NCATS ENACT
Cohort Discovery, Research on De-identified Aggregated Data (RDAD) and Enclave Research
i2b2 Data Repository

1. Purpose and objectives:

The goals of the National Clinical and Translational Science Award (NCATS) Evolving Next-Generation Accrual to Clinical Trials (ENACT) project, formerly Accrual to Clinical Trials (ACT), are to enable the federated network made up of sites from the Clinical and Translational Science Award (CTSA) Consortium, to accelerate clinical research through Cohort Discovery and Research on De-identified Aggregated Data (RDAD) and to utilize the patient-level data collected across the federated network in a safe, secure, Health Insurance Portability and Accountability Act of 1996 (HIPAA) compliant environment. ENACT takes advantage of the widespread implementation of electronic health records (EHRs) and the well-established extensive informatics and regulatory expertise within the CTSA Consortium.

The purpose of this document is to assist CTSA sites in obtaining IRB approval for their i2b2 Cohort Exploration Data Repository (hereafter called the i2b2 repository) which will contain identifiable patient data from electronic health records and from other sources such as clinical billing systems. This paper can also be used as a guide to amend an existing i2b2 IRB-approved protocol to add new data elements (e.g., dates) and NCATS ENACT sites, as necessary.

NOTE: Technology Work Group members have the requisite knowledge about the development of your site's i2b2 repository and the Shared Health Research Information Network (SHRINE) system.

2. Background:

Institutions participating in the NCATS ENACT federated network will use the i2b2 open-source software framework for their data repository. i2b2 is an acronym that stands for "Informatics for Integrating Biology and the Bedside." The i2b2 software was developed at Harvard through an NIH-funded National Center for Biomedical Computing (NCBC) program devoted to translational research. The i2b2 architecture consists of two major components. The first component is the core infrastructure (called the "Hive") that manages data extraction, data security, and the underlying data repository. The second component is an application suite of query and mining tools.

The i2b2 repository at each NCATS ENACT site will be federated by a companion software system developed at Harvard called the Shared Health Research Information Network (SHRINE). Using the i2b2 platform, SHRINE architecture enables user querying across the federated network of CTSA sites. The output is an aggregated patient count from each network site or an aggregated dataset.

Based on the query results, an investigator can:

- assess the feasibility of a study by determining if sufficient number of patients are available
- determine which sites to approach for collaboration in a proposed clinical trial based on the distribution of the proposed subject population

- Perform RDAD
- secure identifiable patient-level data collected across the federated network to be used and analyzed in a secure, HIPAA-compliant environment

When coupled with the EHR data and data from other sources, query results can lead to recruitment of patients into clinical trials, by identifying potentially eligible patients after securing appropriate institutional approvals and to exploration of important health questions by utilizing a patient-level data maintained in a HIPAA-compliant, "safe" cloud environment for a specific, limited time.

3. Procedures:

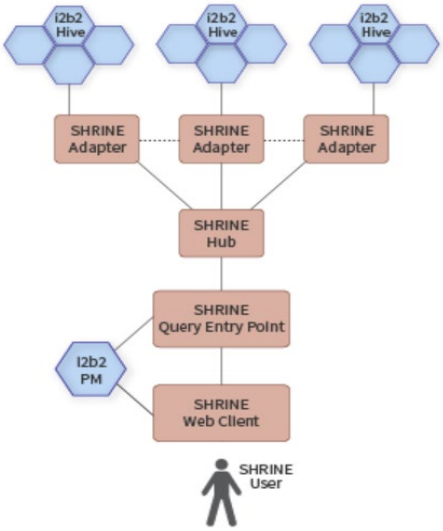


Figure 1: Example of three i2b2 repositories ("hives") as a federated network using SHRINE software

Informatics for Integrating Biology and the Bedside (i2b2) Repository

The i2b2 open-source software is conceptualized as a "hive" and involves two critical concepts. The first concept is the existence of services provided by applications that are "wrapped" into functional units, such that their functionality is exposed as messages that travel to and from the various cells of the hive. The i2b2 software has core cells that provide data extraction, de-identification, data conversion, and management of the "hive." The second concept is that of persistent data storage, which is managed by the cell named the "Clinical Research Chart."

The i2b2 repository is comprised of data obtained directly from the EHR and from other relevant sources. Patient identifiers – primarily medical record numbers – are used to create a single, integrated set of data for each patient. Identifiers are also used to help remove any spurious or duplicate information. Once all procedures used for the extraction, transformation, and loading (ETL) of data are performed the initial ETL process is complete. The second ETL process removes the patient identifiers (except dates) and harmonizes the data values to a common representation (e.g., common values for race). It is during this process of extraction from the EHR that the i2b2 repository is developed and updated.

To allow for the linking of records across multiple data sources, within the i2b2 repository the patient medical record numbers are replaced with an arbitrarily assigned value. This allows the maintenance of a connection across these data sources. The key linking the arbitrary numbers and the patient identifiers is stored in a separate file that is only accessible by the i2b2 systems administrators and informatics personnel.

The i2b2 repository is physically and logically separate from the EHR and other data sources. User access to the i2b2 repository is controlled through the SHRINE web client portal using a unique username and password.

The i2b2 software framework includes the following data security strategies:

a. Data exclusion such as removal of the explicit identifiers.
b. Data transformation, in which an irreversible modification is made to the data that destroys the original values while preserving the relationships of interest; for example, by adding random noise to the data.
c. Data obfuscation, in which the domain of data values is changed while preserving relationships.
d. Small cell size, for queries that results in less than 10 patients per site the phrase, "Less than 10 patients" will be displayed instead of the actual/obfuscated number.
e. Data encryption, in which identifiers are changed using hashing algorithms.

To further minimize the risk of a breach of patient privacy due to loss of data, the following precautions will be put in place at network sites:

a. All servers will be located behind a local firewall in a secure data center.
b. Access to the servers will be limited to approved systems administrators and informatics personnel.
c. The following types of data are excluded from the i2b2 repository:
    - Patient Medical Record number
    - Patient and treating physician's first and last names
    - Phone, fax, and pager numbers
    - Patient address
    - No data from notes (clinician, progress notes, etc.)
    - Genomic data is currently NOT available

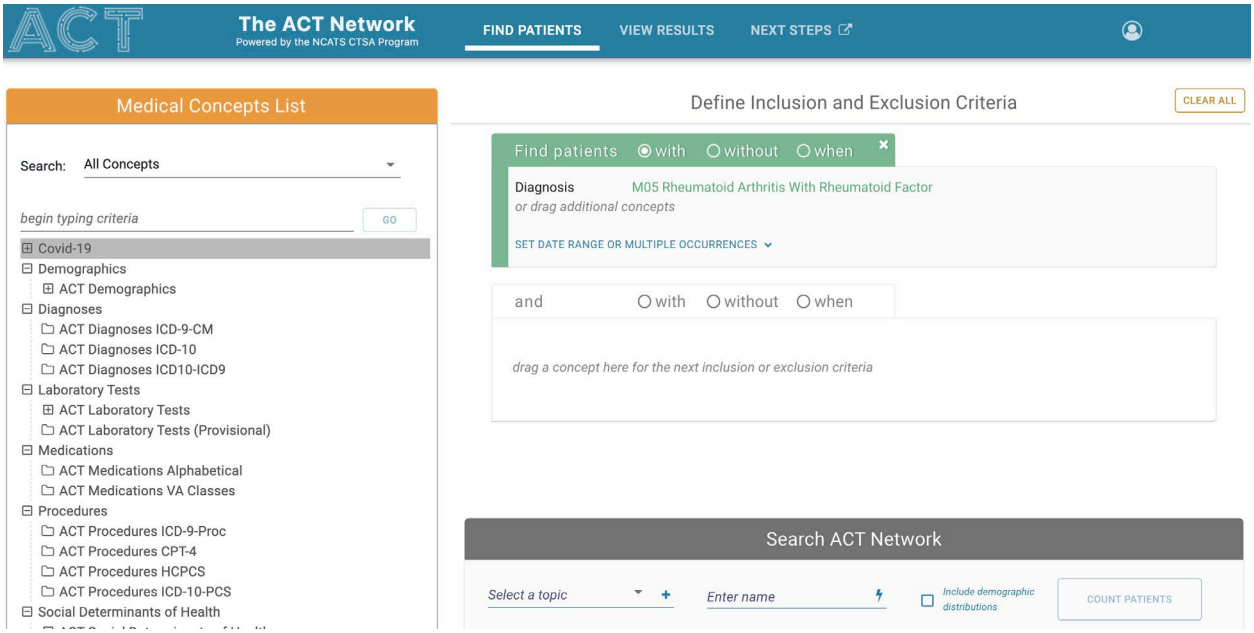
In conducting the information technology security assessment, in addition to the standards set forth by the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), sites may also adhere to security standards set forth by internal site administration, the National Institute of Standards and Technology (NIST), or other security framework. Sites utilizing these alternative standards may be called upon to provide documentation of policies that cover the following twelve security areas:

Risk Management	Security Policy	Organization of Information Security	Asset Management

Human Resources Security	Physical and Environmental Security	Communications and Operations Management	Access Control
Information Systems Acquisition, Development, and Maintenance	Information Security Incident Management	Business Continuity Management	Compliance

Shared Health Research Information Network (SHRINE)

Figure 2: SHRINE User Query Tool



The Shared Health Research Information Network (SHRINE) architecture is comprised of four components: web client, query entry point, hub, and adapter (Figure 1).

The SHRINE web client will be the users' gateway into a NCAT ACT network. The first step in a user's interaction with the NCAT ACT network will be to authenticate using a username and password. Once this authentication occurs, the user's session information is included in all requests that the web client sends to the SHRINE Query Entry Point (QEP).

The Query Entry Point (QEP) is the service entry-point into NCAT ACT network. The QEP authenticates all incoming user requests. After authenticating the user's request, the QEP forwards that request on to the SHRINE Hub for broadcasting throughout the network.

When the QEP receives the collected results back from the Hub, it aggregates the results into a single response to the user (i.e., aggregated patient counts by site).

The Adapter is the component that acts as an entry point to a site's i2b2 hive. The Adapter receives queries from the Hub and translates them into local site terminology.

The SHRINE architecture is configured with the following security features that ensure queries are securely broadcast only to other databases in the federated network:

a. Each user of the system needs to be authenticated at their individual site to verify employment and faculty or qualified designee status.

b. All communications are encrypted using standards approved by the W3C Consortium.

c. Queries return only aggregate counts, rounded to the nearest 5.

d. Aggregate numbers are blurred (or obfuscated) by +/-10, so that the counts returned are an estimate of the number of patients meeting the queried upon criteria at each site.

e. For queries where results include only patients ages 90 or older per site, the phrase, "Age 90 or older" will be displayed instead of the obfuscated number.

f. No personally identifiable patient information ever leaves an individual site.

g. Institution-specific user log-in credentials never leave an individual site.

h. Each site can only communicate with other sites in the NCAT ACT network via the SHRINE Hub.

i. Actual query histories are logged and maintained.

4. NCATS ACT Data Elements:

The primary source of data imported into a site's i2b2 repository is EHRs. Data from other sources is also gathered as appropriate to augment the phenotypic characteristics of each patient. Data elements being stored in the i2b2 repository include:

Demographics

- Birth Date (YYYY-MM-DD) used to generate Current Age data at time of query
- Sex (A = Ambiguous, F = Female, M = Male, O = Other, NI = No information)
- Hispanic (Y = Yes, N = No, NI = No information)
- Race (1 = American Indian or Alaska Native, 2 = Asian, 3 = Black or African American, 4 = Native Hawaiian or Other Pacific Islander, 5 = White, 6 = Multiple race, NI = No information)
- Vital Status (1 = Known Deceased)
- Death Date (YYYY-MM-DD)

Diagnoses

- Diagnosis (Diagnosis code)
- Diagnoses Type (1 = ICD-9-CM version x, 2 = ICD-10-CM version x)
- Diagnoses Date (YYYY-MM-DD)
- Diagnosis Source (AD = Admitting, DI = Discharge, FI = Final, IN = Interim, NI = No information)
- Diagnosis Priority (P = Principal, S = Secondary, NI = No information)

Procedures

- Procedure (Procedure code)
- Procedure Code System (ICD-9-CM, ICD-10-PCS, CPT-4, HCPCS)
- Procedure Type (1 = ICD-9-CM version x, 2 = ICD-10 version x, 3 = CPT-4 version x)

- Procedure Date (YYYY-MM-DD)

Visit Details

- Admit Date (YYYY-MM-DD)
- Discharge Date (YYYY-MM-DD)
- Visit Type (IP = Inpatient Hospital Stay, AV = Ambulatory Visit, ED = Emergency Department Visit, OA=Other Ambulatory Visit, NI = No information)

Medications

- Medication Code (Medication coding systems RxNORM, NDC, HCPCS)
- Medication Source (Dispense, Fill, Admin, Other)
- Order Date (YYYY-MM-DD)

Laboratory Tests

- Lab Code (LOINC laboratory test code)
- Specimen Date/Time (YYYY-MM-DD and HH:MM:SS)
- Result Qualitative (BORDERLINE, POSITIVE, NEGATIVE, UNDETERMINED, NI=No information)
- Result Quantitative
- Result Modifier (EQ=Equal, GE=Greater than or equal to, GT=Greater than, LE=Less than or equal to, LT=Less than, TX=Text, NI=No information)
- Result Unit of Measure
- Abnormal Result Indicator (AB=Abnormal, AH=Abnormally high, AL=Abnormally low, CH=Critically high, CL=Critically low, CR=Critical, IN=Inconclusive, NL=Normal, NI=No information)

Vaccination

- Vaccination Code (Vaccination coding systems CVX, RxNorm, NDC, CPT4, ICD10PCS)
- Vaccination Date

Social Determinates of Health (SDoH)

- SDoH Code (LOINC, SnoMed)

Vital Signs

- Height
- Weight
- Blood pressure
- BMI

Derived Variables and Scores

Future Additions

- Data quality totalnums
- Note metadata (note type, note date)
- Genetic data

5. Subject Population:

The i2b2 repository will use data from the EHRs of patients from the CTSA site and/or a CTSA partner site (medical centers, community hospitals, etc.). At least 3 years' worth of data patient records from the EHR database will be included in the i2b2 repository.

6. Recruitment:

Not Applicable. All patient records will be included in the i2b2 repository.

7. Institutional Review Board Approval:

The i2b2 repository qualifies as Expedited, Category 5 – "Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment or diagnosis)."

8. Informed Consent and Requesting a Waiver of Consent (45CFR46.116(f)):

i. The project is not FDA-regulated.

ii. The development of the i2b2 repository involves no more than minimal risk. Minimal risk is defined as the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life. The primary risk is a breach of privacy. Privacy breaches from malicious "hacking" or erroneous disclosure of private information occur daily. The risk of a privacy breach in this protocol will be minimized through appropriate privacy and security measures described in this paper.

iii. The project could not practicably be conducted without a waiver. Obtaining informed consent from such a large number of patients is impracticable.

iv. The i2b2 repository involves using EHR data and data from other sources, creation of the repository would be unmanageable without using patient identifier to link the information.

v. Granting the waiver will not adversely affect privacy rights and welfare of the individuals whose records will be used.

- Cohort Discovery - Using the i2b2 repository to obtain de-identified patient counts across the federated network does not constitute human subjects research. Users will not obtain data through intervention or interaction with the patient; or obtain identifiable private information about the patient. Also, for queries that results in less than 10 patients per site the phrase, "Less than 10 patients" will be displayed instead of the actual/obfuscated number. For queries that results include only patients ages 90 or older per site, the phrase, "Age 90 or older" will be displayed instead of the obfuscated number.
- Enclave Research – The i2b2 repository will provide access to a patient-identifiable data. The data will be maintained in a HIPAA-compliant, "safe" cloud environment for a specified, limited time. The data cannot be copied, printed, or transferred. Request for an Enclave Research dataset will require separate IRB approval.

9. Requesting a Waiver of Health Insurance Portability and Accountability Act (HIPAA) Research Authorization (45 CFR 164.512 (i)(2)(ii)):

The use or disclosure of the Protected Health Information (PHI) involves no more than minimal risk to the privacy of individuals based on, at least, the presence of the following elements:

i. An adequate plan to protect the identifiers from improper use and disclosure. Identifiers will be stored in a separate database to which access is limited to approved systems administrators and clinical informatics personnel.

ii. An adequate plan to destroy the identifiers at the earliest opportunity consistent with the conduct of the research, unless there is a health or research justification for retaining the identifiers or such retention is otherwise required by law. Maintenance and updating of the i2b2 repository will be ongoing with no specific end date. The project will include a governance structure and documented user agreement. The i2b2 repository contains dates which constitutes a limited dataset under HIPAA regulations, Cohort Discovery users only receive de-identified aggregated patient counts across the federated network. Enclave Research users will have access to a patient-identifiable dataset maintained in a HIPAA-compliant, "safe" cloud environment for a specified, limited time and will be required to secure separate IRB approval.

iii. An adequate written assurance that PHI will not be reused or disclosed to any other person or entity, except as required by law, for authorized oversight of the research study, or for other research for which the use or disclosure of protected health information would be permitted (i.e., under the HIPAA regulations). Cohort Discovery activities do not constitute human subjects research. Enclave Research users will be required to sign a User Terms of Data Access agreement affirming that they will not reuse or disclose PHI.

iv. The project could not practicably be conducted without the waiver. Obtaining consent from such a large number of patients is impracticable.

v. The project cannot practicably be conducted without the use of PHI. The i2b2 repository is comprised of data obtained directly from the I and from other relevant sources. Patient identifiers – primarily medical record numbers – are used to create a single, integrated set of data for each patient. An adequate plan to protect PHI from improper use and disclosure is described in this paper. The i2b2 repository will not contain explicit identifiers (e.g., name, address, medical record number, social security number, health plan number, etc.). It will contain a limited data set (i.e., dates). Medical record numbers will be maintained in a separate database to which only a limited number of project team members have access. Authorized Users will also be required to sign a User Terms of Data Access agreement prior to accessing the i2b2 repository.

vi. The privacy risks are reasonable relative to the anticipated benefits of project. Moreover, as discussed in this paper, established security measures minimize privacy risks.

vii. Whenever appropriate, subjects will be provided with additional pertinent information after participation. Research activities that involve the use of patient identifiable data will undergo separate IRB review.

10. Anticipated Duration of Maintaining the Repository:

The maintenance and updating of the i2b2 repository will be ongoing with no specific end date. The NCATS ENACT award is funded through 2027.

11. Foreseeable Risks:

The primary risk of maintaining the i2b2 repository is a breach of patient privacy due to the exposure of the patient data (including dates) during ongoing maintenance and updating of the repository. No patient identifiable information leaves the i2b2 repository. Enclave Research users will have access to a patient-identifiable dataset maintained in a HIPAA-compliant, "safe" cloud environment for a specified, limited time and be required to secure separate IRB approval.

12. Expected Benefits:

There is no direct benefit to patients. Patients may benefit from enrollment in future clinical trials. It is expected that the federated network will improve the efficacy and quality of clinical and translational research and will accelerate subject accrual into critical clinical trials.

13. Alternative to Study Participation:

All patients within a site's EHR system will have information included in the i2b2 repository. Patients will have the ability to decline participation in future clinical trials.

If you have questions, please contact:

Eric Mah, MPH, PhD
Associate Dean, Clinical and Translational Research
University of California, San Diego
emah@health.ucsd.edu
Phone: 858-822-4700