CTSA Annual Informatics Meeting October 13-14, 2010

Natcher Conference Center Auditorium, NIH Campus

Abstracts
## Podium Presentations

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Presenter</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>REDCap – What’s New and What’s Next?</td>
<td>Paul Harris</td>
<td>Vanderbilt University</td>
</tr>
<tr>
<td>2</td>
<td>Procurement of shared data instruments for research electronic data capture (REDCap)</td>
<td>Catherine McGraw</td>
<td>University of Cincinnati</td>
</tr>
<tr>
<td>3</td>
<td>Medical imaging in i2b2: new tools for integrating clinical images into research studies</td>
<td>Shawn Murphy</td>
<td>Harvard University</td>
</tr>
<tr>
<td>4</td>
<td>Extending i2b2 to Support Health Outcome and Disparity Measurement</td>
<td>William Adams</td>
<td>Boston University</td>
</tr>
<tr>
<td>5</td>
<td>The Health Ontology Mapper (HOM) Project Status</td>
<td>Rob Wynden</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>6</td>
<td>Supporting Security in Applications and Services for Secondary Data Use: UNC TraCS Data Warehouse Experience</td>
<td>Javed Mostafa</td>
<td>University of North Carolina, Chapel Hill</td>
</tr>
<tr>
<td>7</td>
<td>LC Data QUEST: A Technical Architecture for Community Federated Clinical Data Sharing</td>
<td>Ching-Ping Lin</td>
<td>University of Washington</td>
</tr>
<tr>
<td>8</td>
<td>An Architecture for Creating Semantically-enabled Registries through Federated Query of Clinical Databases at a Multi-institution CTSA</td>
<td>Andrew Post</td>
<td>Emory Center for Comprehensive Informatics</td>
</tr>
<tr>
<td>9</td>
<td>Adaptation of a caBIG-based Clinical Trials Management Suite to the National Children’s Study</td>
<td>William Hogan</td>
<td>University of Arkansas for Medical Sciences</td>
</tr>
<tr>
<td>10</td>
<td>Using CaBIG and Semantic Web Technologies for Collaborative Research</td>
<td>Michael Krauthammer</td>
<td>Yale University</td>
</tr>
<tr>
<td>11</td>
<td>Using the Open Metadata Registry (OpenMDR) to create data sharing interfaces</td>
<td>Rakesh Dhaval</td>
<td>The Ohio State University</td>
</tr>
<tr>
<td>12</td>
<td>A Highly Scalable Approach to Extending Functionality of the Research Networking Tool Profiles Research Networking Software with the OpenSocial Standard</td>
<td>Eric Meeks</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>13</td>
<td>Interoperability among research networking platforms: A national pilot</td>
<td>Griffin Weber</td>
<td>Harvard University</td>
</tr>
<tr>
<td>14</td>
<td>Beyond “Web 2.0” as Buzzword: Engaging Researchers to Contribute Online</td>
<td>Mini Kahlon</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>15</td>
<td>The Human Studies Database Project</td>
<td>Ida Sim</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>#</td>
<td>Title</td>
<td>Presenter</td>
<td>Institution</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>16</td>
<td>i2b2 at the University of Michigan - Cohort discovery and beyond</td>
<td>Brian Athey</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>17</td>
<td>Locating Clinical Researchers Based on Expertise Documented in Patient Encounter Records</td>
<td>Victor Brodsky</td>
<td>Weill Cornell Medical College</td>
</tr>
<tr>
<td>18</td>
<td>Automating Electronic Data Capture Application User Management</td>
<td>Linda Carlin</td>
<td>University of Colorado, Denver</td>
</tr>
<tr>
<td>19</td>
<td>Synergies in the Strategic Health IT Research Project (SHARP) Area 4: Secondary Use of EHR Data and the Health IT Pilot Communities through Recovery Act Beacon Community Program</td>
<td>Christopher Chute</td>
<td>Mayo Clinic</td>
</tr>
<tr>
<td>20</td>
<td>Data Standards and Interoperability Affinity Group (DIAG)</td>
<td>Davera Gabriel, Meredith Nahm</td>
<td>University of California, Davis &amp; Duke University</td>
</tr>
<tr>
<td>21</td>
<td>Enabling Research Through Technology: Leveraging Enterprise Data Stores in Support of Cohort Generation and Facilitated Subject Recruitment</td>
<td>Jeffrey Ferranti</td>
<td>Duke University Health System</td>
</tr>
<tr>
<td>22</td>
<td>openSESAME: A new tool for discovering connections between biological phenotypes using public gene expression data</td>
<td>Andrew Gower</td>
<td>Boston University School of Medicine</td>
</tr>
<tr>
<td>23</td>
<td>Collaboration Tools and Resources at the UIC CCTS Biomedical Informatics Core</td>
<td>Denise Hynes</td>
<td>University of Illinois, Chicago</td>
</tr>
<tr>
<td>24</td>
<td>Algorithmic Infrastructure and “Omic” Data Integration for Network-Based Analysis of Complex Phenotypes</td>
<td>Mehmet Koyuturk</td>
<td>Case Western Reserve University</td>
</tr>
<tr>
<td>25</td>
<td>A Federated Virtual Data Warehouse Platform for Translational Research at a Multi-institutional CTSA</td>
<td>Tahsin Kurc</td>
<td>Emory Center for Comprehensive Informatics</td>
</tr>
<tr>
<td>26</td>
<td>Maximizing Data Value From Multiple Clinical Trials</td>
<td>Bernie LaSalle</td>
<td>University of Utah Center for Clinical and Translational Science</td>
</tr>
<tr>
<td>27</td>
<td>Creating A Research Subject Registry That Enables Business Process And System Integration</td>
<td>Michael Lin</td>
<td>Mayo Clinic</td>
</tr>
<tr>
<td>28</td>
<td>Systems biology analyses of gene expression and genome wide association study data in obstructive sleep apnea</td>
<td>Yu Liu</td>
<td>Case Western Reserve University</td>
</tr>
<tr>
<td>29</td>
<td>Institutional Architecture for Secondary Data Use</td>
<td>Aenoch Lynn</td>
<td>Duke Translational Science Institute</td>
</tr>
<tr>
<td>#</td>
<td>Title</td>
<td>Presenter</td>
<td>Institution</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>30</td>
<td>Lessons Learned from Implementing Harvard Catalyst SHRINE which Inform the Building of a Network across Clinical Translational Science Centers to enable Massive Scale Research using Routine Clinical Data.</td>
<td>Douglas MacFadden</td>
<td>Harvard University</td>
</tr>
<tr>
<td>31</td>
<td>Georgetown Database of Cancer (G-DOC): Enabling Personalized Medicine through data integration and informatics best practices</td>
<td>Subha Madhavan</td>
<td>Georgetown University</td>
</tr>
<tr>
<td>32</td>
<td>Research Collaborations and their Temporal Evolution across Multiple Scales</td>
<td>Radhakrishnan Nagarajan</td>
<td>University of Arkansas for Medical Sciences</td>
</tr>
<tr>
<td>33</td>
<td>A Permissions Ontology for the Integration of Electronic Research Authorizations, Consents and other Permissions Data</td>
<td>Jihad Obeid</td>
<td>Medical University of South Carolina</td>
</tr>
<tr>
<td>34</td>
<td>The Awesomely Large OSIM Dataset (ALODS)</td>
<td>David Oppenheimer</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>35</td>
<td>The &quot;5x5&quot; Program: Brief Videos for Informatics Education</td>
<td>Steve Ross</td>
<td>University of Colorado, Denver</td>
</tr>
<tr>
<td>36</td>
<td>High-throughput Microscopy Image Analysis for Deep Integrative In Silico Study of Brain Tumors</td>
<td>Joel Saltz</td>
<td>Emory Center for Comprehensive Informatics</td>
</tr>
<tr>
<td>37</td>
<td>A Research Permissions Management System at the Statewide Scale: Progress</td>
<td>Iain Sanderson</td>
<td>Medical University of South Carolina</td>
</tr>
<tr>
<td>38</td>
<td>CV management in Digital Vita: A formative evaluation</td>
<td>Titus Schleyer</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>39</td>
<td>A new tool for researchers to access de-identified pathology information and paraffin embedded tissue</td>
<td>Linda Schmandt</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>40</td>
<td>SPIRiT: Building the Infrastructure to Support Collaborative Science at 5 CTSA institutions</td>
<td>Linda Schmandt</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>41</td>
<td>Clinical and Research System Interoperability with Research Electronic Data Capture (REDCap)</td>
<td>Robert Schuff</td>
<td>Oregon Health &amp; Sciences University</td>
</tr>
<tr>
<td>42</td>
<td>Semantic Integration of AIDS-related Malignancy Data within the ACTSI Biomedical Informatics Program Framework</td>
<td>Alfredo Tirado-Ramos</td>
<td>Emory Center for Comprehensive Informatics</td>
</tr>
<tr>
<td>43</td>
<td>Utilizing Open Source Clinical Research Tools for Behavioral Studies</td>
<td>Umit Topaloglu</td>
<td>University of Arkansas for Medical Sciences</td>
</tr>
<tr>
<td>44</td>
<td>The CCTR Network : Using Sharepoint for collaboration, management and evaluation of clinical and translational research</td>
<td>Alan VanBiervliet</td>
<td>University of Arkansas for Medical Sciences</td>
</tr>
<tr>
<td>45</td>
<td>Providing Flexible and Integrated Informatics Support to Investigators and Trainees to Improve Clinical &amp; Translational Research</td>
<td>Dongwen Wang</td>
<td>University of Rochester</td>
</tr>
<tr>
<td>#</td>
<td>Title</td>
<td>Presenter</td>
<td>Institution</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>47</td>
<td>Building and standardizing a clinical data warehouse</td>
<td>Brian Wells</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>48</td>
<td>A Unified System for Application Submission, Review, and Post-Approval Tracking and Reporting</td>
<td>Elizabeth Wood</td>
<td>Weill Cornell Medical College</td>
</tr>
<tr>
<td>49</td>
<td>Extending the caGrid Federated Query Capability</td>
<td>Matthew Wyatt</td>
<td>University of Alabama, Birmingham</td>
</tr>
</tbody>
</table>
Abstract:

1) Description of the project, best practice, activity, system to be presented;
REDCap (Research Electronic Data Capture) is a software program designed to assist scientific teams with research data collection and management. Software and support are provided at no financial charge to academic and non-profit institutions through a consortium network (www.project-redcap.org). As of August 2010, REDCap is supporting approximately 3,260 projects and 10,250 research end-users across a large and growing consortium of 139 academic and nonprofit partners (including 43 CTSA). Software and support are provided to consortium partners at no cost and no contributory obligation, but many sites chose to contribute back to the project in numerous ways and this culture of collaboration has substantially informed the evolution of REDCap. We have added significant functional modules during the past 12-months to support research teams in diverse scientific research areas. These enhancements include: enhanced modules for automatic de-identification of data, interoperability modules to enable the exchange of data between REDCap and non-REDCap data systems (e.g. clinical data warehouse, external programs); data visualization tools; customized rendering for mobile devices, embedded training videos; logging functions and locking/e-signature methods to support 21CFR-Part11 requirements; and a REDCap consortium-wide shared data instrument library designed to save time and promote standardization across studies. REDCap is continually evolving and our plans for the next two versions include: additional data interoperability tools; combined tools to collect and manage patient- and study-team- data instruments for individual studies; participant randomization, formal data query and reporting tools; and a supported version of the software that runs independently on a laptop without internet connections. This presentation will include details about the consortium model, current capabilities and also information about ongoing work prioritized to best serve the research community.

2) Why it is important to be presented at the 2010 IKFC meeting
The REDCap program received funding as a CTSA Informatics Pilot Project award in 2008 to support new work focused around increasing capacity and streamlining processes for collecting and managing data in diverse scientific areas of study. A majority of CTSA institutions are already using REDCap as at least a component of their CTSA-supported data management services. This update report will provide existing partners with a strong knowledge of what is coming next and other sites an opportunity to see and evaluate the program.

3) Impact on the CTSA Consortium Strategic Goals
   (1) Building National Clinical and Translational Research Capability (SG 1)
   Data capture is a universal need in any clinical and translational research study. REDCap provides research teams an ‘easy way to do the right thing’ when planning and implementing a secure and centralized data collection system for individual- and multi-center projects.

   (2) Enhancing training and career development (SG 2)
   Designing a study data collection plan is a daunting task for young investigators. Many training programs stress best practices for data collection, but neglect teaching of practical ways for implementation. Coupling best practice teaching with REDCap implementation ensures students receive both ‘why’ and ‘how’ instruction and also have a straightforward path for setting up a centralized data collection system to support single- and multi-institution research studies.

   (3) Enhancing Consortium-Wide Collaborations (SG 3)
   Scientific collaboration is naturally fostered through a growing number of REDCap-supported domestic and international multi-site research studies. In addition, the REDCap consortium includes a large group of informatics, IT and biostatistics professionals. Participation is not limited to CTSA institutions, but the CTSA program is well-represented among partner institutions. We are now routinely seeing studies where research teams are able to cooperate and collaborate efficiently because their sites are already using REDCap for data collection.
Abstract #2

**Procurement of Shared Data Instruments for Research Electron Data Capture (REDCap)**

Catherine McGraw¹, Jihad S Obeid MD², Jose G Conde MD MPH³, Rob Pawluk⁴, Brenda Minor⁵, Janey Wang⁶, Sean Banks⁷, Todd Ferris MD MS⁸, Sheree Hemphill MS⁹, Michael Lin⁴, Rob Schuff MS⁴, Bob Wong PhD², Elizabeth Wood MS¹¹, Rob Taylor, Paul A Harris PhD⁸

¹University of Cincinnati, ²University of Puerto Rico, ³Medical University of South Carolina, ⁴Mayo Clinic, ⁵Vanderbilt University, ⁶University of Texas at Austin, ⁷Stanford University, ⁸Case Western Reserve University, ⁹Oregon Health & Science University, ¹⁰University of Utah, ¹¹Weill Cornell Medical College

REDCap (Research Electronic Data Capture) is a novel and versatile software solution and tool set that allows biomedical researchers to create secure online forms for data capture, management and analysis with little effort and training. Initially developed at Vanderbilt University, REDCap has grown into a collaborative consortium supporting approximately 10,000 end-users across 138 domestic and international academic and non-profit institutional partners. One module recently developed in REDCap is the Shared Data Instrument Library (SDIL), a collaborative project between Vanderbilt and Mayo Clinic. The SDIL was designed to address the limited use of data standards and data sharing across typical research projects due the diversity of interests and needs. The objectives of the SDIL are to: 1) facilitate reuse of data dictionaries and reduce duplication of effort; 2) promote data standards and best practices; and 3) promote data sharing.

The development of the SDIL and procurement of instruments is overseen by a REDCap Library Oversight Committee (REDCOC), which consists of members from various institutions in the consortium. The charter of the REDLOC is therefore to assess the needs of the research community to determine how to prioritize instruments for coding into shared data dictionaries and making them available through the library to REDCap users. During the instrument coding review process, the committee ensures that resulting shared instruments meet best practices for data collection and adhere as much as possible to the original validated static or paper form of the instrument thereby reducing the likelihood of compromising statistical validity during translation. The committee also reviews library metrics for submissions, downloads and usage in an effort to improve the procurement process (Figure 1). In order to ensure adequate representation for the wide diversity of consortium members, the committee is composed of 10 individuals, all from different NCRR-funded institutions, with a rotating annual membership. These individuals work closely with the SDIL development and support teams from Vanderbilt and Mayo Clinic.

Currently, the library includes 60 instruments. REDLOC’s inclusion criteria for shared instruments consist of scientific relevance, ease of translation to REDCap format, status as public domain instrument and direct contact with author or governing source to ensure there are no copyright violations. Examples include: all components of the BRFSS (CDC’s Behavioral Risk Factor Surveillance System), WHO-5 Well-Being Index and RAND-36 for patient outcomes.

The success of this project demonstrates a clear example of leveraging the REDCap consortium to build and oversee tools that directly support individual geographically and scientifically distinct research teams. Phase 2 project plans for improvement include adding additional high-profile instruments, building workflow and REDLOC review mechanisms to facilitate researcher submissions, and possibly developing a model for sharing and consuming copyrighted instruments. The strong success of REDCap in diverse scientific domains coupled with the new SDIL resources has also created new opportunities for collaboration with external groups (e.g. caBIG CTMS group, PROMIS and CDASH) and we look forward to exploring fully these collaborations to best serve REDCap research teams.

**REFERENCES:**

Medical imaging in i2b2: new tools for integrating clinical images into research studies.

Shawn Murphy¹, Daniel Marcus², Christopher Herrick¹, Tim Olsen², Randy Gollub¹, Steven Piper¹, Carl Kesselman³, (¹Harvard Medical School, Boston, MA, ²Washington University, St Louis, MO, ³University of Southern California, CA)

The use of medical images is part of the protocol for many present and future clinical research studies. Images are used in clinical trials to quantify disease burden such as tumor growth, inflammatory changes, hemorrhage, and infarction. They are used to quantify the outcome of interventions, such as changes in tumor size or loss of brain tissue. Furthermore, they guide the way to novel diagnostic approaches to disease, such as the use of diffusion tensor MR imaging in the evaluation of brain trauma. The Informatics for Integrating Biology and the Bedside (i2b2) software suite is used by many CTSC awarded entities to enable the repurposing of healthcare data for clinical research. Through an administrative supplement from the NCRR, it became possible to develop a plug-in software addition (a “cell” in the i2b2 “hive”) that allows clinical images to be pulled from a hospital’s Picture Archiving and Communication System (PACS). The i2b2 plug-in utilizes a widely available, open source imaging informatics platform (XNAT, available at http://www.xnat.org) supported by the Biomedical Informatics Research Network (BIRN) in order to manage the images obtained from the PACS.

The images that become available through XNAT can be navigated through a user interface and accessed by external applications via an open, web-based application programming interface (API). Using this API, research tools have been adapted to compute and extract novel information from the clinical images using XNAT. These tools include Freesufer (developed at Martino’s Center at the Massachusetts General Hospital), which allows brain MRIs to be automatically segmented into anatomic regions, and 3DSlicer (developed by the National Alliance for Medical Image Computing), which allows non-rigid image registration models to be computed from conventional MRI images. In addition, any medical imaging software that complies with the DICOM standard can be interfaced with the XNAT research image repository using the XNAT Gateway application as a secure bridge.

Obtaining images from a clinical PACS can be a complex affair. The four hospital PACS systems in the Harvard Hospitals of Brigham and Women’s, Massachusetts General, Children’s, and the Beth Israel were appropriately concerned with what might be a serious interference with their clinical mission. Indeed, a tool that might pull tens of thousands of images from the clinical PACS could cause a serious clinical performance impediment. Therefore, a contract was put into place with the software that guaranteed only a governed number of images would be pulled at a specific rate at specific (off-peak) times. Furthermore, audit trails and security protocols were added to comply with HIPAA where standard PACS DICOM interfaces would fail to offer appropriate assurances.

Although the software does not need the i2b2 Clinical Research Chart (CRC) to function, the system is much more effective when patient queries that utilized the CRC can direct the download of the imaging studies to specific phenotypes. The image studies can then be specified and transferred to XNAT, where they are analyzed and annotated. The annotations can then be returned to the CRC for further querying and analysis with the rest of the phenotypic and genotypic data. The XNAT plug-in for i2b2 will become available shortly as an open source addition to the software platform, and is supported by both NCRR (1 UL1 RR 025758-01) and the BIRN (1 U24 RR 025736-01) funding.
Abstract #4

**Presenter/Contact:** William G. Adams, MD, badams@bu.edu

**Co-authors:** Kressin, NR, Thomas, S, Thomas, SM, Shanahan, CW, Paasche-Orlow, MK

**CTSA:** Boston University with the BUMC/BMC Health Disparities Research Program

**Title:** Extending i2b2 to Support Health Outcome and Disparity Measurement

Availability of high-quality clinical data and i2b2 offers tremendous opportunities for clinical and translational researchers seeking to use health data for intra- and inter-institutional hypothesis generation, cohort identification, and health services research. i2b2 currently excels in the area of cohort identification but lacks functionality related to health care process and outcome assessment. In this presentation, we describe our efforts over the past 12 months to: 1) integrate clinical and administrative data from an academic medical center (BMC), 5 affiliated community health centers, and the largest Medicaid insurance plan in MA; 2) create new i2b2 hierarchies for insurance, service area, and vital signs; and 3) develop a new i2b2 Cell, the “Health Outcome Monitoring and Evaluation (HOME)” Cell. The first two areas will be discussed during the presentation. This report focuses on Home Cell functionality and architecture.

We developed the HOME Cell to serve as a shared resource within the i2b2 community that will be improved over time. A functional prototype is available and will be demonstrated. The Cell functions within i2b2 without any changes to the core i2b2 schema or software. The HOME Cell reuses queries developed within the i2b2 Query Cell and extends functionality through user-specified “constraints” within the HOME Cell. Three core constraints have been developed and tested to date. The Occurrence Constraint models temporal relationships between facts. Users can specify that only persons with a fact that occurred in a certain temporal relationship with another fact should be included (i.e., to assess lipid screening rates in people with diabetes, only include people with a primary care visit that follow a diagnosis of diabetes). The Value Constraint extends the Occurrence Constraint further to specify a numeric range (less than, greater than, or between) on values of a clinical observation and specifies which value(s) to use -- min, max, average, earliest, most recent or all -- (i.e., to assess blood pressure control in people with hypertension, all systolic and diastolic blood pressure measurements after the diagnosis of hypertension should be less than 130 and 90 respectively). The Age Constraint models the relationship between patients’ age and facts (i.e., age at visit < 65). The constraints can be grouped as logical conjunctions (AND) of disjunctions (OR), just like in the i2b2 Query Cell. The HOME Cell specification of a clinical process or outcome starts with a base i2b2 query. A group of constraints are applied to the base query to define a denominator patient set and a numerator patient set. Existing i2b2 queries are used to specify various strata for analysis (i.e., African American, uninsured diabetics etc.). A reference interval can be applied to any subset of constraints to report the data within a specified date range by year, month or the entire range.

The architecture of the HOME Cell uses the existing i2b2 execution framework. User-specifications are translated into XML and parsed by the i2b2 server into a set of sequential queries which are executed in the order submitted. The query execution begins with an initial patient list generated as a result of the base i2b2 query. The generated SQL queries representing the specified HOME Cell constraints are applied sequentially to this list using joins. A counter is maintained for every patient in the list to track the number of constraints the patient satisfies. This counter is used to identify subgroups of individuals satisfying all constraints within a numerator and denominator population. Results can be reported as counts, averages, or proportions. The final denominator and numerator populations can then be stratified by any existing i2b2 query to assess exposure-outcome relationships for subgroups of individuals. The SQL ‘Group By’ function is used during execution to allow results to be reported by year, month, or the entire date range.

Our experience has thus far demonstrated that a vast and varied number of health services queries (disparity outcomes, comparative effectiveness, and quality reporting) can be supported using this model. Within the rapidly expanding i2b2 collaborative network the HOME Cell has the potential to be a powerful tool for translational research.
The Health Ontology Mapper (HOM) Project Status
Lead institution: UCSF

Objective: To foster CTSA consortium-wide collaborations we need data sharing networks. We have created a general-purpose instance mapper system that can translate raw local data into common standard data models and ontologies so that data can be shared across diverse institutions.

Background: Data sharing is of critical concern to CTSA consortium-wide collaborations. The CTSA would benefit from the widespread deployment of grid connected data warehouses containing all clinical, research and bench science data. To query such a network the CTSA will require both syntactic and semantic interoperability.

Project Aims: 1) define mapping framework for adoption as standard 2) build a user interface for mapping raw data into common standards; 3) test and deploy an on-demand grid connectivity system for both Harvard SHRINE and caGRID.

Accomplishments: By supplying syntactic interoperability and by leveraging the semantic interoperability of components developed for caGRID the HOM system has been successfully integrated with i2b2 for usage on caGRID and SHRINE. HOM mapping standard has been adopted into HL7 CTS2. HOM can normalize all local formulary data to RxNorm and can normalize all local clinical disposition data to the proposed HL7 Discharge standard. The normalization of all clinical lab data is in progress.

Participating CTSAs: UCSF; UCD; URMC; U Penn; Ohio State U; TCH Denver and UW.

WHY IMPORTANT TO PRESENT AT 2010 CTSA IKFC MEETING
HOM can now be widely used and solves enormous problems among data sharers. Without HOM we would be able to share information but we would not be able to query it.

IMPACT ON CTSA CONSORTIUM STRATEGIC GOALS
CSG #1 Enhancing national clinical and translational research capability: A critical data-sharing infrastructure. HOM translates raw local data into common standards to enable data sharing networks.
CSG #3 Enhancing consortium-wide collaborations: HOM is in use on several CTSA based grid data sharing projects, including CICTR, HSDB (Human Studies Database), COHRI (Dentistry), HOMERUN (comparative medicine), DBRD (Rare Disease).
The Carolina Data Warehouse for Health (CDW-H) is a major resource at the University of North Carolina at Chapel Hill which is used for secondary data analysis. The CDW-H draws active users from performance assessment (quality) and research communities. Launched in 2009, the CDW-H system was developed using IBM’s InfoSphere and DB2 platforms and it is co-managed by the Hospital’s Information System Division and the Translational and Clinical Sciences Institute (UNC’s CTSA organization). The CDW-H currently contains about 2 million patient records in 202 tables and 2840 unique fields.

A primary objective for the CDW-H team is to offer end-to-end services such that a user can begin a search for medical records, save the retrieved file, manipulate data, produce derivative data and reports, and share results generated from their analysis while interacting within a single environment. We are developing software components and tools to achieve this end-to-end vision. In parallel, however, a critical dimension we are also addressing is data security both at data governance and at information technology levels.

At the governance level, we have made the most progress and created a multi-tiered structure to review and approve requests for data use. The governance group is a multidisciplinary team consisting of members from the UNC Hospitals Information System Division, campus Information Technology and Security Operation, High Performance Computing Center (Renaissance Institute), Lineberger Cancer Center’s Governance Group, Institutional Review Board, and the Translational and Clinical Sciences Institute. We use the governance structure to also review proposals for new system developments that use our data or may require modification to the data warehouse environment (e.g., data marts). So far about 50+ project proposals have been reviewed and approved through this governance structure ranging from basic data access requests to development of new systems to facilitate data federation.

At the system level, we created two sets of services: 1) pre-research and 2) data provisioning. For pre-research, a software was developed to allow users to query and identify useful data sets based on de-identified and aggregated views of the data, and it is complemented with support from human analysts. For data provisioning, we deployed a service environment which provides users access to their approved data on a secure network-based file server, where they can utilize typical Windows file access techniques, such as mapping a network drive, to open, view, manipulate, and analyze their data. A fundamental gap in the current approach is that it cannot prevent a user to copy data to an unauthorized media or a drive. We recently developed a basic “data leakage” tracking application which can monitor data movements and alert users and administrators regarding unauthorized transfers of data. We are in the process of evaluating the solution and plan to launch it in our production environment soon. When we closely examined the motivation for such copying we found that convenience is a key factor. Data requesters wish to have access from any location, at any time, and they wish to use the data in environments with data analysis and report generation tools that they are familiar with.

In this presentation we will provide an overview of our ongoing efforts to address data security in provisioning of medical data and point out key technologies and practices developed at UNC that are ready for adoption by other CTSA sites. Additionally, we plan to cover certain critical barriers we faced in secure data provisioning. Some of the barriers we plan to discuss are: 1) tracking and auditing the use of data over a long period, potentially across multiple projects and project members, 2) offering virtualization and customization for specific types of applications and use that integrate data analysis software with data in the same environment, and 3) linking the data use environment to compliance environments such as automated IRB tracking systems. We have ongoing work on all of these areas and our goal for this part is to seek input from the audience and encourage potential collaborations.
LC Data QUEST: A Technical Architecture for Community Federated Clinical Data Sharing

Ching-Ping Lin PhD, Kari A. Stephens PhD, Robert A. Black MA, Peter Tarczy-Hornoch MD
Institute of Translational Health Sciences, University of Washington, Seattle, WA

Health data sharing with and among practices is a method for engaging rural and underserved populations, often with strong histories of marginalization, in health research. The Institute of Translational Health Sciences (ITHS) has developed the LC Data QUEST (Locally Controlled Query Extraction Standardization Translation) pilot project to build a technical architecture for supporting clinical health data sharing across networks of primary care practices and American Indian and Alaska Native communities in the Washington, Wyoming, Alaska, Montana, Idaho region. This architecture facilitates translational research by increasing the accessibility to health data captured in electronic medical record systems used within clinics that serve rural populations, thereby accelerating the integration of new findings into care practices.

The LC Data QUEST sites use diverse EMR products without agreed upon data standards and data practices. Discussions with national CTSA colleagues building similar architectures and Practice Based Research Networks (PBRNs) led us to pinpoint vendors based on their experience with delivering ETL services in medical setting, point-of-care tools, and solutions to semantic alignment. These requirements, in addition to our original requirements of including a federated architecture and remote management, comprised the core set of system requirements that we used to evaluate vendors. Our vendor extracts a set of common data elements into individual LC Data QUEST repositories located at each practice site. Once the data is loaded into the LC Data QUEST repository, it can be shared with researchers or analyzed by the clinics for cohort discovery, randomized control trials, or comparative effectiveness research. Supporting this participation is the new ability for individual practices to analyze their own repository data for quality improvement initiatives. A quality-focused program generates a point-of-care report that includes clinical decision support for recommended national guidelines of care. Patients and practitioners review the point-of-care report during visits and can correct data errors thus enhancing data quality. The common reference set of data elements that facilitate decision support comprise the core elements for the data sharing pilot.

The major technical components of the data sharing architecture include: 1) the extraction, transformation, load (ETL) process that transfer and align the health data from the local electronic medical records (EMR) to a separate research-focused repository at each site, 2) a set of end-user applications that can deliver data appropriately to users, 3) data quality management, and 4) metadata management. We selected the best technical solutions to carry out these activities that met our system requirements while staying within the scope of our financial and human resource constraints. Our initial informatics focus was to implement the ETL process. We also developed data quality and metadata management approaches to establish a foundation for a data sharing architecture upon which we can apply end-user applications.

LC Data QUEST is a unique and evolving partnership between clinical researchers, informaticists, community liaisons and community members. Research networks partnering with geographically dispersed, often underserved, diverse and rural populations have enormous potential for increasing recruitment of study participants and introducing new programs and therapies into communities. Our experience has underscored for us the importance of sustainability and growth when building data sharing networks in practices and communities by developing incentives for all partners. Our architecture reflects the work of engaging communities and their priorities to develop common processes and technical interoperability to advance research.
An Architecture for Creating Semantically-enabled Registries through Federated Query of Clinical Databases at a Multi-institution CTSA

Andrew Post, MD, PhD1,2, Tahsin Kurc, PhD1,2, Marc Overcash1,3, Dedra Cantrell, RN1,2,4, Tim Morris1,1, Kristi Eckerson, MSPH1,3, Circe Tsui1,3,6, Terry Willey, RN1,4, Arshed Quyyumi, MD5, Danny Eapen, MD5, Guillermo Umpierrez, MD5, David Ziemer, MD5,6, Joel Saltz, MD, PhD1,2

1Atlanta Clinical and Translational Science Institute Biomedical Informatics Program, Atlanta, GA, 2Emory University Center for Comprehensive Informatics, Atlanta, GA, 3Emory University Research and Health Sciences IT, Atlanta, GA, 4Emory Healthcare Information Services Department, Atlanta, GA, 5Emory Healthcare Department of Medicine, Atlanta, GA, 6Grady Health System, Atlanta, GA

In the CTSA era there is great interest in aggregating and comparing patient populations at multiple clinical sites that may span institutions with different IT infrastructures. The Atlanta Clinical and Translational Science Institute (ACTSI) Biomedical Informatics Program (BIP) serves such a community of researchers. To meet this need, BIP is implementing a virtual data warehouse to provide access for research to clinical databases across its member sites. The first implementation phase is underway and involves building reusable infrastructure for creating semantically enabled, up-to-date and topic-specific data marts or registries from source databases. These registries will support direct query by researcher end-users, integration with other clinical research data, and export to statistical analysis tools.

The Registry Project architecture provides an ontology-based semantic layer that maps underlying database schemas to a shared set of common data elements and controlled terminologies. It permits specifying derived data elements that represent categories and temporal patterns in source data. We initially are developing common data elements based on existing standards for cardiovascular disease, diabetes and co-morbidities. We are mapping those data elements to two source systems, the Emory Healthcare Clinical Data Warehouse and the Grady Health System Diabetes Patient Tracking System. The CDW contains more than 80% of the clinical and administrative data in Emory Healthcare’s Cerner Millennium (Kansas City, MO) EHR and is updated daily. The Grady DPTS is a home-grown database on diabetes clinic patients with data obtained manually from paper chart abstraction and electronically from billing and laboratory systems. We also are developing derived data elements that represent categories at various levels of breadth (e.g., heart failure-related ICD-9 codes, cardiovascular disease-related data) and temporal patterns of common interest (e.g., whether the second of two sequential hospitalizations within 30 days with related diagnoses constitutes a re-admission). These features aim to enable extraction of categories of CDW and DPTS data including computation of derived values while maintaining links to the source data to allow periodic data refreshes. We are developing software to extract the data, compute derived values and import the results into i2b2 for end-user researcher query and analysis. Thus, the architecture supports creating customized, updatable, and semantically enabled registries that are specific to a topic area or a research project.

We initially are targeting the creation of registries for research in cardiovascular disease, diabetes and co-morbidities. Ultimately, the system will provide a broad range of capabilities. These include preparatory-to-research queries on specific topics that will allow researcher end-users to query by concept without having to know the underlying source database schemas. We also plan to support funded research projects that require access to electronic health record data by enabling those projects to create and maintain data marts containing specific data of interest. Providing registries, as opposed to allowing access to institutions’ entire databases, is expected to provide enhanced usability, performance, and security. Future phases will extend the project’s scope to the other member sites and databases in ACTSI.
Adaptation of a caBIG-based Clinical Trials Management Suite to the National Children’s Study

William R. Hogan MD, MS1,2, Umit Topaloglu PhD1,2, Cheryl Lane2, Jiang Bian2, Thomas Kieber-Emmons PhD2, Laura F. Hutchins MD2, Charlotte Hobbs, MD, PhD3

1Division of Biomedical Informatics, University of Arkansas for Medical Sciences, 2Winthrop P. Rockefeller Cancer Institute, and 3Arkansas Children’s Hospital Research Institute

Description
At the University of Arkansas for Medical Sciences (UAMS), our Winthrop P. Rockefeller Cancer Institute has been developing a Clinical Trials Management Suite (CTMS) of applications based on caBIG®. We are adapting this CTMS to the informatics requirements of the National Children’s Study (NCS). The NCS funds 37 study locations at present with the ultimate goal of examining the effects of environmental influences on the health and development of 100,000 children across the United States.

The UAMS CTMS comprises the caBIG® Central Clinical Participant Registry (C3PR), Patient Study Calendar (PSC), and caTissue tools in combination with OpenClinica and LimeSurvey. In addition, we have created a web-based portal or “dashboard” that allows easy access and single sign-on (using one’s UAMS network login) to all the applications. All components of the CTMS are open source and available for adoption by other interested sites.

We created the NCS, its epochs, the eligibility criteria for each epoch, and test subjects in C3PR, as well as registered test subjects there. We created the current draft of the NCS calendar in the PSC. The NCS calendar is complex, requiring different activities at different times for the mother, father, child, and other possible participants such as neighbors and other family members. We were able to match the correct activities with each participant type. We also were able to create the NCS protocol in the caBIG® caTissue application for planned specimen collection. We created case-report forms in OpenClinica, an open-source, electronic data-collection tool. And finally, we used the open-source LimeSurvey tool to implement electronically a draft version of one of the participant surveys for the NCS. Our results thus far indicate that the functionality of the tools will meet most of the required operations and data-collection needs of the NCS. We intend to continue adaptation of the suite for planned enrollment of subjects beginning later this year.

In conclusion, our existing, open-source informatics infrastructure for conducting cancer trials at UAMS appears flexible enough to accommodate a substantial portion of the needs of the NCS, a non-cancer study. The results show that our UAMS CTMS dashboard is promising to standardize our research data collection efforts in compliance with national initiatives. The NCS study center at our institution has decided to proceed with the UAMS CTMS, and to evaluate options for pre-consent recruitment tracking for inclusion into the suite.

Importance to the CTSA IKFC Meeting
Several CTSA universities also host study centers for the National Children’s Study. Many of these sites are adapting informatics solutions developed as part of their CTSA to the NCS. We were invited by Dr. Hirschfeld, the Director of the NCS, to present our solution at the NCS Steering Committee Meeting in August. Participants will learn about how CTSA informatics activities at UAMS can be adapted to the NCS. Our solution set is open source and adoptable by other sites.

Impact on CTSA Consortium Strategic Goals
This research is relevant to the mission of SGC1: to improve all processes related to the development, approval, activation, enrollment, and completion of clinical trials. Also, Dr. Hirschfeld is also co-chair of the CTSA Consortium Child Health Oversight Committee (CC-CHOC). Thus, important links between the goals of the NCS and the Consortium already exist.
Using caBIG and Semantic Web Technologies for Collaborative Research

Jim McCusker¹ and Michael Krauthammer¹,²
¹Department of Pathology, ²Yale Center for Medical Informatics, Yale University School of Medicine, New Haven CT

Abstract:
Our goal is to provide robust means of storing translational research data and making it available for exploration by individual researchers, for analysis by bioinformatics and statistics experts, and to store the data in such a way that it can be seamlessly merged with existing clinical, genomic and proteomic data. Our initial focus is data from translational research in skin cancer (melanoma), but our approach is easily extendable to similar data from Yale CTSA projects.

We make extensive use of tools and resources of the caBIG Tissue Banks and Pathology Tools (TBPT) and Integrated Cancer Research (ICR) workspaces. The core of our infrastructure is caTissue, a specimen tracking system that links experimental results back to samples, specimens and study participants; caArray for storing -omics data; calIntegrator2, a data exploration portal; GenePattern for bioinformatics data processing, and Corvus, a data warehouse for integrative statistical data analysis. Corvus is a hybrid of Semantic Web and relational database technologies.

Corvus leverages linked data resources such as Bio2RDF and can be accessed from R, a statistical analysis package. We have successfully used Corvus for generating analyses across different melanoma omics data sets, such as integrative analyses across the melanoma transcriptome and phosphorylome. We are currently developing a Corvus SPARQL endpoint, for linking Corvus to semantic web inference engines. We are particularly interested in building RDF graphs from Corvus data, and merging this graph with graphs from ontological pathway and gene/protein annotation resources. The merged graph can then be interrogated using pathway-based reasoning, or reasoning across gene and protein hierarchies.

Finally, we are designing a gateway (clinical data portal) to trialDB, a Yale Clinical Trial Management System. The idea is to build a one-stop solution to enter clinical data about research subjects. Currently, we are operating two different study calendars, one for caTissue and one for trialDB. Using the SOA paradigm, the clinical portal would ultimately substitute these calendars with a common caBIG patient study calendar (PSC).
Poster Title: Using the Open Metadata Registry (OpenMDR) to create data sharing interfaces.

Authors:
Rakesh Dhaval, MS, Center for IT Innovations in Healthcare, The Ohio State University; Calixto Melean, MS, Center for IT Innovations in Healthcare, The Ohio State University; David Ervin, Center for IT Innovations in Healthcare, The Ohio State University; Philip R. O. Payne, Ph.D., Center for IT Innovation in Healthcare, Department of Biomedical Informatics, The Ohio State University

Description:
The domains of clinical and translational research are collaborative in nature and team science necessitates using information systems that are both locally relevant and globally interoperable. The Translational Research Informatics and Data Management (TRIAD) grid is an initiative to provide innovative methods for integrating heterogeneous information across institutional boundaries to increase the speed, efficiency, and impact of clinical and translational research. As an extension of the caGrid middleware, TRIAD is a service-oriented infrastructure designed to support translational research by enabling the creation of scalable, secure and knowledge-anchored data-sharing environments.

OpenMDR is a suite of tools that provides grid-compatible semantic metadata management capabilities, including the creation of locally relevant ontology-anchored data elements and conduct of federated queries and retrieval of semantic metadata from repositories across grid-enabled networks, including TRIAD and caGrid. The suite comprises of four different components: 1) MDR Core, 2) MDR Query, 3) MDR Plug-in, and 4) MDR Domain Model Generator.

MDR Core is an ISO11179 semantic repository capable of storing, versioning, and maintaining semantic and representational metadata. MDR Query is a grid service used to search multiple semantic repositories, giving developers the option of using other semantic metadata management tools in addition to those provided by the NCI such as caDSR and EVS. UML modelers use MDR Plug-in within Enterprise Architect to search for semantic metadata from multiple registries. Service developers can use the MDR Domain Model Generator to create semantic metadata for caGrid and TRIAD grid data services. Each of these projects provides functionality that enables federated semantic metadata annotations to be created and used in Grid Service Registration and Discovery. OpenMDR is designed to be locally deployed, populated, and curated. This allows service developers and institutions to maintain locus of control for their data and terminologies while facilitating rich semantic interoperability with other institutions, and maintaining a fast and agile process for annotating and delivering a strongly typed and semantically anchored grid services into production. Such services provide out-of-the-box data sharing functionality and through the use of existing grid tooling and shared services relating to discovery and federated query capabilities to address issues of research networking. Service creators and knowledge seekers across institutions can leverage the OpenMDR query tools to locate terminology relevant to their area of interest and perform queries using the rich semantic knowledge it provides.

OpenMDR 1.0 was released on February 12, 2010 and is undergoing revisions for an upcoming 1.1 release with additional features and functionality.

References:
2: Cancergrid’s cgMDR, UK (http://cancergrid.org)
A Highly Scalable Approach to Extending Functionality of the Research Networking Tool
Profiles Research Networking Software with the OpenSocial Standard

Eric Meeks (UCSF), Leslie Yuan (UCSF), Griffin Weber (Harvard), Mini Kahlon (UCSF).

Research Networking applications are becoming widely adopted as tools for expertise mining and discovery. They are quickly transitioning from “early adopter” to “mainstream” status. Numerous institutions have installed some form of a research networking platform using either a proprietary solution or an Open Source product such as Profiles Research Networking Software or VIVO. At UCSF we have deployed Profiles (http://profiles.ucsf.edu) and are contributing an extension to the Profiles product that will allow rapid expansion of non-intrusive features and functionality through an emerging technological standard known as OpenSocial.

Product Description. The features and functionality we are referring to will be developed as OpenSocial Apps. At UCSF we have extended Profiles to be an OpenSocial Container by utilizing the Apache Shindig code base. This is a common approach for making social networking sites OpenSocial compliant and has been utilized by well known sites such as LinkedIn and iGoogle. We have worked closely with the Harvard Profiles team to dovetail the implementation of OpenSocial within the new modular architecture of the Profiles software. A detail of our first list of OpenSocial Apps follows:

<table>
<thead>
<tr>
<th>Type of Application</th>
<th>Name of Application</th>
<th>What it does</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fosters productivity / collaboration</td>
<td>Mentorship</td>
<td>Allows for addition and searching on mentorship information. Supports CCSC Strategic Goal #2</td>
</tr>
<tr>
<td>Fosters productivity / collaboration</td>
<td>Listserv generation</td>
<td>Create email lists of selected people identified on Profiles</td>
</tr>
<tr>
<td>Integrate information from 3rd party networks</td>
<td>LinkedIn Profile</td>
<td>Shows the public LinkedIn profile of a researcher from within Profiles.</td>
</tr>
<tr>
<td>Social media communication tools</td>
<td>Activity Stream</td>
<td>Add an activity stream to Profiles</td>
</tr>
<tr>
<td>Fosters productivity / collaboration</td>
<td>Huddle</td>
<td>Create online collaborative workspace from within Profiles</td>
</tr>
</tbody>
</table>

In allowing features to be rapidly added to the product, we hope to foster a thriving market of functional innovation. The non-intrusive nature of the OpenSocial Apps will allow institutions to build, borrow, evolve or ignore features as they see fit. Multiple institutions can develop complementary or even competitive Profiles functionality without the need for coordinated timelines or risk of code conflict. As a further benefit, a wealth of OpenSocial Apps can be readily found in Open Source form today. Many of these existing OpenSocial Apps can be used to extend Profiles functionality in a complete plug-and-play manner, while others can serve as source code templates that can be modified to meet a more specific clinical researcher need.

Importance for the 2010 IKFC meeting. Our work does not need to be restricted to one product. Our work to make Profiles an OpenSocial container could be adopted by other products such as VIVO or Loki to extend the same benefit to their platforms. In addition to reducing the workload for such an effort by allowing others to leverage our work, this would also help to insure that OpenSocial Apps written for one platform would also work for other research networking platforms with little if any modification. More importantly, the OpenSocial standard is evolving, and we have an opportunity for influence. In producing our first set of OpenSocial Apps, we will start to learn what form and functionality are most suited to help advance clinical and translational science and we can work with the groups managing the OpenSocial standard to formalize our extensions into the API. This will enable us to lock in interoperability in the expanding field of research networking, and thus directly support CCSC Strategic Goal #3.
Interoperability among research networking platforms: A national pilot
Griffin Weber (Harvard), Mike Conlon (U Florida), David Eichmann (U Iowa), Holly Falk-Krzesinski (Northwestern), Michael Halaas (Stanford), Layne Johnson (U Minnesota), Eric Meeks (UCSF), Donald Mitchell (Stanford), Titus Schleyer (U Pittsburgh), Sarah Stallings (UC Denver) Mini Kahlon (UCSF).

Project Description. Research networking tools use data-mining, social networking and semantic web approaches to enable expertise discovery, matchmaking and more. Several research networking platforms have been built - the commercial product ‘Collexis’, U Pittsburgh’s Digital Vita, Harvard Catalyst Profiles, U Florida’s VIVO, U Iowa’s Loki, Stanford CAP, and others – and many additional institutions have also deployed at least one of these products. On August 6th 2010 the CTSA Research Networking Group facilitated a day-long meeting at the University of California, San Francisco (UCSF) with CTSA and non-CTSA participation including leads of major research networking tools, to discuss interoperability between products and institutions and to design a pilot for a federated national network. As a result of the meeting, initial guidelines for a pilot national network were established and a ‘pilot project’ to create the first test instance of the pilot network was kicked off. This presentation to the IKFC will describe the conceptual and technical architecture of the pilot, point out future challenges, and explain how others can participate in the national network.

The design of the pilot network addresses a simple use case—searching for a potential biomedical research collaborator across multiple institutions in a way that provides value over existing methods, such as Google or Facebook. The pilot was further defined by our belief that individual institutions can provide “cleaner” and more complete data about their own researchers by combining external sources of data with their own local databases. Therefore, we decided that the focus of our pilot network would be to generate buy-in from institutions so that they will be both willing and eager to share their information and encourage their researchers to adopt the tool.

Against a backdrop of allowing individual institutions as much control as possible over the data shared and the user experience, the network, as currently envisioned, has two components that define it – a) a technical architecture and common interface and b) an agreement between participating institutions. The agreement between institutions emphasizes that upon the execution of a federated search, any participating institutional website will only show the number of ‘hits’ a search term produces at each of the pilot participating institutions, and a URL back to each institutions local research networking website. In this way, participating institutions leave the presentation of research expertise to each individual institution, while still allowing for the benefits of a search across institutions. The technical design of the pilot network achieves the requirements of the pilot agreement by using a federated architecture with users initiating searches from within the framework of their local institution’s website. Key aspects of the architecture are that there is no central database, search index, or website; a global ranking algorithm is not needed; institutions can define which populations to load into their databases, what a search “match” is, and how to sort/rank people within their institution; and institutions may remove themselves from the network at any time.

The pilot network is open to any institution that is interested, while the initial pilot project (‘proof of concept’) although also open to any institution has requirements such as having a mature, deployed research networking tool and resources committed to assessments. We recognize that once we reach a critical number, followup meetings will be necessary to discuss governance, scalability, user experience, data quality, and other issues. The CTSA Research Networking group is the facilitator of this process, but participants already include non-CTSA institutions (University of Minnesota).

Importance to the IKFC Meeting. Research Networking has always been an important component of the annual IKFC meetings, and in the past year there has been remarkable growth in both the number of institutions adopting software tools and the sophistication of those products. The fact that we are finally at the point where we can discuss interoperability and are on the verge of having a national network is a major milestone that will be of interest to all CTSA sites.

Impact on Consortium Strategic Goals. This project is the direct result of the Research Networking Group’s ongoing efforts to enhance consortium-wide collaborations and interactions with non-CTSA institutions.
Beyond “Web 2.0” as Buzzword: Engaging Researchers to Contribute Online

Most university public relations and communications departments are now aware of buzzwords like “social media” and “Web 2.0”. But in implementation, the most critical component often gets lost. That is, getting users, especially faculty, to directly add content online.

At UCSF’s CTSI, we have rolled out three different tools as part of our Virtual Home web portal that require user input for their full potential to be achieved. Each has proven successful using very different tactics. In this presentation we will describe prior research that informed our tactics, the three distinct projects and their value to the research community, adoption statistics correlated with individual tactics, comparison to external benchmarks, lessons learned and open questions.

The three challenges we set out to address include:
1. Engaging the largest possible community in collaborating to generate ideas to improve research processes (“crowdsourcing”). We used a tool we call “Open Forums” to engage faculty and staff, with a series of three increasingly large groups participating. Statistics shared will include relative contributions from faculty and staff, ratio of active contributions (posts, comments & votes) to passive reading of content, and ratio of ideas that ended up expanding and improving to those initially proposed.

2. Enabling “curbside consulting” online to increase the efficiency of ethics consultations. A recently launched blog, authored by a national biomedical ethics expert, gained momentum through a carefully crafted strategy of engaging “captive” users.

3. Kickstarting faculty contributions to a research networking tool. We recently launched *UCSF Profiles*, our expertise discovery and networking tool built off Harvard Catalysts’s “Profiles Research Networking Software”. We used “ambassadors”, a media campaign, and a sweepstakes to gain user adoption in contributing personal details such as research narratives to the tool. Statistics shared will describe the correlation of several communications tactics with usage of and contribution to *UCSF Profiles*.

General principles of success gleaned from these initiatives include 1) seeding tools to provide value before users are required to themselves contribute, 2) being as focused as possible in the ‘ask’, ideally connecting the ‘ask’ to a tangible benefit for users, 3) being realistic and focused about interest and user groups and 4) phasing communications initiatives to ensure each audience can advocate to the next broader group of users.

*Importance for 2010 IKFC Meeting*: Engaging users to contribute online is a challenge with few success stories in biomedical research. Web 2.0 is a buzzword with few concrete examples of what it means and how to make related strategies successful. This is a good time to present metrics and best practices on this topic.

*Relevance to CTSA Strategic Goals*: This presentation is directly related to SGC 3 – enabling collaborations and SGC-1 – improving infrastructure.
The Human Studies Database (HSDB) Project

**Lead institution:** UCSF

**Objective:** A federated computable database of the design of all human studies to enable large-scale computational query and analysis of human studies data for clinical and translational research.

**Background:** Human studies (both interventional and observational) are the most important source of evidence for advancing our understanding of health, disease, and treatment options. Study designs and results should be made computable for large-scale data mining, synthesis, and re-analysis.

**Project Aims:** 1) define scientific features of human studies in the Ontology of Clinical Research (OCRe); 2) define the HSDBgrid data federation architecture (Figure 1); 3) collect descriptions of human studies from individual CTSAs and federate over HSDBgrid. We will federate results data later, contingent on acceptable data ownership policies.

**Accomplishments:** We have completed a first version of OCRe, including a study design typology and a statistics ontology. OCRe is in OWL and UML, and is being bound to external standard terminologies. We are working out an end-to-end architecture using LexEVS, BioPortal, and OpenMDR services to define, serve, and federate semantic standards and value sets over caGrid and SHRINE. We are using Dynamic Extensions technology from caTISSUE suite to build data entry and query interfaces, and are exploring ways of integrating HSDB with trial registration and IRB approval workflows. Three HSDB papers won AMIA Distinguished Paper awards, and we are in the 2nd year of an NCRR R01 supporting this work.

**Participating CTSAs:** Columbia; Duke; Emory; Hopkins; Mayo; Ohio State; Rockefeller; Stanford; UC Davis; UCSF, U Colorado; UT Southwestern; UTHSC San Antonio, U Washington; Wash U. St. Louis.

**Why Important for IKFC Meeting**

All CTSAs engage in human studies research. The more CTSAs that share human studies data, the richer the HSDB will be for clinical and translational research. In the coming months, we will be ready to pilot collection and federation of human studies descriptions from individual CTSAs. We will describe our approach to semantically-standardized data sharing, demonstrate tools, and make the value case for CTSA institutions to participate in HSDB.

**Impact on CTSA Consortium Strategic Goals**

**Goal #1** Enhancing national clinical and translational research capability: As a computable inventory of past and ongoing human studies, HSDB will be a critical infrastructure for clinical research. Its scientific details will be useful for scientific query and analysis as well as for scientific portfolio and other clinical research management.

**Goal #2** Enhancing the training and career development of clinical and translational scientists: The study design typology can be used to teach investigators about study designs and to help BERD consultation units manage and direct resources to supporting design types of greatest demand.

**Goal #3** Enhancing consortium-wide collaborations: HSDB has been a leader in using cutting-edge semantic technologies to implement platform-independent data sharing. The scientific detail in HSDB can also be used to facilitate research networking along specific scientific interests (e.g., use of depression rating scales, genomic profiling of drug side effects, or N-of-1 studies in obesity treatment).

**Goal #5** T1 Research: HSDB standardizes human study design terms and descriptions for T1 researchers to query and explore past studies to inform the design of new first-in-human studies.
i2b2 at the University of Michigan  
Cohort Discovery and Beyond

Preliminary Activity at UM
Several groups are exploring or proposing potential use of i2b2 to support health research. They include the National Center for Integrative Biomedical Informatics (NCIBI), the Michigan Institute for Clinical and Health Research (MICHRI) and the UMHS Hospitalist Program (Flanders). NCIBI has completed a “proof of concept” integration of its automated annotation tool (Gene2MeSH). MICHRI (CRIC) completed an in-depth evaluation of i2b2 in the context of the NIH-funded Physio-MIMI project. Flanders submitted a GO grant as part of the 16-member Hospitalist CER/HSI Research Network. This network (HOMERUN) proposes to use the i2b2 federated query capabilities of the CICTR project.

Value Proposition
An institutionally-supported i2b2 initiative will help meet a significant unfilled need of the CTSA at the University of Michigan by providing an integrated research data information framework with appropriate analytical tools that can be used to support clinical and translational research. Specific and near-term achievable opportunities include cohort discovery demonstration project, multi-institutional query capability (HOMERUN) and NCIBI opportunities, including support for many translational research investigators at the Center for Computational Medicine and Bioinformatics (CCMB) (Kretzler, Burant).

Loading Data – 12 years of patient data, 436 million rows
Demographic Observations (3.1 million patients)  
Procedures and Diagnoses (20.4 million)  
UM-Coded Lab Results (206 million)

i2b2 Demonstration Project
A prototype data mart that includes patient demographics, diagnoses, procedures, medications and lab data. Goals of the prototype are to 1) demonstrate i2b2 to groups within UMHS to develop broad support for the project, 2) provide a way for UMHS researchers to assess the capabilities of i2b2 and to identify any missing capabilities that would be necessary for i2b2 to be successful, 3) allow UMHS researchers to assess the way different data types work within the i2b2 framework to support cohort identification so that this information can be used to identify and prioritize data types to be targeted for a production implementation, and 4) participate in i2b2 national demonstration project using i2b2’s SHRINE capability and UMHS Hospitalist program.

It takes a Village . . . the i2b2 team at UM
Brian Athey (MICHRI, Res Admin)  
Joe Bath (MCIT, App Infra)  
David Belenky (MICHRI, Tech Supp)  
Jeff Cowall (MCIT, Tech Supp)  
Moira Dowling (MICHRI, Bus Mgt)  
James Law (MCIT, Tech Supp)  
Steve Gendler (MCIT, Bus Mgt)  
D Hanauer (Researcher, Validation)  
Debra Haslan (MICHRI, Compliance)  
Emma Hung (MICHRI, PM)  
Jim Maszatics (MICHRI, Bus Mgt)  
Jim Monkiewicz (MICHRI, BA)  
Rodney Nelson (MCIT, App Infra)  
Kurt Richardson (MCIT, Bus Mgt)  
Kirk Rupar (MCIT, DBA)  
Kevin Smith (MICHRI, Res Admin)  
Nimi Subramanian (MCIT, Serv Infra)  
KP Unnikrishnan (Research, Validation)  
Kay Wilson (MICHRI, BA)  
Joe Zukowski (MCIT, DBA)
Background: Identifying clinical researchers is a mysterious process with several possible methods. While ideally, each researcher would maintain their own public, up-to-date profile, at this point one must resort to searching publications, grant applications, conferences, and local hearsay to find a specific clinical researcher. Instead, we used existing ICD9-coded diagnoses within the Electronic Medical Record to arrive at physician profiles. The focus was on obtaining a properly ordered list of ICD9 diagnosis codes to correctly describe the professional profile of a given physician that would reveal their expertise and deemphasize the incidental diagnoses associated with their practice.

Methods: The ICD9 codes associated with the encounter note, patient, and the physician ID were extracted from the Epic EMR database. To identify valid approaches, five sample physicians were chosen at random and all of their associated ICD9 codes were manually marked as either “relevant” or “irrelevant” to the physician’s profession.

Before attempting to algorithmically cluster the ICD9 codes into the two groups, multiple strategies of scoring the ICD9 codes to rank them based on relevance were evaluated. Initially, using the number of patients and encounters for each ICD9 code as a score was attempted. The second attempt added the inverse of the percentage of the physicians who have encountered a given ICD9 code to factor in the possible association of a code with a particular physician. The third determining factor was derived from placing each ICD9 code from a particular physician’s code list onto the hierarchical ICD9 tree and considering the number of each code’s neighbors based on the distance from the initial code. Each of the factors needed to be appropriately normalized and integrated to produce a single relevance score for each ICD9 code per physician. To evaluate whether each additional factor improved the sort order of the physician’s ICD9 codes, Hamming distance to the nearest target list was calculated. During the evaluation, a target list was defined as having all of its relevant ICD9 codes (as defined by the control) sequentially listed at the top of the list in any order.

Results: In our preliminary findings for one of the sample physicians, the Hamming distance between the algorithmically produced ICD9 list sorted by the described combined relevance score and the nearest target ICD9 list improved from 198 (when based on sorting by the number of patients seen with a given ICD9 code) to 126 when the combined score was used, indicating a 36% improvement over the simplest method. Additional improvement was demonstrated by the mean, median, and standard deviation of the relevant ICD9 code rankings improving by 52%, 68%, and 26% respectively. 50% of all encountered diagnoses relevant to the given physician’s practice were thereby found in the top 16.5% of the list of all diagnoses rendered by this physician. Using the k-means clustering algorithm successfully identified the amount of ICD9 codes at the top of each resulting list that should be considered as the “clinical profile” of a given physician, up to 97% of which consisted of ICD9 codes considered “relevant” to the profession.

Conclusion: A large number of physicians at major academic institutions participate in clinical research and mining patient encounter notes within the electronic medical record to obtain their clinical experience profile would significantly ease the task of locating specific clinical researchers based on their field of expertise. A more developed version of this technology, where physicians would be able to explicitly label themselves as clinical researchers, should allow other medical researchers to identify the most likely relevant specialists based on up to date records from the EMR.
Automating Electronic Data Capture Application User Management
Linda Carlin, PhD
University of Colorado, Denver

Objective: The goal of this poster is to 1) define informatics application user management goals 2) identify the information and capabilities needed to meet these goals; 3) discuss solutions for improving relevant processes and tools.

We have created three broad user management goal categories: efficiency, data security, and accuracy in reporting to internal and external stakeholders. Examples of each of these goals are shown in the table below.

<table>
<thead>
<tr>
<th>Goal</th>
<th>User management task/info</th>
<th>Reason</th>
<th>Issue</th>
<th>Solution</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Security</td>
<td>Identifying accounts with expired password</td>
<td>Unchanging password increases risk of unauthorized system access</td>
<td>Manually tracking and enforcing password change is not possible</td>
<td>Automate reporting of expired passwords and email to user; lock out if unchanged</td>
<td>Increased data security</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Identify inactive accounts</td>
<td>Opportunity to contact user to determine whether there is a problem that can be addressed; inactive accounts should be closed for security reasons</td>
<td>Manually tracking account activity is not possible with a large user base</td>
<td>Automate reporting of last user activity</td>
<td>Identification of problems; removal of unnecessary accounts</td>
</tr>
<tr>
<td>Accuracy in Reporting</td>
<td>Maintain accurate data about users, projects, and use trends</td>
<td>Detailed reporting of activity is required for annual evaluation</td>
<td>Gathering required data is time consuming and in some cases, not possible</td>
<td>Automate reporting of data required for evaluation</td>
<td>Accurate and timely data for evaluation</td>
</tr>
</tbody>
</table>

Why Important to Present at 2010 IKFC Meeting
User management (e.g. control of access, tracking and maintaining activity) is an important component of any security-critical application. If user management capabilities are not incorporated in the application, attempting to conduct these tasks manually is time-consuming and error-prone. Although requirements for key user management tasks may be similar across all CTSA, how these tasks are handled can vary widely. We hope to raise this issue at the IKFC meeting to initiate sharing of issues and ideas for solutions among various Informatics groups. The desired result would be to catalogue a superset of requirements and a model for implementing them.

Impact on CTSA Consortium Strategic Goals
Goal #1 Enhancing national clinical and translational research capability: Making informatics application user management more efficient by automating tasks and information gathering will free up the administrator’s time to address user issues in a timely manner so that investigators’ work will not be held up by software issues. In addition, sharing ideas for user management requirements among CTSA Informatics groups will enable them to solve the problem once, instead of each organization solving the problem independently. Finally, automating reporting of user statistics will improve the quality of data reported for annual evaluations, while decreasing the time and effort to collect the data. Consistency of reporting across CTSA may also increase, making comparisons more meaningful.
Synergies in the Strategic Health IT Research Project (SHARP) Area 4: Secondary Use of EHR Data and the Health IT Pilot Communities through Recovery Act Beacon Community Program

Christopher G. Chute, MD, DrPH; __________________________ Lacey A. Hart, MBA, PMP
Mayo Clinic, Rochester, MN
Contact email: hart.lacey@mayo.edu

Abstract.

Mayo Clinic, long a leader in the science of health care delivery, is proud to be a recipient of both the Area 4 Strategic Health IT Advanced Research Project award and the Beacon award in collaboration with Southeast Minnesota. Both programs – part of the Office of the National Coordinator for Health Information Technology, are focused on improving quality, safety and efficiency of health care through Information Technology.

Traditionally, a patient’s medical information, such as medical history, exam data, hospital visits and physician notes, are stored inconsistently and in multiple locations, both electronically and non-electronically. The SHARP Program will collaborate to create, evaluate, and refine informatics artifacts that advance the capacity to efficiently leverage EHR data to improve care, generate new knowledge, and address population needs. The goal is to make these artifacts available to the community of secondary EHR data users, manifest as open-source tools, services, and scalable software. In addition, we have partnered with industry developers who can make these resources available with commercial deployment.

A natural synergy for SHARP tool application surfaces in the Southeast Minnesota Beacon community-based program to spotlight a variety of “best practice.” The Southeast Minnesota Beacon Community consists of eleven counties, their public health departments, many health care providers, and school districts working towards creating a unified exchange of information to:

- Improve quality, safety, efficiency, and reduce health disparities
- Engage patients and their families in their health care
- Improve health care coordination
- Improve public health and the health of the community’s population
- Ensure privacy and security protections for personal health information

Specifically, we will apply the SHARP services in medical centers and population-based settings, as well as, examine “best practice” in real-world interventions in the Southeast Minnesota Beacon that advance Clinical and Translational Science Awards (CTSAs), health information exchanges (HIEs), and National Health Information Network (NHIN) connections.
Objective: The objective of this group is to promote information sharing on health data standards and interoperability activities with bearing on or participation by the IKFC membership.

Background: With increasing attention being paid to both research IT as well as issues of interoperability addressed by data standards by the ONC and other agencies at the Federal level, it is critical to CTSA informatics organizations to have a forum or collaborating on standards and standards development activities. Inasmuch as standards are receiving attention at a national level, clinical research has not yet emerged as a focus of national health data standards policies. Therefore it is highly advantageous for CTSA organizations to develop standards awareness and competencies as a component of informatics collaborations.

Project Aims: The DIAG activities aim to provide information regarding standards development organizations or standards-involved projects. This includes providing a forum for the sharing experiences of group membership in standards development and implementation activities; participation in interoperability profile development and demonstrations; dissemination of information regarding standards-concerned grants awarded, and educational sessions pertaining to standards and policy documents with calls for public comment and access to consortium expertise to critique, collaborate, review and generally enhance consortium informatics projects.

Accomplishments: Membership Assessment, Education sessions on standards in comment period, spawned Permissions Ontology DIAG Subgroup (PODS) which is reviewing an ontology to support a South Carolina “GO grant” and is positioned for use in other Federal health programs.

Participating CTSAs: Albert Einstein College of Medicine, Case Western Reserve University, Duke University, Emory University, Georgia Institute of Technology, Harvard University, Indiana University School of Medicine, Johns Hopkins University, Mayo Clinic, Medical University of South Carolina, Morehouse School of Medicine, New York University School of Medicine, Northwestern University, Oregon Health & Science University, Stanford University, The University of Alabama at Birmingham, University of Arkansas for Medical Sciences, University of California, Davis, University of California, San Francisco, University of Cincinnati, University of Colorado Denver, University of Michigan at Ann Arbor, University of North Carolina, University of Pittsburgh, University of Texas at Houston, University of Texas Health Science Center at San Antonio, University of Texas Health Science Center at Houston, University of Texas Medical Branch, University of Texas Southwestern Medical Center at Dallas, University of Utah, University of Washington, Vanderbilt University

Why Important To Present At 2010 CTSA All Hands Meeting: As the DIAG was only just launched in the spring of this year, we feel it is important to provide information regarding the group and its activities to all participants at the IKFC all hands meeting, most importantly members of the newly awarded CTSA organizations. We would also benefit from the opportunity to receive feedback regarding the group and its activities thus far as a means to improve its utility to the consortia membership.

Impact on CTSA Consortium Strategic Goals

CSG #1 Enhancing national clinical and translational research capability: Developing pragmatic competence in the selection and management of content with the goal of achieving standards awareness or compliance supports the research infrastructure by supporting interoperability. Development of best practices around appropriate standards implementation supports adoption of proven research infrastructures.

CSG #3 Enhancing consortium-wide collaborations: Standards application directly impacts data sharing across member CTSA organizations. Wide-spread standards adoption can lower the barrier to participation for new CTSA organizations.

CSG #4: Enhancing the Health of Our Communities and the Nation: Application of standards within CTSA projects can appropriately inform standards development and policies in national and international public health.
Although the adoption of health information technology (HIT) has been identified as critical for improving the nation’s health care, the invaluable clinical data gathered by HIT, which could spur research to an unprecedented degree, has often gone untapped. Translation of mounds of data into meaningful outcomes measures is extremely difficult. In order to harness this valuable enterprise asset, Duke University Health System has developed a suite of tools that interacts with the organizational data warehouse to provide multidimensional research access to nearly 15 years of clinical information for over 3.8 million Duke Medicine patients. Our multifaceted research portal, the Duke Enterprise Data Unified Content Explorer (DEDUCE), is a user-friendly, visual data extraction system used to support grant applications, research projects, and quality improvement activities. Developed in 2008 with support from Duke’s CTSA award, DEDUCE currently has over 500 users and enables cohort generation and data extraction based on a wide variety of clinical parameters. Two distinct environments cater to the differing technological savvy of our diverse user base without requiring any knowledge of SQL or database concepts. A simple Guided Query Tool provides a window into the data warehouse using a wizard-like web interface that filters within up to six subject domains of patient care. Users can export both aggregate reports of counts and detailed extracts that may include protected health information (PHI) in accordance with IRB approval. The more sophisticated Cohort Manager environment allows one to span multiple clinical subject areas by constructing advanced queries using a wide variety of set operations, logical operators, and filters. Users may search all Duke Medicine patients or upload their own patient list. DEDUCE has been particularly useful in identifying patients for a large study to evaluate cardiovascular outcomes after treatment with sitagliptin in patients with type 2 diabetes mellitus. This DEDUCE query was built interactively by an independent clinician to consider 6 inclusion and 12 exclusion criteria to return 3,144 patients (as well as extracts representing their associated 31,000 encounters and 51,000 diagnoses). Once a specific patient cohort is found, the user may browse additional information available in these individuals’ health records by way of electronic chart review. This feature is made possible by integration with the Clinical Context Object Workgroup protocol (CCOW), an HL7 standard that allows applications from disparate systems having distinct viewers to present synchronized information on a patient in real time. From the list of cohort MRNs, CCOW-enabled hyperlinks allow investigators to in one click simultaneously view a patient’s electronic chart notes, radiology results, nursing notes, and ECGs.

In order to round out our vision of enhancing the efficiency and efficacy of Duke Medicine research, we implemented the Duke Integrated Subject Cohorting and Enrollment Research Network (DISCERN) as a novel HIT tool to improve the clinical trial recruitment process. While DEDUCE offers access to an extensive repository of retrospective data, DISCERN uses the open-source MIRTH engine to allow users to combine these data with prospective, real-time HL7 data streams and reason over timely information when attempting to characterize an optimal subject cohort. DISCERN’s capabilities include the automatic notification of study personnel when “trigger” conditions identifying a potential recruit are met, such as scheduling of an appointment, a rise in a lab value, or the presence another set of eligibility parameters. DISCERN enhances clinical trial workflow by automating tasks that would otherwise require labor-intensive activities, such as manual chart review—a feature that can substantially improve the efficiency of recruitment efforts. Integration of DISCERN into the DUHS data warehouse infrastructure allows centralized management of both retrospective and prospective data, potentially yielding economies of scale and freeing the study investigators from important burdens related to data management tasks.

The description of our HIT strategy to enhance clinical research at Duke Medicine would be of great interest to IKFC participants as a case study of how one large organization sought to leverage and creatively enhance existing, local resources to work towards nationally-important goals in translational medicine and clinical trial recruitment. In demonstrating the technology, this presentation also describes how our clinicians, who are commonly patient-oriented in their thinking, grow to harness these tools in an increasingly electronic era. In alignment with the CTSA goal to build clinical and translational research capacity, our applications provide the substrate for data driven exploration of clinical repositories, which when mined and understood, can spur meaningful measurement of patient safety, quality, regulatory compliance, and financial risk.
openSESAME: A new tool for discovering connections between biological phenotypes using public gene expression data

Adam C. Gower¹, Benjamin Bassey², Charles Berg³, Avrum Spira¹,⁴,⁵, Marc E. Lenburg¹,⁴,⁵
¹Bioinformatics Program, Boston University
²Medical Information Systems Unit, Boston University Medical Campus
³Translational Bioinformatics Program, CTSI, Boston University Medical Campus
⁴Section of Computational Biomedicine, Department of Medicine, Boston University Medical Campus
⁵BU-BRIDGE Translational Bioinformatics Core, Boston University Medical Campus

The wide availability and decreasing cost of gene-expression microarrays have spurred the creation of public repositories of genome-wide profiles of gene expression in a variety of biological and experimental perturbations. The largest of these is the Gene Expression Omnibus (GEO), which contains tens of thousands of experiments encompassing hundreds of thousands of microarrays. Several methods have been developed to search such compendia for experiments related to a given gene expression profile. To date, these approaches have used prior knowledge about the phenotypic variables in the deposited data and have relied on comparisons between pre-determined groups. To address this issue, we have developed a method, openSESAME (open-ended Search for Expression Signatures Across Many Experiments), to find sets of microarrays in which the genes in a query signature show a similar pattern of coordinate differential expression as that defined in the query without any prior knowledge of sample annotation or experimental design.

As proof of principle, we used openSESAME to query 45,000 human Affymetrix microarrays from GEO with several gene expression profiles. We first used two previously reported signatures of well-understood biological perturbations. When queried with a signature of estradiol treatment, openSESAME identified experiments involving treatment with estrogens and antiestrogens. When queried with a signature of silencing of the transcription factor p63 (a key regulator of epidermal differentiation), openSESAME identified datasets comparing stratified squamous epithelium to other tissues and comparing primary and metastatic melanoma. We then used openSESAME to perform a query with a profile obtained by comparing the cytologically normal airway of smokers with and without lung cancer, and found that this profile was similarly coordinately differentially expressed between normal tissue and tumor in bladder and head and neck cancers. The strong correlation between the incidence of these cancers and tobacco use suggests that openSESAME has discovered that common tobacco-sensitive pathways may contribute to carcinogenesis in these organs.

In conjunction with the CTSA, we are also completing the development of a web site that allows the scientific community to perform openSESAME queries of our preprocessed subset of GEO using their own gene expression signatures. The site, which is designed to be simple to use and to allow the rapid submission of multiple queries, provides users with PDFs of gene expression heatmaps as well as tables detailing the statistical significance of each GEO experiment. The algorithm has also been optimized for speed, and a typical query takes less than one minute to perform. A feature that is in development is post-hoc analysis of annotation terms to identify phenotypic variables associated with differential expression of the gene expression signature query.

We believe that openSESAME has the potential to be useful for discovering novel relationships between biological phenotypes on the basis of common patterns of differential gene expression. Furthermore, by making this tool freely accessible to the scientific community, the CTSA will provide a resource to researchers who wish to leverage the huge volume of publicly available gene expression data to better understand the molecular mechanisms underlying both normal human biology and its dysregulation in disease.
Collaboration Tools and Resources at the UIC CCTS Biomedical Informatics Core

The Biomedical Informatics Core at the University of Illinois at Chicago’s Center for Clinical and Translational Science (CCTS) has focused efforts in developing collaboration tools and resources to support clinical and translational research capabilities and which ultimately will enhance consortium-wide collaborations.

UICollaboratory Research Profiles. UIC contracted with Collexis to invest in a robust online professional networking tool that will help UIC researchers to identify potential collaborators and scientific experts. Known as the UICollaboratory Research Profiles website, this networking tool includes all Health Science College faculty at UIC. Each faculty profile details their academic appointment and contact information, publications and grant funding information. Initial phase data sources include data from PubMed and NIH RePORTER. Future phases will consider expanding to include publications beyond PubMed, additional grant information and patent data. The strength of using Collexis to implement the UICollaboratory Research Profiles lies not only in the immediate campus needs that can be met, but also in the opportunity for collaboration and communication across the national network of CTSA sites that are also using Collexis. Collexis is currently in beta testing for a community feature that will allow users to explore research networks throughout all Collexis-affiliated organizations; this feature will greatly expand opportunities for professional networking and collaboration between CTSA sites, directly addressing the CTSA strategic goal of enhancing consortium-wide collaborations.

Research Electronic Data Capture (REDCap). In collaboration with our Design & Analysis Core (DAC) we have joined the national REDCap Consortium and are deploying the data collection and distribution program developed by Vanderbilt University. REDCap is available freely to members of the REDCap consortium and is a metadata-driven application that can be useful to help manage data for small and medium sized non-trial research projects. A unique addition and enhancement to the REDCap Consortium, we have new online training tools including user manuals and policies & procedures for REDCap use. As far as we know, we are the first university to create such training materials and plan to share these products at the national level for other organizations using REDCap.

Informatics Support Services and Education Resources. Aimed at raising awareness and knowledge about biomedical informatics resources that can support a broad range of research endeavors, we developed two key efforts. The first is a Biomedical Health Informatics Colloquium (BHIC), a monthly in-person forum for faculty to present current & ongoing research and to discuss work in development and potential collaborations. The second is an online course: Informatics for the Clinical Investigator. This 8-week online course provides an overview of bioinformatics and health informatics for clinical researchers. The inaugural course is slated for summer 2011 and will be easily accessible to those outside the university. By offering an informatics course to researchers outside of UIC, the BI Core is helping to enhance the training of clinical and translational scientists on a national level.

These collaboration tools and efforts are in the initial phases and have great potential to support goals of the UIC CCTS, and to contribute to the national CTSA goals.
Algorithmic Infrastructure and “Omic” Data Integration for the Analysis of Complex Phenotypes*

Mehmet Koyuturk, Sinan Erten, Salim A. Chowdhury, Rod K. Nibbe, and Mark R. Chance
Case Western Reserve University

An important challenge in the study of biological systems, as well as human diseases, is the characterization of complex phenotypes - that is, phenotypes that are based on a set of complex interactions between multiple genetic and environmental factors. In the post-genomic era, availability of various sources of “omic” data enables study of complex phenotypes from a systems perspective. These include genomic, transcriptomic, proteomic, and interactomic data. Genomic association studies provide significant information on the genetic bases of complex phenotypes, however so far have defined a limited set of phenotypes. Large scale monitoring of gene expression reveals underlying cellular mechanisms, however, is limited in capturing the abundance and activity of functional proteins. Protein expression data provides more reliable information on function, but with limited coverage. Protein-protein interactions (PPI) highlight functional relationships between proteins, but PPI data is highly noisy, incomplete, and static.

In this presentation, we will describe our project on integrating these useful, yet limited sources of biological data to gain insights on the molecular mechanisms of complex phenotypes – with clinical and translational applications. We will introduce several computational problems that stem from these applications, including (i) network-based prioritization of candidate disease genes, (ii) integration of protein and gene expression data to identify dysregulated sub-networks in cancer, and (iii) use of sub-network markers for cancer classification. Our recent results on various public and dedicated datasets show that such integrative approaches can provide significant insights into the systems biology of complex phenotypes. In particular, we will show that sub-network markers significantly outperform single-gene markers in predicting metastasis of colon cancer and network-based algorithms can be very effective in filtering out many false negatives in genome-wide association studies of various diseases.

The algorithms and software described here will be of broad interest and applicability among all CTSA sites in that they provide (i) general, (ii) statistically validated, and (iii) algorithmically transparent methods. These methods will enable researchers on all CTSA sites to use to the “omic” data generated in their labs to generate knowledge beyond what can be generated by standard statistical analyses and currently available pathway analysis software. In accordance with CTSA strategic goals, this project contributes to improvement and development of resources for moving discoveries (particularly markers) by providing novel algorithmic infrastructure for disease-oriented studies.

*This work was supported by National Institutes of Health Grant UL1-RR024989 from the National Center for Research Resources (Clinical and Translational Science Awards) to enhance T1 research.
A Federated Virtual Data Warehouse Platform for Translational Research at a Multi-institutional CTSA

Tahsin Kurc\textsuperscript{1,2}, Andrew Post\textsuperscript{1,2}, Tim Morris\textsuperscript{1,3}, Kristi Eckerson\textsuperscript{1,3}, Tony Pan\textsuperscript{3}, Ashish Sharma\textsuperscript{2}, Alfredo Tirado-Ramos\textsuperscript{1,2}, Circe Tsiu\textsuperscript{1,3,4}, Stephen Wheat\textsuperscript{9}, Marc Overcash\textsuperscript{1,3}, Doug Blough\textsuperscript{5}, Mustaque Ahamad\textsuperscript{5}, Apurva Mohan\textsuperscript{5}, Alexander Quarshie\textsuperscript{6}, Elizabeth Ofili\textsuperscript{6}, Arlene Chapman\textsuperscript{7}, Mary Jean Harrod\textsuperscript{5}, Alex Orso\textsuperscript{5}, Mijung Kim\textsuperscript{5}, Lance Waller\textsuperscript{8}, Joel Saltz\textsuperscript{1,2}

\textsuperscript{1}Atlanta Clinical and Translational Science Institute Biomedical Informatics Program, \textsuperscript{2}Emory Center for Comprehensive Informatics, \textsuperscript{3}Emory University Research and Health Sciences IT, \textsuperscript{4}Grady Health System, \textsuperscript{5}College of Computing, Georgia Institute of Technology, \textsuperscript{6}Morehouse School of Medicine, \textsuperscript{7}Emory School of Medicine, \textsuperscript{8}Department of Biostatistics, Rollins School of Public Health, \textsuperscript{9}Emory Office of Information Technology

A wide range of clinical studies gather, reference, and analyze a diverse set of biomedical datasets potentially distributed across multiple sites. The Atlanta Clinical and Translational Science Institute (ACTSI), a multi-institutional CTSA led by Emory University, Morehouse School of Medicine, and Georgia Institute of Technology, supports such studies in collaborating institutions. These studies face technical challenges in 1) discovering relevant data and analytical resources and 2) accessing datasets across institutional boundaries managed in a variety of systems and data formats. The ACTSI Biomedical Informatics Program (BIP) is developing a virtual ACTSI-wide data warehouse platform that will support management, discovery, federated query, and integration of distributed datasets from multiple patient populations and studies. The ACTSI informatics platform draws from existing standards, best practices, tools, and middleware infrastructures from the Web, Web Services, Semantic Web, and Grid Computing communities.

This federated ACTSI data warehouse platform will enable investigators to see a big data warehouse that would contain clinical, omics, radiology imaging, ECG, virtual pathology slide data, information on available biospecimens, and laboratory data. It will host a number of application services, including i2b2 and the Emory Analytical Information Warehouse (AIW) system for clinical information, a Research PACS for radiology images and image annotations, a suite of open source and vendor laboratory information systems for tissue specimens, and a collection of clinical registries, for research teams and shared resources to use as dedicated databases for storing the results and data they generate. The platform will allow queries searching and joining information from multiple databases and data types such as “Give me all gene expression, SNP, and metabolomic data for patients with at least one systolic BP reading > 150 accrued at any Clinical Interaction Network site in 2012” and “Give me all radiology images and virtual slide patient and donor biopsy images for kidney transplant patients whose grafts survived for more than 5 years”.

The platform is designed to provide several architecture capabilities to support a federated environment: 1) management of common data elements and controlled terminologies; 2) security infrastructure for distributed identity management and authentication, XACML based access control policies, and attribute-based authorization; 3) an identifiers infrastructure supporting HL7 OID-based identifiers; 4) federated query and data aggregation across multiple data sources; 5) a testing environment for testing individual data sources and database integrity across the federated system; and 6) semantic resource discovery drawing from the Emory eBIRT system.

Acknowledgement. The project is funded with Federal funds from the National Cancer Institute, National Institutes of Health under Contract No. HHSN261200800001E, 94995NBS23, and 85983CBS43; NIH PHS Grant (UL1 RR025008, KL2 RR025009 or TL1 RR025010) from the CTSA program of NCRR; NHLBI R24 HL085343; NIH U54 CA113001; NLM R01LM009239-01A1, and BISTI P20 EB000591.
Maximizing Data Value From Multiple Clinical Trials
Bernard A. LaSalle\textsuperscript{1}, Richard L. Bradshaw\textsuperscript{2}, Susan A. Matney\textsuperscript{2}

Although Spinal Muscular Atrophy (SMA) occurs in approximately 1 in 6000 births, it is difficult to conduct large-scale clinical trials or multiple trials because of the severity of the symptoms, the age of the target population (mostly young children) and variable expression of the phenotype. Traditional data management practices do not enhance the value of data beyond the conduct of individual studies. Data quality is not the issue but data context is.

Metadata (data describing data) and controlled vocabularies applied to SMA clinical trials can significantly enhance the value of the collected data by increasing access through standardized data structures and provide an extensible knowledge base by organizing information for subsequent retrievals.

The objective of the Center for Clinical and Translational Science (CCTS) SMA Metadata Access Project (MAP) is to create information about the database structures of multiple ProjectCure studies and map demographic and key biomarker data elements to standard ontologies using controlled vocabularies. Once completed, SMA-MAP will provide a web based query tool to retrieve data across multiple ProjectCure research studies. These retrievals can provide a knowledge base for both clinical care and clinical research.

\textsuperscript{1} University of Utah Center for Clinical and Translational Science
\textsuperscript{2} University of Utah Biomedical Informatics Department
Creating A Research Subject Registry That Enables Business Process And System Integration

A centralized and secured institution-wide research subject registry is highly desirable to support effective protocol execution and administration, as well as provide the mechanism and data infrastructure to facilitate seamless integration between research systems. It also makes it possible for study teams to eliminate the need to create and maintain their own study subject registries which, often times do not meet security and confidentiality requirements.

Mayo Clinic is developing a protocol and subject registry based on a sound data model that is aligned with national standards (CDISC/BRIDG), and the functional services layer of the caBIG/C3PR product. Architected application program interfaces (API) implemented in the form of web services will be provided to export and import subject registry data to and from study electronic data capture tools (EDC), as well as other institutional systems.

The key functionalities that will be provided by the first phase release (March, 2011) of this research subject registry are:

- standardized online informed consent and re-consent process
- monitoring the progress of study accrual, enrollment and minority inclusion
- tracking current and historical subject status throughout the study lifecycle
- standardized reason codes where applicable and appropriate (such as reasons for refused consent, withdrawing, screen failures, etc.)
- online consent status verification by clinical staff delivering research intervention (eg. CT scan, muscle biopsy, etc.)
- enforcing regulatory and compliance policies
- facilitating and improving accuracy of research billing
- study specific and cross study reporting, as well as business intelligence and data mining

This project aligned closely with the CTSA Strategic Goal 1 to improve the processes related to the development, approval, activation, enrollment and completion of clinical trials.

The Mayo Research Subject Registry is built on Java/J2EE technologies based on open source methodology, service oriented architecture, and is semantically grounded in BRIDG model. Points of integration between business systems are implemented in well defined web service layer which offers ease of adoption by other institutions. The project is anchored on well defined business requirements, workflow, conceptual and physical data models; all of which can be easily shared and leveraged by the CTSA consortium.

We believe this project addresses many common challenges faced by all CTSA consortium sites. A podium presentation and discussion will provide tremendous benefits for the IKFC meeting attendees.
Systems biology analyses of gene expression and genome wide association study data in obstructive sleep apnea

Yu Liu1, Sanjay R. Patel2, Rod K. Nibbe1, Sean Maxwell1, Salim A. Chowdhury3, Mehmet Koyuturk3, Xiaofeng Zhu4, Naresh Punjabi6, Sarah Buxbaum7, Sina Gharib8, Emma K. Larkin9, Susan Redline2, Mark R. Chance1,5*

1 Center for Proteomics & Bioinformatics, 2 Center for Clinical Investigation, 3 Department of Electrical Engineering & Computer Science, 4 Department of Epidemiology and Biostatistics, 5 Department of Genetics, Case Western Reserve University, Cleveland, Ohio, 6 Johns Hopkins University, School of Medicine, Baltimore, MD, 7 Jackson State University, Jackson, MS, 8 Center for Lung Biology, Division of Pulmonary and Critical Care Medicine, University of Washington, Seattle, WA, 9 Division of Allergy, Pulmonary and Critical Care, Vanderbilt University Medical Center, 1215 21st Ave S., Nashville, Tennessee * Corresponding author

Abstract Obstructive sleep apnea (OSA) is a complex disorder caused by a collapse of the airway during sleep, consequently breathing is interrupted. OSA is the major cause of chronic sleep deprivation and excessive daytime sleepiness. It is estimated that up to 5% of adults in Western countries are likely to have OSA syndrome. After decades of research the molecular mechanism of OSA remains unclear, however recent research indicates that several interconnected aberrant pathways and molecular abnormalities are likely contributors to OSA. The purpose of this project is to identify the genes and pathways associated with OSA, which can help to expand our understanding of the risk factors for the disease as well as provide new avenues for potential treatment. Towards these goals, we have integrated relevant high dimensional data from various sources to define sub-network elements that connect some of the known pathways related to the disease as well as define novel regulatory modules. Two distinct approaches are applied. In the first case, we used a biased approach based on sixty genes/proteins with known associations with sleep disorders and/or metabolic disease to seed a search using a commercial software (Ingenuity Pathway Analysis, IPA) to discover networks associated with disease followed by mutual information scoring of the sub-networks using gene expression data. In the second case, we used an unbiased approach and generated an interactome constructed from publicly available gene expression profiles and protein-protein interaction (PPI) databases, followed by novel scoring of the network with GWAS data derived from OSA patients to uncover sub-networks significant for the disease phenotype. A comparison of the approaches reveals numerous proteins that have been previously known to be associated with OSA or sleep. In addition, our results indicate a novel association of Phosphoinositide 3-kinase (PI3K) and its related pathways, and STAT protein family with OSA disorder.

Our systems approach to study OSA shows the method's power to identify both precedent and novel molecular mechanism of diseases. High-throughput genome scale technologies provide a rich source of quantitative biological information about diseases. Systems approach that integrate multiple types of molecular data with phenotypic observations can be used to improve our understanding of complex diseases (like OSA), thus providing new strategies for diagnosis and therapy. All the software tools used in this project are publicly available (e.g. including commercial products, like IPA, or free software, like Cytoscape). Thus, other investigators can easily adopt these approaches and disseminating these methods through the venue of the CTSA IKFC annual meeting will encourage others to apply similar approach to other diseases.

This project was made possible only through collaboration among researchers from multiple fields (bioinformatics, statistics, biology and clinical studies). The development of methodology for systems approached requires data from multiple sources and its integration is challenging in many cases. Thus, the collaboration is a key to success for projects that study complex diseases. The developed tools are also valuable for enhancing translation and leveraging –omics data for all complex diseases and are thus of broad potential interest and applicability to all CTSAs. In addition, the tools bridge the interests of the IKFC and the Translational Technologies KFC and will stimulate cross-core CTSA collaboration.

This research was supported by the National Institutes of Health Grant UL1-RR024989 (PI: Davis, Pamela) from the National Center for Research Resources (Clinical and Translational Science Awards).
Institutional Architecture for Secondary Data Use

Aenoch Lynn, Howard Shang, Swati Chakraborty, Brian McCourt, Meredith Nahm, Jeffery Ferranti, James Tcheng, and Jessica Tenenbaum

Secondary use of data has immense potential to facilitate clinical and translational research through answering new research questions, through cohort identification, and through facilitating data collection to name a few. The Duke Data Repository (DDR) is the integrated data repository at Duke University, combining clinical research data, clinical care data, and results from -omics assays in support of secondary data use.

The DDR is an institutional effort drawing together groups from across our institution (Duke Comprehensive Cancer Center, Duke Center for Human Genetics, Duke Institute for Genome Sciences and Policy, Duke Clinical Research Institute, Duke Biobank, Duke Health Technology Solutions). Our CTSA, the Duke Translational Medicine Institute (DTMI) is spearheading this effort with the goals of lowering the barriers to secondary use of clinical research data and of enabling cohort identification using both clinical care and clinical research data. We present key components of existing and planned components of the DDR, including integration with the Duke Biobank, an institutional-wide subject consent database, an institutional ontology or terminology server, and a metadata driven data transformation engine. The foundation of the DDR effort is an institutional process for annotating and managing metadata about studies, and about data collected and used in care and research. Although we are in the early stages, with a pilot integration of four studies complete, learning has been significant and we present our findings to date.
Lessons Learned from Implementing Harvard Catalyst SHRINE which Inform the Building of a Network across Clinical Translational Science Centers to enable Massive Scale Research using Routine Clinical Data.

We present lessons learned from implementing Harvard Catalyst SHRINE as guidance towards implementing a research network across CTSCs for sharing of routine clinical data for research purposes. This discussion is informed by our experience implementing SHRINE at Harvard’s Beth Israel Deaconess Medical Center (BIDMC), Children’s Hospital Boston (CHB), Partners HealthCare System (PHS), and Dana-Farber Cancer Institute (DFCI). The Harvard implementation includes approximately 5 million patients, over 1 billion facts and over 17,000 query terms for demographics, diagnosis, medications and laboratory results. This implementation required approval from four distinct IRBs and the necessary controls for each participating institution to agree to provide access to all of their routine clinical data.

We will address the topics of implementation of informatics systems, establishing secure network connectivity, mapping of query terminology to local vocabularies and various controls that can be implemented in order to improve local approval.

The Harvard SHRINE implementation is based on Informatics for Integrating Biology & the Bedside (i2b2) and The Shared Pathology Information Network (SPIN). The Harvard implementation is a peer to peer network which allows data to stay under the control of each institution. The SHRINE query client communicates, using the SPIN protocol, with SHRINE adapters at each site in order to execute queries against the local clinical data in i2b2.
Cancer is projected to become the leading cause of death worldwide in the year 2010.

With the sequencing of the human genome and availability of high power computational methods and various high throughput technologies, cancer research and care are poised to undergo revolutionary change. These new technologies and approaches have spawned the field of systems biology; the new field of **systems medicine** is the application of systems biology approaches to biomedical problems at the bedside. In medicine, complex computational tools are becoming essential for deriving personalized assessments of disease risk and management including individualized diagnosis, prognosis, and treatment options. This change, involving the use and analysis of ~petabytes of data, will require new types of tools and a new type of physician - one with a grasp of modern computational sciences, “omics” technologies (e.g. genomics, transcriptomics and metabolomics), and a systems approach to medicine. Using 2 story boards in the fields of breast and GI cancers, we are demonstrating the use and extension of open source and commercial tools for collection and integration of clinical and research data that enables what we call the systems medicine based clinical practice.

**Best Practices and System**

This systems, named as G-DOC, Georgetown Database of Cancer uses many best practices and standards from the caBIG community. These include methods to integrate with other 3rd party tools (open and commercial) to solve key data integration issues that we face at the medical center. Novel technologies surface on a day to day basis in the scientific research realm. A key area where standards and methods need further development is metabolomics. Metabolomics provides important tools for disease diagnosis and prognosis based on body fluids and could become a cost effective method for these assessments compared to other tissue-based analysis. We are developing methods for analyzing and integrating metabolomics with other data types in biomedical research. This system will be available for public access during the end of 2010.

**Why is this important to be presented at the IKFC meeting**

As a new CTSA organization, we would like to present our capabilities, resources and design ideas to the broader CTSA informatics community and obtain feedback and lessons learned from other similar efforts. Also, our goals at the local CTSA are aligned with the CTSA strategic goals around data sharing infrastructure. At our local Georgetown-Howard University CTSA, we will adopt and adapt this G-DOC platform for non-cancer projects, specifically in the area of endocrinology.
Relevance and Summary of Conclusions One of the objectives (Theme 1) of CTSA is to support multi-disciplinary collaborations (Theme 3) that can transform translational research (Themes 2) and facilitate novel scientific initiatives (Theme 5). The present baseline study investigates temporal evolution (2000-2009) of research collaborations across multiple scales (Staff, Department and College) at the University of Arkansas for Medical Sciences (UAMS) using social network analysis. The distribution of the centrality measures were found to be positively-skewed with dominant collaborations persisting across multiple scales and time justifying their critical role in orchestrating collaborations. The fact that these collaborations were immune to perturbations (e.g. institutional policy changes, expansion, faculty recruitment, funding changes) may reflect a universal mechanism underlying collaboration networks and the broad potential of the proposed study. Identifying the dominant actors across multiple scales (Staff, Department, College) with a successful history of collaboration, grant funding and team-science (Themes, 1-3, 5) can especially be useful developing transformational scientific initiatives by targeted cross-fertilization between these actors. The dominant actors (Staff, Department and College) may also serve as effective dissemination points for various CTSA related activities. Removal of these dominant players will de-stabilize the collaboration network and adversely affect research productivity. The analyses were carried out using custom scripts in R open-source language and available upon request.

Materials and Methods A multiscale social network analysis of collaborations at the University of Arkansas for Medical Sciences was conducted based on the research portfolio retrieved from our internally-developed grants management application, the Automated Research Information Administrator (ARIA) over the past ten years (2000-2009). The attributes retrieved from ARIA were: (Grant Number (numeric), Awarding Agency (text), Staff Name (text), Staff Role (text), College (text), Department (text), Start Date (date), Stop Date (date), Total Cost (numeric)). Collaborations and their temporal evolution (2000-2009) were investigated across three hierarchically related scales, namely: (a) Staff (b) Department and (c) College (Staff ⊂ Department ⊂ College) using a battery of social network analysis.

Results The degree distributions (in-degree, out-degree) were found to be positively-skewed, indicative of only a handful of dominant actors in the social network with a significant number of collaborations. This behavior was persistent across scale (Staff, Department, College) and time (2000-2009). A possible generating mechanism of positively skewed distributions such as power-law is preferential attachment, where new actors have a tendency to collaborate with the dominant actors. Statistical justification of power-law can be challenging given the family of positively-skewed distributions. However, the dominant actors showed considerable persistence with time and scale. In order to investigate this in detail, the centralities of the actors in the network were ranked in descending order across each of the years. Since the number of actors (Staff, Department and College) may change considerably across a period of ten years, we chose to investigate the evolution of the upper quartile (i.e. top 25% of the actors ranked using a particular centrality measure in the year 2000) as a function of time (2001-2009). Staff out-degree centrality revealed that (~28%) of the principal investigators in the upper quartile at 2000 to persist across (2001-2009). This ~28% included researchers and clinicians with successful extramural grants that span several years. Analysis of the Staff in-degree centrality also revealed that (~19%) of the dependents/support staff to persist across the ten years. This included statistical, informatics and administrative personnel who are members of grants on a regular basis and indicated the critical role of support personnel and team-science for renewed research funding. A similar analysis of the Department out-degree revealed that a significant proportion (~64%) of the highly ranked departments based on out-degree (i.e. principal investigators) persisted across the ten years. A considerable overlap was observed between the Staff out-degree ranking and the Department out-degree ranking with a majority of the successful principal investigators belonging to Departments with a history of collaboration and successful funding. This need not necessarily be true since a Staff out-degree is a function of inter-departmental (between departments) as well as intra-departmental (within department) collaborations. This in turn justifies the need for multiscale analysis. The Department in-degree (~64%) also persisted with time (2000-2009). As with the out-degree, there was a clear overlap between the Staff and Department in-degree rankings with a majority of the Staff belonging to Departments that provide support services for grants. Similar behavior was observed in collaboration networks across Colleges.
A Permissions Ontology for the Integration of Electronic Research Authorizations, Consents and other Permissions Data

Jihad S Obeid MD1, Lisa J Macdonald BS MT(ASCP)2, Douglas Owan BSCS2, Steven E. Reynolds MSCS2, Matvey B Palchuk MD MS3, Davera Gabriel RN4, Iain C Sanderson FRCA4

1Medical University of South Carolina (MUSC), 2Science Applications International Corporation (SAIC), 3Recombinant Data Corp, 4University of California at Davis

Health Sciences South Carolina (HSSC), a collaborative of three principal research universities and four major health systems across the state of South Carolina, is developing a Research Permissions Management System (RPMS) that will provide a comprehensive mechanism for managing informed consents and privacy authorizations. This work is funded by a GO grant and is a collaborative effort between several academic institutions: HSSC, Medical University of South Carolina, Clemson University and Duke University with industry consultants: SAIC and Recombinant Data corp. An essential component in this effort is the development of a Research Permissions Ontology (RPO). This will enable the research permissions information from multiple institutions to be combined into a single computable data representation. It will provide the semantic foundation for representing and validating permissions data in a variety of data capture forms, relational databases and portlets that could be expressed via a web-based system or embedded in point-of-care clinical and research applications.

The initial version of the ontology was developed by SAIC with guidance from the Informatics team at MUSC. The work included analysis of the permission processes at four HSSC member medical institutions. The terminology and language in the various hospital forms, IRB templates, and hospital privacy practice notices were reviewed along with government websites, HHS, FDA and the Office for Civil Rights in order to ensure a more comprehensive analysis of HIPAA and other federal regulations. Moreover other bodies of standards work were examined with particular attention to permission and consent data standards where applicable including HITSP (Healthcare Information Technology Standards Panel), HL7 (Composite Privacy Consent Directive Domain Analysis Model) and the IHE Basic Patient Privacy Consents (BPPC). Other known ontologies such as SNOMED and NCI thesaurus were investigated to look for existing concepts related to permissions and consents that could be leveraged for this effort. For example many of the concepts in RPO are rooted in NCI thesaurus concepts such as document types and HIPAA authorization. After creating a list of potential terms and concepts that related to the research permission process, Protégé was used to layout the hierarchal order of the ontology (figure 1).

Many key concepts, synonyms and hierarchal “isa” relationships that were already defined in the NCI thesaurus, were imported directly into the Protégé ontology and added to the new concepts and classes that specifically address research permissions. Moreover some local HSSC institutional terms from the exploratory work mentioned above were mapped to concepts in the ontology with equivalent semantics.

The classification of permissions was also reviewed by experts in the regulatory domain on the Ethical, Legal, and Social Implications (ELSI) team from Duke on the RPMS grant. This work is also being validated by the newly formed subgroup of the Data Standards and Interoperability Affinity Group (DIAG), the Permissions Ontology DIAG Subgroup (PODS).

A crucial component of RPMS, the RPO will standardize collection, sharing and retrieval of research permissions across institutions, make permissions and consent assumptions explicit, and open the door for potential semantic reasoning. The RPO is designed with the intention of being open sourced and reused in other research projects in the CTSA community and beyond. Standardizing collection of research permissions will facilitate research by ensuring that patients’ intentions and inclusion/exclusion in research projects are met while ensuring privacy and confidentiality.
The Awesomely Large OSIM Dataset (ALODS)

Lead institution: UCSF

Objective: To support cross institutional development of queries and other data mining tools it is necessary to have a large synthesized dataset with known features and artifacts. This will accelerate development and publishing of test results based on repeatable and controlled data removing concerns of PHI exposure.

Background: Testing and evaluation of research tools such as I2B2 require that data results be consistent and reproducible. When using authentic patient datasets even when de-identified, results cannot be shared cross institutionally without IRB approval and concern of privacy. To overcome these concerns, UCSF is developing a synthetic large patient dataset containing more than five million patient records. This synthesized dataset will allow the publishing and distribution of data features including full patient demographics normally requiring de-identification and IRB control.

Project Aims: 1) define and create five million synthetically generated master patient records; 2) assign and distribute a randomly generated health information history for each patient ID; 3) link lab values, payment history, visits, and EMR data through synthetically generated patient ID; 4) distribute controlled datasets with a data dictionary to all interested intuitions.

Accomplishments: Using a variety of tools and manual processing techniques we have created a large pool of Master Patient Records containing 8 Million person ID’s. Using the FNIH’s Observational Medical Data Simulator (OSIM) we have generated over five million patient records.

Participating CTSAs: UCSF

Figure 1. The relationship between the pseudo master patient records and supporting datasets

WHY IMPORTANT TO PRESENT AT 2010 CTSA IKFC MEETING
A widely distributed and controlled dataset will allow results of query tools to be independently verified and shared rapidly without impact to PHI

IMPACT ON CTSA CONSORTIUM STRATEGIC GOALS
CSG #3 Enhancing consortium-wide collaborations: The ability to perform rapid ad hoc evaluation of research tools and publish results without regulatory approvals will enhance institutional cooperation.
The “5x5” Program: Brief Videos for Translational Research Informatics Education

Description of the Project:

One of the major aims of Informatics Core of the University of Colorado CTSA is to develop curricula to support the educational needs of trainees and faculty who wish to use bioinformatics and clinical research informatics tools, and supporting course masters who wish to incorporate actual bioinformatics tools and educational resources in their courses. We have developed a variety of “Translational Informatics Video Learning Resources” which are publicly accessible at http://cctsi.ucdenver.edu/RIIC/Pages/TranslationalInformaticsVideos.asp. Presentations fall into one of three categories: (1) videotaped presentations of up to an hour, (2) brief presentations of up to 20 minutes, and (3) “5x5” presentations of approximately 5 minutes long.

The “5x5” format has been especially well received. Translational research investigators who are seeking “just enough information” to meet an immediate need appreciate these very short, highly-focused, single topic learning modules. The “5x5s” cover a variety of topics, but follow a common format, explaining (1) fundamental or conceptual issues for the topic/technology, (2) why the topic/technology is important in translational research, (3) practical operational issues, (4) for whom and when the topic/technology applies, and (5) where to find more resources on the topic.

Currently we have posted seven “5x5s”, covering both bioinformatics topics (such as “mRNA expression technology” and “Epigenomics”) and translational research informatics topics (such as “trustworthy information sources for patients” and “health literacy.”) We are actively developing additional content, including 5x5s on key data management best practices, a comparison of Microsoft Access, Excel and REDcap and statistical and data quality issues.

After evaluating a variety of formats for developing and streaming these presentations, we found the Adobe Presenter platform hosted by our Educational Support Services to be ideal. Lecturers inside and outside the CTSA have developed content using the Adobe Presenter add-in for Microsoft Powerpoint and a microphone. These presentations are converted to Flash format files, which can be streamed smoothly to multiple simultaneous users using a dedicated Adobe Presenter server.

Why It Is Important to be Presented at the 2010 IKFC Meeting:

This structure for brief educational videos is well accepted by translational researchers and could easily be adopted by other CTSA sites. We are eager to share our content with other CTSAs, and to collaborate with other CTSA sites on developing a rich library of succinct videos on a variety of translational informatics topics.

Impact on the CTSA Consortium Strategic Goals

By providing a well-received method of providing translational informatics education, this relates to CTSA consortium strategic goal #2: “Provide training and improving the career development of clinical and translational scientists.” Collaborations to develop a rich library of content relate to strategic goal #3: “Enhance consortium-wide collaborations.”
High-throughput Microscopy Image Analysis for Deep Integrative In Silico Study of Brain Tumors

Joel Saltz¹, Fusheng Wang¹, David Gutman¹, Jun Kong¹, Lee Cooper¹, Tahsin Kurc¹, Sharath Cholleti¹, Carlos Moreno², Ashish Sharma¹, Tony Pan¹, Patrick Widener¹, Daniel Brat²

¹Center for Comprehensive Informatics, Emory University
²Department of Pathology and Laboratory Medicine, Emory University

Histological analyses play key roles in the characterization of brain tumor progression, classification, and response to treatment. The morphology of brain tumor nuclei, for example, is a pivotal attribute used to classify brain tumors; the classifications (e.g. astrocytoma, oligodendroglioma) are closely associated with patient survival, response to treatment, and are also closely linked to genetic and gene expression classifications. The analysis of even a moderate number of whole virtual microscopy slides, however, can lead to challenging problems. These problems stem from the fact that analysis pipelines consist of complex algorithms applied on large images and even a single analysis may generate large numbers of spatial structures (e.g., nuclei and angiogenesis regions), features (e.g., area, eccentricity), and classifications (tumor cells, endothelial cells, macrophages, etc). In an effort supported in part by the NCI cancer Biomedical Informatics Grid program and the Atlanta Clinical and Translational Science Institute, we are developing a microscopy image analysis protocol and a suite of supporting informatics tools for more effective use of computer-aided analyses on large volumes of image data in deep integrative translational studies. The parent project of this effort develops novel multi-scale, integrative in silico experiments for study of brain tumors that correlate information from complementary platforms and data types. The research effort leverages molecular, pathology, radiology, and clinical data obtained in The Cancer Genome Atlas, Rembrandt, and Vasari studies.

The main steps of the image analysis protocol includes running multiple analysis pipelines with different parameter values and algorithm variations (to account for and reduce analysis sensitivity to algorithm parameters, histological structures being examined, and tissue characteristics), managing results from multiple computer analyses and reviews by Neuropathologists, and comparing and evaluating results from multiple analyses. For analysis of image datasets, our tools partition large images into multiple tiles and distribute tiles and processing operations (implemented as Matlab scripts) across a computational cluster so that multiple tiles and operations can be executed concurrently. We have developed a data model, called PAIS, to manage and query spatial structures, features, and classifications generated from computer algorithms and human reviews as well as to capture limited provenance information on how the results were generated. The model also supports foreign key relationships via de-identified patient ids so that molecular analysis results, clinical data, and imaging results for the same patient or group of patients can be joined. For human markup and annotation of images, we employ a graphical user interface that captures image markups and annotations in XML documents. Both human reviews and computer analysis results are loaded to a relational database built on the PAIS model. Our implementation supports a variety of common metadata and spatial queries that 1) select and retrieve data based on image metadata, anatomic objects, object features, classifications, and provenance information; 2) select and compare segmented objects, features, and classification results from multiple analysis pipelines; and 3) compare summaries from multiple images, multiple pipelines, and multiple studies. These queries are useful for exploring subsets of analysis results for further analysis, mining, and integration with genomic and clinical data, as well as for algorithm comparison and validation. For example, a query on spatial objects that are classified by an expert observer but not detected by an algorithm can provide information about potential weaknesses of the algorithm.

Acknowledgement. The project is funded with Federal funds from the National Cancer Institute, National Institutes of Health under Contract No. HHSN261200800001E, 94995NBS23, and 85983CBS43; NIH PHS Grant (UL1 RR025008, KL2 RR025009 or TL1 RR025010) from the CTSA program of NCRR; NHLBI R24 HL085343; NIH U54 CA113001; NLM R01LM009239-01A1, and BISTI P20 EB00591.
A Research Permissions Management System at the Statewide Scale: Progress.
Iain C Sanderson FRCA, Anand Gramopadhye PhD, Kevin Weinfurt PhD, Jihad S Obeid MD, Jay Moskowitz PhD.
1Medical University of South Carolina, 2Clemson University, 3Duke University, 4University of South Carolina.

Health Sciences South Carolina (HSSC), a collaborative of three research universities and four major health systems across the state of South Carolina, is developing a Research Permissions Management System (RPMS) that will provide a comprehensive mechanism for managing informed consents and privacy authorizations. This work is funded by a GO grant and is a collaborative effort between several academic institutions: HSSC, Medical University of South Carolina, Clemson University and Duke University with industry consultants: Science Applications International Corporation (SAIC) and Recombinant Data corp.

Progress in this project has taken a number of tracks, consistent with its specific aims, namely: 1) an analysis of the current status of research permissions data collection across the HSSC hospitals, 2) the examination of the currently-used set of permissions, and those anticipated by advances in biomedical science, by an Ethical, Legal, and Social Issues (ELSI) team from Duke with lay community involvement that will propose recommendations in terminology and data collection process, 3) the creation of a Research Permissions Ontology (RPO) and its subsequent review by a CTSA- sponsored Affinity Group, 4) a clinical study with an IRB-approved protocol to examine the effect of touch-screen and other hardware and software design interfaces on usability, satisfaction and comprehension for general research permissions collection in a controlled clinic setting and 5) the development of software that would be influenced by all of the above and would provide a complete, open source, solution for the collection of research permissions, including a patient portal displaying an individual’s collection of permissions and the representation of research permissions as an attribute in an i2b2 data warehouse (Figure 1). Discoveries in the project to date, include the finding that the electronic documentation of research permissions at clinical institutions in South Carolina clearly lags behind other advances in Electronic Medical Records, and is nearly entirely manual and paper-based. This includes the collection of consent for clinical studies and other types of general permissions like authorizing the use of deidentified, discarded blood for biobanking, that are usually sought at patient registration. There are also significant opportunities for better subject education and for improving the semantic and ethical consistency of the text and user experience presented during this phase the research cycle. The ongoing ELSI process, combined with the results of the research study into user comprehension and interface design, and the use of an ontology, will lead to a best practice approach for research permissions collection and its reflection in the software application. The functional requirements for the software application with the architecture and scope illustrated in figure 1 have already been completed, and coding has started with a delivery date of Q4 2010. A pilot implementation is planned in the registration clinic areas of Palmetto Health, with full implementation across HSSC hospitals by the end of the calendar year. The collection of research permissions in the hospital environment will be supplemented by the ability of individuals to view and manage their permissions from a patient portal. The final manifestation of this feature will be guided by the ELSI process, but should lead to a significant advance in consumer choice and new educational opportunities for subjects to engage in the research process.
**Introduction:** Digital Vita (DV), an application developed by the University of Pittsburgh, integrates CV management functions with academic social networking and basic research team management and collaboration functions. With DV, researchers can: manage their complete academic CVs; output CV information in a variety of formats, such as online profiles and NIH biosketches; build their social network through collaborations on publications and grants automatically; group their colleagues into research teams of their choosing and request biosketches; fulfill biosketch requests with a few mouse clicks; and maintain and selectively update multiple versions of NIH biosketches. Digital Vita V.1.0, incorporating CV information management and output functions, has been implemented as a production application at the University of Pittsburgh since the fall of 2009.

**Objective:** We present the results of an initial formative evaluation of DV, the first step of a series of systematic evaluations, focused on the adoption and use of CV and profile management features in V.1.0.

**Methods:** From the outset, we instrumented DV to capture most user interactions with the system. We used these logs, as well as monthly snapshots of the database, to generate utilization data. Our analysis focused on key variables of system adoption and use, such as user accounts, data stored and documents created in DV.

**Preliminary results:** As of July 30, 2010, DV had 417 users, with between 20 and 50 new accounts being added each month. The 417 users accounted for 10,726 publications, 3,655 presentations, 1,636 grants, and 1,279 mentoring and 1,701 teaching activities. Users also created a total of 310 NIH biosketches. In addition, about 20% of all updates to data are made through automated and semi-automated functions. These findings show that DV is being adopted within the University of Pittsburgh Health Science schools and that the distinctive functionality present within the system is also being used.

**Conclusion:** DV is accomplishing more than simply replicating the capabilities of basic research directory. While still in its early stages, the user focused strategy behind the design of DV appears to have lead to features and capabilities that strengthen DV’s use as a basic research networking system (RNS).

**Why is it important to present this project at this meeting and its impact:**

DV speaks directly to one of the goals of Strategic Goal Committee 3, “Enhancing Consortium-Wide Collaborations.” This goal seeks to facilitate researcher networking across national institutions and across topic domains, creating a virtual community where collaborations across institutions can arise easily, where expertise is mapped and can be located easily, and where matchmaking between collaborators or others can expand beyond an institution’s walls as desired. A presentation at the 2010 IKFC meeting is important because DV is now adopted by over 20% of its target population, with a minimal investment in marketing and training. More information about design and functions of DV is available in the paper [http://www.jmir.org/2008/3/e24/](http://www.jmir.org/2008/3/e24/), as well as the project information site [http://di.dental.pitt.edu/orc/](http://di.dental.pitt.edu/orc/).
A new tool for researchers to access de-identified pathology information and paraffin embedded tissue
Girish Chavan, MS, Kevin Mitchell, MS, Karma Edwards and Rebecca Crowley, MD, MS

The TIES System is a CTSI and NCI funded project to make available highly annotated and de-identified clinical reports for use in biomedical research. The TIES system leverages natural language processing algorithms and query visualization methods, and provides secure, easy to use and highly accurate access to research data and associated tissue. It allows researchers from different institutions to collaborate on studies from within the system, build cross-institutional case sets and order related tissue. TIES uses role based authorization, an integrated honest broker framework, grid-based communication and security frameworks that are based on studies conducted on institutional IRB policies and federal regulatory requirements to ease acceptance and adoption by any institution.

The TIES deployment at the University of Pittsburgh provides researchers access to over 1.7 million surgical pathology reports across all UPMC hospitals over the last 15 years. The TIES system directly supports CTSA Consortium Strategic Goals 1 and 3 by providing a mature, HIPAA compliant, collaborative research system for institutions that is ready for deployment today. The IKFC attendees are the targeted end users of the TIES system and will greatly benefit in learning more about the benefits of TIES. The system is already deployed at a number of institutions, including some CTSA funded institutions, and is creating a potential network of TIES systems that could facilitate multi-center research projects across the country. It is important to present this system at the IKFC meeting to enlist support for the development of a national TIES network.
Theme 3: Projects, best practices, activities, or systems that seek to catalyze or support collaboration and team science

SPIRiT: Building the Infrastructure to Support Collaborative Science at Five CTSA Institutions

We present the data Sharing Partnership for Innovative Research in Translation (SPIRiT), a virtual consortium of CTSA institutions that will share informatics tools, regulatory best practices and research data to further CTSA Strategic Goal 3, Enhance Consortium-Wide Collaborations. The five institutions will participate in SPIRiT initially -- Johns Hopkins University, University of Pennsylvania, University of Pittsburgh, Washington University, and Yale University -- are already engaged in collaborations that involve data sharing. Over the next five years SPIRiT will create a robust infrastructure for sharing data about investigators, research tools, regulatory knowledge bases, and biospecimens. SPIRiT will demonstrate both the feasibility and the scientific value of sharing data and biospecimens among CTSA institutions.

SPIRiT will initially be organized around deployment and use of informatics software developed at the University of Pittsburgh. The Text Information Extraction System (TIES) will be deployed at all participating centers to provide federated query against de-identified pathology reports that describe millions of specimens in institutional paraffin archives. The Regulatory and Informatics Cores will develop Materials Transfer agreements and common workflows to establish a Collaborative Tissue Network, where those specimens will be exchanged to support two SPIRiT pilot studies. In addition, the Digital Vita application and the Resource Discovery System (RDS) developed at Pitt will be integrated and deployed to all sites to allow investigators to find collaborators and research tools at all SPIRiT sites. Digital Vita provides users with the ability to search for people with similar or complementary research interests, review detailed profiles, and establish groups of colleagues to exchange biosketches during preparation of grant proposals. RDS provides an intuitive user interface to a database of scientific tools and technologies that are used at CTSA sites.

To lay the foundation for SPIRiT, the Regulatory and Informatics Cores at all participating institutions have already contributed content to RDS, and the Digital Vita team will release version 2.0 of the software later this year. The PIs from all five sites have met together with the Informatics Core Directors to develop the plans for SPIRiT, and the Directors of the Regulatory Cores have also participated in kickoff teleconferences.

Within five years all of the SPIRiT software applications, along with best practices and document templates developed by the initial members, will be available to all members of CTSA for use in an expanded SPIRiT consortium or to establish collaborations beyond CTSA institutions.
Clinical and Research System Interoperability with Research Electronic Data Capture (REDCap)

Robert Schuff MS1, Paul Harris PhD2, Jeffrey Horn BS1, Rob Taylor2, Jon Scherdin2
1Oregon Health and Science University, 2Vanderbilt University

Research Electronic Data Capture (REDCap) is a suite of easy to use software tools providing clinical and translational researchers robust and secure web-based electronic data collection forms. REDCap is developed and maintained by Vanderbilt University and is currently in use by over 11,000 researchers 140 academic and nonprofit institutions world-wide with over 3,500 research project in either active or development status.

Although REDCap provides an extremely intuitive and facile environment for data collection and management, researchers often manually enter research-related data contained in local clinical systems into their research databases (e.g. lab results, demographics). This tedious process adds no value to the research project, wastes time, and can have a negative impact on data quality through transcription errors. While REDCap traditionally has partially addressed this issue with a manual import utility, this approach requires a specially formatted data file and manual process to initiate each data load. To more completely address this issue in a scalable way, a collaborative project between Oregon Health and Science University and Vanderbilt University to enable secure, automated, and flexible interoperability of REDCap with external systems such as enterprise data warehouses, lab systems and other systems containing data of interest to researchers was recently undertaken and has produced a new module called REDCap Data Transfer Services (DTS).

DTS is a process based on system-to-system data transfer that eliminates most of the human intervention required to move data from source systems into REDCap databases. Recognizing the diversity of computing environments across the REDCap Consortium, DTS incorporates key design features (e.g. layered architecture, plugins for dataset formats) that enable it to efficiently function in a wide range of environments while still providing a consistent and intuitive interface for informaticists who set up the “backend” DTS processes and a familiar and intuitive experience for researchers who consume the externally acquired data.

The DTS workflow starts with tasks performed by the local REDCap administrator who creates a “connector” which provides the means to establish a connection with a data source such as a text file or a SQL-based query result-set. Next, the informaticist describes the structure of the dataset by creating a “transfer definition”. Finally, the variables in the transfer definition are mapped to variables in one or more REDCap projects. Once these configurations are set, processes are invoked that acquire the data and hold it in a local cache for processing by a “recommender” which interrogates REDCap metadata and any existing data to develop recommendations for matching instances of data (e.g. HDL) to a given patient at a given timepoint. Finally, the data and recommendations are presented to the researcher for adjudication before being written to the REDCap project database. This “two-phase commit” process minimizes work for researchers yet still gives them full control over which data are transferred into their database.

DTS version 1.0 will be deployed at Oregon and Vanderbilt in September 2010 and is simultaneously being tested and deployed to select REDCap Consortium members and will be made generally available to all consortium members in October 2010.

In addition to the technical capabilities conferred to REDCap through DTS, a successful implementation establishing interoperability with REDCap requires policies and procedures to ensure regulatory compliance and patient privacy. Methods of data acquisition and handling that engender trust of stewards of enterprise data are also essential. Examples of these essential technical and non-technical “ingredients” have been developed at Oregon and Vanderbilt is leveraged as we work with Consortium members to deploy DTS.
Semantic Integration of AIDS-related Malignancy Data within the ACTSI Biomedical Informatics Program Framework

Alfredo Tirado-Ramos, PhD1,2, Tahsin Kurc, PhD1,2, Andrew Post, MD, PhD1,2, Jingjing Gao, PhD2, Alexandra Anghelescu2, Mary Jo Lechowicz, MD3, Christopher Flowers, MD3, Joel Saltz, MD, PhD1,2

1Atlanta Clinical and Translational Science Institute Biomedical Informatics Program, Atlanta, GA, 2Emory University Center for Comprehensive Informatics, Atlanta, GA, 3Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA

Clinical and Translational research on HIV-associated malignancies is made increasingly difficult by the separation of information on HIV/AIDS and cancer and the incremental expansion of each field as scientists discover new drugs and interventions. Recent data on cancer rates among HIV-positive patients stress the need for the development of new methodologies and drugs that restore immune function more effectively than currently available processes and treatments. While clinical trials and outcomes data on AIDS-defining Malignancies (ADM) and non-AIDS-defining Malignancies (nADM) have been produced in the last few years, there is still much work to do in researching tools and technologies for the integration of such information. The HIV-K project investigates technologies and tools for integration and semantic annotation of ADM and nADM malignancy data. HIV-K aggregates findings at a semantic level, using the Biomedical Informatics Program (BIP) as the framework to integrate state of the art resources from the Atlanta Clinical and Translational Science Institute (ACTSI) and a data of a community of researchers. Within the BIP, the HIV-K provides a translational use case for the search, identification, classification and semantic annotation of relevant datasets, modeling of the analyses and queries in identified studies, in order to increase the understanding of available clinical trial data and outcomes of ADM such as NonHodgkin lymphoma, as well as nADM such as anal cancer, Hodgkin lymphoma, or liver cancer.

One of the aims of this project is to leverage state of the art technologies and processes, such as the caBIG framework, while exploring possibilities of enhancement and reuse offered by related efforts in semantic data warehousing, like Semantic Extraction, Transformation and Loading (SETL), which could be of interest to other CTSA collaborators and partners. Our initial approach is to provide insights into the availability of clinical trial and outcomes data, and the applicability of semantic annotation and ontology development for such data. Our focus lays on immunologic, malignancy, pharmaceutical, pathologic, infection history, raw sequence, as well as general statistical and demographics data. To this end, we work at the outset with a lymphoma database called the Lymphoma Enterprise Architecture Data-system (LEAD), which integrates pathology, pharmacy, laboratory, cancer registry, clinical trials, and clinical data from institutional databases. This allows us to leverage previous work on data elements and structures within LEAD, used to manage clinical research data from phase 1 clinical trials, cohort studies, and registry data. Our approach is to have users first be able to perform preliminary queries following pre-set templates, and later to be able to perform full queries using such preliminary results of interest, with the added ability to pick and choose their parameters of interest freely. A workflow engine provides the templates for the first stage experiments and the process to request and perform the full queries. Future work will include research on technologies and tools for de-identification, potential incorporation of data from metagenomics and publicly-available knowledge repositories such as PubMed, provenance, as well as presentation of results in more advanced visualization and navigation environments. Finally, one interesting potential outcome of this project is the investigation of scientifically interesting statistical methods for modeling and analyzing HIV disease progression, evaluating the effect of treatments from the point of view of the semantic aggregation of available data.
Utilizing Open Source Clinical Research Tools for Behavioral Studies
Umit Topaloglu Ph.D., Samuel Jackson, William R. Hogan M.D., MS., Cheryl Lane, Zhidan Feng, Ildeniz Yayla, Laura F. Hutchins MD, Warren Bickel Ph.D., Catherine Stanger Ph.D.

University of Arkansas for Medical Sciences

Abstract
At the University of Arkansas for Medical Sciences, we have implemented an open source clinical research informatics infrastructure that addresses registration, calendar, survey, and clinical data management requirements. In addition to successful adoption of the systems for many biomedical research projects, the same toolset is used for studies conducted at the UAMS’ Psychiatric Research Institute. One such study is an adaptive randomized trial to improve adherence to diabetes management among adolescents with poorly controlled type I diabetes mellitus. The goal of this study is to adapt a previously developed, family-based, 14-week Contingency Management (CM) treatment to target a specific behavioral component of adherence to proper diabetes management, blood glucose monitoring.

System Design and Implementation
Our infrastructure consists of National Cancer Institute’s caBIG-based clinical trials management suite (CTMS) along with LimeSurvey, an open source assessment tool. Applications used in this study include caBIG Central Clinical Participant Registry (C3PR), OpenClinica, and LimeSurvey. C3PR is being utilized to manage participant and family registrations, including screening, inclusion/exclusion criteria, and randomization. Currently one family is registered to C3PR and accrual continues. OpenClinica is being utilized to manage electronic case-report forms (CRFs) and the creation of the study-specific CRFs was completed. The Clinical Research Coordinator on the study uses both applications to enter data. With LimeSurvey, we have developed electronic versions of pen-and-paper psychosocial assessments. Parents and teens complete these assessments at intake and closeout on LimeSurvey that saves staff time comparing previous double data entry. The assessments include: the Self Care Inventory – Revised Parent form, the Parental Monitoring of Management – Parent Self form, the Diabetes Family Conflict Scale – Parent form etc. Prior to the implementation, in addition to some paper based data collection, the institute was using a FileMaker database for the portion of the data collection and it was time consuming and required an FTE developer. Use of existing research infrastructure eliminates a FTE requirement.

Conclusion
Our existing, open-source informatics infrastructure for conducting cancer trials at UAMS appears flexible enough to accommodate the needs of behavioral studies ongoing at the PRI at UAMS. The results show that our UAMS CTMS is promising to standardize our research data collection efforts in compliance with national initiatives.

Importance to the CTSA IKFC Meeting
As behavioral research is an important program for the institutions within the CTSA, having an informatics solution is vital for these programs. We’d like to share our experiences so that other CTSA institutions may choose to invest in such an open source infrastructure.

Impact on CTSA Consortium Strategic Goals
This work addresses the mission of SGC1 to …improve all processes related to the … enrollment and completion of clinical trials.
The CCTR Network: Using Sharepoint for collaboration, management and evaluation of clinical and translational research

Alan VanBiervliet, PhD and Aaron Jundt, BA University of Arkansas for Medical Sciences

Problem: The newly funded Arkansas Center for Clinical and Translational Research (CCTR) needed to create ways to enable 15 key function areas to share information, ideas and collaborate, to manage the key function activities to support translational research, and to gather and store information that could be used for formative and summative evaluation.

Challenges and Restraints: Key components of the system must be operational within 60 days of start date; budget for software and staff extremely limited; system must be scalable to meet expanding needs, user training requirements must be minimal, and the system must be sustainable over a multi-year period.

Solution: Following an extensive review of the literature and discussions with professionals both within and outside the UAMS community, a decision was reached to use Microsoft Sharepoint for the basis of the CCTR Network. Factors that influenced the decision included: 1. Sharepoint is a robust, mature product that is capable of meeting all of the needs that were identified; 2. The technology is server based and web accessible thus facilitating information storage and protection, while providing a wide range of users easy access; 3. Permission to store information and rights to modify components is granular to the individual object level; 4. Prebuilt functionalities are robust and additional feature can be easily created; 5. Sharepoint is well supported by Microsoft, user groups and third party vendors; 6. Sharepoint is a key product in the Microsoft enterprise solutions line; 4. UAMS had been providing faculty and staff access to Sharepoint for many years; 5. UAMS IT had a track record of successfully providing the backend support to the product; 6. Single sign on to the UAMS Sharepoint environment from any web site provides easy access for all users; and 7. Interviews of UAMS users demonstrated they were able to effectively use Sharepoint for several important campus functions. However, the interviews also indicated that support for customization of Sharepoint sites was a serious need.

Development Planning and Implementation: The project was planned using the key principles of user-centered design and evidence based design (Usability.gov). To facilitate the adoption of the CCTR Network approaches encompassed in the diffusion of innovation theory were used, including involving champions and users in design, and selecting high profile initial projects. CCTR Network plans and governance procedures were approved by the CCTR executive board. Within 60 days the basic structure of the CCTR Network was available for use including mechanisms for electronic submission and review of pilot grant proposals. A Protocol Review Dashboard was created to manage multitask database views from a single location. Dashboards provide custom hyperlinks and automated workflows. Online data entry forms have been customized to enable larger data entry fields, form based instructions, and custom formats. These forms have been used for Core technology reviews, pilot grant reviews, and monthly for SAC protocol reviews. Hundreds of documents have been uploaded into the central database and are in use by many CCTR work groups. The Network includes tools for use by all members of the CCTR as well as all UAMS staff and students. A three tiered architecture (everyone, team stakeholders, and project sites) has been created for each key function area. These are the everyone site available to all CCTR members and staff as well as UAMS community, a Team site with collaboration tools for the use among the team stakeholders, and project or work group sites with project management and collaboration tools focus on achieving specific functions. Several tools have been created to facilitate the workflow for managing and using information. An online process to grant access to the CCTR Network for external community members has been created. A major advantage of the CCTR Network is that using the same tools users can perform many of the tasks that are typically managed with multiple tools and different user interfaces. The CCTR Network is also being used for individual research project planning and management.

The Web Strategies plan for the CCTR includes the CCTR Network and a web portal with features similar to the other funded CTSA organizations. Sharepoint is widely used in medical research settings and the lessons learned and innovations developed by the CCTR can be readily adopted at other sites.
Providing Flexible and Integrated Informatics Support to Investigators and Trainees to Improve Clinical & Translational Research

Dongwen Wang, PhD

Introduction. The University of Rochester (UR) was one of the first batch CTSA institutions. Since the establishment of the UR Clinical & Translational Science Institute (CTSI) in 2006, the biomedical informatics key function (BIKF) has been playing a critical role through innovative use of informatics methods and technologies for improvement of research workflow, discovery and dissemination of knowledge, and collaboration and integration across the boundaries of research disciplines. An important mission of the BIKF is to provide flexible and integrated informatics support to investigators and trainees.

Methods. To better serve investigators and trainees, we have initiated multiple programs delivered in flexible formats, including (1) consulting service through meetings or walk-in consultation clinics, (2) direct support to projects in data management, protocol implementation, online resource development, modeling and simulation, data mining, and behavioral intervention, (3) collaborative research and grant writing, and (4) various educational activities. In particular, we emphasize integration of systems, tools, and online resources with human supporting systems as well as collaboration with the other UR CTSI programs, such as pilot project, biostatistics, education and career development, and regulatory support.

Results. Starting with 20 faculty and 4 trainees from 11 academic units in year 1, coverage of informatics service is now expanded to 103 faculty (415% increase), 11 trainees (175% increase), and 16 staff (220% increase) from 36 academic units (227% increase). Informatics service is now the 2nd most frequently used among all the UR CTSI services, with 89% of users having a positive response to recommend the service to others. The BIKF faculty and staff have participated as key personnel (PI, co-investigator, staff, etc.) or served as a critical institutional resource in various grant applications, with 32 grant proposals submitted to 14 federal, state, and other funding agencies. These proposals have addressed a wide range of problems (from lymphoma, HIV, and autism, to influenza, hypertension, and suicide prevention) in an array of settings (from emergency room, ICU, and inpatient unit, to ambulatory care, community setting, and online social network), formulated as different types of applications such as R01, R24, R34, P30, P50, U19, U48, K08, K23, T32, TL1, D43, RC2, and state contract. Of the 32 applications that were submitted, 14 have been funded (44% success rate), with a total of $29 million awarded. In addition, the BIKF is engaged in various educational activities, including teaching informatics course, sponsoring informatics seminars, participating in advising students and fellows in degree and career development programs, organizing conferences, as well as directing and contributing to education programs.

Discussions. The flexible formats of informatics support and the integration with other CTSI services are important factors for our success. Within our infrastructure, the investigators and trainees can have access to support at any stage of their research and training. To conceptualize a research project, an investigator can schedule a consultation meeting or simply walk in to our regular consultation clinics. For projects with promising ideas, we provide continuous support through collaborative research; additional resources can be obtained through CTSI pilot programs and joint grant applications. For investigators with ongoing projects, we provide direct support tailored to the project needs and integrated with the other UR CTSI programs. For example, we have bundled informatics and biostatistics support and offered an integrated package that includes study design, protocol implementation, data collection, and data analyses. Similar approaches are used to integrate data management and regulatory support services. To better serve trainees, we have integrated informatics courses and seminars with local informatics resources that a trainee can leverage. For example, many trainees have selected a component of their ongoing research as their informatics class projects, which addressed a wide range of real-world research topics that can have direct access to informatics service. Evaluations have shown that this part is consistently selected by the trainees as the most important and interesting component of their training. More important, for all service provided, we are not only building systems but also having our staff to work with the clients to develop a human support system that is synergistically integrated with the computer systems. Evaluation has shown this is one of the most important values of our support that are appreciated by investigators and trainees.

Conclusion. Providing informatics support to investigators and trainees is an important mission of CTSA. Flexibility of service format, service integration, and human support integration are the keys to success.
Griffin Weber, MD, PhD, Nick Benik, Paul Cappelluzzo, Paul Gomez, Ken Huling, Shashank Jain, Melissa Kenny, Kevin Laitinen, Kellie Lucy, Rob Piscitello, George Rakauskas, Jeff Rosen, Marlon Violette, Steve Wimberg, John Halamka, MD

Profiles Research Networking Software is an open source tool, developed at Harvard and now used by several CTSAs (UCSF, HSSC, etc.), which directly addresses the goals of three National CTSA Interest Groups:

The Research Networking Group (co-chairs Griffin Weber, Mini Kahlon) was created to facilitate the discovery of collaborators across topic domains and institutions. When Harvard received its CTSA award, we created the “Profiles” website, which we loaded with more than 20,000 faculty across 33 institutions affiliated with our CTSA using a variety of external data sources including PubMed and ISI Thomson Web of Knowledge, and internal Human Resources and administrative databases. These profiles are linked together through “Passive Networks”, which are automatically generated based on information known about investigators. For example, publications in PubMed are associated with MeSH terms, which we can use to map keywords to individuals. By comparing the keywords of each person to every other faculty member, we can automatically create networks of people with similar interests. Users can also create “Active Networks”, by looking up people they know and manually describing their relationships to them, such as “collaborator” or “past advisor”. Over the past year, we turned this novel concept of Passive and Active Networking into an ontology-driven freely available open source product, now used in several CTSAs and other institutions. A web service API allows us to perform federated queries across multiple instances of the software. Recombinant Data Corp is an Authorized Service Provider, offering commercial support options for institutions adopting the Profiles Research Networking Software.

The Bibliometrics Interest Group (co-chairs Griffin Weber, Julie Earnest) was recently formed to identify tools and methods for obtaining publication data for CTSA investigators, exploring trends in biomedical literature, and evaluating the impact of CTSA. Profiles includes an automatic probabilistic author disambiguation algorithm that uses names and article titles, keywords, affiliations, and co-authors to search PubMed and determine the likelihood that a publication is matched to an individual. Multiple tools are available in Profiles to allow researchers or their proxies to manually correct any mistakes or enter missing publications. The latest version of Profiles has new visualizations that illustrate the distribution of expertise in a topic across an organization and how an individual’s research focus changes over time.

The Social Network Analysis (SNA) Group (chair Noshir Contractor) uses SNA as a tool to assess communication and collaboration and visually depict how people and groups are working together. Profiles automatically calculates a variety of SNA metrics and uses these to construct numerous interactive visualizations. We use SNA in Profiles to fund pilot projects that result in new collaborations, to identify faculty who aren’t advancing in their careers, to understand gender and race inequalities, and to build interdisciplinary teams in novel ways.

Importance to the IKFC Meeting. There is a great deal of interaction among these three CTSA Interest Groups. This presentation will illustrate this synergy in the context of a successful open source software tool developed by a CTSA Informatics Program.

Impact on Consortium Strategic Goals. Profiles was created to encourage new collaborations (a core component of SGC#3, Enhancing Consortium-Wide Collaborations), but we have found numerous uses for the bibliometric and SNA components of the software (e.g., as a mentoring tool used to support the Training and Career Development of C/T Scientists).
Abstract

Penn Data Store (PDS) is the Penn Medicine retrospective data repository serving the analytical needs of physicians, researchers, quality managers and strategic analysts. It currently contains over 1.4 billion rows of primarily clinical data in an operational data store (ODS) organized into these subject areas:

- **Patient** (all UPHS patients)
- **Encounter** (all type of encounters & visits)
- **Diagnosis and Procedures** (based on registration information)
- **Orders and Results** (all type of orders and results entered into SCM and EPIC)
- **Infection** (based on data entered into TheraDoc – new subject area).

PDS is supported by a team of 5 developers and 1 manager. The goals of the Penn Data Store project over the next five years are as follows:

- Define a formal governance structure led by the newly hired CMIO
- Map all clinical concepts to national standards
- Continue to add important clinical data sources to the existing subject areas based on customer needs
- Incorporate revenue cycle information
- Position Penn Data Store as the initial point of contact for research data needs across Penn Medicine.
- Continue to grow current staffing as needed to support plan.

This podium presentation will cover the following:

- Scope of Penn Data Store
- Architecture
- Content
- Progress to date
- Lessons learned
- Future plans and challenges

This chart depicts the depth, breadth and usage of the Penn Data Store to date.
CTSAs offer a wide variety of resources to their home institutions and institutional partners, ranging from didactic training and various award types to use of research patient care and core facilities. Annual reporting, long-term tracking and local reporting needs necessitate a system that unifies these activities, allowing for differences while providing a core set of data elements that can tie together disparate data. Ideally, tracking begins at the time of an investigator’s first encounter with the CTSA through data collected in an application or “resource request” form. Several challenges are faced by CTSAs when designing electronic systems for this purpose:

- Application content may vary significantly depending upon the specific resources being requested.
- Regulatory requirements may differ depending upon factors such as whether a study involves human subjects, animals, or neither.
- Application and review processes will also vary, ranging from rolling review of simple resource requests to once-a-year calls for applications with clear-cut deadlines, as is common for award and didactic training applications.

In its 3+ years as a CTSA, Weill Cornell has recognized several opportunities to address these needs through a series of enhancements to its existing electronic protocol authoring and review (ePAR) system. ePAR was originally developed at Weill Cornell and has been in use since 2005. It is one of several integrated modules of the nationally shared WebCAMP system. Other modules include the Census and Protocol Tracking and Core Utilization Tracking modules, which track progress of approved projects, and a CTSA Annual Reporting module. Each module can operate in a standalone fashion, but combining modules results in a synergistic gain. WebCAMP is distributed and supported by the Weill Cornell CTSA and is currently in use at over half of all CTSAs nationally.

In September of 2007, at the start of Weill Cornell’s CTSA funding, ePAR was a functional system that supported the entire application and review lifecycle, including application initiation, authoring and submission by the study team; management of the review process by CTSA administrative staff; on-line submission of reviews by reviewers; on-line response to review issues by the study team; iteration of the review process until final approval (or non-approval); support for submission and review of amendments after approval of a new application; and a tool for monitored transfer of data on approved ePAR applications to the Census and Protocol Tracking module.

Since September of 2007, the Weill Cornell CTSA has addressed CTSA-related challenges through several key enhancements:

1. Introduction of “application types” -- multiple, pre-defined configurations, each with customizable application content, review processes, and post-approval requirements.
2. Addition of a customizable “Trainee Application Form”, offering an alternative to the traditional “Protocol Summary Page”, appropriate for education program applications.
3. New support for post-approval required documents, which can be configured to be required at specific time points relative to approval (e.g., quarterly after approval, annually after approval, annually on specific dates, etc.). This functionality is used at Weill Cornell for the collection of trainee progress reports.
4. Improved support for administrative functions, including automation of the process for transferring approved ePAR application data to Census and Protocol Tracking, and tools for obtaining quick information about documents requested/due but not yet submitted.
5. Addition of a post-approval follow-up page that supports budget and regulatory post-approval follow-up, as well as tracking of outcomes (publications, new grants, etc.) over time.

The Weill Cornell CTSC now uses ePAR to support all of its award and education applications, as well as most traditional resource requests. “Short form” options, along with expedited review cycles, have been developed for simple resource requests (e.g., biostatistical consultation, REDCap utilization) and for resource requests connected with externally peer-reviewed and approved studies.
Extending the caGrid Federated Query Capability
John Osborne, MS; Harsh Taneja, MS, M.Tech.; Matthew Wyatt, MSHI, University of Alabama at Birmingham

Introduction:
UAB believes the caGrid infrastructure is a foundation upon which collaboration across boundaries within UAB and across the country can thrive. There are numerous activities and projects that are utilizing caGrid. However, we have found that the typical use case is limited in terms of result set size, in complexity or variability of source data. The caGrid federated query use case generally performs on small data sets, returns metadata only, or performs parallel queries on equivalent data models with limited results.

If the caGrid tooling is to be scalable and broadly applicable for federated query it will need to be able to return very large results, regardless of the type of source. The design and implementation of the UAB Aggregate Cohort Estimator (ACE) has approached this problem with success. ACE provides investigators an estimation of patient volume to determine if cohort size is sufficient for research proposals. This is a typical data warehousing use case, but our architecture and our separation of data sources are novel. Underlying ACE is the caGrid Federated Query Processor (FQP) and caGrid data services (one for patient demographics, diagnoses and procedures, and one for lab tests and results).

The Problem: Query and results are too large
We experienced scaling issues with caGrid when dealing with both large data sets and in constructing associations between the data sets. The issues associated with large data sets include the reliance on XML for message transport which is problematic for large scale queries because of the overhead of serialization/de-serialization for large messages. This was partially addresses through various adjustments to hardware, time out settings, and the data model. Additionally we have in progress the addition of WS-Enumeration in FQP, which allows partial results to be paged to FQP which reduces the overhead of large messages and result sets. Currently FQP implements WS-Enumeration on the client side, but does not yet interrogate data services to determine if they are WS-Enumeration capable so that it can enumerate results. A more severe limitation was the inability to construct large associations between data services; specifically, foreign associations with more than approx. 20 thousand identifiers. For instance, the version of hibernate used in caGrid 1.3 (3.2.0ga) relies on recursion for a number of functions that process logical operators like "OR". If FQP sends off a criteria group of a large number of these criteria, a function call is generated for each logical operator which will overflow the stack at the data service level for reasonable JVM stack size settings. Some of these issues have been fixed in more recent hibernate releases but other hibernate functions are not fixed to our knowledge. Regardless of hibernate, the use of a large number of identifiers by necessity creates large queries that must be processed by the database. Even after adjusting JVM stack size settings and replacing recursion with iteration in hibernate, our database (both Oracle and MySql) could not process such large queries. There are limits on the size of the SQL query that databases are willing to accept, including total query byte size, maximum number of operators permitted (inside or outside of a clause), number of acceptable bind variables and perhaps other problems.

The Solution: Make the query smaller
Ultimately the issue with a large number of identifiers is fixed by splitting the identifier bloated CQL queries sent to data services into smaller queries nearly identical to each other, differing only in having a distinct subset of identifiers for their criteria group. CQL queries generated by FQP are scanned to determine if the criteria group which contains identifiers from a foreign association is above a certain threshold. Then the CQL query is split into queries of a user-defined chunk and sent in series to the data source. Results are combined into a single CQL object.

Other Non-Standard Configuration
- WS Enumeration enabled on data services
- Disabled lazy loading in data services
- Increased row limit from 10K-1 million rows
- Changed default JVM sizes for data services and FQP (currently 15G and 6G respectively)
- Oracle JRocket for 3 services (grid/data/FQP)

Importance to CTSA Consortium and IFKC
- Increases the ability to use the caGrid infrastructure for federated data sharing, cross institution collaboration, and translational research
- Allows the deployment of larger and more diverse data sources while providing more flexibility in the types of queries that can be performed via caGrid
- Proves the utility of grid based federated cohort identification, thus mitigating requirements for data sources at individual institutions to be standardized.
- Increases the scalability of federated query such that data size is less problematic