DataSecTech

Secure Compute

Use case: Health care

DST SecureCompute

Context

In many cases, patient data is scattered across multiple health care organizations. This could especially occur for chronic diseases such as asthma where patients may receive care at multiple institutions within a region. Single site studies may provide inaccurate results due to data inaccuracies. For example, due to certain selection biases, the number of patients from certain race groups may be underrepresented in one location. In addition, severity information of diseases may not be complete if all the emergency care visits are not recorded. Without proper record linkage and data duplication, many of the disease specific conditions may be over-represented. For example, it is reported that after cross-institution deduplication, the number of records related to diabetes reduced 24.0%, asthma reduced 28.0%, and myocardial infarction reduced 10.9%[1]. Therefore, it is critical that patient data scattered across multiple health organizations need to be linked before analysis.

Privacy-preserving Data Sharing, Linkage and Analytics Opportunity

The data belonging to multiple healthcare organizations could be linked, cleaned and analyzed to build machine learning models, get useful statistics and insights into the underlying disease. For example, such infrastructure could be used to build a machine learning model (e.g., logistic regression) to understand the relationship between air pollution to asthma attacks for each race group. Furthermore, linked and cleaned data could be used to understand health care and disease trends and can be used to provide statistical insights.

Although there have been efforts to implement health information exchanges to facilitate data integration and exchange, linking patient records across multiple health care organizations creates significant security and privacy challenges. For example, sharing common identifiers such as social security numbers without any protection (e.g.,

[1] A. N. Kho et al., "Design and implementation of a privacy preserving electronic health record linkage tool in Chicago," Journal of the American Medical Informatics Association, vol. 22, no. 5, pp. 1072–1080, 06 2015. [Online] Available: https://doi.org/10.1093/jamia/ocv038

encryption) may increase privacy risks such as identity theft. In addition, the Federal Health Insurance Portability and Accountability Act (HIPAA) outlines required procedures for securing and protecting protected health information, and the HIPAA requirements need to be addressed before sharing and linking patient data. Due to these security and privacy concerns, practical and privacy-preserving record linkage and analytics tools are needed.

Our solution

At DataSecTech, we provide a cloud based secure record linkage and analytics solution that could be used to address this important use case. Using our solution, different healthcare organizations can link their data even if they do not have a unique identifier such as social security number and records that have errors (e.g., typos) to get the complete history of the patient.

Using our privacy-preserving secure computation solution, users can choose different attributes (e.g., names, surnames address etc) and linkage algorithms. Furthermore, the linkage accuracy would be as close as possible to state-of-the-art record linkage.

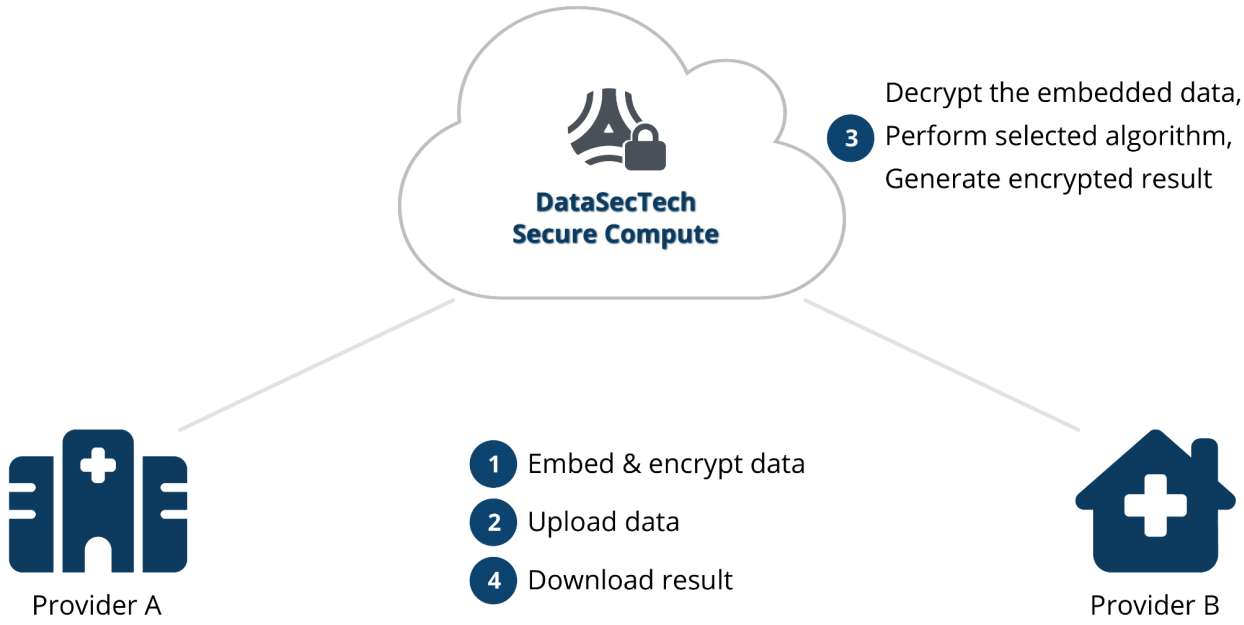

Figure 1: Overview of DST Secure Compute service

The data that is used for linkage and analytics purposes is sanitized and encrypted. Using the existing confidential computing cloud and hardware support available, the entire

process is end-to-end encrypted, and plaintext data will not be accessible to us and/or the cloud provider.

Once the records are linked, our solution allow users to do further processing such as outputting required statistics, learn a machine learning model (e.g., learn a new model that predicts asthma attacks), or apply a given ML model on the linked dataset (e.g., use a previously developed model to detect patients with high likelihood of suffering from a certain disease).

Finally all the results will be sent back encrypted to the users.

Benefits

1. Cloud based and deployable on existing cloud infrastructure
2. Entire data is sanitized encrypted end-to-end.
3. Protects privacy and complies with existing privacy regulations.
4. Resilient to errors and typos in the attributes such as names, addresses, date of births etc. used for linking.
5. Further processing is feasible on the linked data including learning ML models, testing the linked data with existing ML models, and getting useful statistics based on the linked data.
6. More efficient than heavy crypto solutions such as secure multi-party computation.