Title

Implementation of a deidentified federated data network for population-based cohort discovery

Permalink

https://escholarship.org/uc/item/4222f43r

Journal

Journal of the American Medical Informatics Association, 19(e1)

ISSN

1067-5027

Authors

Anderson, Nicholas
Abend, Aaron
Mandel, Aaron
et al.

Publication Date

2012-06-01

DOI

10.1136/amiajnl-2011-000133

Peer reviewed

Implementation of a deidentified federated data network for population-based cohort discovery

Nicholas Anderson,[1] Aaron Abend,[2] Aaron Mandel,[2] Estella Geraghty,[3] Davera Gabriel,[3] Rob Wynden,[4] Michael Kamerick,[2] Kent Anderson,[3] Julie Rainwater,[3] Peter Tarczy-Hornoch[1]

► An additional material is published online only. To view this file please visit the journal online (www.jamia.bmj.com/content/19/e1.toc).

[1]Department of Biomedical Health Informatics, University of Washington, Seattle, Washington, USA
[2]Recombinant Data Corporation, Newton, Massachusetts, USA
[3]University of California Davis, Davis, Sacramento, California, USA
[4]University of California San Francisco, San Francisco, California, USA

Correspondence to
Dr Nicholas Anderson, Biomedical and Health Informatics, Department of Medical Education, Biomedical Informatics, University of Washington, Box 358051, 815 Mercer St, Seattle, WA 98109, USA; nicka@uw.edu

Received 23 January 2011
Accepted 7 July 2011
Published Online First
26 August 2011

ABSTRACT

Objective The Cross-Institutional Clinical Translational Research project explored a federated query tool and looked at how this tool can facilitate clinical trial cohort discovery by managing access to aggregate patient data located within unaffiliated academic medical centers.

Methods The project adapted software from the Informatics for Integrating Biology and the Bedside (i2b2) program to connect three Clinical Translational Research Award sites: University of Washington, Seattle, University of California, Davis, and University of California, San Francisco. The project developed an iterative spiral software development model to support the implementation and coordination of this multisite data resource.

Results By standardizing technical infrastructures, policies, and semantics, the project enabled federated querying of deidentified clinical datasets stored in separate institutional environments and identified barriers to engaging users for measuring utility.

Discussion The authors discuss the iterative development and evaluation phases of the project and highlight the challenges identified and the lessons learned.

Conclusion The common system architecture and translational processes provide high-level (aggregate) deidentified access to a large patient population (>5 million patients), and represent a novel and extensible resource. Enhancing the network for more focused disease areas will require research-driven partnerships represented across all partner sites.

INTRODUCTION

Identifying potential subjects for clinical trial recruitment and for power calculations is a significant challenge for researchers, who use a range of techniques to prospectively identify subjects based on a range of eligibility criteria. Electronic patient health data collected during the normal course of care are a critical resource for identifying potential subjects, but are not typically captured to allow answering of research queries. They are typically stored in a variety of proprietary data systems with interfaces in a wide range of formats, and require independent institutional regulatory approval if they involve protected health information (PHI) data on patients.[1] The increasing availability of electronic medical record (EMR) systems or multi-source integrated data repositories (IDR) are advancing our ability to screen for patients,[2 3] but individual EMR or IDR systems often cannot provide the population sizes necessary to identify sufficiently large cohorts of patients with rare diseases and/or restrictive eligibility criteria.

The research community sees potential for IDR systems to provide research access to EMR systems and associated data within institutions.[4] Currently, however, IDR systems have not advanced inter-institutional data collaborations because of the challenges of coordinating inter-institutional semantic data alignment and regulatory approval that ensures patient and institutional privacy and providing usable services to end-user researchers.[5–7] The Cross-Institutional Clinical Translational Research (CICTR) project thus sought to examine and pilot approaches to overcoming these difficulties by giving researchers access to a large population of deidentified (and thus human subjects exempt) potential subjects across multiple, geographically separated and distinct IDR systems.

BACKGROUND AND SIGNIFICANCE

The CICTR project was formed to pilot systems and processes for information exchange across three partner sites of the Clinical Translational Science Award (CTSA) consortium, located at the University of Washington (UW), the University of California, San Francisco (UCSF), and the University of California, Davis (UCD). The CTSA consortium includes 55 research institutions, each with a large and heterogeneous patient population, and has a mandate to enhance cross-institutional collaborations.[6]

Although the two University of California (UC) schools were academically affiliated through the UC system, none of the sites had affiliated healthcare systems or a history of working together to share clinical data for research. Each site was independently developing an IDR by implementing or considering implementing Informatics for Integrating Biology and the Bedside (i2b2) as a platform for providing research access to clinical data. Leveraging this work and the collaborative opportunity provided by the CTSA consortium, the CICTR project implemented a federated query environment based on the i2b2 platform.[2] There were three phases over 18 months.

- Phase 1: establish a common technical foundation to federate queries and exchange test patient data without requiring regulatory approval.
- Phase 2: pilot a cohort discovery service against real deidentified demographics and disease diagnosis PHI data.
- Phase 3: extend the resolution of cohort data by implementing common queryable mappings

to selected medications, laboratory data, and discharge dispositions.

METHODS

Meeting the incremental goals of these phases was highly dependent on coordinating strategy, expertise, process, and resources with a range of stakeholders in both research and clinical environments and across all three organizations. The project had fixed milestones as conditions of funding, and required a flexible and reactive coordination approach to maintaining momentum and supporting experiences at each site. Of the 12 core individuals involved across all sites and including the independent subcontractor, Recombinant Data Corporation, none had more than ~25% of funded time available to dedicate to the project. To coordinate these activities, team members were organized into one or more of four thematic teams: software deployment, semantic interoperability, policy oversight, and evaluation of end-user experience. All teams met weekly or biweekly throughout the project. The project also hosted three open symposia (held at each site in turn at the conclusion of each phase) which provided additional opportunities to disseminate, seek input, and enhance the strategy for subsequent stages. These interdisciplinary meetings offered the project team opportunities to coordinate and share experiences on deploying an IDR for cross-site data sharing, and also provided useful opportunities for involving and recruiting future stakeholders to the i2b2/Shared Health Research Information Network (SHRINE) community. Throughout the project period, the group maintained a central website, wiki, and source archival system that tracked issues, documentation, use cases, test cases, and configuration requirements.

The project had a primary focus of diabetes to limit project scope, and developed a set of diabetes-focused use cases early in the project and used these throughout all phases. The use cases had two purposes: (1) to facilitate coordination and communication of the common end-user expectations of research functionality of the resulting network IDR systems; (2) to define measurement criteria to ensure that the technical deliverables and data mapping work at each site were met. These use cases were developed through a literature review of published and completed studies with keywords of diabetes, population, public health, and epidemiology, and included representative study design methods: cross-sectional, cohort, and randomized controlled trials (online supplementary material).

Phase 1: establishing a common technical foundation

This phase focused on coordinating approaches to two work areas: (1) identifying and accessing heterogeneous site-specific clinical data environments; (2) establishing and testing local and network system security in advance of exchange of real clinical data. Instrumental to this work was team building and information dissemination to support the technical infrastructure for the three site-federated query capabilities.

Assessing infrastructure expectations and dealing with inconsistent back-end systems

At project initiation, each site had various locally developed solutions for providing clinical data for research and was engaged in developing more robust access to enhanced resources as part of each individual CTSA site's research mission. Existing capability was unevenly implemented, and, in all cases, the existing systems were designed primarily to meet institutional needs such as Quality Assurance/Quality Improvement (QA/QI), operations, or administration, with use for research, either within or across institutions, as a secondary priority, with limited funding support or resources. UCD did not have an institutional IDR beyond the Epic Clarity reporting database, and relied on manual queries performed by informatics staff to access Epic data for research. UCSF had no ready access to clinical data for research, and research results were typically provided to researchers after long delays. Additionally, medical center data were not pre-linked and therefore researchers often required access to the PHI data in order to link multiple data sources, causing potential institutional exposure to legal disclosure issues. UW was in transition from a MIND (a locally developed IDR in use for 15 years using INGRES[8]) to a Microsoft Amalga-based IDR platform, but did not have a service workflow for managing research access and data delivery in place. Each site had committed to supporting clinical data access for this project in advance of funding, and the project stakeholders had identified resources necessary to access the IDR and source clinical systems.

Establishing i2b2 as a common platform

Each site had an existing licensing arrangement with different commercial database systems and was evolving different research computing environments. Since i2b2 is implemented as part of a multi-tier architecture that allows the use of a range of database systems, each site retained its preferred database environment: UCD used Oracle Enterprise server 10.2g, UW used Microsoft SQL Server, and UCSF used Sybase IQ. All institutions leveraged virtualization technologies using VMWare vSphere Hypervisor to minimize the cost of the deployment, and all sites hosted the application in either Centos (UCD) or Ubuntu Linux (UW, UCSF). Sites first deployed i2b2 in secure test environments, and then in production environments.

i2b2 uses a plug-in architecture to allow customization and extension of the capabilities of the core platform. Developers refer to these extensions as 'cells'. A collection of cells is referred to as a 'hive', and a core collection is present and required in every deployment of i2b2. Each site implemented the minimum necessary cells (the hive cells) to form a queryable IDR—specifically the Clinical Research Chart, the Ontology Management cell, and the Project Manager cell. Initially, the CICTR project had planned to develop a new cell for i2b2 that would allow federated querying,[9] but became aware of a parallel effort in the form of the SHRINE project. SHRINE was a specific set of two i2b2-compatible cells that supported a basic approach to allowing a standard i2b2 installation to recognize and query remote installations with minimal site architecture modification.[10] At this point in the project, SHRINE had been successfully deployed in a prototype form across four partner hospitals within the Partners HealthCare environment, but had not been shared or evaluated beyond this environment. Interested in gaining experience and supporting dissemination of these components, the SHRINE and i2b2 teams made pre-release code available to the CICTR project and collaborated closely throughout the implementation and testing phases.

The primary function of the two SHRINE cells is to provide secure interfaces for managing communications based on common taxonomies between internal i2b2 hives comprising a site IDR and requests or messages originating from outside institutional firewalls. Collectively, the minimum i2b2 hive cells to support the IDR functions and the two SHRINE cells were referred to as a 'node'. The CICTR project defined common security requirements for establishing the SHRINE network for

CICTR, and developed and implemented the CICTR-focused taxonomies for testing this functionality.

Network and security architecture using SHRINE

The two primary technical requirements for a SHRINE network are: (1) each site must be able to securely communicate with trusted peer nodes; (2) all sites must have a way to ensure semantic consistency when communicating with peer nodes. Trust is established between SHRINE nodes through the exchange of secure Public Key Infrastructure certificates among all nodes. SHRINE is a peer-to-peer network, so security across the nodes is only as good as its weakest link. The team defined the following requirements for each institutional participant (figure 1).

- Cell traffic between local hives occurred exclusively over a 128-bit encrypted Secure Sockets Layer.
- Site SHRINE components resided within a DMZ ('demilitarized zone'—a secure network independent of a site's production operations). If a machine housing the SHRINE application stack were to be compromised, the primary firewall of the institutional data center could still provide security.
- Network traffic to each SHRINE node was ignored unless it originated at one of three white-listed IP addresses.
- All i2b2 nodes required password-secured accounts to access data. Users with permission to federate queries were only granted access to obfuscated results.
- External access to the systems was via Virtual Private Network (VPN) from a locked IP range, and was individually audited by the security staff at the institution.
- Every query run at one node was logged in an audit trail that was accessible from the peer nodes.

At the conclusion of phase 1, the project demonstrated that common test acceptance cases could be queried from each site and return identical expected results.

Phases 2 and 3: piloting a deidentified cohort discovery service

With a common network infrastructure in place, the focus of the next phases was to increase capacity of the system across four initial data domains. Data domains include: demographics, diagnoses, medications, and select laboratory results. The project anticipated challenges to gaining institutional approval to query and return individual patient-level data, and focused instead on each site returning aggregate counts of patients matching query criteria, which could be independently managed under site-specific human subjects exempt institutional review board (IRB) protocols. This approach allowed each site to apply their own site-specific data permutation deidentification processes within the extraction, transformation, and loading (ETL) work required to migrate data into i2b2, but using the same aggregation and display processes at each site through the common i2b2 web query interface. Aligning and providing each of these data sources for ETL involved four loosely coupled stages: (1) extracting data from back-end clinical systems and moving these data to the i2b2 repositories; (2) applying data deidentification rules; (3) developing 'maps' to these data that translated between local

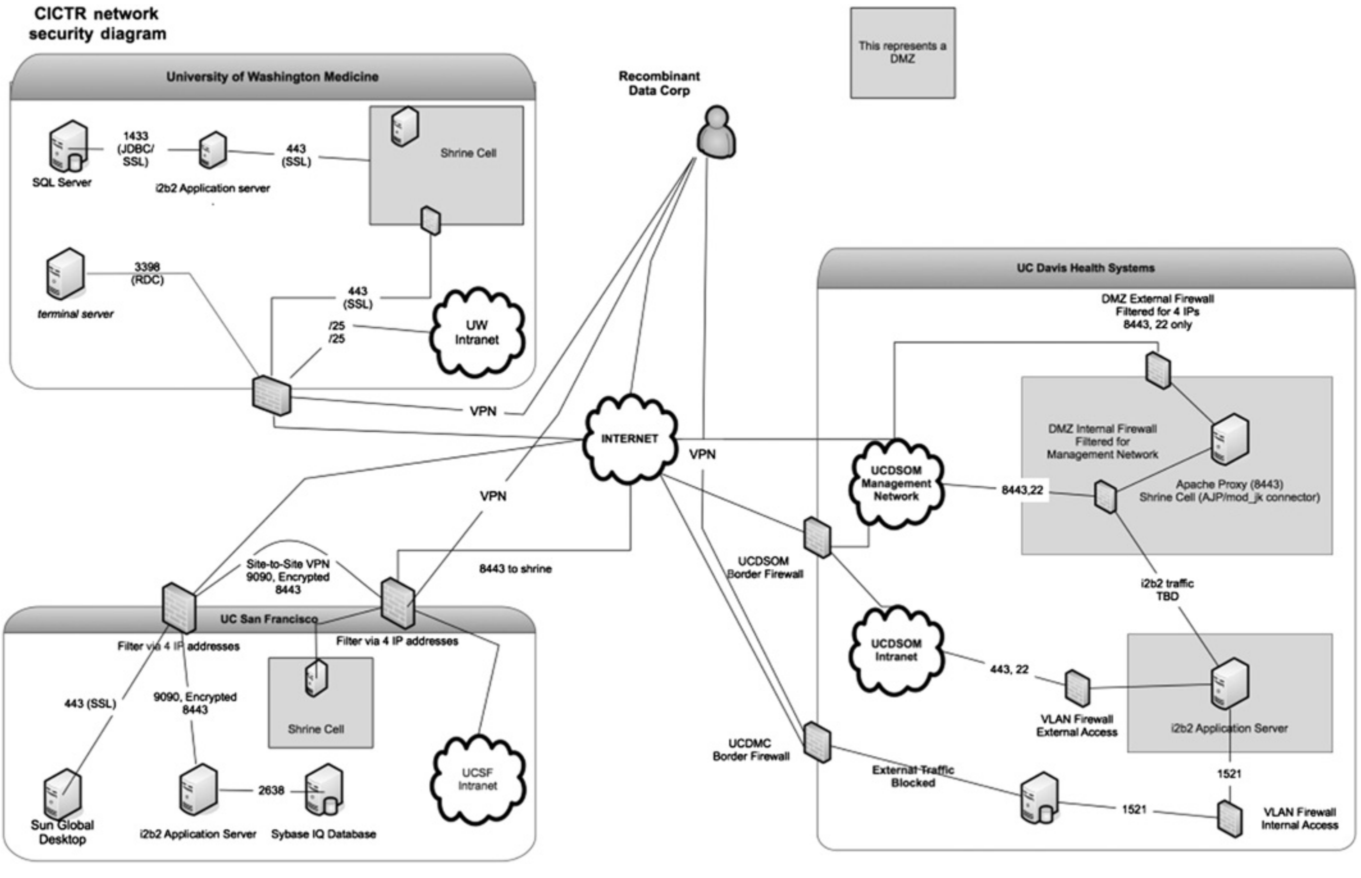


Figure 1 Cross-Institutional Clinical Translational Research (CICTR) network security diagram (simplified) describing isolation of Shared Health Research Information Network (SHRINE) cells from direct access to i2b2 data repositories. CICTR, Cross-Institutional Clinical Translational Research project; DMZ, Demilitarized Zone; IP, Internet Protocol; JDBC, Java Database Connectivity; SSL, Secure Sockets Layer; UCSF, University of California, San Francisco; UW, University of Washington; VLAN, Virtual Local Area Network; VPN, Virtual Private Network.

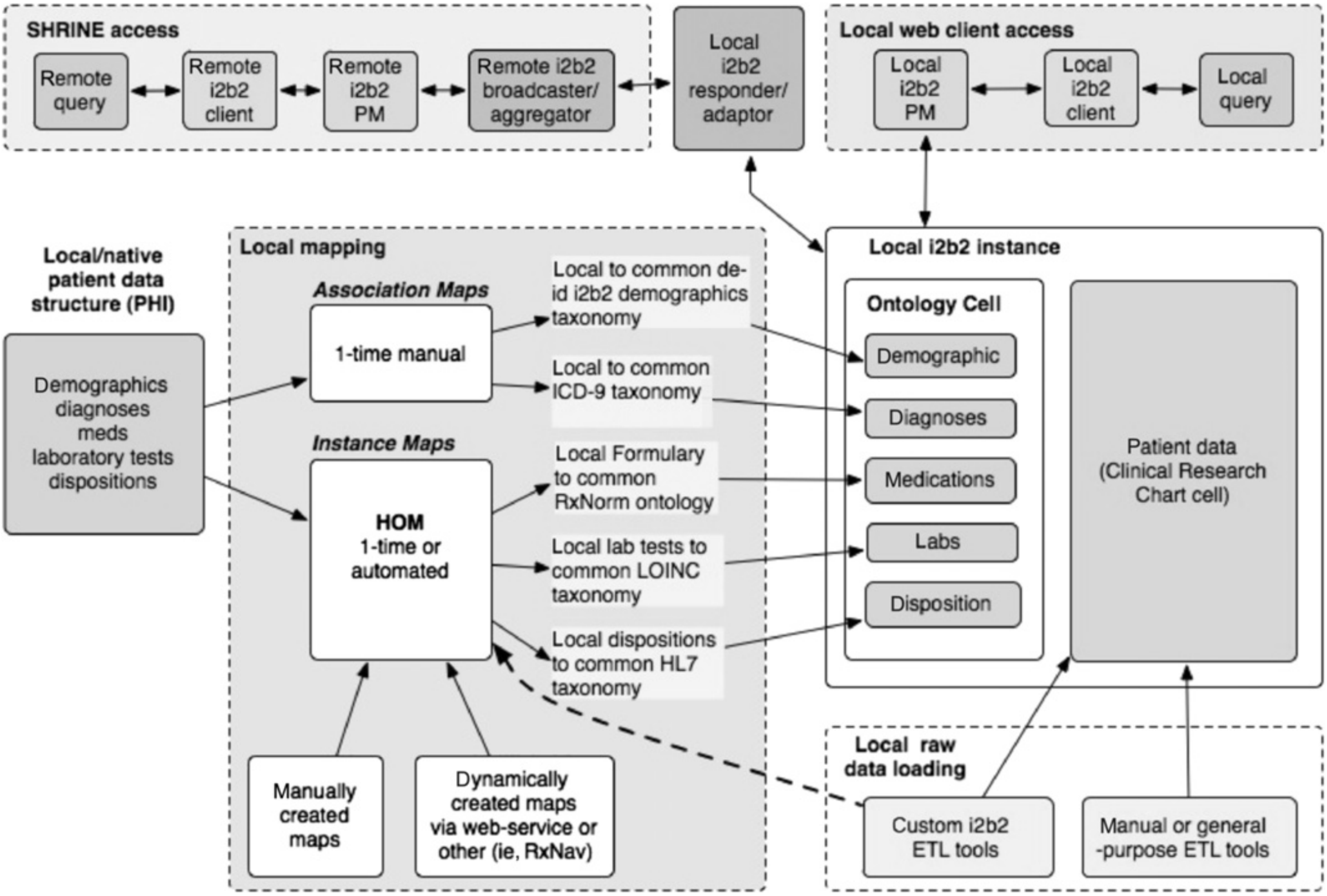

SHRINE access
Remote query
Remote i2b2 client
Remote i2b2 PM
Remote i2b2 broadcaster/ aggregator
Local i2b2 responder/ adaptor
Local web client access
Local i2b2 PM
Local i2b2 client
Local query
Local/native patient data structure (PHI)
Demographics diagnoses meds laboratory tests dispositions
Local mapping
Association Maps
1-time manual
Instance Maps
HOM 1-time or automated
Local to common de-id i2b2 demographics taxonomy
Local to common ICD-9 taxonomy
Local Formulary to common RxNorm ontology
Local lab tests to common LOINC taxonomy
Local dispositions to common HL7 taxonomy
Manually created maps
Dynamically created maps via web-service or other (ie, RxNav)
Local i2b2 instance
Ontology Cell
Demographic
Diagnoses
Medications
Labs
Disposition
Patient data (Clinical Research Chart cell)
Local raw data loading
Custom i2b2 ETL tools
Manual or general -purpose ETL tools

Figure 2 Local site-coordinated data loading, deidentification, and mapping processes.

structure and the common taxonomies; (4) testing resulting mappings through use-case-based queries locally before testing SHRINE-based queries across the three sites. In practice, steps 1 and 2 were combined (figure 2).

Extracting, deidentifying, and moving data from back-end clinical systems

Deidentification at each site treated patient demographics differently from patient event data (diagnosis, medication orders, laboratory tests, admit/discharge dispositions). Each site acquired human subjects exemption and followed permissible rules for deidentifying demographics associated with the 18 Health Insurance Patient Accountability Act (HIPAA) identifiers, as well as local requirements for any additional IRB-required obfuscations or redactions (eg, protection of potentially vulnerable patient classes that may be locally sensitive) (table 1). Processing of each patient demographic record generated unique patient keys, which were maintained internally to support linkage and updates for patient event data. Each unique patient key was associated with random date offsets for patient event data. All patient event data were deidentified by adding or subtracting a random number of days to each event, consistently for all events for each given patient. Each site applied their own choice of this random offset within their deidentification workflows.

Table 1 Common demographics deidentification and mapping methods

Element	Function	Method
MRN	Obfuscated	Site-dependent
DOB	Obfuscated	1/1/YYYY
Gender	Mapped	HL7 v2, Table 001
Age	Calculated	Relative to DOB
Race/ethnicity	Mapped	OMB R/E
Geocode	Calculated	First 3 digits and all zip codes <20 K patients clustered into State:00000
Vital status	Mapped	HL7 Entity.LivingSubject.deceasedInd
Marital status	Mapped	HL7 v2, Table 002
Language	Mapped	ISO-639.2

DOB, date of birth; HL7, health level 7; MRN, medical record number; OMB R/E, office of management and budget racial/ethnicity
State: 00000 represents the state (eg, WA) associated with a blank zipcode of 00000) HL7 v2, Table 002 refers to table 002 of the HL7 standard which defines marital status.

Table 2 Patient event data scope and mapping methods

Data source	Scope	Mapping method	Terminology
Diagnoses	001-999 V01-V89 E800-E999	Association	ICD-9
Medication orders	All local formularies	Instance	RxNorm
Laboratory orders	Diabetes-focused common LOINC subset	Instance	LOINC
Discharge dispositions	All	Instance	HL7

HL7, health level 7; ICD-9, International Classification of Disease version 9; LOINC, Logical Observation Identifiers Names and Codes; RxNorm, A standardized nomenclature for clinical drugs and drug delivery.

Mapping of diagnostic, medication, and laboratory data

Each site's academic medical center maintains multiple distinct diagnostic, formulary, pharmacy, laboratory, and admit/discharge information systems as part of their electronic patient care systems. Each of these systems uses a correspondingly unique (and often locally modified) descriptive terminology, and contains a different historical range of historical data. Data we sought to extract from each of these sources were evaluated through an assessment of local data availability and functionality so as to assess what commonalities and expected utility could be established to support a common mapping process (table 2). On the basis of these assessments, the team developed data-source-specific guidelines to establish what semantic and syntactic mapping approaches would need to be used, and how these would be implemented through either an association approach (essentially one-to-one element mapping) or an instance mapping approach (many-to-many, one-to-many, or many-to-one mappings).

Medications mapping to RxNorm

RxNorm was selected as a common target taxonomy to support common queries of medications between the sites.[11] We chose to focus on mapping ordered medications only so as to limit scope (administered or dispensed medication events would have required reconciliation with outpatient systems). There was agreement that the presence of active pharmaceutical ingredients would be instrumental to the discovery of patient cohorts; however, we recognized that this approach has limitations because of variability in physiological actions in multiple dose forms and in combination drug preparation. Using RxNorm, we developed common mappings for a set of medications from three different institutional inpatient formularies that focused on (1) clinical drug form, (2) dose form, (3) ingredient, and (4) precise ingredient.

The UW and UCSF used local installations of the Health Ontology Mapper (HOM) application and a common HOM script to process local formulary data against the RxNorm web service APIs to build i2b2 taxonomies. HOM is a general-purpose open-source tool that uses a common scripting process to generate instance maps between local terminologies to formal data encoding standards.[11] [12] UCD used a manual process following the same processing rules as the HOM script:

1. i2b2 is loaded with a raw table of local medication formulary
2. For each medication in this list that is ordered for a patient in the i2b2 clinical research chart cell:
   a. parse medication names against the RxNorm web service to generate normalized medications based on the RxNorm concept unique identifier
   b. populate a new instance of each text-readable category (clinical drug form, dose form, ingredient, and precise ingredient) in a new i2b2 medication if one does not already exist.

The resulting medication taxonomy structure is thus built on top of, and does not transform, existing formulary data descriptions—and is maintained independently of the primary 'raw' data mappings between medications and patients at each site. This resulted in an end-user-focused taxonomy that is common and queryable across the three sites, and can be updated through reapplication of the mapping script as RxNorm content evolves.

Laboratory mapping to Logical Observation Identifiers Names and Codes (LOINC)

Developing a methodology to map results of local patient laboratory tests to a mechanism to query for results of these tests was an initial project goal, but was evaluated to be beyond the scope of this initial pilot project. We confirmed that each institution had multiple laboratories serving their health systems, each of which possessed different reference ranges associated with local practice of care and aligned with normal or abnormal results. Owing to these factors, we made the decision to focus on a specific set of laboratory test orders rather than test results in order to test the effectiveness of mapping within the core use cases. We recognize that ordered laboratories do not show diagnostic utility, but do suggest clinical intent for a patient. We initially used the use case criteria to select 17 'classic labs' from chemistry and hematology (vs other clinical physiological measurements such as body mass index and blood pressure) from a review of 58 completed trials in ClinicalTrials.gov, a registry of federally and privately supported clinical trials in the USA, provided by the National Institutes of Health. All 58 trials focused on type 2 diabetes studies with results. We determined best-fit LOINC codes for eight individual laboratories from this set (discarding the remainder as unevenly implemented at each site) that also fit within the common LOINC value set, and could reasonably be of utility to cohort discovery across sites.[13] Each site manually developed mappings of their local codes to the project-identified LOINC codes.

Admit/discharge disposition data mapping to modified HL7 taxonomy

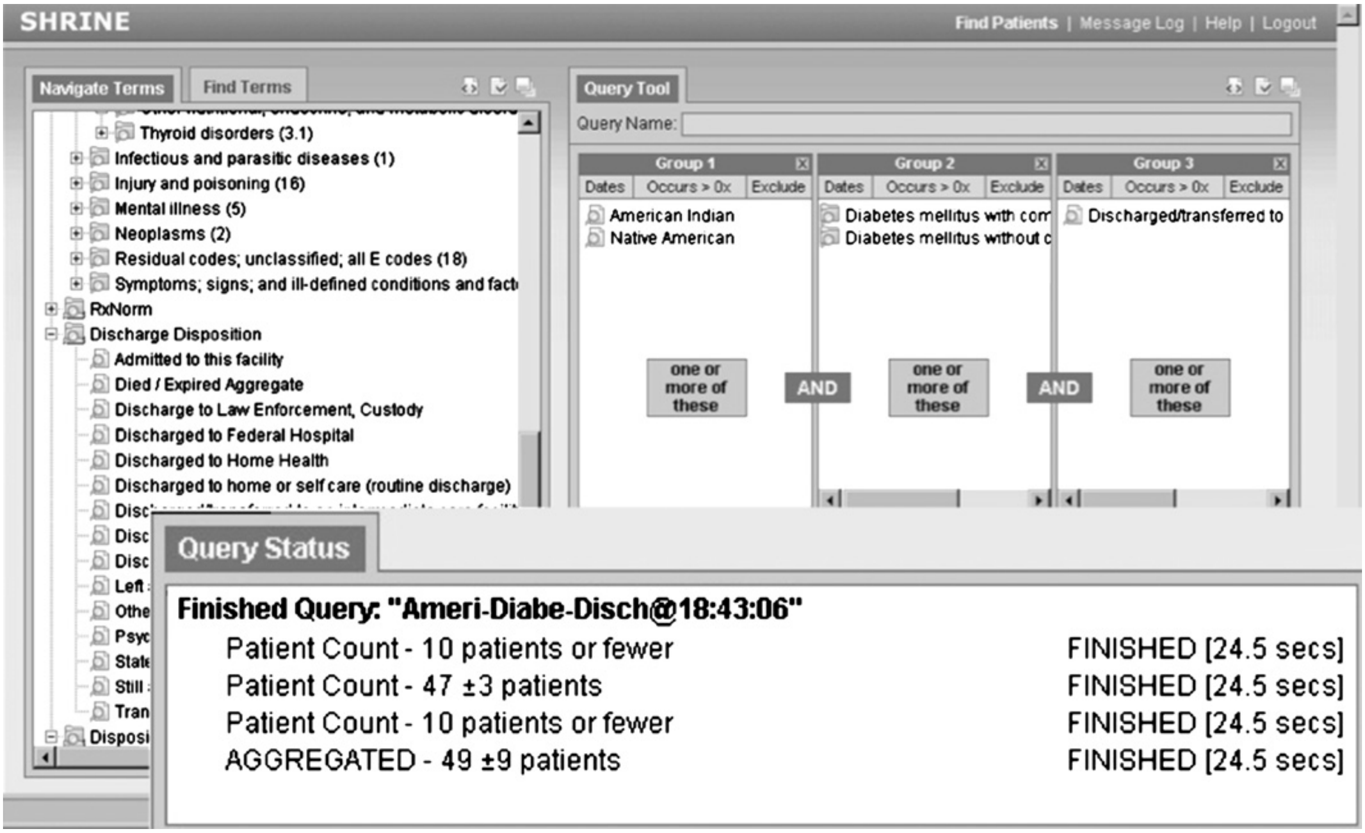
As we approached the end of the final phase, we used the experience of developing common mappings in the previous data sources to complete a 2-month project of providing access to patient admit/discharge dispositions. This required engagement with a new set of stakeholders comprising comparative effectiveness researchers who sought to use population-level data to look for rates of patient readmission across sites, and then to use the other data sources to stratify these results. New use cases were developed that paralleled the existing diabetes-focused cases. We used the HL7 v2.7 'user defined' value set of discharge dispositions to develop a target set to which we mapped local institutional sources (table 3).

Aggregating and delivering deidentified data

A condition of gaining IRB approval for human subjects exemption at each site was that queries of the network would

Table 3 HL7 v2.7 target discharge disposition table

HL7 description	HL7 code
Discharged to home or self-care (routine discharge)	1
Discharged/transferred to another short-term general hospital for inpatient care	2
Discharged/transferred to skilled nursing facility	3
Discharged/transferred to an intermediate care facility	4
Discharged/transferred to another type of institution for inpatient care or referred for outpatient services to another institution	5
Discharged/transferred to home under care of organized home health service organization	6
Left against medical advice or discontinued care	7
Admitted as an inpatient to this hospital	9
Discharge to be defined at state level, if necessary	10
Law enforcement custody	12
Other, undetermined	13
Psychiatric facility	14
Discharge to federal hospital	15
Expired (ie, dead)	20
Still patient or expected to return for outpatient services (ie, still a patient)	30


SHRINE
Find Patients | Message Log | Help | Logout
Navigate Terms
Find Terms
Thyroid disorders (3.1)
Infectious and parasitic diseases (1)
Injury and poisoning (16)
Mental illness (5)
Neoplasms (2)
Residual codes; unclassified; all E codes (18)
Symptoms; signs; and ill-defined conditions and fact
RxNorm
Discharge Disposition
Admitted to this facility
Died / Expired Aggregate
Discharge to Law Enforcement, Custody
Discharged to Federal Hospital
Discharged to Home Health
Discharged to home or self care (routine discharge)
Disc
Disc
Disc
Left
Othe
Psyc
State
Still
Tran
Disposi
Query Tool
Query Name:
Group 1
Dates
Occurs > 0x
Exclude
American Indian
Native American
one or more of these
AND
Group 2
Dates
Occurs > 0x
Exclude
Diabetes mellitus with com
Diabetes mellitus without c
one or more of these
AND
Group 3
Dates
Occurs > 0x
Exclude
Discharged/transferred to
one or more of these
Query Status

Finished Query: "Ameri-Diabe-Disch@18:43:06"

```
Patient Count - 10 patients or fewer          FINISHED [24.5 secs]
Patient Count - 47 ±3 patients                FINISHED [24.5 secs]
Patient Count - 10 patients or fewer          FINISHED [24.5 secs]
AGGREGATED - 49 ±9 patients                   FINISHED [24.5 secs]
```

Figure 3 Return of aggregated and blurred results across Cross-Institutional Clinical Translational Research (CICTR) network.

only return numeric counts for each site—and each site name would remain further obfuscated, and that each of these counts would be further obfuscated by adding a random number (between −3 and 3), or reset to '<10' if the count was <10.[14] An aggregate of these counts was summed and provided after successful completion across all sites. This functionality is a core capability of the SHRINE and web client interfaces (figure 3).

RESULTS

There were two primary results: (1) a functional implementation of a federated query tool on deidentified data for over 5 million patients and five data sources across three independent institutional IDRs (table 4); (2) the development and testing of a generalizable iterative process model for assessing the data workflow processes and expected use of a research network (figure 4).

The project defined specific milestones associated with the completion of each phase before project start, and used common use cases and derivative test acceptance cases to evaluate completion of each project milestone. On the basis of these milestones, the project met its technical ambition to establish a functional federated query network for cohort discovery from each site, but was challenged in engaging end users using the full network for research use. UCD had success in testing their local pilot implementation, and won a California UC school award for innovation.[15] All sites predominantly experienced requests for local uses versus requests for cross-site network data. Despite these challenges, the project has engaged with additional CTSA partners, which, at the time of this submission, includes eight national partners (including the original three sites) in varying stages of design and implementation.

Table 4 Cross-Institutional Clinical Translational Research population as of April 15, 2011

Site	Archive history	Update frequency	Patient count in thousands
UW	5 years	Daily	1460
UCSF	4 years	Daily	3090
UCD	7 years	Daily	706
Total			5256

UCSF, University of California, San Francisco University of California, Davis; UW, University of Washington.

The iterative process model was based on the spiral model of software engineering and supported the project in evaluating that the work of the four thematic groups was coordinated.[16] The model defined four major interdependent quadrants: building partnerships, system requirements, technical architecture, and evaluation/promotion. This model was used in the final phase of the project to assess the complexity and manage the risk of adding an additional data source across all sites in the form of disposition data—the results of which were new comparative effectiveness-focused use cases and a short-term subproject which was completed successfully in approximately 6 weeks.

DISCUSSION

This pilot project evaluated how a common technical platform, data governance approach, and data mapping expectations could be coordinated to support the building of a large-scale distributed data discovery system. Both in design and implementation, the CICTR pilot network followed a spiral process which managed these themes through multiple coordinated iterations of development and evaluation (figure 4). As the project

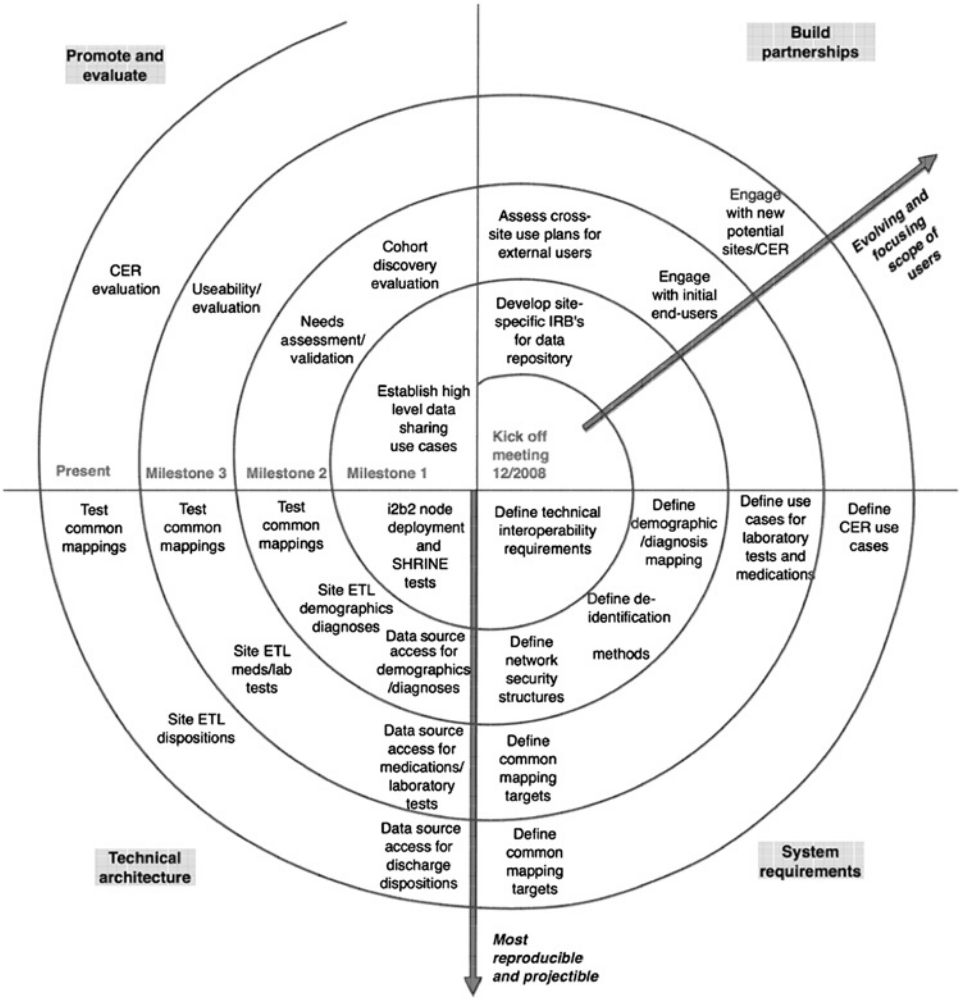

Promote and evaluate
Build partnerships
Evolving and focusing scope of users
Engage with new potential sites/CER
Assess cross-site use plans for external users
Cohort discovery evaluation
CER evaluation
Useability/ evaluation
Engage with initial end-users
Needs assessment/ validation
Develop site-specific IRB's for data repository
Establish high level data sharing use cases
Kick off meeting 12/2008
Present
Milestone 3
Milestone 2
Milestone 1
Test common mappings
Test common mappings
Test common mappings
i2b2 node deployment and SHRINE tests
Define technical interoperability requirements
Define demographic /diagnosis mapping
Define use cases for laboratory tests and medications
Define CER use cases
Site ETL demographics diagnoses
Define de-identification methods
Data source access for demographics /diagnoses
Define network security structures
Site ETL meds/lab tests
Site ETL dispositions
Data source access for medications/ laboratory tests
Define common mapping targets
Data source access for discharge dispositions
Define common mapping targets
Technical architecture
System requirements
Most reproducible and projectible

Figure 4 Descriptive spiral model of implementation phases and stakeholders. CER, Comparative Effectiveness Research; ETL, Extraction, Transformation and Loading SHRINE - Shared Health Research Information Network; IRB, Institutional Review Board; i2b2, Informatics for Integrating Biology and the Bedside

matured, this model-driven management approach proved helpful in coordinating new use cases to support new stakeholders and data sources to the project. This model and the project experiences identified that evaluating and aligning stakeholders' data needs with semantic data capabilities of distributed data query systems remains complex, although specific technical components can become more clearly specified if these needs are clearly defined in advance.[17 18]

Assessing expected user utility within privacy policy frameworks and clinical/research environments

Integrating heterogeneous IDRs and hospital IT systems across multiple sites will remain challenging, since the ownership of such systems and responsibility for providing high-quality data in forms that can be shared are often unclear even within an individual site. The heart of this issue is that data creation in these environments does not anticipate or accommodate secondary data use considerations, whether for QA/QI or research. Local data quality is typically driven by non-research institutional operational requirements, whereas research or other data-sharing activities are typically driven by specific funding initiatives. Without specific incentives to provide access to clinical data for research (financial or otherwise), it will remain a challenge as well as an opportunity for individual sites to consider how the methods and results of research data sharing could contribute value or quality metrics back to the IDR itself.

This project identified that there remain considerable challenges to effectively anticipating and measuring user utility to justify research investment in distributed data discovery systems.[19] Our experience of having to deidentify, obfuscate, aggregate, and blur data to meet all sites' IRB requirements created a resource that was perceived to be 'over-sanitized' from an end-user utility perspective and challenged the ability to advertise the service for use. Feedback of this form posed ethical challenges for the project team—as not being able to equally or effectively represent rural, minority, and rare conditions (the original mission of the project) because difficulties in common data-sharing expectations posed questions of how the project could have better shepherded the development and advocacy of the system within the current framework of national and local privacy policy.

To address this, we envision a process by which end-user data requirements are defined on the basis of common terminology standards and common user-driven data-sharing expectations within specific disease or analysis domains, but include the recognition that exemptions for 'what is shared' will necessarily be left to local sites where differences in data use policy can be implemented. Having these end users as full and engaged partners in the development and utility evaluation processes should be established at project outset. If inter-institutional research networks are to be based on strict HIPAA deidentification, other approaches may be needed that explore both higher-utility deidentification and methods to enhance trust between institutions and communities.[20–22] Given the national privacy policy landscape, site stakeholders (both users and gatekeepers) will invariably remain the primary arbiters of how their institution views or supports the ability to balance discovery of information about patients against institutional exposure—and will need to weigh the institutional utility for supporting such a service.

Evaluating i2b2 as an open architecture for federated networks

The ability of i2b2 to benefit an individual site before the complex issues of network federation are engaged is a demonstrated strength.[3] Although i2b2 deployment as an IDR requires a considerable amount of organizational and resource commitment—in particular, ETL experience—the barriers to implementing and testing are typically locally manageable and reach a higher level of evaluative utility more quickly than other grid-based data-sharing environments or commercial products. This is demonstrated by the considerable number of national and international i2b2 sites that are 'home grown' and do not receive funding or assistance from the i2b2 center itself—although many are presently research funded.[23] However, the ability to quickly create a new stand-alone i2b2 system does not directly advance the building of a collaborative network that meets broad user needs—this requires essential coordination of semantic, security, quality, technical, and domain resources.

The i2b2 center presently lacks the resources or mission to support coordination of the multiple site implementations necessary for collaborative data-sharing networks. The software represents an increasingly powerful set of tools with a strong community user base, each of whom are independently beginning to build both technical and process capacity for large-scale data discovery and sharing. It would be of significant benefit to the research data warehousing community to advocate assessing and communicating best practices of effective clinical data translation workflows that can be coordinated and result in centralized and shareable protocols—such as are beginning to become available through coordination with the NCBO bioportal.[24]

CONCLUSION

Coordination across three geographically diverse clinical and research environments to align technical, semantic, and policy issues required considerable consensus-building and interdisciplinary education, and reaffirmed that informatics implementation projects such as this remain as much a social as a technical problem.[24] The challenge of coordination will increase with projects that bridge multiple technical scientific domains—both within and across institutions—particularly with the relatively common short time frames driven by current funding support and the rapidly moving research enterprise. This project overcame multiple obstacles to piloting a novel federated discovery environment for research, and identified lessons learned in terms of collaboration, management, technical implementation, policy, and utility. Our experience suggests that a generic and reflexive approach to managing expectations for building and using federated networks and measuring utility throughout implementation should be established at project outset. We anticipate that the continued sustainability of this pilot network, or networks of this form, will require enhancement of the ability to support iterative and inclusive development, evaluation, education, and validation approaches, and that coordination and partnership with scientific end users who are stakeholders at local sites will be critical to overall success.

Acknowledgments We would like to thank Susanne Churchill, Isaac Kohane, Shawn Murphy, Mike Mendis and Andrew McMurry of the i2b2 center for their extensive support in this project. John Gennari and Fred Wolf at the University of Washington assisted in ontology analysis expertise and evaluation throughout the project as well as in review of this manuscript. We would also like to thank Elaine Collier of the NCRR for her insight and support throughout this project.

Funding This work was funded by the National Center for Research Resources, National Institutes of Health, Department of Health and Human Services, under Contract No HHSN268200700031C, as well as through NIH/NCRR UW-ITHS grant number UL1 RR025014, UCSF-CTSI grant number UL1 RR024131, and UCD-CTSC grant number UL1 RR024146.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Anon.** Data's shameful neglect. *Nature* 2009;**461**:145.
2. **Murphy SN,** Weber G, Mendis M, *et al*. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *JAMIA* 2010;**17**:124–30.
3. **Deshmukh VG,** Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med Res Methodol* 2009;**9**:70.
4. **Kaplan B,** Harris-Salamone KD. Health IT success and failure: recommendations from literature and an AMIA workshop. *JAMIA* 2009;**16**:291–9.
5. **Chilana P,** Fishman E, Geraghty E, *et al*. Needs in clinical translational science. *JOEUC*, Special issue on Scientific End-user Computing 2010.
6. *Clinical and Translational Science Awards*. http://www.ctsaweb.org/ (accessed 16 Jan 2011).
7. *HIPAA Privacy Rules for Researchers*. http://privacyruleandresearch.nih.gov/faq.asp (accessed 9 Sep 2009).
8. **Tarczy-Hornoch P,** Kwan-Gett TS, Fouche L, *et al*. Meeting clinician information needs by integrating access to the medical record and knowledge resources via the Web. *Proc AMIA Annu Fall Symp* 1997:809–13.
9. *A Data Warehousing Strategy for Translational Medicine. American Medical Informatics Association Translational Bioinformatics Summit*. San Francisco, CA, 2008.
10. **Weber GM,** Murphy SN, McMurry AJ, *et al*. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *JAMIA* 2009;**16**:624–30.
11. *Health Ontology Mapper*. http://www.healthontologymapper.org/ (accessed 15 Jan 11).
12. **Wynden R,** Weiner M, Sim I, *et al*. Ontology mapping and data discovery for the translational investigator. *AMIA Summit on Clinical Research Informatics*. San Francisco, 2010.
13. *Common Lab Orders LOINC Value Set Version 1*. http://loinc.org/news/common-lab-orders-loinc-value-set-version-1-available.html (accessed 15 Jan 2011).
14. **Murphy SN,** Chueh HC. A security architecture for query tools used to access large biomedical databases. *Proc AMIA Symp* 2002:552–6.
15. *University of California 2010 Larry L. Sautter Award for Innovation* http://www.ucdmc.ucdavis.edu/welcome/features/2010-2011/08/20100818_Sautter_Award.html (accessed 1 Jun 2011).
16. **Boehm B.** A spiral model of software development and enhancement. *SIGSOFT Softw Eng Note IEEE Computing Society* 1986;**11**:14–24.
17. **Ash JS,** Anderson NR, Tarczy-Hornoch P. People and organizational issues in research systems implementation. *JAMIA* 2008;**15**:283–9.
18. **Anderson NR,** Lee ES, Brockenbrough JS, *et al*. Issues in biomedical research data management and analysis: needs and barriers. *JAMIA* 2007;**14**:478–88.
19. **Anon.** Evaluating technology in healthcare: testing the usability of a clinical trial query tool using think-aloud methods. *American Evaluation Association Annual Meeting*. San Antonio, Texas, 2010.
20. **Brickell J,** Shmatikov V. The cost of privacy: destruction of data-mining utility in anonymized data publishing. *4th ACM Special Interest Group on Knowledge Discovery and Data Mining*, Las Vegas: ACM, 2008:7–78.
21. **Malin B.** A computational model to protect patient data from location-based re-identification. *Artif Intell Med* 2007;**40**:223–39.
22. **Anderson N,** Edwards K. Building a Chain of Trust: using policy and practice to enhance trustworthy data discovery and sharing. *Proceedings of Association of Computing Machinery Workshop on Workshop on Governance of Technology, Information, and Policies*. Austin, Texas, 2010.
23. *I2b2 National and International Sites*. http://healthmap.org/i2b2/ (accessed 30 Jun 11).
24. **NCBO Bioportal.** http://bioportal.bioontology.org/ (accessed 15 Mar 2011).