© Journal for Modeling in Ophthalmology 2019;3:115-132
Original article

Increasing protected data accessibility for age-related cataract research using a semi-automated honest broker

Samaikya Valluripally, Murugesan Raju, Prasad Calyam, Mauro Lemus, Soumya Purohit, Abu Mosa, Trupti Joshi

University of Missouri-Columbia, Columbia, MO, USA

Abstract

Ophthalmology researchers are becoming increasingly reliant on protected data sets to find new trends and enhance patient care. However, there is an inherent lack of trust in the current healthcare community ecosystem between the data custodians (*i.e.*, health care organizations and hospitals) and data consumers (*i.e.*, researchers and clinicians). This typically results in a manual governance approach that causes slow data accessibility for researchers due to concerns such as ensuring auditability for any authorization of data consumers, and assurance to ensure compliance with health data security standards. In this paper, we address this issue of long-drawn data accessibility by proposing a semi-automated "honest broker" framework that can be implemented in an online health application. The framework establishes trust between the data consumers and the custodians by:

1. improving the efficiency in compliance checking for data consumer requests using a risk assessment technique;
2. incorporating auditability for consumers to access protected data by including a custodian-in-the-loop only when essential; and
3. increasing the speed of large-volume data actions (such as view, copy, modify, and delete) using a popular common data model.

Correspondence: Prasad Calyam, 201 Naka Hall, University of Missouri-Columbia, Columbia, MO 65211, USA.
E-mail: calyamp@missouri.edu

Via an ophthalmology case study involving an age-related cataract research use case in a community cloud testbed, we demonstrate how our solution approach can be implemented in practice to improve timely data access and secure computation of protected data for ultimately achieving data-driven eye health insights.

Keywords: common data model, honest broker, precision medicine, protected data access, semi-automated compliance

1. Introduction

Health care big data being collected today for patients typically comprises heterogeneous data sets collected as: electronic health records (EHR) of patient history, wearable and other sensor data, genetics, environmental factors, medical imaging, clinical diagnosis of signs/symptoms/outcomes, and laboratory results.[1] With the increased push to promote data-driven methods in healthcare, there are massive data collection and archival efforts underway. In fact, biomedical big data has become one of the critical thrust areas for the US National Institutes of Health[2] due to the potential of multi-source data sharing and analysis for discovering rare patterns. Particularly, EHRs have transformed the availability of patient data as well as disease information to researchers and physicians. The American Medical Informatics Association (AMIA)[3] Genomics and Translational Bioinformatics Working Group has identified knowledge discovery and data mining as important components of clinical research informatics and next-generation clinical decision support.

Building upon these advances, researchers and clinicians in ophthalmology and other fields of medicine can enhance existing knowledge relating to studies of disease management (diagnosis, prevention, early prediction, personalized treatment)[4] for quality health care.[1] They can potentially analyze/visualize any accessible (protected) data sets to pursue medical breakthroughs in the areas of personalized medicine,[5] and big data knowledge discovery.[6] Thus, they can investigate novel data-driven methods in, *e.g.*, identifying latent associations of patient data sets to determine risk factors for diseases and test hypotheses with relevant heuristics as part of ongoing clinical research studies.[7]

However, there is an inherent lack of trust in the current health care community ecosystem between the data custodians *(i.e.*, health care organizations and hospitals) and data consumers (*i.e.*, researchers and clinicians). This typically results in an approach that causes slow data accessibility for consumers due to concerns such as ensuring auditability for any authorization of data access, and information assurance to ensure access compliance with health data security standards. The governance to authorize data access requests from researchers and clinicians includes several data custodian tasks. The tasks include:

1. ensuring privacy preservation of data owners (*i.e.*, patients) by checking for compliance of health regulations such as the Health Insurance Portability and

Accountability Act (HIPAA);[8] and

2. ensuring data consumers have the appropriate Institutional Review Board (IRB) protocol approvals when accessing the data.

These essential tasks introduce additional steps that cause delays in timely handling of the user data query requests. Figure 1 illustrates the current "manual honest broker" governance that is used by data custodians to manage data consumers' access of multiple data sources with heterogeneous data sets that are multi-domain, sensitive, and guarded by multiple access regulations. Consequently, any data consumer user request involving various data sets access and their corresponding secure computation resource requirements are subject to governance tasks and community cloud infrastructure configuration. These actions often take several months before approval is granted for data consumption.

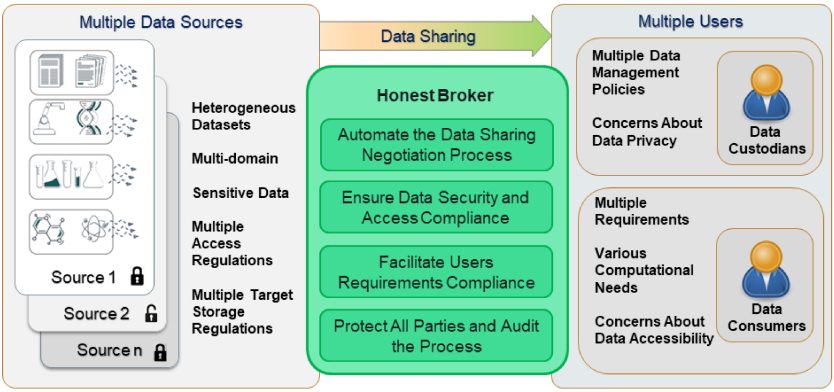


Fig. 1. Illustration of an honest broker facilitation for health big data sharing.

In this paper, we address this issue of long-drawn data accessibility by proposing a semi-automated "honest broker" framework that can be implemented in an online health application. Figure 1 shows the various features that our honest broker solution must ideally support. It should help in the automation of the data sharing negotiation process, as well as ensure data access compliance to protect all parties involved (*i.e.*, data owners, data custodians, and data consumers). The design of our honest broker framework aims to foster trust establishment between the data consumers and the custodians in several ways. Firstly, it improves the efficiency in compliance checking for data consumer requests using a risk assessment technique that uses a natural language processing scheme to automate portions of the compliance checking (based on the NIST SP 800 guidelines)[9] given a data consumer's filled-out request form, and a data custodian's access policy document. Secondly, it incorporates auditability for consumers to access protected data by including a custodian-in-the-loop only when essential. Logging, notification, and manual approval confirmation

are performed for data actions (such as view, copy, modify, and delete) depending upon the assessed risk and the sensitive nature of the data being accessed. Lastly, it increases the speed of large-volume data actions using a popular common data model (CDM)[10] on a global data catalogue (a.k.a., centralized data repository) that is hosted in a community cloud infrastructure with high-performance computing, high-throughput storage, and virtual desktop access resources.[11] We implement our semi-automated honest broker solution in the form of an online health application for an ophthalmology case study involving an age-related cataract research use case involving the assessment of risk factors (*e.g.*, race, age, and diagnosis) that can possibly cause the cataract diseases.[12] We also demonstrate our implementation benefits for the ophthalmology case study in a real-world community cloud testbed to improve timely data access and secure computation of protected data for ultimately achieving data-driven eye health insights.

The remainder paper organization is as follows: Section 2 lists the related work and discusses the novelty of our contributions. Section 3 presents a background regarding our ophthalmology case study and challenges experienced when accessing protected data sets. Section 4 details our solution approach for the semi-automated honest broker framework development. Section 5 describes our solution implementation and benefits demonstration in a real-world community cloud testbed for protected health data management. Section 6 concludes the paper.

2. Related work

2.1 Brokering solutions for health applications

To expedite the processing of data requests, existing works[13,14] adapt the functionalities of a brokering approach that includes either retrieving the data faster or checking the access compliance based on data custodian policies to identify the risk of sharing the data. Prior work[15] proposed an integrative clinical and genomic data framework called SPARKS, featuring a user interface that requires low levels of training on how to query/access the diverse data. However, these works[13–15] fail to address the latency caused by the manual governance approach to process the related data requester queries. In order to address such data access issues in a diverse data environment, our approach employs a semi-automated honest brokering approach (automation is performed when possible based on risk factors) to the process of compliance checking that ultimately enhances data access speed and its secure computation in a community cloud infrastructure.

2.2 Data model approaches for integrating diverse data

Due to the diverse nature of the data sources handled in existing works,[13,14] the data retrieval process for each user request can cause overheads for the users, such as knowing the query language for data retrieval, getting to know the internal structure

of data sources, or dealing with the burden of combining multiple data sources. To resolve such challenges, a data repository known as the common data model (CDM)[10] is employed, which includes standardized data elements, reduced data discrepancies, combined handling of multiple data sources within standard views, and facilities for reproducibility and consistency of data retrievals.[16] The fundamental part of using a CDM[10] is that it supports data collections for different purposes and from multiple sources. For instance, providers, clinical research, patient care, and financial information can be combined within a common structure. The OMOP-CDM[10] has been widely adopted to structure clinical research data,[17] and multiple open-source tools have been published to bring disparate data sources into the CDM.[18] Given the advantages of using a CDM, we use the OMOP-CDM version 5 in our work to improve query performance of data retrievals from the original data sources and establish a CDM-based common data repository.

2.3 Risk assessment approaches

Health data access and availability of information from different data sources are essential factors to take a clinical decision for the data consumers (*i.e.*, researchers and clinicians) involved in a health care ecosystem. Ensuring trust between the data consumers and the data custodians to facilitate data accessibility is one of the key components for many data-oriented transactions involved in health care applications. To establish mutual trust or auditability in cloud-based software deployments, the risk of each data transaction and its impact needs to be analyzed.[19] As part of establishing trust between the users using our honest broker solution approach, we employ a two-step process which includes the NIST-based risk assessment approach.[19,20] We also adapt a semi-automated compliance checking building upon the work in[20] in order to bring a custodian-in-the-loop only when necessary.

2.4 Compliance-checking solutions

Industry efforts in existing works[21–23] deploy various computation capabilities along with HIPAA compliance to support the development of health care applications. Similarly, to efficiently utilize cloud-based services, consumers have to continuously monitor and manage the Service Level Agreements (SLA) and also ensure compliance among the services requested by the users. Moreover, a NIST-based compliance mechanism in our prior work[20] elaborates how to align and check compliance for different computing service policies for a given user request with security and performance requirements for a bioinformatics use case. Our proposed honest broker solution features semi-automated HIPAA[8] compliance checking[20] based on Natural Language Processing (NLP) methods to establish the trust and auditability of data transactions in health care big data applications.

3. Background of age-related cataracts case study

3.1 EHR database

Recent advancements in health care community using big data provide unique opportunities to investigate risk factors for the development of age-related cataract diseases. As part of a cataract research study,[12] we utilize the data in a computation-friendly EHR database such as the Cerner HealthFacts® database. This EHR database captures and stores de-identified, longitudinal EHRs, including information on demographics, type of encounter, diagnosis, medications, procedures, laboratory tests, hospital information, and billing details with more than 60 million patient records. In our work, we particularly utilize a cataract data set research study as a use case for our proposed honest brokering solution development.

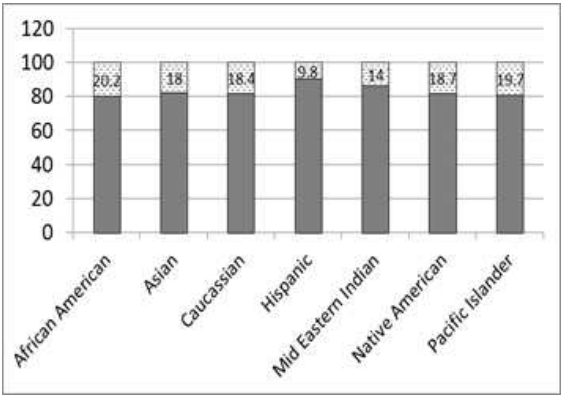


Fig. 2. Demographics of the ophthalmological data in the Cerner HealthFacts® database.

This EHR database provides us with a unique opportunity to study more variables (patient demographic, gender, weight, diagnosis, etc.) to analyze the risk factors, as shown in Figure 2 of age-related cataract diseases.[12] As part of accessing the data for this risk factor analysis,[12] the data consumers (*i.e.*, clinicians and researchers) need to run large-scale queries similar to a query shown in Figure 3. In our experiments, it took approximately 24 minutes to run a basic program using about 1% of the data available in the HealthFacts database. For more complex programs using more of the data, it took 3 or 4 days to run the query, as shown in Figure 3. Such performance outputs are extremely slow for data consumers who would have to wait long periods to test bold data-driven hypotheses and derive analytical insights for the future illness treatment or drug discovery.

3.2 Governance process challenges in the current model

In addition to the above discussed practical problems, new bottlenecks in utilizing large health-related databases can arise due to the manual governance employed in

```
PROC SQL;
    CREATE TABLE PREJOINED AS
        SELECT IDSK.PATIENT_ID, IDSK.PATIENT_SK, IDSK.RACE, IDSK.GENDER,HF_F_ENCOUNTER.ADMITTED_DT_TM, HF_F_ENCOUNTER.ENCOUNT
        FROM IDSK AS L INNER JOIN HFFACTV3.HF_F_ENCOUNTER AS R
        ON L.PATIENT_ID=R.PATIENT_ID
        WHERE HF_F_ENCOUNTER.PATIENT_ID IN (SELECT PATIENT_ID FROM EYEVIS.EYE_ID);
QUIT;

PROC SORT DATA=PREJOINED NODUPKEY;
BY PATIENT_ID PATIENT_SK ADMITTED_DT_TM ENCOUNTER_ID WEIGHT HOSPITAL_ID;
RUN;

PROC SQL;
    CREATE TABLE JOINED AS
        SELECT PREJOINED.PATIENT_ID, PREJOINED.PATIENT_SK, PREJOINED.RACE, PREJOINED.GENDER, PREJOINED.ADMITTED_DT_TM, PREJOI
        FROM PREJOINED AS L INNER JOIN D_DIM_V3.HF_D_UNIT AS R
        ON L.WEIGHT_UNIT_ID=R.UNIT_ID;
QUIT;
```

Fig. 3. Query execution of an ophthalmology health big data application case study involving HealthFacts and the Statistical Analysis System (SAS) analytical tool.

the current HealthFacts, as shown in Figure 4. The challenges caused by the manual governance approach can be grouped under issues pertaining to: approval for extracting data, latency in extracting data, and storing the data for analysis. To elaborate, the data request process in the current HealthFacts system includes the following steps. Firstly, a data consumer sends his/her data request to the HealthFacts database custodian involving various data types such as: aggregated, de-identified, identified, and limited. Secondly, each of the user requests are reviewed by a data custodian-appointed governance committee to check for HIPAA compliance along with the policies employed across different data sources. The time taken for obtaining consent from each of the parties involved in the data access transaction typically takes several days or even months due to the manual processes. This results in queued data requests, query building to use for data visualization, latency in accessing the data (slow query response), and delayed patient care decisions, which we term as a case of *Loss of opportunity*.

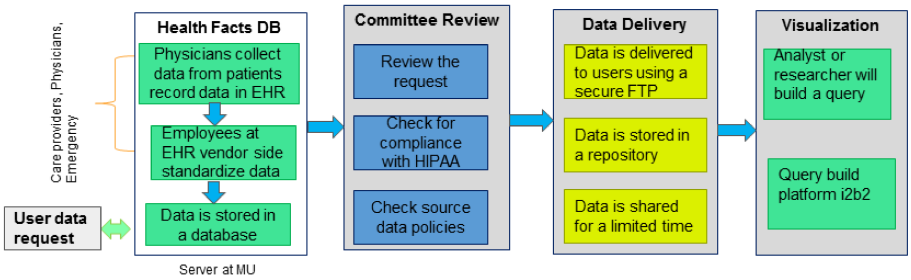


Fig. 4. Data sharing pipeline subject to a governance process for compliance assurance and auditability of protected health big data applications.

The usability is affected for data consumers due to the latency in accessibility of the data specifically to run large-scale queries for analytical purposes. The end users, who are typically not experts in high-performance computing, require automation for handling queries that involve performing a data lookup at the high scale. Moreover,

lack of auditability mechanisms for data transactions in a health care ecosystem and other related trust issues among the data consumers and custodians leads to the fear of *Loss of value* for data sets being requested. Furthermore, slow disk mechanisms used for data lookup, inadequate memory provisioning, and low-scale processing back-ends can all cause additional latency in data accessibility. To scan billions of health records from the HealthFacts database using a query look up, relevant user interfaces and appropriate community cloud system configurations need to be designed to obtain reasonable response times for retrieval of relevant records.

4. Semi-automated honest broker framework

In order to address the challenges shown in Figure 4 relevant to the case study, we propose a semi-automated brokering solution that features a pipeline to speed up the data request, access, and sharing process to help research and day-to-day clinical applications. We mainly categorize our proposed solution approach into three modules:

1. User interface (UI) module - for the user requests;
2. Honest broker module - for the brokering service to process the requests involving the governance committee; and
3. CDM - to integrate the disparity among the multiple sources of data for speedy query/analysis, as shown in Figure 5.

Thus, through our approach, multiple and disparate data sources can be integrated and the queried data can be shared without compromising the access policy compliance of the sources. Our approach also expedites data accessibility by improving performance while handling users requirements, and making the data sharing process trustworthy.

4.1 UI design

The UI module features a web-based interface that allows users to request/review data relevant to the age-related cataract research study. To develop the UI, we used an open-source platform known as HumHub,[24] which is an open-source social network development kit based on the Yii2 Framework. HumHub uses the Model-view controller (MVC) architecture that allows adding new features or allows changing existing core features by means of custom modules. We implemented our UI equipped with both front-end technologies (*e.g.*, HTML, CSS, Javascript) and back-end technologies (*e.g.*, PHP, PostgreSQL, Python). The UI allows data consumers (*i.e.*, researchers and clinicians) to create a new data request by filling out a questionnaire that helps the administrators in governance committee reviews, and initiates the processing of user requests to make a (approve/deny) decision. The decisions are

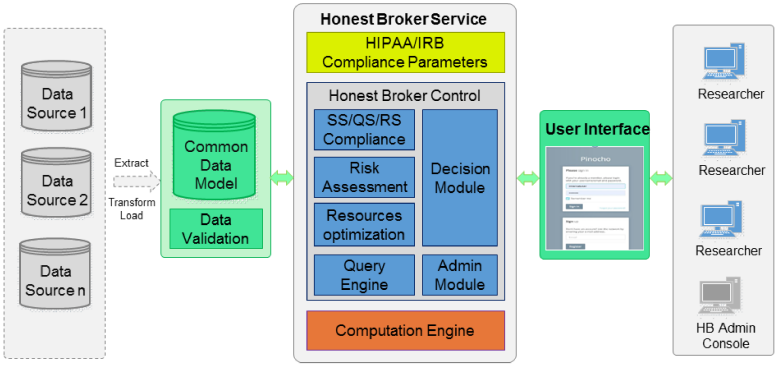
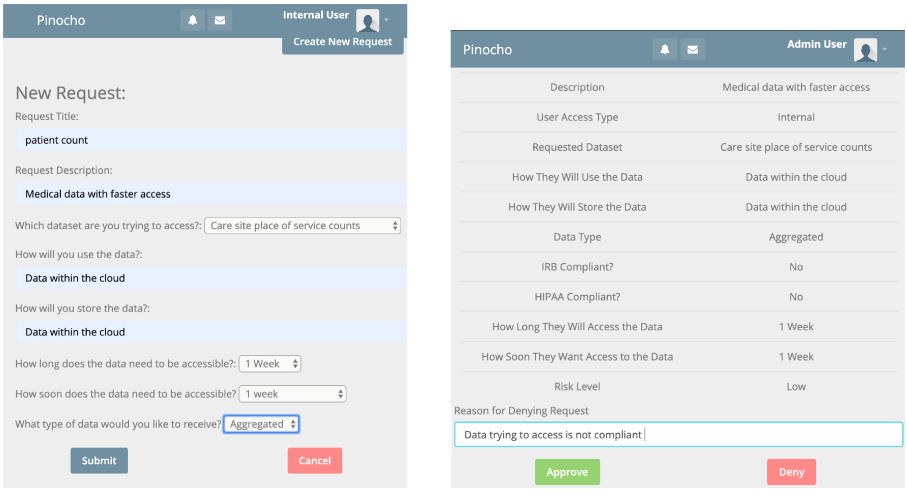


Fig. 5. Layout of the functional modules of the honest broker that include the: UI for the data custodians and data consumers, the honest broker service for semi-automated assurance and auditability of the protected health data access, and the CDM for supporting high-volume data actions.

made based on the compliance and risk values obtained for each of the user requests considering tradeoffs in the *Loss of opportunity* and *Loss of value* criteria. The interactions between the server and the web applications that are part of the data request and authorization review process in the UI are communicated via JSON formatted messages. We assume two different types of data consumers, *i.e.*, local-user (Intra-domain) and external-user (Inter-domain), who can submit their requests. These users can have access privileges with, *e.g.*, default access for Intra-domain user, and limited access for Inter-domain user. An example of the user data request form and the admin request review form to access diverse data types (*e.g.*, aggregated, de-identified, limited, identified) are included in Figure 6.

4.2 Honest brokering design

The honest brokering module serves as a monitoring system that acts upon the tasks of a governance committee, such as HIPAA and data source policy compliance, risk assessment of each data request, and subsequent decision making on each of the user requests. The honest broker module in Figure 5 illustrates the compliance check functionality, risk assessment[19] relevant to a user request, and user access privileges; decision module that pertains to the approval/denial of the user requests. The *compliance* functionality, allows the admin users to classify the requests into several bags of words that are categorized as performance and security requirements (Fig. 7). The performance and security requirements obtained, as well as the policies of the requested data sources are checked for HIPAA compliance adapted from the work in Dickinson *et al.*,[20] whose compliance levels are classified as low, medium, and high.[19]



(a) UI for the data consumer. *(b)* UI for the governance committee.

Fig. 6. Honest broker UI screenshots from the HumHub-based implementation of the user data request forms, and the logging/notification/approval-seeking for assurance and auditability.

```
analysis a08.5 imaging data, stage pertain material modeling biological specimen analysis generate large amount raw pr
ocessed data set image files. require different set data standardized analysis within week.

['analysis', 'a08', 'imaging', 'data', 'stage', 'pertain', 'material', 'modeling', 'biological', 'specimen', 'generate
', 'large', 'amount', 'raw', 'processed', 'set', 'image', 'files', 'require', 'different', 'standardized', 'within', '
week']
data is performance
analysis is performance
standardized is security
```

Fig. 7. Conversion of a user request into performance/security requirements using an NLP tool.

The honest broker module handles various parameters related to the requested data, and analyzes the risk associated with each user request, as shown in Figure 8. A risk assessment evaluates the risk score associated with each of the user requests, including data type requested, user type, and data source. Due to the independent event nature of these user request entities, a maximum function is used to determine the overall likelihood (f1). For accurate risk assessment, we model the compliance score (level) of the user request as input into the risk assessment module to compute the overall risk score (f2), which is an average function of (f1 and compliance score). The overall risk score is categorized on a uniform scale of 1-10 such as: low (1-3), medium (4-6), and high (7-10). To take an effective decision based on all the computed parameters, such as compliance and risk associated with the user requests, a decision process is integrated. This enables the admin to give a decision to approve/-deny that considers tradeoffs in *Loss of opportunity* and *Loss of value* criteria. For ex-

ample, consider an Intra-domain user (local user) requesting aggregated data whose data source risk input is minimum. The risk assessment in this case determines the overall risk score as low. Consequently, data access is allowed by logging the data transaction and additionally notifying the admin about the data transaction. With such a risk-analysis based decision, a long-drawn manual governance committee review is avoided. However, if an Inter-domain user (external user) requests, *e.g.*, identified data involving a large number of days range, then the risk assessment directs the admin to initiate a manual governance committee review.

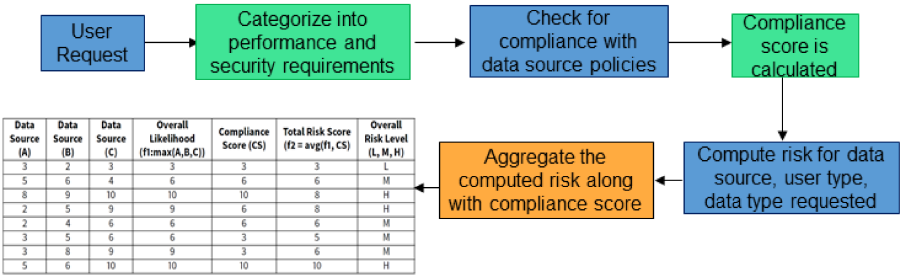


Fig. 8. Steps to calculate the compliance score during risk assessment for a given user request.

4.3 Data query using the CDM

The CDM[10] helps in the creation of a global catalogue that can provide a standardized repository of the data. This standardized data is accessible to the honest broker system for sharing data analytics pipelines to Intra/Inter-domain users. As part of the CDM module, we adapt the MOP-CDM v5.0 to provide a common view of the data and control the access pertaining to different types of data and resources allocated to fulfill a particular request.[10] Our CDM module improves performance by exposing a standard way to access any data request. This module combines the data from multiple data sources into a single standard model via Extraction, Transformation, and Load (ETL)[18] functions, as shown in Figure 9. Data is recorded on a day-to-day basis as part of the regular healthcare operations in the EHR at the data source sites. An ETL process from EHR in a CDM implementation can be automated by scheduling incremental refresh of data on a nightly, biweekly, monthly, or even quarterly basis. Data can also be refreshed in real-time using HL7 messages. In the case that a specific data attribute is queried across a given time and day range by user(s), the corresponding data can be retrieved if necessary metadata has been added by the data sources involved. Note that the metadata information can be processed as new attributes in the data model as part of the CDM implementation. To validate our solution approach, we considered data sources containing age-related cataract patients data in the DE-SynPUF format stored in text files.

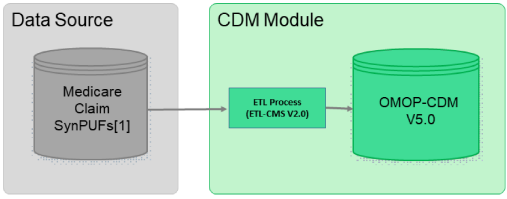


Fig. 9. Illustration of CDM module handling multiple data sources through an ETL process to convert to the standard OMOP-CDM v5.0 format and to load into central data repository.

An offline ETL process extracts the data from the text files, transforms the data into the OMOP-CDM v5.0 data model[10] format, and loads the standardized data into the CDM repository built on a PostgreSQL v9.5 database. With data available in the CDM repository, the honest broker is able to process approved/authorized user requests by running predefined queries to handle the user-requested data sets.

The sample data sets as shown in Table 1, are utilized for the data retrieval process where the user requests are processed by the honest broker, as shown in Figure 10. Once the user request is processed and an approval decision is made automatically (with notification to the admin), or manually by the admin or the governance committee, the related predefined queries are executed against the CDM repository, where the time taken for data retrieval is logged. The data flow process shown in Figure 10 includes a set of text files from data sources comprising patient data in the DE-SynPUF format. The data is extracted from those files, transformed into the OMOP-CDM V5.0 format,[10] and loaded into the central data repository. When user data requests are processed through data retrieval from the central repository, the result-sets are presented to the users for analysis/visualization.

If a request is semi-automatically approved by the honest broker, then the total processing time for that request is just limited to the time taken to execute the query (on the order of a few seconds). Our approach is thus based on the fact that the decision about approving or denying a request is automated based on the data access factors of the request and the additional information provided by the user for risk assessment. If further authorization is needed for data access, then a custodian-in-the-

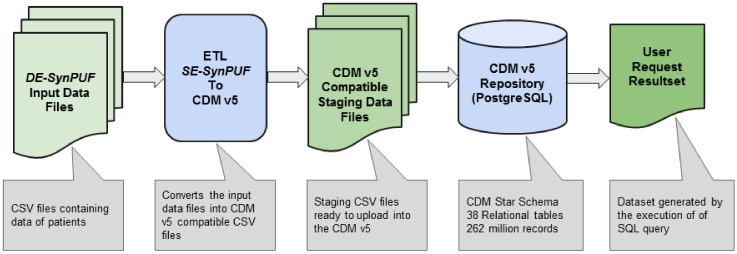


Fig. 10. Illustration of the data flow process handled by the honest broker from multi-domain sources to the target CDM-supporting computing location.

loop intervention is introduced (on the order of a few days or months). Regardless, the semi-automated honest broker along with the CDM simplifies user data access requests, and shortens the time for checks (*e.g.*, HIPAA fields) and data transformation (*e.g.*, de-identification) that are currently performed entirely manually. Thus, the semi-automated honest broker aids the admin as well as the governance committee to formally approve data access from multiple data sources with varied compliance requirements. Correspondingly, it also helps in streamlined and automated handling of data queries based on the fields, as well as their associated metadata in the CDM.

5. Implementation and evaluation results

5.1 Testbed setup

To evaluate our proposed honest broker solution, we implemented our solution approach using a community cloud testbed on the NSF GENI cloud infrastructure,[25] as shown in Figure 11. In this community cloud testbed setup, we included a host dedicated for the honest broker service on a HumHub instance[24] and customized our UI functionality. An ExoGeni host allows Layer 3 access to the CDM module. Virtual machines are deployed for the two hosts dedicated to the original data-source repositories (located locally and externally), and the two hosts to serve as the users (Intra-domain and Inter-domain). All of these components are connected via a network switch that emulates the functionalities of each of these entities across multiple network domains.

Each of the components in the testbed setup as shown in Figure 11 are equipped with different networking speeds based on the user type (Inter-domain, Intra-domain). An Intra-domain user requests the data access via a LAN (Local Area Network) switch to the data repository, whereas an Inter-domain data access occurs over the Internet. To process these requests, the honest broker administrator (HB-ADMIN) who is a local user, will have the option to review the request based on the responses from the honest broker component used in the testbed setup. The decisions are sent to the HB-ADMIN from the honest broker module via a high-speed network to avoid the latency in the request process. If the request is approved, then the requested data is fetched from the ETL-CDM component used in the community cloud-testbed setup shown in Figure 11. Evaluation results for the functionalities of each

5.2 Risk assessment results

As part of the honest broker service, we performed compliance checking along with risk assessment based on the scheme detailed in Section 4. As the main three entities of a user request (*i.e.*, data source risk, data type requested, and user type) are given as the input parameters as shown in Figure 8, we term these entities as A,B,C with eight combinations of user request scenarios. As defined in Section 4, we utilize a scale of 1-10 for the related risk assessment calculations. As function f1 is modeled as

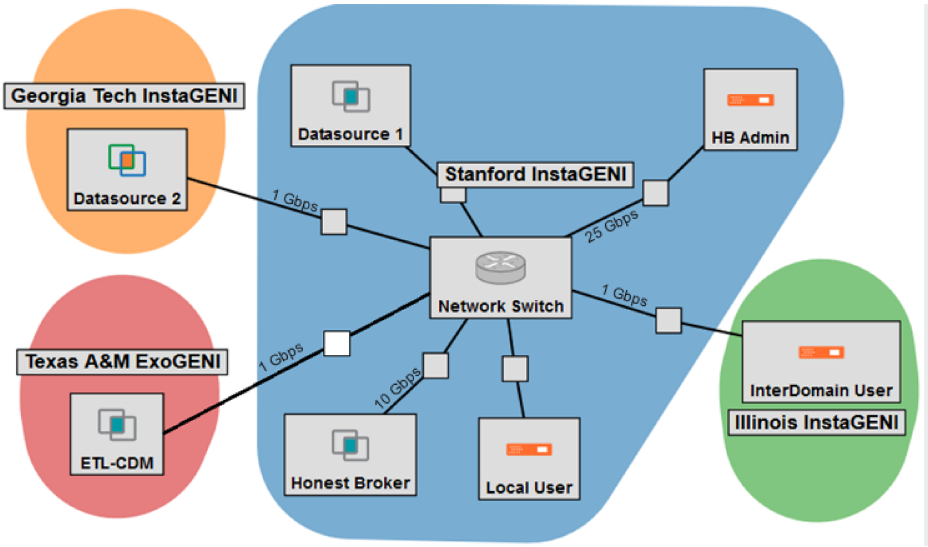


Fig. 11. Community cloud testbed topology including the honest broker node, the ETL-CDM Exo-Geni node, the data sources, an Inter-domain user external to the data custodian organization, and an admin user corresponding to the data custodian organization.

a maximum function, it is irrelevant which entity among A,B,C is high, as each of these A,B,C parameters are independent of each other in an online health application. Thus, the combination of user request scenarios are reduced to eight and used to compute f1, as shown in Table 1. Furthermore, as shown in Figure 8, once the compliance module identifies that the given user requests are compliant with HIPAA privacy rules, a compliance score is assigned. The values of the compliance score are assigned as low (L), medium (M), and high (H) scale based on the user request being processed.[9,19,20].

This compliance H/M/L score is sent as an input to compute the risk score, which is the average of the three entities. The compliance score for each user request is then output as shown in Table 1. To elucidate, consider the scenario when a local user re-

Table 1. Risk assessment for exemplar user requests handled in the online health application.

Data source (A)	Data source (B)	Data source (C)	Overall likelihood (f1:max(A,B,C))	Compliance score (CS)	Total risk score (f2 = avg(f1, CS)	Overall risk level (L, M, H)
3	2	3	3	3	3	L
5	6	4	6	6	6	M
8	9	10	10	10	8	H
2	5	9	9	6	8	H
2	4	6	6	6	6	M
3	5	6	6	3	5	M
3	8	9	9	3	6	M
5	6	10	10	10	10	H

quests (compliance level: Low) for an aggregated data (Low) and data requested is "Return the count of condition type (diagnosis)" which is compliant with HIPAA policies (as the data does not have any PHI data); the corresponding overall risk score level is Low. In this case, the admin approves the user request and shares the data based on *Loss of opportunity* (data availability) priority considerations. Similarly, for the other seven user request scenarios, we enlisted the risk scores as shown in Table 1, with corresponding data access authorizations involving logging, notification, or formal (manual) approval from admin or governance committee.

5.3 CDM-enabled query results

After the (approve/deny) decision process has been completed, the response for the requested data access is shown on the UI dashboard of the data consumer. If the request is approved, the data consumer is granted access to the data via a data delivery mechanism that can include links to, *e.g.*, archived repositories or shared folders, which have an expiration period. Each of these data delivery mechanisms is linked to the data retrieval from the CDM implementation, and logged on the data custodian system side for auditability.

Table 2. CDM-enabled data query times in the community cloud testbed implementation of the semi-automated honest broker solution for high-volume patient data requests.

Requested data set	Query time
Counts of drug types	4.1 sec
Counts of persons with any number of exposures to a certain drug	328 ms
Distribution of age across all observation period records	709 ms
How long does a condition last	783 ms
Number of patients by gender, stratified by year of birth	846 ms
Number of people continuously observed throughout a year	178 ms
Number of people who have at least one observation period that is longer than 365 days	183 ms
Patient count per care site place of service	321 ms

We validated the functionality of our proposed CDM module by uploading a cataract data sample that includes health claim-related information. The query time shown in Table 2 is the parameter we use to evaluate the implementation of our CDM module in the honest broker implementation. Based on the query time shown in Table 2, we can observe that most of the resultant data set is retrieved in a time period of less than a second, whereas, the user request related to drug type took approximately four seconds. Thus, from the results shown in Table 2, we show that the complexity of the queries is impacted when performing joins and retrieving the required data sets from the relational schema. Based on these results, we

plan to further evaluate the effectiveness of our honest broker implementation in the future by considering additional usability-related metrics relevant to the user requests for the age-related cataract study, including: query time, data request time, and latency/delay for handling a data request. We also plan to compare the results for different community cloud infrastructure configurations and data source policy variants to investigate CDM-based designs that improve the user experience.

6. Conclusion

Health care data comprises heterogeneous data sets collected from multiple sources, such as patient information, health claims, billing info, and user demographic data. To pursue new trends and medical breakthroughs, researchers and clinicians are inclined to analyze or visualize any protected data sets. In this paper, we addressed the issue of long delays in data accessibility using a semi-automated honest broker solution within an online health application implementation. Our solution approach is targeted for an ophthalmology case study involving an age-related cataract research study that is supported via an actual community cloud testbed. Our honest broker solution facilitates the data request process in the case study based on the following steps. Firstly, a user can send a data request via the user interface module we developed. Secondly, these user requests are checked for compliance with HIPPA and data source policies as part of a compliance module that uses NLP. Thirdly, the risk associated with each user request is analyzed based on the risk assessment calculation that follows NIST SP 800 guidelines. Finally, once an approve/deny decision is taken based on the tradeoffs of *Loss of opportunity* and *Loss of value* criteria, the response to the user request is sent back to the user. If the decision associated to the user request results in an approval, then the requested data is retrieved using a CDM that transforms the disparity in any multi-source data as a common representation (terminologies, vocabularies, and coding schemes). In the other cases, our honest broker solution approach handles data accessibility with a limited custodian-in-the-loop intervention using logging, notification, or formal approval from admin or governance committee.

Our future work is to extend the CDM implementation with additional features related to an automated ETL process to bring multiple data sources into the standard model on a regular basis. In addition, we also plan to compare the effectiveness of our current semi-automated honest broker solution to the current manual governance considering other exemplar case studies. Using such detailed analyses, we seek to establish a fully automated honest broker solution with minimal dependency on the totally manual governance processes in current practice through advanced features that include user interfaces for non-experts to easily query within online health applications, and parallel computation services for scalable data processing in a variety of analysis contexts of health big data sets.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under award number OAC-1827177. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the NSF.

The authors would like to thank the following students at the University of Missouri, who contributed to this work: Matthew Chisholm, Matthew Gambino, Zachary Hess, Felipe Costa, and Joshua Westbrook.

References

1. Andreu-Perez J, Poon Y, Merrifield R, Wong S, Yang G. Big Data for Health. IEEE Journal of Biomedical and Health Informatics, 2015;19(4): 1193–1208.
2. What is Big Data?, https://datascience.nih.gov/bd2k/about/what, [Last accessed 05/28/2019].
3. Genomics and translational bioinformatics trending advancements and their working groups, https://www.amia.org/programs/workinggroups/genomics-and-translational-bioninformations, [Last accessed : 05/28/2019].
4. Aronson S J, Heidi L R. Building the foundation for genomics in precision medicine. Nature 526.7573, 2015;336. doi: 10.1038/nature15816.
5. Suh K, Sarojini S, Youssif M, Nalley K, Milinovikj N, Elloumi F, et al. Tissue Banking, Bioinformatics, and Electronic Medical Records: The Front-End Requirements for Personalized Medicine. Journal of Oncology, 2016;
6. Fayyad U, Piatetsky-Shapiro G, Smyth P, Pecora A, Schecter E, Goy A. Knowledge Discovery and Data Mining: Towards a Unifying Framework. Association for the Advancement of Artificial Intelligence, 1996;
7. Bergner M. Quality of Life, Health Status, and Clinical Research. Advances in Health Status Assessment, 1989;27(3):
8. An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule. NIST Special Publication 800-66 Revision 1, 2013; Available from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-66r1.pdf.
9. Security and Privacy Controls for Federal Information Systems and Organizations. NIST SP800-30 Technical Report. NIST Special Publications, 2013;
10. OMOP Common Data Model (CDM) V5.0. Observational Health Data Sciences and Informatics (OHDSI), 2019; Available from: https://www.ohdsi.org/data-standardization/.
11. Valluripally S, Murugesan R, Calyam P, Chisholm M, Sivarathri S, Mosa A, et al. Community Cloud Architecture to Improve Use Accessibility with Security Compliance in Health Big Data Applications. ICDCN '19 Proceedings of the 20th International Conference on Distributed Computing and Networking. ACM, 2019; 377–380.
12. Raju M, Chisholm M, Mosa AS, Shyu C, Fraunfelder FW. Investigating Risk Factors for Cataract Using the Cerner Health Facts® Database. Journal of Eye and Cataract Surgery, 2017; doi: 10.21767/2471-8300.100019.
13. Dhir R, Patel A, Winters S, Bisceglia M, Swanson D, Aamodt R, et al. A multidisciplinary approach to honest broker services for tissue banks and clinical data. Cancer, 2008;7, 1705–1715. Available from: https://doi.org/10.1002/cncr.23768. doi: 10.1002/cncr.23768.
14. Boyd A, Hunscher D, Kramer A, Hosner C, Saxman P, Athey B, et al. The Honest Broker Method of Integrating Interdisciplinary Research Data. AMIA Annu Symp Proceedings, 2005;

15. Orechia J, Pathak A, Shi Y, Nawani A, Belozerov A, Fontes C, et al. OncDRS: An integrative clinical and genomic data platform for enabling translational research and precision medicine. Applied & Translational Genomics, 2015;6, 18–25. doi: 10.1016/j.atg.2015.08.005.
16. Zhao Y, Yan B, Rocca WA, Wang Y, Shen F, Sauver J, et al. Annotating Cohort Data Elements with OHDSI Common Data Model to Promote Research Reproducibility. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018;1109(10): 1310–1314.
17. Sia Y, Wenga C. An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria. PMC Stud Health Technol Inform, 2017;245(1): 950–954.
18. Lambert GCA, Kumar P. Transforming the 2.33M-patient Medicare synthetic public use files to the OMOP CDMv5: ETL-CMS software and processed data available and feature-complete. 2016; Available from: http://www.ohdsi.org/web/wiki/doku.php?id=resources:ohdsi_symposium_2016_posters.
19. Ronald R. Guide for Conducting Risk Assessments. 2012; Available from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf.
20. Dickinson M, Debroy S, Calyam P, Valluripally S, Zhang Y, Antequera R B, et al. Multi-cloud Performance and Security Driven Federated Workflow Management. IEEE Transactions in Cloud Computing, 2018; Available from: https://ieeexplore.ieee.org/document/8392768.
21. Oh S, Cha J, Ji M, Kang H, Kim S, Heo E, et al. Architecture Design of Healthcare Software-as-a-Service Platform for Cloud-Based Clinical Decision Support Service. IEEE Healthcare Informatics Research, 2018;
22. Getting your data ready for precision medicine - https://www.ibm.com/blogs/insights-on-business/healthcare/getting-data-ready-precision-medicine.
23. Community cloud architecure for salesforce health care applications - https://www.salesforce.com/products/community-cloud/faq.
24. HumHub: Open-source Social Network Development Kit. Available from: https://humhub.org/en.
25. Berman M, Chase J, Landweber L, Nakao A, Ott M, Raychaudhuri D, et al. GENI: A Federated Testbed for Innovative Network Experiments. Elsevier Computer Network, 2014;61(14): 5–23. Available from: https://ieeexplore.ieee.org/document/8392768.