1. STUDY TITLE

ENACT Hub Ephemeral Enclaves for De-identified data

2. PRINCIPAL INVESTIGATOR

Michael Hogarth, MD; Division of Biomedical Informatics (DBMI), Department of Medicine, School of Medicine

3. STUDY RATIONALE

The goal of the ENACT Hub is to build upon ENACT, a federated network of institutional clinical data repositories, to enable the performance of in silico trials that inform clinical care. The ENACT federated data network currently provides cohort discovery through distributed querying of participating sites for patient/case counts based on search criteria. The goal of the ENACT Enclave Hub is to enable secure transfer of study-specific patient data as a limited data sets(LDS) to an ENACT Hub "study enclave". ENACT Hub enclaves are "air-gapped" HIPAA compliant private virtual clouds hosted by UCSD using a commercial cloud service (AWS). ENACT Enclave Hub users can analyze the data using standard analytical tools but since they do this within the enclave, they are blocked from connecting to resources outside of its internal network boundary (i.e., the Internet). The ENACT Hub enclaves are isolated by study and are ephemeral as they can be archived or deleted after the conclusion of research projects.

4. SPECIFIC AIMS/HYPOTHESES

Aim 1: Implement an air-gapped virtual data cloud (ENACT Enclave Hub), for secure computation on sensitive data.

Aim 2: Demonstrate secure transfer of ENACT study-related limited data sets (LDS) to the ENACT Hub for multi-site studies.

Aim 3: Use the ENACT Enclave Hub to conduct study-specific analysis for a multi-site collaborative *in silico* trial

5. BACKGROUND AND SIGNIFICANCE

As of 2017, 96% of hospitals in the US had implemented certified Electronic Health Record (EHR) systems[1]. Nearly all organizations implementing EHRs have also deployed clinical data warehouses. These data warehouses can significantly enhance the speed, cost effectiveness and breadth of clinical research [2]. Connecting multiple data warehouses provides the opportunity to leverage a very large data set with attendant geographic and demographic variability for a number of endeavors including on research on rare disorders as well as pragmatic trials in which it is important to minimize biases from practice variation.

Currently, there are two architectural models for multi-site clinical data warehouses: (1) the centralized data warehouse, and (2) the federated "virtual" data warehouse. The federated model is more expedient from a regulatory standpoint but requires data harmonization to be reliably done at each site. The central data warehouse enables complex analysis involving iterative computation (i.e., logistic regression), but requires data to be moved from each site to a central location, which is often viewed as less ideal by participating sites. Furthermore, due to its requirement to move data to a central location, the central data warehouse also inherits a fixed data latency and cannot be 'real time'.

Since January 2018, the Clinical and Translational Science Award (CTSA) program has supported a multi-site federated "virtual" data warehouse, the Accrual to Clinical Trials (ACT) network [3]. All ACT sites store data in their respective ACT research data warehouse node according to the Informatics for Integrating Biology and the Bedside (i2b2) common data model [4]. Querying of the federated ACT network is done using through a secure web-based application and query aggregator system, the Shared Health Research Information Network (SHRINE), which sends the query simultaneously to each data warehouse and displays returning aggregate counts of the number of matching patients [5]. This "cohort discovery" function has been used extensively across the ACT network sites for clinical trial planning. However, a large number of investigators have desired a further expansion that adds the ability to bring together de-identified row-level data from the cohort records into a single temporary study-specific repository for collaborative analysis.

In 2022, the CTSA funded an expansion of ACT, which became the "Evolve to Next-Gen ACT" (ENACT) network. The expansion adds the ability to use the ACT cohort discovery system to select records and transfer their de-identified data into a centralized secure enclave within the ENACT Enclave Hub. UC San Diego currently operates the ENACT Enclave

Hub, which is hosted on Amazon Web Services (AWS). ENACT Enclave Hub users access their study-specific data by using a virtual research desktop (VRD) provisioned within the study's enclave. All systems inside each study-specific enclave, including user virtual research desktops, are blocked from connecting to resources outside of its internal network boundary. The ENACT Hub enclaves are ephemeral meaning they can be either archived or deleted after the conclusion of the research project.

The significance of this expansion is a dramatic increase in research capability of the ENACT network, which will enable observational research from across 50 sites and nearly 150 million patient records. It will support observational studies that could uncover important predictors of disease, hospitalizations while doing so with minimal risk to privacy. The potential breadth of the network could also lead to the development novel machine learning (ML) based prediction models.

6. RESEARCH DESIGN AND METHODS

The ENACT Enclave Hub has been established for non-therapeutic correlative "in silico" studies designed and conducted by researchers from participating ENACT network sites.

Single IRB:

This protocol is submitted to establish a central IRB for the ENACT Hub.

Data Transfer Agreement:

ENACT intends to use a multi-party data transfer agreement (DTA) to govern the sharing of data into ENACT Hub study-specific enclaves. It describes terms and conditions related to the data submission from ENACT study sites. It describes ENACT data to be transmitted as HIPAA defined Limited Data Sets (LDS). The multi-party DTA would be signed by the clinical data contributor sites and University of California, San Diego, which hosts the ENACT Enclave Hub. The document stipulates that participating institutions and their investigators agree to follow all relevant local security policies regarding the use of EHR data for research. Furthermore, any researcher using ENACT enclaves shall report in writing to UCSD and unauthorized access and/or disclosure of data.

The ENACT DTA also states that data may be accessed by ENACT investigators from participating institutions to conduct pre-specified, study-specific research projects. Additionally, all ENACT enclaves are transient, existing initially for up to 24 months from the date of the initial data transfer, with the ability to extend the enclave for an additional 12 months at a time pending approval by the ENAT Protocol Review Committee.

Research using De-identified data:

The ENACT Hub was established to provide a highly secure environment (data enclave) for non-therapeutic "in silico" studies designed and conducted by investigators from participating ENACT sites using data Limited Data Sets (LDS) as defined in the Health Insurance Portability and Accountability Act (HIPAA).

An amendment to the original ACT Network Agreement in place with ENACT network sites permits research to be conducted across the network's deidentified data. This amendment enables access to all network deidentified data for purposes of cohort exploration, analysis, and publication. Each institution is responsible for their investigators who are provided access to ENACT. The ENACT Terms of Query Access must be signed by each user. Sites are required to have data stewards regularly monitoring network activity.

De-identification

All data in ENACT enclaves will house data de-identified in accordance with the LDS definition in HIPAA. Each site contributing data in a study is responsible for ensuring data transmitted to the ENACT Hub enclave system is de-identified in this fashion.

Anticipated Cohorts (clinical phenotype):

We anticipate study cohorts that will have data housed in their study-specific enclave to include patients of any age, gender, cultural background, and health status.

Study Data:

The ENACT Data Harmonization Workgroup (DHWG) has defined a set of data domains of data being transferred into study-specific enclaves. The data domains currently include demographics, diagnoses, procedures, and medications. As the project evolves, there may be the need to both modify existing domains or add new data domains (e.g. laboratory test orders and results, e-prescriptions, vital status). ENACT has defined a

standard operating procedure (SOP1) for adding or modifying data domains through DHWG review and subsequent approval by ENACT Network sites.

Monitoring and Auditing:

ENACT will establish a central query metadata archive for monitoring and auditing of SHRINE cohort discovery queries. SHRINE queries initiate the process of patient record selection at each study site.

ENACT Enclaves

Data enclaves are computing environments constructed with security measures that isolate them from other networks, such as the public internet. Users access a data enclave through a secure process, and once within can analyze the data, but are unable to connect to an external network to move data in or out of their enclave. The proposed system is based on the design of the UCSD Research Health Cloud on AWS, a data enclave in use since 2018. Unlike the UCSD Research Health Cloud, the proposed ENACT platform adds two key features: (1) that the enclaves are temporary (ephemeral) and persist until terminated at the conclusion of a study, and (2) the study specific ephemeral enclaves can be securely accessed by collaborators who may be from multiple CTSA Hubs.

The ENACT enclaves are instances of “air-gapped” HIPAA compliant private clouds hosted in Amazon Web Services (AWS). Each “data and computing enclave” is blocked from connecting to resources outside of its internal private network boundary. The environment provides access to elastic and scalable secure compute resources for a broad range of research computing. Capabilities within the enclave include a “Virtual Research Workstation” (VRDs) for each user. VRDs are specially configured virtual desktop infrastructure (VDI) instances with an array of tools for analysis on biomedical data. Sensitive data sets are securely submitted to the UCSD hosted ENACT Hub by participating CTSA sites and UCSD staff coordinate moving the data set into the study-specific enclave.

Enclave Security

The enclaves use AWS Control Tower to help ENACTHUB govern their resources and monitor compliance across groups of AWS accounts. When a guardrail is deployed to an organizational unit (OU), every AWS account within the OU will be affected by the guardrail. Therefore, when ENACTHUB users perform work in any AWS account in the landing zone, they're always subject to the guardrails that are governing the account. guardrails.

All users of the enclave access internal resources through a “bastion”. System administrators use an AWS virtual server (EC2) as a bastion, whereas ENACT study investigators use an Amazon WorkSpace virtual desktop (VRD). Access is controlled using security groups.

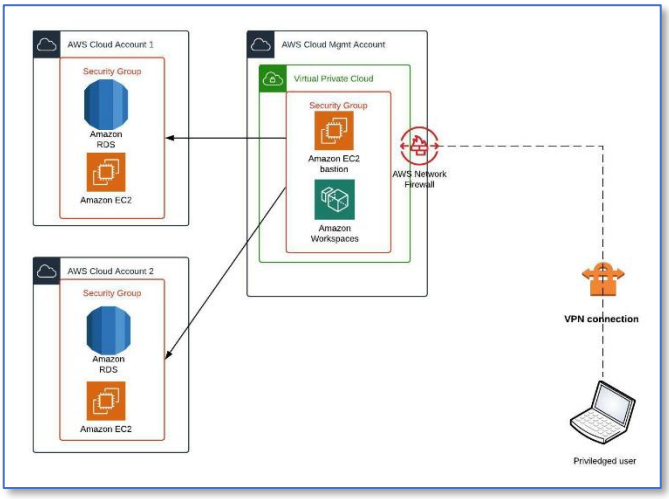

Figure 1: Use of Bastion with Amazon Workspaces to access Enclave resources

Data Transfers

AWS Transfer will be the service used for transferring data into and out of the ENACTHUB secured cloud environment. AWS transfer offers fully managed support for the transfer of files over SFTP, FTPS, and FTP directly into and out of Amazon S3 or Amazon EFS. The AWS transfer service will be enabled in the shared-services account services VPC. Data can be stored in AWS buckets within the same account or other accounts. IAM roles/ cross account IAM roles are used to isolate and restrict users access to study-specific buckets and folder.

Logging of Activity

ENACT Hub uses Amazon CloudWatch to gain system-wide visibility into resource utilization, application performance, and operational health. CloudWatch tracks the metrics so that ENACT Hub system administrators can visualize and review them. CloudWatch gives visibility into resource utilization, application performance, and operational health.

7. RESEARCH PARTICIPANTS

We anticipate ENACT enclaves having de-identified data from patients of any age, gender, cultural background, and health status.

8. RECRUITMENT

Studies to use the ENACT Hub enclaves will use existing data from electronic health record systems for correlative non-therapeutic studies. We do not anticipate recruitment efforts.

9. INFORMED CONSENT

We are requesting a waiver of consent as the ENACT Enclave Hub is intended for non-therapeutic *in silico* studies and we believe there is substantial benefit through research and discovery to ensue from the use of patient data for ENACT studies.

We are requesting waiver of HIPAA authorization as:

1. The anticipated research studies cannot be conducted without the protected health information

2. The anticipated studies cannot be practicably conducted without the waiver.

3. The use of PHI will involve no more than minimal risk to privacy.

4. Granting the waiver will not adversely affect the privacy or welfare of those individuals whose records will be used.

Risk to privacy is minimal as:

1. Patient data submitted to ENACT enclaves be de-identified and not contain the following personal identifiers: names, addresses, phone/fax numbers, email addresses, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers driver/certificate license numbers, vehicle identifiers/plates, device identifiers, URLs, IP addresses, biometric identifiers, and face photos or comparable images. A unique numeric context-free identifier will be assigned to each unique patient record.

2. The linkage between study-identifier and personal identifier will not be available to study investigators. Data stewards at ENACT sites submitting data may retain their internal linkage, if necessary to allow for subsequent data transfers on the same study participants as part of the study protocol.

3. The DTA signed by each participating institution disallows reuse of PHI for other purposes.

4. The ENACT study-specific enclaves are transient and the data within will be deleted once the study is concluded.

10. BANKING OF INFORMATION/BIOSPECIMENS FOR FUTURE USES

We do not anticipate banking of data in ENACT Hub enclaves. Enclaves are transient and are deleted upon conclusion of the study.

11. MINIMIZATION OF RISK

_**X**_ Check here if the only risk of the research is a breach of confidentiality. If so, no additional information is

required in this section.

12. PRIVILEGES/CERTIFICATIONS/LICENSES AND RESEARCH TEAM RESPONSIBILITIES

Dr. Michael Hogarth is a UCSD Professor of Biomedical Informatics and co-PI of the UCSD CTSA grant. He will supervise the ACTRI Biomedical Informatics (BMI) team, which will manage the ENACT Enclave Hub. He has participated in the design and development of the ENACT Enclave Hub.

13. REFERENCES

1. Office of the National Coordinator for Health Information Technology. 'National Trends in Hospital and Physician Adoption of Electronic Health Records. https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records
2. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, Bauck AE, Cifelli D, Smerek MM, Dickerson J, Laws RL, Madigan RA, Rusincovitch SA, Kluchar C, Califf RM. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc. 2013 Dec;20(e2):e226-31
3. Morrato EH, Lennox LA, Sendro ER, Schuster AL, Pincus HA, Humensky J, Firestein GS, Nadler LM, Toto R, Reis SE. Scale-up of the Accrual to Clinical Trials (ACT) network across the Clinical and Translational Science Award Consortium: a mixed-methods evaluation of the first 18 months. J Clin Transl Sci. 2020 Jun 30;4(6):515-528
4. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30. doi: 10.1136/jamia.2009.000893
5. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):624-30

14. BIBLIOGRAPHY

See References