Project Title: *Safeguarding Personally Identifying Information While Aggregating and Querying Medical Data*

Charles Henderson, chasehenderson@smu.edu
Nicole Bartholow, nbartholow@smu.edu

Project Problem:

Our goal for this project is to be able to aggregate and query non-private information while securing the personally identifying information.

One of the most sensitive types of personally identifying information (PII) databases are medical records. In the digital age, we are now able to aggregate patient data and compare it with other relational databases to determine intelligent disease clusters and anomaly detection.

Therefore it would be highly inefficient to protect patient PII through purely carbon data storage but even with the most sophisticated network security; PII is vulnerable to exposure when it is being pulled by databases and joined with other datasets.

Research Methodology:

Our goal for this project is to be able to aggregate and query non-private information while protecting the secure information. We will aggregate specific non-identifiable information such as diagnosis, date and location from two separate secure databases that hold this in combination with the secure patient data.

The initial steps require us to flesh out encryption offerings and databases to ensure we have free/test offerings for all of the moving pieces. While all substantial databases have some form of native encryption to protect it from external threats, we anticipate that aggregation between disparate databases may require a standardized encryption solution. Then we will create a data model that will reflect a simplified medical record for a patient, including a subset of data we will seek to protect. We will then define, for this project, what data is necessary to identify disease clusters or patterns and to establish the line between PII and that data.

We must also define basic security roles that allow for differentiation of the information (private versus non-private) to be viewed directly within each database. We will then define the aggregation methods and architecture for the query-able data. As part of this question, we also need to establish the timeliness required for the data to be fresh and relevant for disease clusters or other patterns we determine imperative to identify.

We will install these databases and implement the same patient record schema in both. We will internally be able to change between a secure all seeing view with PII and a secure third party readable PII omitted version.

Related Works:

Mohammed, Noman, et al. "Secure and private management of healthcare databases for data mining." *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. IEEE, 2015.

Dubovitskaya, Alevtina, et al. "Secure and trustable electronic medical records sharing using blockchain." *AMIA Annual Symposium Proceedings*. Vol. 2017. American Medical Informatics Association, 2017.

Guo, Cheng, et al. "Fine-grained database field search using attribute-based encryption for e-healthcare clouds." *Journal of medical systems* 40.11 (2016): 235.

Tang, Huanrong, Ning Tong, and Jianquan Ouyang. "Medical Images Sharing System Based on Blockchain and Smart Contract of Credit Scores." *2018 1st IEEE International Conference on Hot Information-Centric Networking (HotICN)*. IEEE, 2018.

IOM (Institute of Medicine). 2010. Clinical data as the basic staple of health learning: Creating and protecting a public good: Workshop Summary. Washington, DC: The National Academies Press

Boris Lubarsky , “Re-Identification of “Anonymized” Data,”1 GEO. L. TECH. REV. 202 (2017)

Research Plan and Schedule:

Date	Milestone
June 9	Complete Initial Research - Group meeting to compare notes on initial articles and identify additional roadblocks.
June 10	Finalize Database Selection and Encryption Methodology Selection
June 14	Design and Create data model for patient - including separation of private versus needed data
June 16	Databases Installed Encryption Installed
June 19	Schemas Created, Data Loaded, Roles Created
June 20	Establish Slide Overview for initial presentation
June 23	Complete Deck for Project Initial Presentation
June 25	Project Initial Presentation
June 30	Data Encryption Checkpoint - Is this working between databases?
July 6	Cross Database Query Established
July 14	Project Draft
August 6	Project Presentation
August 18	Project Final Paper

Resources Needed:

Our first step requires us to flesh out encryption offerings and databases to ensure we have free offerings that work together. Professor mentoring will be sought after for additional resources that can be utilized for achieving this desired task. An exploration a various SMU licenses will be necessary in the hope of gaining additional database offering resources.

Other questions:
We must determine the amount of data needed, horizontally and vertically, to prove the point. Laws establish (and change) what qualifies as *Personal Information*. Ideally, our solution is easy to change, but it isn't yet clear how authentic our data needs to be. Could Generic columns of information work? At the most basic level, we have columns to share and columns to hide, so authenticity may not matter.
On the other hand, Identifying a disease cluster might rely on, at least, a diagnosis, a timestamp range (100 cases of the flu is more pressing if the time period is 1 day versus 1 year), and some approximation of a zip code. Being able to replicate a data buildup of diagnosis within a time period sounds like a fun show to put on, but seems like a lot of work that misses the overall point of data security.