating PII prior to transmitting it beyond organizational boundaries. This reduces the risk profile for and trust requirements in a third-party performing record linkage.

This panel will explore approaches for achieving PPRL using techniques that allow for probabilistic matches. It will then shift to applications of these approaches. This includes a presentation on the use of PPRL in PCORnet, which enables the construction of longitudinal records at scale. It also includes a presentation on the CODI project, which tailors the PPRL techniques to pediatric populations. The panel will conclude with a discussion on this material with some thoughts on how audience members may leverage the artifacts being developed within the CODI project.

Privacy Preserving Record Linkage

Record linkage is a family of methods to identify shared patients between healthcare institutions [1,2]. Most record linkage methods rely on commonly shared identifiers, also referred to as linkage variables, such as first name, last name, date of birth, or social security number. It is common practice that patient data from the sources are consolidated and linked by a trusted third party [3,4]. Traditional RL uses clear-text identifiers, which have significant privacy concerns and typically cannot be shared for research across data partners without patient consent [5]. New RL methods, called Privacy-Preserving Record Linkage (PPRL), are being developed using encrypted identifiers to protect patient identity [6,7]. Dr. Toan Ong will present on what PPRL is and how it is applied.
PPRL methods can be classified into two major categories: deterministic and probabilistic. Deterministic PPRL methods link records based on the exact agreement/disagreement of one or more hash values of the linkage variables [8]. Deterministic PPRL methods often have very low number of false positives and are scalable because of the simplicity of value similarity computation (equality). However, deterministic PPRL methods often fail to match records with discrepancies in their values.

Probabilistic PPRL methods determine the likelihood that two records refer to the same entity. The most widely used probabilistic PPRL is the Fellegi-Sunter (FS) method, which uses conditional probabilities to estimate match and non-match numeric scores for each value of a linkage variable [9,10]. Based on these scores, a normalized summation weight is assigned to each linkage variable indicating its significance in contributing to an overall similarity score [11].

The similarity between the values of two linkage variables in each pair of records is quantified by distance. PPRL techniques use hashing functions to obfuscate sensitive information. While this approach is effective when information contained within patient records is an exact match, it does not address situations such as typos, nicknames, or other possible variations in data. These variations in data can be addressed by constructing n-gram sequences from identifiers and using the resulting sequences to construct Bloom filters [12]. The resulting filters can be compared to establish distance. The overall similarity score between two records is computed as the weighted sum of the distances of all linkage variables. This similarity score will then be compared to a match threshold to determine if a match between two records is a true linkage.

Finally, deterministic and probabilistic PPRL methods can be fine-tuned to improve linkage performance, dependent on the characteristics of the source data and constraints on linkage accuracy [13].

**Use of Privacy Preserving Record Linkage in PCORnet**

Dr. Tom Carton represents PCORnet, the National Patient-Centered Clinical Research Network, which consists of nine Clinical Research Networks (CRNs), two Health Plan Research Networks, two national Coordinating Centers, and one overarching, unifying foundation. While PCORnet has touted its large reach across the country and speaks of a network with over 100 million patient records, the actual number of de-duplicated, unique patients is still unclear. However, PCORnet is currently developing and will have in place by Spring 2020 a network-wide PPRL approach to data linkage with the technology and governance to link on-demand for specific research projects, many of which will involve pediatric populations.

PCORnet is drawing on the work of individual networks in developing the network-wide approach. This presentation will describe the history of within-CRN linkage, how lessons learned translated into a generalizable PCORnet-wide approach, and the methods used to select and implement a PPRL method. This is an active project and the presentation will cover research queries that will be deployed in the Spring 2020 timeframe.

**The CODI Project**

Childhood obesity impacts almost 14 million U.S. children and is associated with serious and costly immediate and future health risks, lower academic achievement, and poor mental health. Research that assesses and compares childhood obesity interventions is hindered, however, by limitations in researchers’ ability to link patient records across information systems where pediatric health-related data are stored.

To address this challenge and to improve data capacity for childhood obesity research, the U.S. Centers for Disease Control and Prevention, funded by the Patient-Centered Outcomes Research Trust Fund, is leading the Childhood Obesity Data Initiative (CODI). CODI will use and modify existing tools in innovative ways to facilitate access to patient-level, linked, longitudinal data that includes risk factors, clinical and community settings' weight management interventions, and health outcomes across health systems and sectors.

Andy Gregorowicz will discuss how CODI is using PPRL with a focus on a pediatric population. This will create a longitudinal data set that can answer research questions such as “What are the characteristics of weight management interventions associated with effectiveness?” CODI is focused on developing this approach with publicly available tools so that the methods developed for childhood obesity can be repurposed for other research domains.
A panel of Dr. Carton, Dr. Ong and Mr. Gregorowicz will be moderated by Nedra Garrett. The panel will discuss how the CODI project is generating a synthetic population to test publicly available PPRL tools. This discussion will encompass the behavior of PPRL solutions when encountering variations of data observed in real world data. It will also cover the differences in using PPRL for research as opposed to a patient care context.

Panel Discussion
- What are some of the differences between identity matching for patient care as compared to matching for research?
- As research continues on PPRL techniques, how will it improve in the future?
- What were some of the hurdles encountered in scaling record linkage beyond regional collaborations?
- What are some of the areas CODI researchers are exploring?
- How can other organizations participate in the CODI collaboration?

Conclusion
Building longitudinal records across organizations for research is difficult under typical circumstances. Use of PPRL reduces the risk of exposing PII when linking records. When this barrier is lowered, it becomes easier to gather a larger group of collaborators. Work is happening to take this approach to a national scale. Further, the CODI project is demonstrating that it is possible to use this approach on pediatric populations, where linking is more difficult.

Participation
All participants have agreed to take part on the panel.

References