KUMC HSC IRB Protocol

Title
Establishing a Biomedical Informatics Data Repository: Project HERON (Healthcare Enterprise Repository for Ontological Narration)

Principal Investigator: Lemuel Russ Waitman, PhD
Co-Investigators: John D. Keighley, PhD, Gregory Ator, MD

Purpose of the research

The Division of Medical Informatics, Department of Biostatistics staff works on the repository project but they are not conducting research as part of this activity. They are serving in a technical role as an honest broker to provide data to the investigator so that limited data sets can be created.

We are requesting permission from the IRB to deploy a data repository which will be based on the i2b2 database architecture. The repository would be accessed primarily through a web-browser, allowing users to identify and analyze patient cohorts after they have signed a System Access Agreement. We also will allow researchers to define specific projects and download extracts from the repository for off-line analysis and hypothesis generation, assuming they have signed a Data Use Agreement and submitted their request to the Data Request Oversight Committee. Finally, if the researcher requests identified data, those requests follow current procedure and require prior approval by the IRB.

Background/Literature Review

i2b2 is an acronym that stands for "Informatics for Integrating Biology and the Bedside." It is an NIH-funded National Center for Biomedical Computing (NCBC) devoted to translational research (http://www.i2b2.org and Murphy et al).

More specifically, it is a scalable, open-source informatics framework and architecture that can be used to host a research data warehouse. This architecture consists of two major pieces. The first is the back-end infrastructure (the "Hive") that takes care of things like security, access rights, and managing the underlying data repository. The second piece is an application suite of query and mining tools that allows users to ask questions about the data (the workbench). The system was first developed within the Partner's HealthCare system in Boston at Massachusetts General Hospital. It served as the architecture for their Research Patient Data Registry.

Study Duration

This project is ongoing, with no specific end date.

Hypotheses/Specific Aims

Designed around cohort identification to facilitate translational research, this system will allow researchers to query patient populations to identify subsets based on certain inclusion and exclusion criteria. Included data sources will initially focus on data from electronic health records, lab results, and billing records. We will subsequently integrate data from public sources such as social security death indices. This information will be linked, aggregated and cleaned. There are two major data sources underlying the repository as shown in Figure 1 below. The first, available to the general i2b2 user, will be a general repository containing de-identified information on all the patients in our database and will be query-able through the web-based workbench tool. The second will maintain the identified data from source systems. The identified server is mainly used to link and clean the data from the source systems for loading into the de-identified server. However, if investigators require identified data for their research, the informatics staff can execute queries against the identified repository after IRB approval.

In coordination with the hospital (KUH), clinics (UKP), and university's executive teams, oversight is provided by a Data Request Oversight Committee (DROC) composed of representatives from each organization (currently including KUH CMIO, Director of the Human Subjects Committee, Director of Frontiers CTSA, KUH Director of Organizational Improvement, Chief Operating Officer of UKP). Oversight of requests for patient contact information is provided through our Clinical and Translational Science Award (CTSA), Participant and Clinical Interactions resources Program (PCIRP) (see the fourth use below). All users will sign a system access agreement (Exhibit A) to view data within HERON and a data use agreement (Exhibit B) to obtain data from the system. The system access agreement process is outlined in Figure 2 and the data use agreement process is outlined in Figure 3. There are four methods for interacting with the system:

1. View only: Using i2b2, investigators can conduct cohort identification queries and visualize patient population distributions. In this mode, all data still resides within HERON. Such viewing will require signing a system access agreement and user activity will be logged but do not require prior approval. After consultation and approval from the KUMC, UKP, and KUH privacy officials it was determined that we will expand functionality to allow investigators to view line item data in addition to performing counts. For example, the investigator will be able to see visually when the potassium lab result occurred relative to a drug exposure in a timeline view or see demographic breakdowns for populations using other plug-ins supported by the i2b2 application, Qualified faculty may also sponsor view only access by fellows, residents, students, and staff working under their supervision by submitting a sponsorship request to the DROC.
2. De-Identified: Requests for de-identified patient data which meet limited data set criteria are not deemed human subjects research but are reviewed by the DROC.

If the request deviates from standard practice the investigator is informed and the DROC convenes to review the request weekly.

3. Identified: Identified data requests require IRB review. After approval, HSC personnel coordinate requests with the medical informatics. Initially, identified data retrieval will only be performed by medical informatics personnel but will plan to implement retrieval tools with logging to facilitate HICTR personnel authorizing access and simplified retrieval of data by researchers.
4. Patient Contact: Researchers who request contact information for cohorts from the Frontiers Clinical Translational Science Unit (CTSU) Participant Registry have their study request analyzed for overlap with other requests by the Frontiers Data Request Committee. This is ensures proper coordination among investigators and evaluation of patient safety, overburdening, and the potential threat of cross-study contamination.

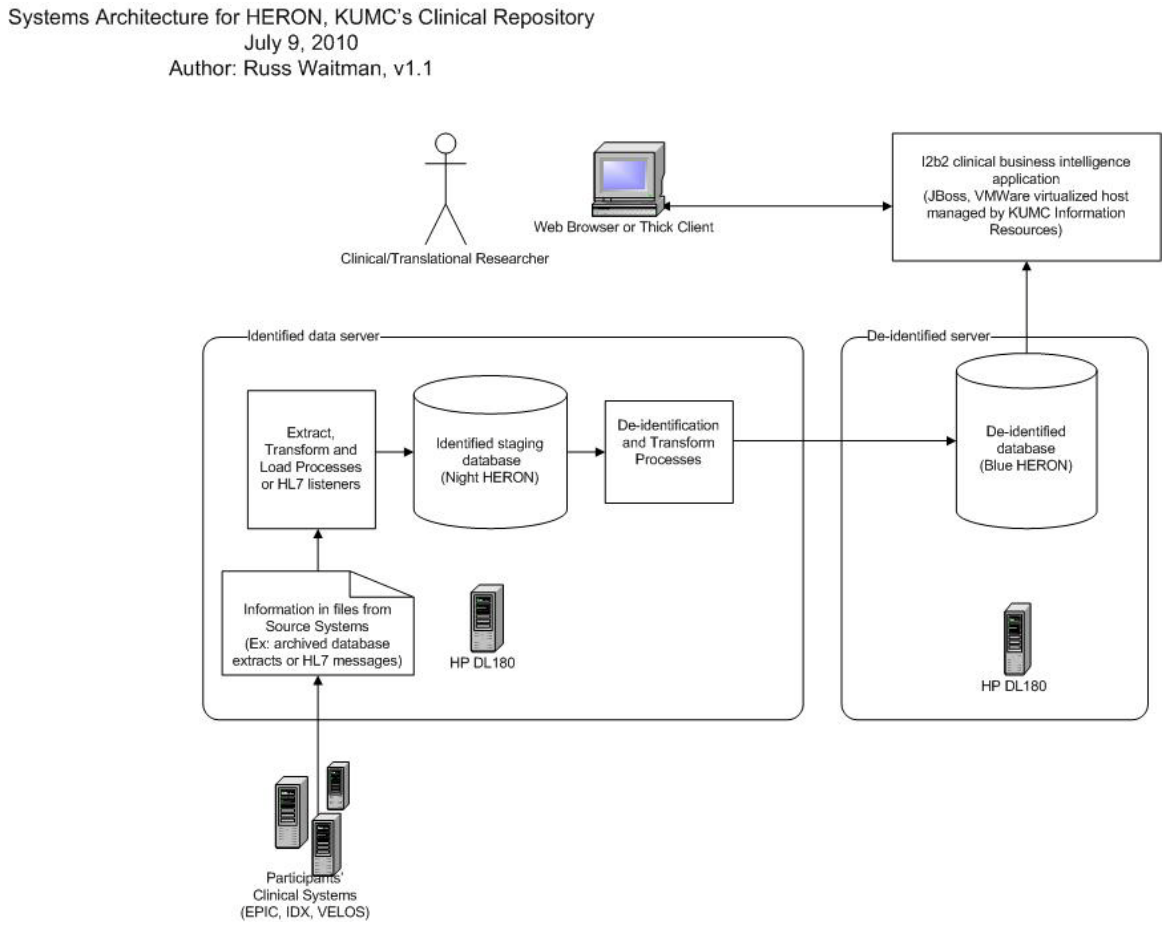

Figure 1. HERON Systems Architecture

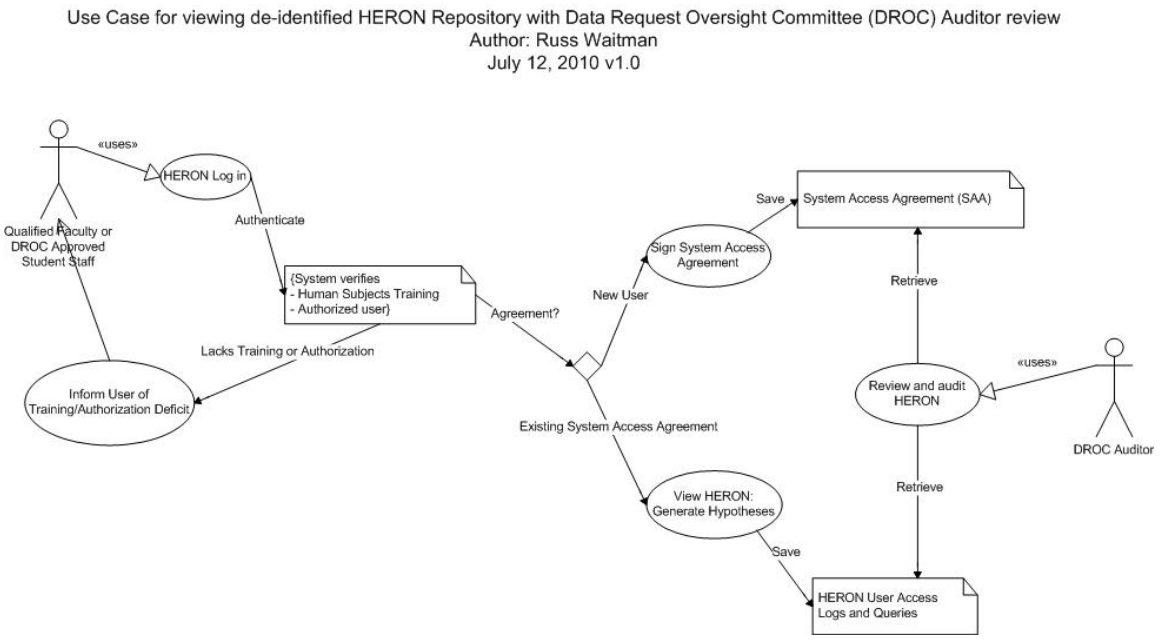

Figure 2. HERON Systems Access Agreement Process

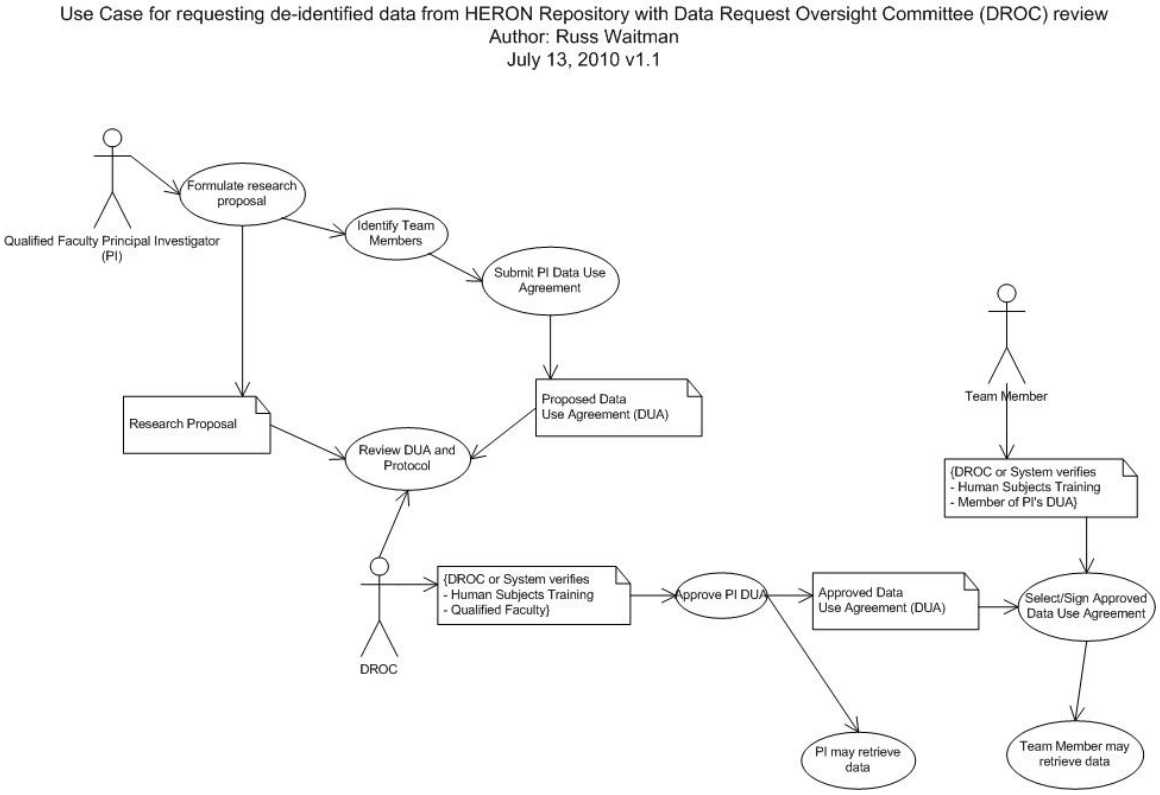

Figure 3. HERON Data Requests Process

Recruitment and Informed Consent Process (if applicable)

Kansas University Hospital alone sees over 25,000 hospitalizations, 40,000 emergency room visits, and 330,000 clinic visits each year. That number will only grow as the project progresses and our records are updated with new patient information. Obtaining consent from these individuals would be difficult, if not impossible. We believe that removing the need to obtain consent will not have any adverse affect on the patients. Only a limited dataset is present to the end user, ensuring that patient privacy is protected and the end user must sign a data use agreement. Identifiable information will only be provided to an investigator after approval from the IRB.

Methods and Measurement Tools

Data Sources:
Data from clinical and research information systems will be included in the repository. The data sources listed in the table below represent the initial sources of information which will be incorporated into the repository and our preliminary understanding of the methods we will use to obtain the data. As time progresses, other sources will be identified and added and the methods to receive data refined. The informatics team will be responsible for storing and maintaining all data. The data will be stored on an ongoing, long-term basis, with no pre-defined end date. The i2b2 framework allows for collaboration with outside institutions through federated queries, but we will seek further IRB approval before allowing such action.

Data Source	Contents	Transfer Method
IDX	Patient demographics, diagnoses, and procedures for ambulatory encounters and HICTR Participant Registry.	Extract performed as query from UKP SQLServer
O2, Epic	Patient demographics, diagnoses, procedures, medication orders, laboratory results and other patient observations for inpatient and (eventually) ambulatory encounters.	Initial approach will likely be received monthly from KUH as part of the Clarity back-up process.
Seimens/SMS	Master Patient Index Demographics, billing codes, and charges	Extract provided by KUH
Cerner CoPath	Pathology Reports	Extract or evaluate whether data can be obtain from EPIC or by picking off real-time HL7 feed to EPIC

Athena	Patient demographics, diagnoses, and procedures for ambulatory encounters managed by KUH clinics ("Jayhawk Clinics")	Likely an extract approach in coordination with KUH.
ORSOS	Perioperative scheduling, case carts, and timing information	Lower priority: Evaluate extract versus real-time feed.
CRIS, VELOS	KUMC Clinical research patient demographics and observations.	Extract from CRIS Oracle instance.
CAS	KUMC Central Authentication Service, maintains user identities	Published data sets and/or real-time authentication service.
REDCap	KUMC Clinical research registries containing patient/subject observations	Extract from REDCap's MySQL database or file transfer using REDCap's Application Program Interface (API)
Biospecimen Shared Resource	KUMC database characterizing research specimens	Monthly extracted file from KUCC personnel
Tumor Registry	KUH tumor registry with outcomes data for cancer patients	Monthly extracted file following the NAACR format provided by KUH Tumor Registrar's office
Social Security Death Master File	Publically available file from US Government of when people died who have social security numbers	Monthly file updates of SSDMF from ntis.gov

HIPAA Identifiers:

We will be transforming identified data into a form that addresses all of the 18 de-identification criteria. We will shift all in the EMR 1–365 days into the past; the shift is different across records but constant within the records of each patient, thereby allowing temporal analyses such as the development of adverse effects after a drug. We have listed the identifiers specified by HIPAA and whether they will be included in our data sources and the general i2b2 repository. While de-identified, we will be requesting that investigators treat released data with the same sensitivity as a limited data set.

Included in Source Data	**Included in de-identified i2b2 repository**	**Identifier**
Yes	No	1. Names
Yes	No	2. Postal address information. Zipcode has been requested as the predominant method for bundling

		cohorts of patients (ex: all zipcodes in Kansas City Metropolitan Area) but we will bundle search criteria into regions defining populations greater than 20,000. Example: we will allow users to search for patients within a 5 mile radius of KUMC but not the zip code 64111
Yes	No	3. Social security numbers
Yes	No	4. Account numbers
Yes	No	5. Telephone & fax numbers
Yes	No	6. Elements of dates for dates directly related to an individual, including birth date, admission date, discharge date, date of death. We will preserve the relationship between care encounters but randomly shifted dates, not actual dates, will be stored in the de-identified respository. The data stored may be up to 365 days before the actual date of service.
Yes	No	7. Medical record numbers
No *	No	8. Certificate/license numbers
No *	No	9. Electronic mail addresses
Yes	No	10. Ages over 89 and all elements of dates indicative of such age
Yes	No	11. Health plan beneficiary numbers
No *	No	12. Vehicle identifiers & serial numbers, including license plate numbers
No *	No	13. Device identifiers & serial numbers
No *	No	14. Web Universal Resource Locators (URLs)
No *	No	15. Internet Protocol (IP) address numbers
No (see note)	No	16. Biometric identifiers, including fingers and voice prints. Clinical molecular diagnostic results may be present in clinical laboratory results. We do not intend to incorporate large scale microarray expression data or full genome sequencing in HERON. If that was requested, we would submit a separate IRB application.
No *	No	17. Full face photographic images & any comparable images
No *	No	18. Any other unique identifying number, characteristic or code that is derived from or related to information about the individual

Identifiers marked with a '*' are not believed to be captured in any of our data sources, but they may be added without our knowledge.

Data Processing

The data that we receive from our sources are fully identified. We use these identifiers – primarily medical record number – to create a single, integrated set of data for each patient. We also use these identifiers to help remove any spurious or duplicate information. All procedures used for the extraction, transformation and loading (ETL) of data are performed on identified server (reflected on the lower left side of Figure 1). Once the initial ETL process is complete, the second ETL process removes the identifiers before loading data into the general i2b2 repository.

To allow for the historical linking of records across multiple data sources, patient and physician numbers are replaced with an arbitrarily assigned value through a process called a "one-way hash". This allows us to maintain a connection across records without revealing any identifying information. Additionally, we will not incorporate free text notes until we are able to implement de-identification scrubbing technology to remove names and other identifiers from free text (see Roden et al)

The de-identified repository is deployed on a database (Oracle) that is physically and logically separate from the identified database (which includes PHI and HIPAA identifiers). Access to the production database is controlled by username and password. The i2b2 workbench accesses the de-identified database through the middleware server that connects (read-only) using a unique username and password. The identified database is controlled by a separate set of usernames and passwords that are only available to members of the informatics team.

Access Methods:

There are three primary routes for users to access data in the i2b2 repository:

1. Using the web-based Workbench, which handles user authentication and provides automated query, export and analysis tools. Figure 4 provides a screenshot of the web-based i2b2 workbench used for cohort identification.
2. Using the Java client of the i2b2 Workbench. This provides similar capabilities as the web-based i2b2 workbench but has some additional methods for working with de-identified datasets.
3. By consulting with the informatics team to create specialized queries and reports.

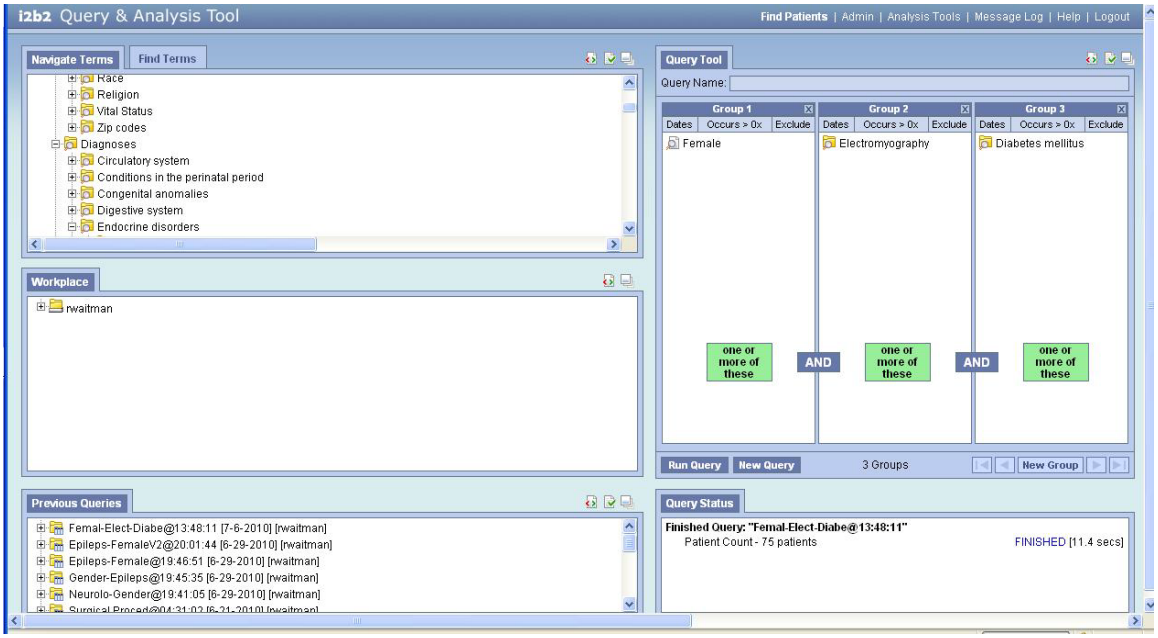

Figure 4. The i2b2 web-based workbench illustrating patient cohort identification for female patients with an ICD9 code for diabetes mellitus and a CPT procedure codes for electromyography.

The i2b2 workbench provides allows the user to query the repository after logging in with their unique username and password. We will integrate the user authentication method with the existing KUMC Central Authentication Service. Using the workbench (see picture above), a user can drag-and-drop search terms into a Venn diagram-like interface (upper right) Once executed, the query will return the aggregate number of patients meeting the specified criteria (lower right). If there are more than 10 patients in any age, race and gender grouping, that value is presented to the user. If the value is less than 10 (but greater than zero), we denote it with a '<10'. The user is able to view their previous queries (lower left). After approval by the Data Request Oversight Committee, the informatics team can configure the user access so that the user may generate a limited, de-identified export of the patient cohort.

Users may need the ability to work directly with informatics team to meet their data needs. The team would typically handle complicated queries, queries involving specialized data sources, or those involving identified data after the investigators has obtained IRB approval for their study.

Potential Benefits

For the first time, researchers and investigators will be able to directly query a large data repository for the purposes of cohort identification and hypothesis generation. We will also be able to augment many existing research databases, such as the Clinical Research Information System, with a wealth of clinical information that was previously

unavailable, allowing investigators to draw new insights and conclusions as they perform their data analysis.

Federal regulations state that the creation of a research repository is a research activity that requires prior IRB approval. However, we are constructing the repository to capitalize on regulatory flexibility that allows the repository to be constructed such that the end users are not engaged in human subjects research. The repository will meet the criteria outlined by the federal Office for Human Research Protections in its October 2008 document "Guidance on Research Involving Coded Private Information or Biological Specimens," i.e., the data were originally collected for a clinical purpose and the end user cannot readily ascertain the identity of individuals in the database. Because these criteria are met, the end user is not conducting human subjects research and does not need IRB approval for his/her individual project.

By constructing the repository in this way, the HERON system offers an invaluable resource to investigators who can rapidly assess the feasibility of future research and perform initial hypothesis testing. Further, the repository supports the research goals of the medical school and residency programs who aim to expose trainees to basic concepts of research through activities such as retrospective chart reviews that must be completed in an interval that rarely allows sufficient time for IRB submission and approval.

Funding

This study is investigator-initiated and is not affiliated with any other IRB-approved study. The research and funding will be conducted as part of the initial establishment of the division of medical informatics and we have adequate resources to complete the project. As described in the attached governing legal agreement, organizational leadership is provided by an executive committee with representation from the University of Kansas Medical Center, the University of Kansas Hospital Authority, and the Kansas University Physicians, Inc.

Statistical Analyses

Creation of the repository is not by itself a scientific hypothesis but instead a tool which will facilitate future hypotheses for the greater research community. Dr. John Keighleywill be engaged with the team to choose the proper statistical methods for data characterization and identification of aberrations which may occur and need correction as part of the ETL processes.

Data Security

We are taking the utmost care to ensure that there are no unauthorized releases of protected information. Patient identifiers are stored in a separate, secured database. In the de-identified repository, patient identifiers are randomly assigned, meaning there is no way to trace them back to an individual. Steps will be been taken to ensure that demographic information is not displayed for queries returning small patient counts (i.e.

less than 10), ensuring that a malicious user cannot try to uniquely identify a patient through the workbench tool. Finally, no data will be released to the end user unless they have submitted a Data Use Agreement and their request reviewed by the Data Request Oversight Committee. Because of the risk of re-identification using other publically available datasets (ex: Benitez and Malin), researchers must treat disclosed limited datasets with sensitivity. All use of the i2b2 data repository is logged for subsequent analysis by the Data Request Oversight Committee.

Physical location of warehouse:
The identified and de-identified servers are housed in the KUMC/KUH data center. Physical access to the data center is controlled by locked doors and the servers are accessible only to those who a) are logged on to the KUMC internal network and b) have been explicitly granted access to the servers. All patient identifiers and any associated PHI are kept in a separate database on a separate server accessible only to those on the informatics team.

User Access
Initially, access to the i2b2 workbench is obtained by submitting an i2b2 access request form to the informatics team. Subsequently, we will develop methods that integrate i2b2 login and authentication with KUMC Central Authentication Service and Chalk to determine if the user is a faculty member and has current human subjects training. After access is granted, the user will be assigned a username and password. After initial cohort identification, when a principal investigator wishes to create a de-identified dataset, a data use agreement will be submitted for the project identifying additional members of the investigators team. Access for other research staff will be granted, but a supervising faculty member must first approve the request. These staff members will be placed in an i2b2 project with the supervising faculty, who assumes all responsibility for the collective actions of the group.

References

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.

Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84:362-369.

Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010 Mar 1;17(2):169-77.PMID: 20190059 [PubMed - indexed for MEDLINE]Related citations

Karp DR, Carlin S, Cook-Deegan R, Ford DE, Geller G, Glass DN, Greely H, Guthridge J, Kahn J, Kaslow R, Kraft C, Macqueen K, Malin B, Scheuerman RH, Sugarman J.

Ethical and practical issues associated with aggregating databases. PLoS Med. 2008 Sep 23;5(9):e190. No abstract available.