Abstracts

2011 CTSA Annual Informatics Meeting
October 12-13, 2011
Natcher Conference Center
NIH Campus
### Podium Presentations

**CTSA Informatics Annual Meeting**

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Presenter</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Novel Methods for Measuring Impact of Informatics Tools</td>
<td>Paul Harris</td>
<td>Vanderbilt University</td>
</tr>
<tr>
<td>2</td>
<td>Patterns of Search: Analyzing and Modeling User Behavior in a Scientific Professional Network (WorkWeb)</td>
<td>Chunhua Weng</td>
<td>Columbia University</td>
</tr>
<tr>
<td>3</td>
<td>PROMPTR: Progress Reporting and Outcomes Measurement to Promote Translational Research</td>
<td>Elizabeth Wood</td>
<td>Weill Cornell Medical College</td>
</tr>
<tr>
<td>4</td>
<td>A Logic Model and Metrics for Evaluation of the Biomedical Informatics Component of the Arkansas Translational Research Institute</td>
<td>William Hogan</td>
<td>University of Arkansas</td>
</tr>
<tr>
<td>5</td>
<td>Integrating Electronic Health Record (EHR) and Community Information System (CIS) Data to Support Population Research and Practice</td>
<td>Brian Dixon</td>
<td>Indiana University</td>
</tr>
<tr>
<td>6</td>
<td>WICER: Building a Community-Centered Translational Research Infrastructure with Electronic Data</td>
<td>Adam Wilcox</td>
<td>Columbia University</td>
</tr>
<tr>
<td>7</td>
<td>The Analytic Information Warehouse: a Platform Supporting Comparative Studies of EHR Data</td>
<td>Andrew Post</td>
<td>Emory University</td>
</tr>
<tr>
<td>8</td>
<td>Explorys - A New Paradigm in Integrated Data Repositories: Overview and Case Study</td>
<td>David Kaelber</td>
<td>Case Western Reserve University</td>
</tr>
<tr>
<td>9</td>
<td>TruData: Computing Truth Through Real-World Transactions</td>
<td>Curtis Cole</td>
<td>Weill Cornell Medical College</td>
</tr>
<tr>
<td>10</td>
<td>Plumbing and Politics: A Governance and Operating Model for the Northwestern Medical Enterprise Data Warehouse</td>
<td>Andrew Winter</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>11</td>
<td>A retrospective research model enabled by an open-source integrated research data repository</td>
<td>Greg Hruby</td>
<td>Columbia University</td>
</tr>
<tr>
<td>12</td>
<td>Harvest: An Open-Source Biomedical Data Integration and Discovery Platform</td>
<td>Jeffrey Pennington</td>
<td>Children’s Hospital Of Philadelphia</td>
</tr>
<tr>
<td>13</td>
<td>Using Electronic Health Records to Identify Cohorts for Drug-Induced Thrombocytopenia, Neutropenia and Liver Injury</td>
<td>Jyotishman Pathak</td>
<td>Mayo Clinic</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>14</td>
<td>TimeFlow - iPads in clinical research: time-motion and secure offline data collection</td>
<td>Warren Kibbe</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>15</td>
<td>eagle-i: development and expansion of a scientific resource discovery network</td>
<td>Doug MacFadden</td>
<td>Harvard University</td>
</tr>
<tr>
<td>16</td>
<td>The ITHS Central Database: A Customer Relationship Management System for CTSAs</td>
<td>Jim Piper</td>
<td>University of Washington</td>
</tr>
<tr>
<td>17</td>
<td>The Human Studies Database (HSDB) Project</td>
<td>Ida Sim</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>18</td>
<td>Electronic Capture and Management of Informed Consent for Research</td>
<td>Aziz Boxwala</td>
<td>University of California, San Diego</td>
</tr>
<tr>
<td>19</td>
<td>Harvard Catalyst Profiles: Research networking, bibliometric analysis, and social network analysis based on linked open data and VIVO ontology</td>
<td>Griffin Weber</td>
<td>Harvard University</td>
</tr>
<tr>
<td>20</td>
<td>Demonstration of the i2b2-based Health Outcome Monitoring and Evaluation (HOME) Cell</td>
<td>William Adams</td>
<td>Boston University</td>
</tr>
<tr>
<td>21</td>
<td>G-CODE: Enabling systems medicine through innovative informatics</td>
<td>Subha Madhavan</td>
<td>Georgetown University</td>
</tr>
<tr>
<td>22</td>
<td>Integrating Research Networking Tools with Broader Social Networking Products to Create an Improved Environment for Accelerating Science</td>
<td>Eric Meeks</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>23</td>
<td>A Research Participant Tracking System that Supports Business Process through System Integration</td>
<td>Robert Gehrke</td>
<td>Mayo Clinic</td>
</tr>
<tr>
<td>24</td>
<td>Enhancing Collaborative Science Using Shared Query Tagging/Searching Plugins for the i2b2 Web Client: The WO Apps</td>
<td>Rajani Sadasivam</td>
<td>University of Massachusetts</td>
</tr>
<tr>
<td>25</td>
<td>HUBzero Frameworks for Institutional and Project Scale Information Management</td>
<td>William Barnett</td>
<td>Indiana University</td>
</tr>
<tr>
<td>26</td>
<td>Building Robust Research Capabilities through Integrating Clinical Study, Patient, and Biospecimen Data Repositories</td>
<td>Leslie McIntosh</td>
<td>Washington University</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>27</td>
<td>VIVO: A Tool for Collaboration and Research Discovery</td>
<td>Kristi Holmes</td>
<td>Washington University</td>
</tr>
<tr>
<td>28</td>
<td>Case Studies and Meaningful Research Use of the Shared Health Research Informatics Network (SHRINE)</td>
<td>Andrew McMurry</td>
<td>Harvard University</td>
</tr>
<tr>
<td>29</td>
<td>Omics Data Standards Working Group: update on activities</td>
<td>Jessie Tenenbaum</td>
<td>Duke University</td>
</tr>
<tr>
<td>30</td>
<td>DIRECT2Experts – Distributed Interoperable Research Experts Collaboration Tool</td>
<td>Griffin Weber</td>
<td>Harvard University</td>
</tr>
<tr>
<td>31</td>
<td>Use-case Driven Development of a Federated Research Data Warehouse</td>
<td>Monika Ahuja</td>
<td>University of Iowa</td>
</tr>
<tr>
<td>32</td>
<td>Cross CTSA Institutions Cohort Discovery Approach</td>
<td>Adil Alaoui</td>
<td>Georgetown University</td>
</tr>
<tr>
<td>33</td>
<td>Exploration of Two Approaches for Determination of Recipients of Targeted Communication in a Scientific Collaborative Network (WorkWeb)</td>
<td>Suzanne Bakken</td>
<td>Columbia University</td>
</tr>
<tr>
<td>34</td>
<td>Improving Robustness and Automating Workflow of the Resource Discovery System with iBIOFind</td>
<td>Michael Baldonieri</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>35</td>
<td>An Informatics Course Based on the CTSA Informatics Competencies</td>
<td>Eta Berner</td>
<td>University of Alabama at Birmingham</td>
</tr>
<tr>
<td>36</td>
<td>Finding Collaborators: Towards Interactive Tools for Research Social Networks</td>
<td>Charles Borromeo</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>37</td>
<td>REDCap Goes to the Rodeo</td>
<td>Linda Carlin</td>
<td>University of Colorado Denver</td>
</tr>
<tr>
<td>38</td>
<td>TIES-SPiRiT Collaborative Tissue Network</td>
<td>Girish Chavan</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>39</td>
<td>A System Architecture for Federating Standardized Clinical Brain Images Across Hospitals</td>
<td>Ann Chervenak</td>
<td>University of Southern California</td>
</tr>
<tr>
<td>40</td>
<td>Ranking CTSA Labs: Metrics for Biomedical Resource Identification, Utilization and Impact Assessment</td>
<td>Jeff Cromwell</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>41</td>
<td>Modeling Electronic Tools to Improve Source Document Verification Audits</td>
<td>Stephany Duda</td>
<td>Vanderbilt University</td>
</tr>
<tr>
<td>42</td>
<td>Engaging campus biobanks in the selection of an enterprise-wide biobanking informatics platform</td>
<td>Helena Ellis</td>
<td>Duke University</td>
</tr>
<tr>
<td>43</td>
<td>Secondary Data Reuse in Comparative Effectiveness Research: A Model Framework and Proof of Concept</td>
<td>Dan Fort</td>
<td>Columbia University</td>
</tr>
<tr>
<td>44</td>
<td>Maintaining Cohort Discovery coded content with the LE Access Portal (LEAP)</td>
<td>Davera Gabriel</td>
<td>University of California, Davis</td>
</tr>
<tr>
<td>45</td>
<td>i2b2 at the University of Michigan: Cohort discovery and beyond . . .</td>
<td>Stephen Gendler</td>
<td>University of Michigan</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>46</td>
<td>Withdraw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Identifying duplicate clinical records using administrative and clinical data</td>
<td>Susan Guerrero</td>
<td>University of Texas, Houston</td>
</tr>
<tr>
<td>48</td>
<td>Achieving Clinical Research Management System (CRMS) Adoption Through Coordinated Changes to Clinical Research and Billing Practice</td>
<td>Diana Gumas</td>
<td>Johns Hopkins University</td>
</tr>
<tr>
<td>49</td>
<td>Withdraw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>An Open-source Informatics Management System and Community for the National Children's Study</td>
<td>William Hogan</td>
<td>University of Arkansas</td>
</tr>
<tr>
<td>51</td>
<td>Identifying and Assessing the Impact of Biomedical Research</td>
<td>Kristi Holmes</td>
<td>Washington University</td>
</tr>
<tr>
<td>52</td>
<td>Informatics Education: Partnerships and Opportunities</td>
<td>Kristi Holmes</td>
<td>Washington University</td>
</tr>
<tr>
<td>53</td>
<td>Informatics and Community Engagement: Collaboration Efforts at UIC</td>
<td>Denise Hynes</td>
<td>University of Illinois at Chicago</td>
</tr>
<tr>
<td>54</td>
<td>eNOTIS - an open source participant tracking system used to enhance patient safety and compliance and Registrar, an integrated registry management system</td>
<td>Warren Kibbe</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>55</td>
<td>NCS Navigator MDES Warehouse - a pragmatic approach to building a semantic data warehouse from semi-parsable, mandated data elements</td>
<td>Warren Kibbe</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>56</td>
<td>FacultyConnect - connecting faculty with award opportunities and tracking interest, follow-through, and outcomes</td>
<td>Warren Kibbe</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>57</td>
<td>LatticeGrid - an open source tool assessing collaboration patterns and tracking interest, follow-through, and outcomes</td>
<td>Warren Kibbe</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>58</td>
<td>A Paradigm Shift: Electronic Health Records Data in Clinical Practice</td>
<td>Ketan Mane</td>
<td>University of North Carolina</td>
</tr>
<tr>
<td>59</td>
<td>Towards Integrated Solutions for Virtual Drug Screening</td>
<td>Jarek Meller</td>
<td>University of Cincinnati</td>
</tr>
<tr>
<td>60</td>
<td>Central Indiana Innovation Network (CI-Net)</td>
<td>Michael Murray</td>
<td>Indiana University</td>
</tr>
<tr>
<td>61</td>
<td>ResearchIQ: Semantic Search for the Research Community</td>
<td>Philip Payne</td>
<td>Ohio State University</td>
</tr>
<tr>
<td>62</td>
<td>StudySearch: A Tool For Connecting Potential Participants with Locally Recruiting Studies</td>
<td>Philip Payne</td>
<td>Ohio State University</td>
</tr>
<tr>
<td>63</td>
<td>High-throughput Informatics Infrastructure for Biorepositories</td>
<td>Susanne Ragg</td>
<td>Indiana University</td>
</tr>
<tr>
<td>64</td>
<td>Large-scale Analytics for In Silico Study of Brain Tumors using High-resolution Digital Pathology</td>
<td>Joel Saltz</td>
<td>Emory University</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>65</td>
<td>Integration of drug classification information with EMR data in the utCRIS Data Warehouse for adverse drug event association analysis</td>
<td>Richard Scheuermann</td>
<td>University of Texas Southwestern</td>
</tr>
<tr>
<td>66</td>
<td>DV Docs: Generating CVs and NIH Biosketches from VIVO Data</td>
<td>Linda Schmandt</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>67</td>
<td>Enabling Research Through Technology: Leveraging Enterprise Data Stores in Support of Cohort Generation and Facilitated Subject Recruitment</td>
<td>Howard Shang</td>
<td>Duke University</td>
</tr>
<tr>
<td>68</td>
<td>Integration of Clinical Data for an ICU Registry</td>
<td>John Sharp</td>
<td>Case Western Reserve University</td>
</tr>
<tr>
<td>69</td>
<td>A Collaborative Support Model for Investigator-Initiated Data Management</td>
<td>Denise Snyder</td>
<td>Duke University</td>
</tr>
<tr>
<td>70</td>
<td>A Web-based Tool for Cataloging Primary Care Electronic Medical Record Federated Data: FlnDiT</td>
<td>Kari Stephens</td>
<td>University of Washington</td>
</tr>
<tr>
<td>71</td>
<td>Warehousing of Trials Data for Data Submission and Reporting</td>
<td>Shariq Tariq</td>
<td>University of Arkansas</td>
</tr>
<tr>
<td>72</td>
<td>Budgeting Issues and Solutions for a Clinical Research in an Academic Health Center</td>
<td>Umit Topaloglu</td>
<td>University of Arkansas</td>
</tr>
<tr>
<td>73</td>
<td>Towards and Oncology Database (ONCOD), a Case Study Using Data Warehousing Technology</td>
<td>Xiaoming Wang</td>
<td>University of Chicago</td>
</tr>
<tr>
<td>74</td>
<td>A Machine Learning Based Approach for Retrieval and Organization of Translational Research Articles</td>
<td>Firas Wehbe</td>
<td>Vanderbilt University</td>
</tr>
<tr>
<td>75</td>
<td>A Method for Standardized Patient Reported Outcome Data Collection at the Point of Care</td>
<td>James Willig</td>
<td>University of Alabama at Birmingham</td>
</tr>
<tr>
<td>76</td>
<td>Using a MeSH-based Index to Demonstrate the Multidisciplinary Propensity of KL2 Scholars</td>
<td>GQ Zhang</td>
<td>Case Western Reserve University</td>
</tr>
</tbody>
</table>
Abstract: Novel Methods for Measuring Impact of Informatics Tools

1) Description of the project, best practice, activity, system to be presented:
Evaluation is essential for measuring the impact for any CTSA program, system or tool. The null hypothesis for any new initiative should be to assume no impact until proven otherwise. Adding ‘measure as you go’ logging and metrics into informatics tools can lead to real-time evaluation of impact. This presentation will focus on methods we have developed across multiple informatics platforms to measure usage and overall impact for informatics tools and systems. Examples will be provided from our local researcher portal (StarBRITE), a CTSA-wide national recruitment registry (ResearchMatch), and a data management platform serving over 28,000 end-users across 6 continents (REDCap). See Figure 1 for examples.

2) Why it is important to be presented at the 2011 IKFC meeting
Best practice evaluation concepts will be relevant to any CTSA informatics team. Example methods will be applicable to a wide range of informatics tool classes.

3) Impact on the CTSA Consortium Strategic Goals
Evaluation of programs and interventions is vital for accomplishing the mission of all CTSA Consortium Strategic Goal Groups.
Patterns of Search: Analyzing and Modeling User Behavior in a Scientific Professional Network (WorkWeb)


The Irving Institute for Clinical and Translational Research, Columbia University, New York, NY

Research design: The Biomedical Informatics Resource of the Irving Institute for Clinical and Translational Research (IICTR) has developed WorkWeb, a collaborative platform with several major applications. One allows users to submit requests for IICTR services. Another is a wiki for online collaboration. The Profiles Directory allows users to search for collaborators based on their scientific interests, publications, and grants. Compared to related work such as BiomedicalExperts, Collexis, and ResearchGate, the WorkWeb Profiles Directory is unique in that it automatically integrates data from both Columbia University and external databases, such as PubMed, and performs accurate person name disambiguation. The WorkWeb Profiles Directory was officially launched at Columbia in March 2011. To understand the usage patterns and information needs of researchers, we designed a query log to track user queries and generate usage statistics in real time. To date, we have collected four months of usage data. We developed a method to model user search behaviors by correlating user and query characteristics. Our analyses address the following questions: (1) what are the characteristics of the frequent users of an institutional scientific professional network? (2) what are the typical uses of the scientific professional network? (3) what information is being searched for? (4) what are common adjacent actions for a search? (5) how do usage patterns differ between navigational queries and informational queries? (6) what is the average, minimum, and maximum duration of a user session? would this piece of information suggest the potential threshold for a system to deliver answers within a certain number of user actions? We also analyze query terms and sequences of search features used.

Results: A typical user session contains mixed iterative keyword search and profile navigation. Consistent with typical general web search queries, most search keywords are short, containing one phrase and occasionally Boolean operators “AND” and “OR”. The most frequently used keywords are person and disease names. The most common user feedback is to request self-additions and deletions of profiles and publications. The most frequent user logged into the system for 41 times with the 4 months to search for researchers with grants or papers on multiple topics. The most frequent actions are searching for people and browsing their profiles; therefore, WorkWeb is currently used mostly to view a federated profile of a scientist’s academic title and contact information, publications, and grants.

Importance to CTSA

Many CTSA institutions have created or implemented scientific professional network platforms that provide rich researcher profiles for facilitating scientific collaborations and for supporting the career development of clinical and translational scientists. However, there are few methodologies for modeling user search behaviors and understanding user needs for large scientific professional networks. We will share our usage tracking and modeling method. We hope these methods will help other CTSAs better understand the information search behaviors of clinical and translational scientists and elicit their tacit information needs to enable interdisciplinary team science.

Impact on the CTSA consortium strategic goals

This research has direct impact on the following two CTSA consortium strategic goals: goal 2 (improving career development of clinical and translational scientists) and goal 3 (enhancing consortium-wide collaborations).

Acknowledgement: The project described was supported by Grant Number UL1 RR024156 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH).
**PROMPTR: Progress Reporting and Outcomes Measurement to Promote Translational Research**

Elizabeth Wood, MS; Geraldine Amera, MPA; My Linh H Nguyen-Novotny; Kaarel Laev, MBA, MPH; James Holahan; Tim Baker; and Stephen B Johnson, PhD

Measuring the impact of CTSA resources on translational research and, ultimately, health outcomes requires long-term follow-up with investigators and trainees. Resources such as PubMed’s Entrez Utilities and the new NIH ExPORTER system increasingly facilitate mining of public databases to obtain information on publications and grants and link this information to specific investigators. However, the need for direct input from the investigator remains critical to understanding the relationship of outcomes (publications, grants) to specific research projects and to obtaining qualitative, free-form information about progress and impact.

WebCAMP is an open-source system supporting a variety of functions important to CTSAs, including application submission and review, education program tracking, research participant scheduling, core utilization tracking, billing and annual reporting, and is currently in use at approximately two-thirds of CTSA sites. The most important aspect of WebCAMP is its unification of these diverse functions in a set of common database tables, allowing for reporting across a variety of separately administered areas.

The current project was begun in late 2010 to leverage WebCAMP’s open database architecture and publicly available data sources to simplify the reporting process for investigators and reduce data re-entry by CTSA staff. Toward this end, we developed a system of customizable, interactive Web-based forms (PROMPTR) with the following key characteristics:

- Pre-population of forms with data already in WebCAMP
- Search tools for selecting and importing data from publicly available data sources (currently, PubMed and the NIH ExPORTER system). 
- Direct and immediate flow into WebCAMP of new or updated data.

While fully interoperable with WebCAMP, PROMPTR is architecturally distinct. Privacy and security are achieved without a need to assign WebCAMP passwords to progress report respondents, and PROMPTR has been carefully designed for broad cross-browser compatibility. These features are critical to ensuring quick and problem-free access for respondents. To facilitate progress report oversight, an Administrative Dashboard was created, permitting CTSA staff to obtain real-time data about responses and generate e-mail reminders to non-responders as well as notes of thanks to early responders.

PROMPTR was piloted at the Weill Cornell CTSC in February 2011 to collect data for the CTSA annual report and ongoing evaluation activities. Progress and outcomes data were solicited from PIs of 264 CTSC studies and 60 scholars/trainees. A total of 292 (90%) responded by connecting at least once to the system and 98% of these responders proceeded to complete and submit their data. Of note, 9% of the completed reports were received within a day of solicitation and 70% were received within 3 weeks. Compared to past approaches, our staff noted cleaner data, fewer redundant requests to investigators, immediate availability of data in WebCAMP, and the consistency enforced by the use of a streamlined system as primary improvements. The pilot was also an opportunity to identify areas for future work. We learned that when we allow a free-form text box for entry of the citation for a meeting presentation or abstract, some investigators will paste an entire biosketch into one field. Improved text validation will reduce this in the future. We also identified a need for clearer identification of relationships of outcome measures (grants and publications) to projects. For example, grants can be associated with a project by either supporting the project or resulting from it. Finally, in the future we will build an optional “exit survey” into the system to collect input from the investigators themselves, ensuring ongoing dialog with this user group and allowing us to track satisfaction with the system as it evolves over time.

PROMPTR’s main strength lies in the elimination of redundant data collection and support of meaningful, seamless linkage to other project-related data. PROMPTR will be released this fall as part of WebCAMP v.3, allowing other CTSAs to configure, generate and administer PROMPTR progress reports.
A Logic Model and Metrics for Evaluation of the Biomedical Informatics Component of the Arkansas Translational Research Institute

William R. Hogan, MD, MS, Jennifer L. Bufford, MS, Teresa J. Hudson, PharmD, Ellen P. Fischer, PhD, Kelly W. Bulloch, JD

The Arkansas Translational Research Institute (ATRI) created a logic model and evaluation metrics for its Comprehensive Informatics Resource Center (CIRC), the component responsible for carrying out its biomedical informatics objectives. This work was a collaborative effort between CIRC and the ATRI evaluation team to define specific outcomes that would measure CIRC’s achievement of specific aims, as well as its impact on the clinical and translational research enterprise at the University of Arkansas for Medical Sciences and statewide.

Considering the degree of evolution that typically occurs while a CTSA is in its infancy, this activity provided an opportunity to document how informatics aims and objectives evolved since the time of initial grant submission. Although our logic model is dynamic in nature, it currently includes 7 aims, 25 key activities, greater than 20 outputs, and 4 short-term metrics. In the first quarter of grant year 3, CIRC submitted its first official outputs report containing both quantitative and qualitative data. Future quarterly reports will provide ATRI leadership with decision-relevant evaluation information and enable rapid feedback to CIRC.

Recognizing the importance of developing an informatics evaluation strategy across the consortium, the Informatics Key Function Committee (IKFC) proposed to develop a CTSA Informatics Core Metrics and Evaluation (ICME) affinity group. The proposed ICME would have been charged with (1) reviewing metrics for managing and evaluating informatics achievements by individual CTSA institutions and subsequently (2) developing a common evaluation framework for adoption within the Consortium. Despite the awareness of this common need, the ICME failed to develop. With many institutions at the brink of renewal, it is even more essential that the IKFC establish an evaluation strategy for determining the impact of its efforts on translational research. As an initial step in this direction, CIRC proposes to share its evaluation experiences at this year’s IKFC with the goal of reigniting interest in forming an ICME affinity group or workgroup.
**Introduction**

One major aim of the Indiana Clinical and Translational Sciences Institute (CTSI) Biomedical Informatics Program (BIP) is “to expand and refine the data available to researchers in an integrated knowledge and data environment.” In support of this goal, entities on the campus of Indiana University-Purdue University Indianapolis (IUPUI) partnered to create geospatially augmented clinical records to support innovative public health informatics and health geographics research.

Analysis of the relationships between geospatial, clinical, and community (e.g., median income, housing) data currently requires the linking of various data sets manually, in batch fashion, project-by-project. The Regenstrief Institute, an internationally recognized medical informatics research organization, and the Polis Center, a geoinformatics and health geographics applied research organization, set out to find a way to create linked, operational data sets that could facilitate more efficient research and use by academics and professionals in public health, health geographics, as well as biomedical/health informatics.

**Methods**

The team developed methods to enhance and link patient records from the Indiana Network for Patient Care (INPC), an operational health information exchange, with geospatial and community data housed in the Polis Center’s SAVI, a nationally recognized community information system. Patient records typically include basic geospatial data captured in clinical electronic health record (EHR) systems, such as address, city, state, and zip code. A component-based approach was implemented whereby the INPC invokes a secure web service, developed and managed by the Polis Center, to geocode patient addresses when they are a) first sent to the INPC or b) updated as a result of a healthcare encounter. The service returns longitude and latitude as well as county, census block, and a unique link to geospatial reference files, community information, and metadata maintained in SAVI.

**Results**

More than 20 million patient records, linked to more than 840 million clinical encounters, from the INPC were geocoded using the Polis Center web service. First, the records in the INPC were enhanced with additional geospatial data not typically provided to an HIE by provider EHR systems, including census block. The enhancements enable the HIE, for example, to more accurately report notifiable disease results to local public health departments. Second, the INPC patient records now contain persisted links to non-clinical, community level data maintained in SAVI. Permanent links allow health professionals and researchers to examine disease patterns, co-morbidity, medication usage, and other population-level clinical characteristics of INPC patients within the context of patients’ neighborhoods, socio-economic regions, health service areas, and other community-level characteristics.

**Discussion and Conclusions**

The enhanced clinical records available to the Indiana CTSI provide an ideal infrastructure for biomedical/health informatics and health geographics research studies as well as development of innovative applications for clinical and public health practice. This infrastructure, for example, could be leveraged to develop new approaches to the examination of the social determinants of health outcomes.

Geospatial enhancement of EHR/HIE repositories and data warehouses in other communities is possible and may be desirable as other CTSA informatics programs collaborate with multidisciplinary colleagues to better understand health care access, quality, and costs. Our approach demonstrates not just what the art of the possible might be with respect to integrating EHR and CIS data but also how to efficiently and effectively integrate these two disparate data silos using information resources that are likely under development or available today within a CTSA institution.
The Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) is a project funded by the Agency for Healthcare Research and Quality to create an infrastructure to support prospective studies using electronic data systems. The WICER project is focused on an underserved and minority community population in northern Manhattan, local to the Columbia University Medical Center. Data from multiple health care organizations and sites of care (inpatient, outpatient, home care, long-term care) are collected together to create a more comprehensive view of the clinical care provided to the patient. In addition, the project is collecting information directly from 12,000 patients or community residents outside the care setting, including vital statistics, demographics, social information, and patient assessments. Biological samples are also collected from residents for genetic sequencing. Data from all sources are merged together for each individual. In this way, WICER creates an electronic data repository that is patient-centered, rather than institution- or care setting-centered. We foresee multiple benefits of a patient-centered data infrastructure. First, data are more complete and allow studies of factors and cofounders that can be related to multiple diseases. Second, data from various sources representing similar meaning can be tested for accuracy and consistency between data sources, and secondary data content can be assessed. Third, subjects can be recruited from this population already having detailed historical data that can be relevant to prospective studies. And fourth, with the link between genetic and phenotypic information; genes can be studied for their effect on individual expressions. In addition to the data infrastructure, we are creating data access tools for researchers to more efficiently perform ad-hoc queries on the data infrastructure and identify criteria for patient cohorts.

This project is of interest to the Clinical and Translational Science Award (CTSA) Informatics Key Function Committee (IKFS) because of its dependence on the CTSA infrastructure and its support of CTSA informatics goals. The WICER project was created through multiple collaborations that already existed within the Columbia CTSA, especially in informatics and community engagement. The WICER project also directly supports the informatics goals of CTSA, in supporting translational research. No other AHRQ data infrastructure project is as tightly connected to a population local to an institution and CTSA site, which makes the WICER and CTSA collaboration at Columbia particularly interesting. In addition, since the WICER project was funded separately and substantially, it can be a model of what informatics support can be created within CTSAs but may not be done at other institutions from lack of funding. Finally, the WICER project is interesting to IKFS because it leverages informatics functions to more efficiently support research. With its community focus and survey data, WICER might be considered a more technically-supported Framingham Heart Study, except that it supports research on multiple diseases and studies an inner-city immigrant population with multiple health care disparities.

The WICER project has already had a large impact on the local activities of the CTSA at Columbia University, especially within the Biomedical Informatics Research core and the Community Engagement core. In addition, it was central to two funded CTSA supplements at Columbia - one to use the infrastructure to study disparities, costs and comorbidities more deeply, and another to expand a biobanking effort to the WICER research population. We expect WICER to help with the CTSA Consortium Strategic Goal to build clinical and translational research capability.
The Analytic Information Warehouse: a Platform Supporting Comparative Studies of EHR Data

Andrew Post, MD, PhD1,2, Sharath Cholleti, PhD2, Jingjing Gao, PhD2, Xia Lin, PhD2, Tahsin Kurc, PhD1,2, Terry Willey, RN3, Danny Eapen, MD4, Tim Morris1,5, Dedra Cantrell, RN1,2,3, William Bornstein, MD, PhD6, Arshed Quyyumi, MD4, Marc Overcash1,5, Joel Saltz, MD, PhD1,2

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

Comparative and clinical effectiveness studies increasingly involve analysis of clinical event profiles and outcomes across institutions and against national benchmarks using data from electronic health records. The Atlanta Clinical and Translational Science Institute (ACTSI), Emory University and Emory Healthcare are addressing these needs through the joint creation of the Analytic Information Warehouse (AIW). The AIW is a software platform for creating integrated data marts from one or more source databases by transforming data into a shared representation that enables comparative studies. The AIW was deployed in production in March, 2011. We have generated data marts in support of an NIH-funded research study, an Emory hospital readmissions project, and comparative analyses of data from Emory Healthcare’s Clinical Data Warehouse (CDW) and the University HealthSystems Consortium (UHC) database (contains inpatient data from 113 academic centers and affiliated hospitals). We expect the AIW to be highly relevant to other CTSA with similar goals of working with their institutions to make EHR data available to investigators and link EHR data to national and other local data sources.

The AIW implements a semantic extract, transform and load (ETL) process. The extract step generates SQL to query source databases and outputs retrieved data in a shared representation that is specified as an ontology. The transform step computes selected derived variables that represent disease groups, medication classes, thresholds in numerical test results, and temporal patterns that are specified in the same ontology. Presently the load step outputs data and derived values into delimited files, or it outputs data, derived values and relevant data element definitions into an instance of i2b2. The AIW has supported the NIH-funded GeneBank project, a study of cardiovascular disease, by generating an i2b2 instance containing EHR data on the enrolled subject population (over 4000 patients) from Emory’s CDW. It has supported analyses of readmissions for Emory Healthcare using 5 years of inpatient data in the CDW. Analyses have included computing relative risk of 30-day readmission for fine-grained subpopulations that are defined by groups of diagnosis and procedure codes, thresholds in laboratory test results, and temporal patterns. Analyses also have included data mining to identify patterns that are associated with and predictive of readmits. The AIW also has supported comparative analyses of readmissions at Emory with a 5-year extract of data from the UHC database (9 million patients and 13 million encounters). Because the AIW extracts data into a shared representation, the Emory and UHC data are more easily comparable.

The AIW has had early successes in supporting research and quality improvement and in linking EHR data to a national dataset. It is a model of collaborative development between a CTSA and academic and healthcare IT that supports both research and quality improvement use cases. We are in the process of developing AIW-based data marts for the Minority Health GRID (which is developing an EHR-linked repository of phenotype, genotype, and biospecimen cardiovascular data from over 2000 subjects across four institutions), integrating AIW into the multi-institutional CardioVascular Research Grid (which is funded by NHLBI to provide software resources to the cardiovascular research community) for integrated clinical data management, implementing support for cancer studies within ACTSI including a Lung P01, creating robust tools for deploying and configuring the system, and releasing the software as open source.

This work has been supported in part by PHS Grant UL1 RR025008, KL2 RR025009 and TL1 RR025010 from the CTSA program, NIH, NCRR; NHLBI grant R24 HL085343; NIH/ARRA grant 325011.300001.80022; and M01 RR-00039 from the GCRC program, NIH, NCRR.
Explorys – A New Paradigm in Integrated Data Repositories: Overview and Case Study

David C Kaelber\(^1\), GQ Zhang\(^2\), and Anil Jain\(^3,4\)

\(^1\)Departments of Information Services, Internal Medicine, Pediatrics, and the Center for Clinical Informatics Research and Education, The MetroHealth System; \(^2\)Division of Medical Informatics, School of Medicine, Case Western Reserve University; \(^3\)Department of Internal Medicine, University of Cincinnati College of Medicine; \(^4\)Explorys, Inc.

Objective: We used the Explorys platform to perform a “non-human subject” retrospective study on approximately 1 million patients looking at patient characteristics associated with venous thromboembolic events (pulmonary embolism and deep venous thrombosis). We compare this study to the analogous prospective cohort study already published in the literature to demonstrate the potential and limitations of this research informatics tool.

Methods: The Explorys platform is an enterprise-strength, cloud-based large-scale clinical data repository and structured search engine. Built on the Unified Medical Language System (UMLS) terminology infrastructure, it provides integrated, standardized, normalized and de-identified digital data from disparate health information systems in several non-affiliated healthcare organizations across the U.S (see figure). The Explorys platform currently contains health information on more than 4 million unique patients.

We used the Explorys platform to perform a series of comparative cohort queries based on demographic information (age, gender, race/ethnicity), diagnosis information (pulmonary embolisms and deep venous thrombosis), medication information, and vital signs (height and body mass index). R (open source statistical software) was used to determine the statistical significance of the differences between the comparative cohort queries.

Results: In a cohort of 959,030 eligible patients and 21,210 patients with VTE, comparing obese, tall subjects to normal BMI, short subjects, the VTE Odds Ratio (OR) was 1.83 (95\% CI: 1.76–1.91) for women and 1.21 (1.10–1.32) for men. Compared to Caucasians, Hispanic/Latino subjects had much lower risk of VTE [female OR = 0.47 (0.41 – 0.55), male OR = 0.24 (0.20– 0.28)] and African-Americans substantially higher risk [female OR = 1.83 (1.76-1.91), and male OR = 1.58 (1.50-1.66)]. Over approximately 125 hours in 11 weeks we replicated, with a 50-fold increase in patients, a prospective 13-year study demonstrating similar results, but with less statistical power.

Conclusions: With the right clinical research informatics tools, integrated data platforms, such as Explorys, can transform some types of cohort studies, incorporating many more patients and requiring substantially fewer resources. Limitations include the ability to only look at health information that can be standardized and normalized, as well as methodological and analytics limitations related to having access only to totally protected health information de-indentified information.
TruData: Computing Truth Through Real-World Transactions
Curtis Cole, MD, Tru Tran, BS, Adam Cheriff, MD, J. Travis Gossey, MS, MD, Victor Brodsky, MD Stephen B. Johnson, PhD Weill Cornell Medical College, New York, NY

Data repositories and warehouses are typically the downstream recipients of millions of data elements received from multiple sources. The issue of matching conceptually identical data is complex and urgent if valid inferences are to be drawn from the data. Manual mapping is slow and costly. This problem is compounded in multi-facility research centers as is typical in CTSA institutions. Ideally, these data are filed appropriately in real time within the source systems and therefore align downstream in the repository obviating the need for cleaning the repository data prior to use. Absent universal use of a complete, standard, reference terminology by data providers, downstream systems must constantly cross-map data to maintain consistency across sources. The goal of this project was to develop and implement a Data Dictionary (TruData) to establish and maintain standardization, mapping integrity, and grouping across disparate clinical data sources to facilitate EHR usability, data exchange, and data warehousing. The mappings produced are exportable to other institutions and TruData itself can be used as a service to normalize multi-institutional data sources.

Laboratory tests provide key data for physicians and clinical researchers. In an ideal world, all data providers would use a complete reference terminology that defined the test/result code a priori according to a standard definition of “truth.” Unfortunately, resulting agencies often use different codes and varying descriptors for the same test/result - even within their own libraries. In addition, commonly deployed ontologies such as LOINC are inadequate for mapping results from multiple agencies because of inconsistencies and/or lack of assignment of codes, incomplete test coverage, and delays in releasing new codes.

At our quaternary medical center, clinicians can order over 12,465 studies from 19 different agencies. TruData determined 3000 of these are distinct while the remaining 9,465 can be mapped to 3,100 common entities.

We designed a system that establishes and sustains a system-wide equivalency map via employment of web crawlers and profiling engines to parse, iterate, infer, and process raw data – yielding mathematically ranked equivalency proposals. As new or updated test/result components are detected, the system parses and automatically generates optimal concept representations to propose LOINC codes for results, and collectively employ to generate ranked equivalent tests across the various resulting agencies. Each proposal is scored by a Bayesian probability and an Interaction Binding Strength (IBS) score/ratio that measures the strength of the proposed matching pairs and its relevance to the overall mapping cluster. To establish the best fit, a lattice of equivalencies across codes (known as the Data Dictionary mapping cluster) from n different sources can be graphed in n-dimensional space, ideally producing a single point in that “DD Space”, called TruPoint, indicating mutually affirmed equivalency (Figure 1). This calculated unity is, in effect, derived “truth.”

TruData now contains over 63,500 local result components from over 19 systems, with 45,000 assigned LOINC codes, close to 9,300 of which are active. We are currently working with other schools to export these mappings or potentially integrate their data metadata directly into TruData to assist with local and national interoperability or clinical and clinical research data.
The Northwestern Medical Enterprise Data Warehouse (EDW) has developed a very successful governance and operating model that supports the combined missions of an academic medical center, including translational research. Administratively housed in the Northwestern University Clinical and Translational Sciences (NUCATS) Institute, the EDW supports research, medical education, clinical quality, and healthcare operations. It is jointly funded and governed by the Northwestern Memorial Hospital, Northwestern Medical Faculty Foundation, and Northwestern University.

When developing a strategic plan to address the analytic needs of the campus, Northwestern elected to create a single data warehouse instead of having a data warehouse dedicated solely to the research mission. In doing so, Northwestern has enjoyed a number of benefits. (1) The cost to build and maintain the EDW is shared by the partner organizations. (2) By having a single data repository for reporting and analytics on the campus, users are no longer faced with the dilemma of receiving conflicting data from multiple repositories that must be reconciled through multiple processes. (3) If data quality issues are encountered, there is a single process for documentation and resolution. Finally, in an academic medical center environment where research discoveries are translated into clinical practice, having a single data warehouse is critical; it greatly reduces the time and complexity for implementing research discoveries in the clinical setting because the infrastructure does not change.

To accomplish this goal, Northwestern tackled the dual topics of “plumbing and politics” in parallel. The EDW team created a set of fundamental principles to address how data would flow into the database, ensuring data quality and utility without sacrificing the speed with which data are made available to end users. Simultaneously, the EDW team worked with its sponsors and leadership to create a governance structure that was lightweight enough to be agile while still retaining effectiveness through the roles required for membership. The EDW governance structure has oversight for the strategic plan, changes to the revenue model, budgeting and funding, and policies. Through this infrastructure, the EDW has developed policies and procedures for data use and quality, which has empowered the contributors of the data to retain control of how the data are used and engender trust across the campus partners.

Data warehouses and data analytics play a crucial role in the CTSA strategic goal of enhancing the research infrastructure in the US. Without access to data, research is all but impossible. Data warehouses enhance the ability to access large data sets and accelerate the analysis component of research. The dual roles of “plumbing and politics” are critical to the success of a data warehouse. Without appropriate data management, faith in the EDW is sacrificed. Without an appropriate degree of governance, an EDW can be crippled by politics. Northwestern has had great success with both plumbing and politics and has a model that can assist other institutions encountering challenges with their data warehouses.
A retrospective research model enabled by an open-source integrated research data repository  

Gregory W. Hruby, James McKiernan, Susanne Bakken, Chunhua Weng  
The Irving Institute for Clinical and Translational Research, Columbia University, New York, NY

Research Description  
At many academic centers, retrospective outcomes research represents a sizable if not the majority of contributions to the academic literature. In an effort to increase the capacity and quality of retrospective outcomes research, we implemented an integrated central research data repository (CDR), CAISIS\(^1\), in the Department of Urology at Columbia University. CAISIS is an open source web-based data management system that integrates research with patient care. The implementation of the CDR was a catalyst for transforming the department’s retrospective research methodology. Prior to the CDR, retrospective research was a manual process and the majority of effort was directed at dataset generation for specific research questions. The reuse of these datasets in future work was limited. In order to promptly respond to increased demand on the CDR, a research model was established (Figure 1). It contains six steps: idea generation, feasibility assessment, statistical consultation, dataset generation, data analysis, team review, and manuscript preparation. Additionally, figure 1 shows the core team players involved at each step. Of note the CDR enables feasibility assessment and data enrichment through an iterative data request process.

Results:  
After implementing this model, research project completion time decreased from an estimated average of 3-12 months using the old model to 1-6 months currently. Outcomes research productivity increased from 25 papers to 28 papers per year. This corresponds to research capacity enhancement from supporting 3-4 researchers to 9-11 researchers. We will also present a case study that compares two research projects testing the same research hypothesis before and after the implementation of the new research model respectively but resulting in different answers. The success in one academic department easily led to the adoption and implementation of this new research model in other interested departments in our CTSA. We conclude that such a new research model can improve efficiency and quality for retrospective outcome research in academic institutions. We will also discuss the best practice for building quality research data set and ensuring data integrity in a federated research data repository.

Importance to the CTSA  
This experience shows the effect of a high-fidelity CDR on a department’s research methodology. The CDR can drive the augmentation of practical research methodologies to more closely emulate the ideal. This unexpected benefit of a CDR that allows for the ideal to become practical should be further explored and developed in other clinical settings.

Impact on the CTSA Consortium Strategic Goals  
The above research model impacts three CTSA strategic goals: (1) Build National Clinical and Translational Research Capacity, (2) Provide Training and Career Development of Clinical and Translational Scientist, and (3) Improve the Health of our Communities and the Nation. A CDR that is integrated with the clinical information systems mitigates the daunting tasks of data management and collection that many new clinical scientists fear. As such, their attention can be turned to study design and methodology towards high quality training of the next generation of research scientist. Finally improved best clinical practices are then disseminated within practicing communities and thereby improves the quality of care in our communities.

Acknowledgement: UL1 RR024156 from the National Center for Research Resources (NCRR).

\(^1\) http://www.caisis.org/
Harvest: An Open-Source Biomedical Data Integration and Discovery Platform

Jeffrey Pennington, Byron Ruth, Michael Italia, Jeffrey Miller, Peter S. White
Center for Biomedical Informatics, Children’s Hospital Of Philadelphia, Philadelphia, PA

Description

Biomedical researchers share a common challenge of making complex data understandable and accessible. This need is increasingly acute as investigators seek opportunities for discovery amidst an exponential growth in the volume and complexity of research data, derived both from advances in experimental technology and the availability of large data stores relevant to research. To facilitate data discovery, we have developed a novel, open-source toolkit called “Harvest”, which provides a data integration platform and intuitive interface for exploring, querying, and reporting complex clinical, genomic, and imaging data. The Harvest interface promotes a rapid, iterative approach to data discovery by using visual representations of multidimensional data types. These visualizations present data in a context and format easily recognized by clinicians and researchers, hiding the underlying complex data models required to organize and store biomedical data. The Harvest architecture addresses multiple functional needs critical to a research and development environment, including domain-specific data modeling, abstraction of complex data models, a modular, customizable web client, and export to popular statistical analysis software file formats. Harvest leverages modern, standards-based open-source Web technologies. We have demonstrated the utility of Harvest by developing and deploying multiple applications across several research domains, including audiology, imaging, genetics, and cardiology. Harvest is available for adoption as an open-source project under the BSD license with a growing development community at http://www.github.com/cbmi

Significance To The IKFC

Many CTSA informatics groups are faced with the need to develop targeted, domain-specific repositories integrating highly complex, multidimensional biomedical data. An open-source solution for construction of these repositories provides Consortium members with an opportunity to collaboratively address shared challenges.

Impact On CTSA Goals

Harvest impacts the national clinical and translational research capability through it’s use in support of multiple NIH-funded, national-scale data repository projects including NIDCD’s Audiology and Genetics Database (http://audgendb.chop.edu) and NHLBI’s PCGC Data Hub.

Consortium-wide collaborative development and adoption of Harvest is facilitated by 1) licensing under the unrestricted Open Source Initiative-compliant BSD license and 2) project hosting on GitHub, a code-hosting site popular for facilitating cross-organization development workflows such as forking projects, sending and pulling requests, merges, and monitoring development.
Abstract 13

Using Electronic Health Records to Identify Cohorts for Drug-Induced Thrombocytopenia, Neutropenia and Liver Injury

Jyotishman Pathak, PhD1  Abel N. Kho, MD2  Joshua C. Denny, MD3  Kevin C. Bruce, BS1  Sean P. Murphy, BS1  Matthew J. Durski, BS1  Christopher G. Chute, MD, DrPH1

1Mayo Clinic, Rochester, MN  2Northwestern University, Chicago, IL  3Vanderbilt University, Nashville, TN

Objective: Develop and evaluate algorithms to high-throughput phenotyping from EHRs (Electronic Health Records) for symptoms and findings associated with drug-induced thrombocytopenia, neutropenia and liver injury.

Background: EHRs are a rich source of longitudinal data about patient’s health and wellbeing. Developing techniques and algorithms for extracting data captured through routine clinical care in EHR can facilitate secondary uses of EHR data for clinical research, healthcare quality, and comparative effectiveness. In particular, our objective is to establish a library of EHR-based algorithms that can identify patients with symptoms and findings, who may have had specific drug exposures. The potential to establish whether genomic characteristics modulate the frequency or intensity of these manifestations could be explored using these modular phenotypes.

Project Aims: 1) To work with the CTSA and pharmacogenomics communities to demonstrate the utility of EHR for identifying drug-induced phenotypes; 2) To develop and validate EHR-based algorithms with high sensitivity and specificity for identifying phenotypes commonly associated with drug exposures; and 3) To evaluate the generalizability and performance of algorithms developed at collaborating CTSAs: Northwestern and Vanderbilt.

Accomplishments in 2010-11: We have developed algorithms for drug-induced thrombocytopenia, neutropenia and liver injury. These algorithms are currently being implemented at Mayo, Northwestern and Vanderbilt CTSAs. The major determining criteria include lab measurements, billing, diagnoses codes, and procedure codes, as well as medications, comorbidities and signs and symptoms. Depending on the institutional EHR environment, such data can be extracted via structured queries (e.g., SQL, SAS) or using natural language processing (NLP) techniques. For instance, medication data was extracted using the cTAKES1 NLP engine at Mayo. Preliminary validation of accuracy and specificity for these algorithms at Mayo identified issues that led to refinement of phenotyping criteria (e.g., classifications of larger drug categories versus specific drugs with a specific chemical mechanism that targets certain cells which, in turn, are a greater risk of developing neutropenia). Our objective is to perform a more comprehensive validation at Northwestern and Vanderbilt, perform further algorithm refinements based on validation results, and eventually create a public library giving access to these algorithms.

Participating CTSAs: Mayo Clinic, Northwestern University, and Vanderbilt University.

Why Important for IKFC Meeting
Our project is a strong collaboration of multiple CTSAs developing novel approaches for high-throughput phenotyping from EHR data for secondary use. Related projects, such as eMERGE2 and SHARP3 funded by DHHS/NIH (of which Mayo, Northwestern, Vanderbilt are members) are pushing the bounds using such approaches for genomics research and healthcare quality measurements. In this poster, we will present preliminary results in algorithm development, validation, and lessons learned. These algorithms are being implemented with participating CTSAs, and will made publicly available to the entire CTSA community.

Impact on CTSA Strategic Goals

Goal #1 Enhancing national clinical and translational research capability: A library of robust EHR-based algorithms for phenotype identification and associated infrastructure will be critical to support clinical research, healthcare quality measurements, and comparative effectiveness research.

Goal #2 Enhancing the training and career development of clinical and translational scientists: With Meaningful Use paving the path for widespread adoption of EHRs within US, the methods developed in this project can be used to train investigators in using complex EHR systems for their own clinical work and research studies.

Goal #3 Enhancing consortium-wide collaborations: EHR-based phenotyping algorithms are of little value unless they have been implemented and validated across multiple EHR environments and institutions. Hence, by definition, this project requires CTSA-wide collaborations in developing robust, generalizable and scalable algorithms.

Goal #5 T1 Research: As evidenced in eMERGE, projects such as this can greatly facilitate genome-wide association studies in identifying novel genotype-phenotype relationships as well as replicating the existing ones.

References


iPads in clinical research: time-motion and secure offline data collection

Northwestern University Biomedical Informatics Center has developed apps that enhance data collection in clinical research: TimeFlow and Surveyor iOS. We are eager to share these widely available and easily adopted tools that leverage hardware innovation for clinical research infrastructure at the 2011 IKFC meeting.

While time-motion studies are the gold standard for work-flow observation, observers are often hampered in their choice of tools to measure multi-tasking. Paper, laptop, and even handheld tools fall short of the scope and variation of activities of a typical observation, especially in the clinic. In collaboration with Hospital Medicine at Northwestern Memorial Hospital, we created an iPad time-motion tool that captures nonlinear data in a clinical setting. Our development was guided by ethnographic observation of user experience conforming to the agile programming methodology, and validated by extremely high inter-rater reliability. TimeFlow is an essential component of process research infrastructure, and extends the gold standard for measurement of clinical processes and capability.

TimeFlow will be widely available via the Apple App Store, and adoption is a simple matter of acquiring the tablet and purchasing the app. Researchers may begin data collection immediately upon customizing activities and categories for their protocol directly on the device.

The benefit of electronic data collection is already being leveraged through web-based assessments, surveys, and case report forms. But the lack of secure, offline collection tools, both for in-clinic and remote use, often motivates a regression to paper-based instruments. Surveyor iOS is a mobile, offline, FISMA-compliant electronic data collection tool that securely authenticates and syncs data back to the server when a network connection is re-established. Coordinator work flow and participant management tools are also being integrated into the app. Electronic data collection is an essential part of research infrastructure, and Surveyor extends its reach to secure offline use in the community.

Surveyor iOS and the Surveyor server are open-source and available today via github at https://github.com/NUBIC. Instruments may be imported from a variety of other tools or written in a Domain Specific Language (DSL), shared with other researchers, and used immediately upon device sync. The open-source nature of the tools and instruments promotes sharing and standardization throughout the clinical research community.
Abstract. *eagle-i* is a web-based application suite enabling scientists to discover resources across a distributed network. The cornerstone of the *eagle-i* software is an ontology-driven data model that supports powerful search methods while maintaining flexibility and interoperability. We discuss our architecture, approach, and expansion.

**Introduction.** The majority of existing research resources cannot be found easily. This longstanding problem involves technical, motivational and data quality considerations; *eagle-i* has been developed with these complexities in mind. By removing barriers to resource discovery, *eagle-i* reduces expensive resource duplication while accelerating the development of diagnostics, treatments, and prevention strategies.

**Workflow** at each of the *eagle-i* consortium institutions started with a local Resource Navigator who promoted *eagle-i* and helped investigators to identify and document their resources. Since May 2010, nearly 46,000 resources have been collected—including Organisms and Viruses, Reagents, Instruments, Biological Specimens, Services, Software, Protocols, Core Laboratories, Human Studies, and Research Opportunities. Collected resource records were then handed over to the curation team to ensure data quality and consistency. All of these curated published resources are now freely searchable at www.eagle-i.net.

**The *eagle-i* architecture** comprises a set of ontology-driven software components deployed at each institution, as well as a central search application that communicates with these federated components. At the core of each institutional deployment is a Resource Description Framework (RDF) repository. Built around semantic web technologies, *eagle-i* follows linked open data (LOD) principles.

**The *eagle-i* data collection tool** produces well-structured resource descriptions that include text, annotations with ontology concepts, and links to other resource instances. The tool also provides a workspace for updating and managing resource descriptions individually and in bulk. Forms and menu choices are dynamically generated for each resource type based on its corresponding ontology class.

**The *eagle-i* search application** back-end is based on a semantic search framework that leverages Solr for rich functionality and inferencing. It also accommodates search plugins for external databases such as NCBI and NIF. The search front-end is a dynamic web application offering categorical search and synonym expansion. In addition, an autosuggest feature provides real-time visibility into available resources and ontology terms.

To promote interoperability and best practice **the *eagle-i* ontology** has been developed in ongoing collaboration with numerous other groups, including LAMHDI, VIVO, NIF, StarMetrics, and Biositemaps. Because ontologists model knowledge as it evolves, *eagle-i* has been built to gracefully adapt. Additionally *eagle-i* makes use of application-specific ontology annotations that bridge between ontology concepts and user.

A recent survey of 259 users showed that 91% of users were satisfied overall with the *eagle-i* search. 71% felt that *eagle-i* would be particularly useful to them if scaled to institutions nationally. The vast majority of respondents (97%) learned about resources they did not know were available. 63% said they would be likely to contribute their lab’s resources. The search/browse functions, auto suggest, and filtering features were all rated as “excellent” or “good” by most respondents. In addition to user feedback, we have received practical suggestions from two usability studies and an expert review; changes and features have been planned accordingly.

Although critical to the advancement of science, resource sharing is under-attributed and not easy to measure systematically. The BRIF workshop group has proposed to implement a **bioresource research impact factor as an incentive to share** human bioresources (*Nat Genetics 2011*). Because each resource in *eagle-i* has a globally-unique URI that can be referenced, *eagle-i* would be a natural platform for implementing a BRIF.

**eagle-i goes open source.** On August 31, 2011, the *eagle-i* prototype was made available through an open source (BSD) license. Currently, there are a two ways to deploy *eagle-i* from scratch (source code) or from installable files. To make deployment even easier, we anticipate offering *eagle-i* as a pre-configured virtual machine and via a hosted (cloud) model. Because *eagle-i* is designed as a federated system, all of these deployment options provide institutions with technical and administrative autonomy. Each institution can choose to make their resources locally discoverable, or globally discoverable at www.eagle-i.net. For information on joining the network, visit http://open.med.harvard.edu/display/eaglei.
The ITHS Central Database: A Customer Relationship Management System for CTSAs

Jim Piper

The Institute of Translational Health Sciences (ITHS), like other CTSAs around the country, is faced with the challenge of collecting and managing data needed to produce: Annual Progress Reports, Competitive Renewals, Bills for services rendered, and internal reports regarding researchers, publications, and grants being assisted by the various cores within the ITHS. The challenge also includes the need to securely provide this tool to users with differing roles from a variety of institutions from all over Alaska, Idaho, Montana, Washington, and Wyoming. The ITHS Central Database (CDB) is a web-based computer system developed by the University of Washington to tackle these challenges using free open source software: Shibboleth, java, jee, jboss, mysql and could be replicated at other CTSAs.

Conceptually the CDB meets these challenges by defining five top level data categories: Persons, Projects, Publications, Grants and Service Descriptions. It then defines the relationships that these categories have with one another. With this schema the system stores the profiles of several thousand people, and for each person can store details about publications authored (CDB incorporates PubMed searching), grants received, and the projects/studies participated in. To tie all these together the system allows ITHS staff from all cores to help ‘tell the story’ of a researcher’s involvement with the CTSA by regularly creating Service Description records documenting the services they provide to individual researchers and studies.

Over time ‘telling the story’ of a researcher’s involvement with the ITHS has several benefits. It allows the ITHS to determine not only ‘who is reportable for the APR?’ but to also show ‘why has this person been marked reportable?’ Also, because all ITHS staff have access to the CDB it helps staff from all cores stay informed as to what the other cores have been doing. In addition, because CDB allows ITHS Staff to specify a particular project/study that they assisted a researcher with, combined with the list of publications that came out of that project/study, enables CDB to provide information regarding the numbers of publications assisted per ITHS core.

Another benefit of the CDB is in collecting and confirming the required investigator roster details needed for the APR such as: eRAComomns Name, Area of Expertise, list of publications assisted, list of grants assisted, demographics, and T1/K2 scholar details. It does this by allowing researchers themselves from all over the Pacific Northwest to securely login to CDB and update their own profiles. The Shibboleth federated single sign-on software the CDB is able to allow users to login with their existing institutional Shibboleth logins or if not already Shibboleth to acquire a Shibboleth login from the publicly available and free ProtectNetwork.
The Human Studies Database (HSDB) Project

Lead institution: UCSF

Objective: A federated searchable 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 (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 data mining, synthesis, re-analysis, and research networking.

Project Aims: 1) define and validate scientific features of human studies in the Ontology of Clinical Research (OCRe); 2) define the HSDBgrid data federation architecture; 3) collect descriptions of human studies from individual CTSAIs and federate over HSDBgrid. We will federate results data later, contingent on acceptable data ownership policies. Accomplishments in 2010-11: We have extended OCRe to include a detailed modeling of outcomes, are continuing to evaluate our study design typology, and have harmonized OCRe's administrative data with BRIDG. OCRe is available on BioPortal as an OWL 2.0 ontology. For data federation, we have moved from an approach centered around CDEs and UML models to one primarily centered around XML, RDF, and semantic web standards (Fig.1). This project drives and uses many tools from other projects, including BioPortal, LexEVS, Dynamic Extensions from the cаTISSUE Suite, VIVO, VITRO, and Query Manager to define, serve, and federate semantic standards and value sets. The priority in the upcoming year is to deploy and test an end-to-end approach in more than one institution.

Participating CTSAIs: Duke; Hopkins; Mayo; Ohio State; Rockefeller; Stanford; UC Davis; UCSF; U Colorado; UT Southwestern; UTHSC San Antonio, U Washington; WUSTL.

Why Important for IKFC Meeting

Our project is a collaboration of multiple CTSAIs pushing the bounds on a workable approach to semantically standardized data sharing of valuable but human studies complex information. We will present OCRe and an end-to-end demo of how we bind OCRe to permissible value sets, acquire instances using a web interface, and perform distributed queries using OCRe semantics. These tools are being tested with participating CTSAIs, and will soon be available to all.

Impact on CTSA Strategic Goals

Goal #1 Enhancing national clinical and translational research capability: A computable HSDB inventory of past and ongoing human studies that will provide critical infrastructure for clinical research, scientific query and analysis, and 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 is at the forefront of using semantic technologies to implement scalable, general, platform-independent data sharing across institutions. We have also begun to explore integrating HSDB with research networking projects.

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.
Electronic Capture and Management of Informed Consent for Research
Aziz A. Boxwala, MD, PhD, Joanne Barker, MS, Elizabeth Johnstone,
Amarnath Gupta, PhD, Xufei Qian, Mona Wong, Lucila Ohno-Machado, MD, PhD
Division of Biomedical Informatics, University of California San Diego

There is increasing reuse, in research studies, of specimens and data obtained previously, either during clinical care or in other research studies. While this type of reuse can make research more efficient, the practice has created challenges in adhering to the autonomy principle of research ethics\(^1\). The types of future studies that might reuse the data or specimens cannot be anticipated at the time these were obtained and the subject initially consented for their use. This could lead to a conflict between the subject's autonomy and privacy, and the public interest in the benefits of the research\(^2\).

We are developing a consent management system comprising a consent registry, a consent broker, and a consent elicitation tool. The consent broker, currently being designed, will make it simpler for investigators to use previously acquired data and specimens without conflicting with subjects' wishes. The investigator can use the consent broker to determine if a proposed use of data or specimen in a study complies with consent recorded in a registry. We are creating an extensible scheme for codifying the consent within the registry. The coding scheme is multidimensional to reflect the many aspects of a subject's concerns, representing, for example, the disease category being researched, confidentiality of the subject's identity, the sensitivity of research, the researchers or institutions that may use the data or specimen, and the desire for and frequency of future contacts. The major benefit of this codification scheme will be in providing a common language to capture consented uses of data and specimens, enabling sharing of these across repository boundaries. The electronic informed consent elicitation tool that has been developed offers subjects' options and elicits their choices in the uses of data and specimens and their desire to be contacted for future research or for learning about research findings. This web-based application is accessed via a touchscreen tablet device. The study information (i.e., the material traditionally included in a consent form) can be presented in different languages and can be navigated by the subject in a non-linear manner. As the subjects review the study information, they can express their choice from tiered options\(^3\). Subjects complete an interactive assessment after they have reviewed the information to help us evaluate their comprehension. Once the comprehension of the information is demonstrated, the subject and the investigator sign the consent form using the touch screen interface. The consent choices and signatures are recorded in a codified manner in a consent registry, and on a Portable Document Format (PDF) form, digitally certified by the Institutional Review Board, which is then printed out for the subject.

The system has the potential to impact clinical and translational research studies at our institution and others by supporting utilization of specimens and data that adheres to the desires of individual subjects. We intend to make the software and the consent ontology widely available for use at all institutions.

Acknowledgment: The project is supported by grant U54 HL108460 from NIH.

References
Harvard Catalyst Profiles: Research networking, bibliometric analysis, and social network analysis based on linked open data and VIVO ontology.

Griffin M Weber, MD, PhD; Nicholas Benik; Kenneth Huling; Shashank Jain; James Norman; Marlon Violette; Steve Wimberg; Harvard Medical School, Boston, MA

Introduction. Harvard Catalyst “Profiles” (http://profiles.catalyst.harvard.edu) is an open source research networking (RN) website developed by Harvard’s Clinical and Translational Science Center, which uses semantic web technologies. It creates research profiles for more than 22,000 faculty at Harvard, and the software is used at several dozen other institutions across the country. Profiles is the first widely available RN tool (other than VIVO itself) to adopt the VIVO ontology for linked open data (LOD). In this presentation, we will provide an overview of the functionality and benefits of Profiles, and discuss the potential impact of moving towards the VIVO ontology standard on both the developers of research networking products and the researchers using these systems.

Profiles Features. The Profiles software creates investigator profiles using a variety of external data sources, such as PubMed, and multiple internal systems. These profiles are linked together through Passive Networks, which are automatically generated based on information known about investigators. For example, we extract keywords from publications and use this to build networks of people with similar interests. Users can also create Active Networks, by looking up people they know and describing their relationships to them, such as “collaborator” or “advisor”. Profiles contains several features that make it an “out-of-the-box” solution for institutions: 1) Automatic Author Disambiguation. Profiles uses sophisticated multi-factorial matching algorithms to build a publication history automatically for each researcher in an institution. Users can easily correct mistakes, and Profiles learns from these changes to improve the results of the next automated literature search. 2) Bibliometric Analysis. Profiles analyzes publications to identify important concepts and trends, to discover connections between people and different areas of research, and to evaluate the performance of individuals or entire programs or institutions. 3) Social Network Analysis. Profiles contains numerous ways of visualizing and exploring networks, and it automatically calculates “centrality” metrics that show how many collaborators a person has, how strong those collaborations are, and whether those collaborations connect different parts (e.g., departments, disciplines, etc.) of an organization. Using this information, Profiles can provide personalized search results that recommend people who not only have expertise in a particular subject area, but would also likely be good collaborators.

VIVO Ontology. An important aspect of Profiles is that it is now ontology based, utilizing LOD and the research networking ontology developed by the VIVO Consortium. LOD provides a standard model for publishing structured data and exchanging it openly, thus forming a “semantic web”. We faced several challenges in converting Profiles to use LOD: 1) Ontology. We had to adapt to changes in the VIVO ontology over time and extend it with new classes and properties. 2) SPARQL. Numerous open source SPARQL engines exist, and we had to select one that would best integrate with Profiles. 3) Scalability. The Passive Networks in Profiles causes an explosion in the number of triples. We also noticed some immediate benefits to using LOD: 1) Domain expansion. We initially built Profiles specifically for Harvard’s medical school. That data model made it difficult to expand beyond biomedicine. Converting to LOD enabled us to extend Profiles to the rest of Harvard’s faculty, including their courses, activities, projects, and websites. 2) Search. With data stored as triples, it was easy to implement a search engine that indexed different types of content and enabled faceting.

Importance to the IKFC and CTSA. Most CTSA institutions either have research networking software or are planning to implement them, and SGC#3 lists research networking as a top priority for CTSA. This presentation serves two purposes: 1) To introduce new people to the capabilities of research networking using Profiles as an example. 2) To demonstrate that VIVO/LOD are becoming true standards in research networking by the fact that another software product used at several CTSA institutions has adopted the ontology.
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
Theme: Integrated Data Repositories
Title: Demonstration of the i2b2-based Health Outcome Monitoring and Evaluation (HOME) Cell

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 currently lacks functionality related to health care process and outcome assessment. We have developed a new cell capable of evaluating a wide variety of health outcomes with potential for use in comparative effectiveness and other health services evaluations. The Health Outcome Monitoring and Evaluation (HOME) Cell has been developed as a shared resource within the i2b2 community that will be improved over time. The Cell functions within i2b2 with no changes to the core i2b2 schema and only modest changes to core i2b2 software. The HOME Cell reuses queries developed within the i2b2 Query Cell and extends functionality through user-specified “constraints” within the HOME Cell.

During this presentation we will provide an in-depth demonstration of the core architecture and software functionality of the HOME Cell as well as preliminary results from ongoing analyses.

Why this presentation is important for the 2011 IKFC Meeting:

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. Presentation of the software to experts could serve to inform more potential collaborators and users as well as support ongoing improvements of this shared clinical and translational research resource.

Impact on the CTSA Consortium Strategic Goals:

Our efforts support multiple strategic goals including: 1) support for building national clinical and translational research capacity by enhancing a broadly adopted informatics tool (i2b2); 2) enhanced consortium-wide collaborations by offering a tool that can be easily shared within the CTSA Network to support multi-institutional collaboration; and 3) improving the health of our communities by offering a tool that has the potential to provide new insights into health care processes and outcomes that could drive innovation and improvement activities.
G-CODE: Enabling systems medicine through innovative informatics

Subha Madhavan, Yuriy Gusev, Michael Harris, David M. Tanenbaum, Robinder Gauba, Krithika Bhuvaneshwar, Andrew Shinohara, Kevin Rosso, Lavinia A. Carabet, Lei Song, Rebecca B. Riggins, Sivanesan Dakshanamurthy, Yue Wang, Stephen W. Byers, Robert Clarke, Louis M. Weiner

The new and emerging field of Systems Medicine, an application of Systems Biology approaches to biomedical problems in the clinical setting, leverages complex computational tools and high dimensional data to derive personalized assessments of disease risk. Systems Medicine offers the potential for more effective individualized diagnosis, prognosis, and treatment options. Achieving this goal requires the effective use of petabytes of data, which necessitates the development of new types of tools. The Georgetown Clinical & Omics Development Engine (G-CODE) is a generic and flexible web-based platform that serves to enable basic, translational, and clinical research activities by integrating patient characteristics and clinical outcome data with a variety of high throughput research data in a unified environment to enable Systems Medicine. Through this modular, extensible, and flexible infrastructure, we can quickly and easily assemble new translational web applications with both analytic and generic administrative features. New analytic functionalities specific to the needs of a particular disease community can be easily added within this modular architecture. This generalizable approach to disease-centric functionality is designed specifically to permit the rapid and cost-effective generation of new portals to satisfy the varied needs of the scientific community. Through G-CODE, we hope to help enable the widespread use of biomedical informatics tools by basic, clinical, and translational researchers by providing powerful analytic tools and capabilities within easy-to-use interfaces, customized to the needs of each research community.

This infrastructure was first deployed in the form of the Georgetown Database of Cancer (G-DOC®; http://gdoc.georgetown.edu), which includes a broad collection of bioinformatics and systems biology tools for analysis and visualization of four major “omics” types: DNA, mRNA, microRNA, and metabolites. While several rich data repositories for high dimensional research data exist in the public domain, most focus on a single data type and do not support integration across multiple technologies. It contains data for more than 2,500 breast cancer patients and almost 800 gastrointestinal cancer patients, all handled in a manner that enables maximum integration. We believe that G-DOC will help facilitate systems medicine by providing easy identification of trends and patterns in integrated datasets, and hence facilitate the use of better targeted therapies for cancer.

One obvious area for expansion of the G-CODE/G-DOC platform infrastructure is for the support of next generation sequencing (NGS), a highly enabling and transformative emerging technology for the biomedical sciences. Nonetheless, effectively utilizing this data is impeded by the substantial handling, manipulation, and analysis requirements that it entails. We have concluded that these are gaps that cloud computing is well positioned to fill, as this type of infrastructure permits rapid scaling with low input costs. As such, the Georgetown University team is exploring the use of the Amazon EC2 Cloud and the Galaxy platform to process whole genome, RNA-seq, and ChIP-seq NGS data. The processed NGS data will be integrated within the Georgetown Database of Cancer to ensure that it can be analyzed in the full context of other “omics” data. Likewise, all G-CODE projects will simultaneously benefit from these advances in NGS handling. Through technology re-use, the G-CODE infrastructure will accelerate progress for a variety of ongoing programs in need of integrative multi-omics analysis, and advance our opportunities to practice effective systems medicine in the near future.
Integrating Research Networking Tools with Broader Social Networking Products to Create an Improved Environment for Accelerating Science

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

Research Networking Tools have experienced widespread adoption across the CTSA consortium. The technologies that power these tools range from open source applications built within our community such as VIVO and Profiles, to commercial products like Elsevier SciVerse, to locally built applications such as Iowa’s Loki and Stanford’s CAP. We have been quick to recognize the advantages of our research networking tools as compared to more commonly known social networking products such as Facebook, Google+ or LinkedIn which lack the data provenance and domain specific functionality that our tools provide. However, there is valuable functionality in many of the mainstream social networking products and services available today, particularly around collaboration, mobile presence and the ability to connect individuals across the Internet regardless of institutional boundaries. Functionality to create ad-hoc groups with forums and document repositories, and to “follow” communication threads centered on a person or topic, applies as well to the process of authoring a grant proposal amongst researchers as it does to hosting a volleyball tournament for Facebook friends. Delivering this type of functionality to our researchers as an extension of our research networking tools, which already contain the personnel, their professional data and interconnections, is an attractive choice. But we should not attempt to reinvent common social networking functionality in our research networking tools, as that is expensive and unnecessary. Existing products which target broad domains and thus have large technical resources have solved these problems. We have successfully integrated Profiles with off-the-shelf products, both open source and commercial, to meaningfully extend the functionality of our research networking tool beyond what is available in any single product.

Our first integration was with Apache Shindig, which allowed our installation of Profiles to become a platform for hosting OpenSocial gadgets. Being a gadget platform has allowed us to create a number of extensions to Profiles to address needs specific to mentorship, email list generation, enhanced search, and more, many which can be seen in our production server at http://profiles.ucsf.edu. Our second integration is to use Salesforce Chatter as a collaboration platform and for mobile integration. With Salesforce Chatter and OpenSocial, we were able to implement “activity streams” to showcase events occurring within the site such as who has uploaded a photo, when the disambiguation algorithm has found new publications, and how many times a researcher’s profile has been viewed. The ability to selectively follow a researchers’ activities and to receive notifications (email, mobile, desktop), is handled completely by the Salesforce Chatter product. Activity streams are a common component of many social networking sites and can be seen on Facebook, Google+ and Twitter as well as Stanford’s CAP. An additional integration is with Drupal, a widely used content management system that can be extended by modules to layer in numerous features. We are developing a Drupal module to pull in content from our Profiles system to be displayed in a number of Drupal-powered web sites at UCSF. This Drupal module will work at any institution running Profiles; when we adopt the VIVO RDF based version of Profiles, it will be modified to work with VIVO content, possibly in conjunction with Miles Worthington’s work at Cornell.

Online collaboration is a natural fit for networking based sites and has become a proven means for advancing business; it should also be used to advance science. Despite our unique domain needs we do not need to re-invent the mechanisms required for powerful online collaboration. We can extend our current research networking tools to integrate with proven products to create “best of breed” solutions that provide a seamless experience for our researchers, target their specific needs, have institutional provenance, and can cross institutional boundaries.
A Research Participant Tracking System that Supports Business Process through System Integration

Robert Gehrke, Thomas Johnson, Michael Lin, Robert Pawluk, Rodney Brusse, Irina Alexandru, Christopher Chute M.D., Dr. P.H., George Klee M.D., Ph.D.

Institution-wide research participant tracking is highly desirable to support effective protocol execution and administration. It also makes it possible for study teams to eliminate the need to create and maintain their own participant tracking mechanisms, which often times do not adequately take into account security and confidentiality requirements.

Mayo Clinic has developed a Research Participant Tracking application based on a data model that is aligned with national standards (CDISC/BRIDG/ISO), and the functional services layer of the open source caBIG/C3PR product. Architected application program interfaces (APIs) implemented in the form of web services are provided to export and import research participant tracking data to and from study electronic data capture (EDC) tools, as well as other institutional systems such as Mayo’s electronic IRB.

The key features that will be provided by the first phase of the Research Participant Tracking are:

- standardized online informed consent and re-consent processing
- monitoring study progress using metrics such as accrual and enrollment progress to IRB targets, as well as ability to measure screen failure, withdrawal, and decline consent rates
- tracking current and historical subject status throughout the study lifecycle
- definition and application of standardized reason codes where appropriate (such as reasons for declined consent, withdrawing, screen failures, etc.)
- online verification of consent status by clinical staff delivering research intervention (eg. CT scan, muscle biopsy, etc.)
- enforcement of regulatory and compliance policies
- facilitating and improving the accuracy of research billing
- study specific and cross study reporting, as well as business intelligence and data mining

This project aligns 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 Participant Tracking system is based on Java/J2EE technologies, open source methodology, service oriented architecture, and data models semantically grounded in national standards. Integration is facilitated by providing a well-defined web service layer which can be loosely coupled to business systems and end user interfaces. 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. An application demo and discussion will provide tremendous benefits for the IKFC meeting attendees.
ENHANCING COLLABORATIVE SCIENCE USING SHARED QUERY TAGGING/SEARCHING
PLUGINS FOR THE I2B2 WEB CLIENT: THE WO_APPS

Rajani S. Sadasivam, Wayne Chan, Rebecca Kinney,
Paul Ranauro, Ralph Zottola, Thomas K. Houston
University of Massachusetts Medical School (UMMS), Worcester, MA.

Background: UMMS has expended significant effort to develop a clinical data warehouse (CDW). We use the i2b2 (Informatics for Integrating Biology and the Bedside) software to store clinical data on all patients served by UMMHC’s hospitals and ambulatory care clinics. This data currently represent more than 50 percent of central MA residents. We have over 80 investigators actively using the system; they have made requests for augmentation in functions.

The challenge: Writing comprehensive, useful queries in our i2b2 CDW is challenging even for domain experts and takes considerable time and effort. The i2b2 web client provides a drag-and-drop interface to write queries; and allows sharing of queries in the web client workspace. Investigators have asked: What happens when there are hundreds or thousands of shared queries? How does an investigator search for a query that is most relevant to them from among similar searches? Currently, investigators can organize shared queries via folders, but a query can be relevant to multiple concepts. The folder organization does not allow investigators looking for queries from a different perspective to easily find them. Current labeling functions are limited.

Innovation: We have conceptualized a Web 2.0 collaborative tagging approach to enhance the sharing and searching of queries using the i2b2 web client. Collaborative tagging is an unstructured process by which groups of individuals can add any word or phrase as a tag to identify an object (a weblog entry, a picture, or a query in this case). A group of unstructured tags is often called a folksonomy. Folksonomies can be expansive as they are unrestrained by structure, they can evolve rapidly to represent new information as new concepts are created, and they are crowd-sourced representing the multiple viewpoints of the investigators creating the tags. Because taggers create whatever they deem as the most important tags, and are not required to select from a complex tree of predetermined tags, folksonomies can be a more palatable form of data entry than selected from a complex, predetermined list of tags.

Prototype development and refinement: We have developed collaborative tagging enhancement functions, the Workspace Object Applications (WO_Apps), as a set of web plugins (WO_Tagger and WO_Searcher). WO_Tagger provides an interface to tag and describe the shared queries. Multiple investigators can look at existing tags, and add new ones. Investigators can search tags and find queries using the WO_Searcher. The resulting list/tabulation of queries can be readily exported into an Excel spreadsheet for easy references.

Presentation Objective: We will demonstrate WO_Apps and present preliminary feedback from UMMS investigators.

Importance for Presentation: WO_APPS will be of interest to the CTSA KFC community of developers that are interested in the enhancement of the i2b2 web client’s usability. Feedback on our approach from the CTSA KFC informaticians will be invaluable as we plan the dissemination of WO_Apps thorough the CTSA consortium.

Fit with the CTSA Strategic Goals: The WO_APPS project fits with the Research Infrastructure development of CTSA Strategic Goal 1 (Enhancing National Clinical and Translational Research Capability), and the social networking tools development of Strategic Goal 3 (Enhancing Consortium-Wide Collaborations).
**HUBzero Frameworks for Institutional and Project Scale Information Management**

William Barnett, Indiana University and Ann Christine Catlin, Purdue University

In this session, the presenters will cover the basic architecture and features of HUBzero and discuss a distinct use case relevant to translational information management – the Cancer Care Engineering HUB (cceHUB.org). The cceHUB is a project based HUB that accelerates cancer research by integrating workflows among 6 distributed research teams, including clinical sample collection, mass spectrometry analysis, and visual analytics.

HUBzero is unique among web portal technologies as it combines features that support information management and analytical workflows. It uses a model – view –controller architecture that allows modular applications development. HUBzero is unique in that it supports user development, sharing and collaboration of data analysis tools, with seamless access to high performance computational back ends and grids. Today, over 30 HUBs support a broad variety of disciplines with worldwide usage by over 300,000 scholars. It is a free, community source product supported by 4 institutions with new, community driven capabilities appearing regularly.

cceHUB has developed data technologies for sample tracking and management that follows the sample through transfer, storage and distribution to laboratories for preparation and analysis. Properties and protocols associated with samples and their handling can be uniquely configured for each study, and data entry forms are dynamically generated and web-accessible. Clinical teams at any number of collection centers can enter sample tracking data and annotations from their desktop browser, with data integrity assured through access control provided by the HUBzero platform. A key element of cceHUB sample tracking workflow enforces linkage of the datasets generated by laboratory instruments directly to the samples used in the analysis, and datasets fed to analysis tools also carry the sample identification. This process guarantees the permanent electronic linkage of analysis results to the original samples.

Patient data encompasses clinical information (diagnosis, treatment, surgeries, medical history, follow-up, final outcome) and demographics (diet, risk factors, environment, lifestyle). cceHUB has developed specialized data technologies for the contribution, management and linkage of patient data. Data elements and categories are defined by researchers in conjunction with the clinical teams and physicians, and the elements are used to dynamically generate data entry forms for clinicians to enter patient-related data. The forms are available for study-specific data entry at any cceHUB registered clinical center. De-identified patient data can also be transferred directly to cceHUB from electronic records stored in hospital databases.

Sample and patient databases are equipped with methods for validating and approving data contributions, annotating data and tracking provenance. Patients and samples are linked by identifiers that are carried throughout the research workflow, so that laboratory analysis can use patient information for phenotyping. The data viewer can be used to browse, sort, search, filter, plot, and download data and metadata from any cceHUB database, including the sample and patient databases. Reference links (to images and citations, for example) are available from the viewer, as well as access to statistical correlation or cceHUB tools for user selected data.

HUBzero provides flexible and customizable data and analytical capabilities that can accelerate a broad array of translational research activities. These support the strategic goal of advancing translational research, are generally available to all CTSA institutions, and have helped accelerate translational research at the Indiana CTSI.
Building Robust Research Capabilities through Integrating Clinical Study, Patient, and Biospecimen Data Repositories
Leslie D. McIntosh, Mary Uhlmansiek, David Mulvihill, Bijoy George, Mark Watson, Rakesh Nagarajan

The goal of the clinical and translational research is to rapidly convert novel advances in basic science to improvements in patient care, and relay findings from clinical studies employing such new and increasingly customized diagnostics and therapeutics back to the bench for further refinement of the disease management process. However, information on participants can be diverse and dispersed making the acquisition, management, sharing, and analyses difficult and tedious to accomplish. To address these issues, we have developed and integrated three data repository systems: ClinPortal, for web-based, clinical studies data management; CIDER (Clinical Investigation Data Exploration Repository), for patient medical record data; and, caTissue, for biospecimen inventory management. In doing so, we are addressing the CTSA Consortium Strategic Goals of “Building National Clinical and Translational Research Capability”.

All tools are web-based with graphical front-ends: ClinPortal is a clinical research data management system that allows for dynamically-built case report forms within the context of a user-defined study calendar; CIDER is a research patient data warehouse that contains comprehensive inpatient and outpatient information from Washington University and BJC HealthCare practice sites; and, caTissue is a caBIG® biospecimen informatics system for inventory tracking and clinical and pathology annotation. Integration of the three systems occurs as follows:

**Improve data entry through participant registration and user-interface (UI) integration with caTissue and ClinPortal:** Clinical studies with biospecimen data can be linked between caTissue and ClinPortal either through participant registration in either system, or through navigation across associated events for data entry in both applications using single sign-on. Through the UI, the end-user can effortlessly navigate from one platform to the other to access either clinical study or biospecimen data. Both methods facilitate data entry and reduce error in clinical research.

**Facilitate participant recruitment through CIDER and Rules Engine:** A Rules Engine has been designed to set conditions, processes, and alerts based on data housed within CIDER. Complex predicates may be set as individual steps, which then can be orchestrated within the Rules Engine to alert researchers when a patient meets criteria for a clinical study. Criteria may be set based on any available data within CIDER such as gender, age, diagnosis code, laboratory results, procedures, and prescribed medications.

**Auto-load data from CIDER into clinical report forms in ClinPortal:** Allowing for seamless data loading, electronic case report forms (CRFs) are developed within ClinPortal to match variables within CIDER, which currently supports discrete medical record data from laboratory results, vitals, billing procedures and diagnoses, and medication orders. Individuals are matched using an electronic master patient index (eMPI) to link patients/participants in each system, while variables within the CRFs are semantically mapped to medical codes used within CIDER. ClinPortal and CIDER are integrated through the Rules Engine to load the appropriate medical record data into clinical studies within ClinPortal.

Full support is given through the Center for Biomedical Informatics (ClinPortal & CIDER) and the caBIG® Tissue/Biospecimen Banking and Technology Tools Knowledge Center (caTissue). ClinPortal and caTissue Suite are available freely for use by other organizations.
**VIVO: A tool for collaboration and research discovery**

Mike Conlon, PhD¹, Kristi L. Holmes, PhD²; Michele R. Tennant, PhD, MLIS¹; Leslie McIntosh, PhD³; Rakesh Nagarajan, MD, PhD²; Curtis Cole, MD³; and VIVO Collaboration

¹University of Florida, Gainesville, FL; ²Washington University School of Medicine, St. Louis, MO; ³Weill Cornell Medical College, New York, NY

**Abstract:**

VIVO (http://vivoweb.org) is an open source, open ontology, research discovery platform for hosting information about scientists, their interests, activities, and accomplishments. VIVO supports open development and integration of science through simple, standard semantic web technologies. A number of institutions and organizations from around the world - across many disciplines - have implemented VIVO as a way of highlighting the activities and expertise of their researchers and faculty members. VIVO, originally developed at Cornell University, was expanded to support national-level networking through a $12.2m grant from NCRR to the University of Florida and six partner institutions.

Institutional VIVOs are populated with information about their researchers, allowing them to highlight areas of expertise, display academic credentials, visualize academic networks and display information about publications, grants, teaching, service, and more. Data can be programmatically imported from authoritative data sources such as institutional records, and bibliographic and grant databases. VIVO can help identify potential collaborators, resources, events, programs, and facilities on the local campus and beyond. VIVO offers a way to stay up to date in a research area and on the efforts of colleagues and competitors, alike. VIVO can even play a role in recruitment by serving as a resource to attract new students, trainees, faculty and even community groups and potential partner organizations.

VIVO provides network analysis and visualization tools to maximize the benefits afforded by the rich data available in VIVO, including co-author and co-investigator visualizations, temporal representations of publications and grants by groups within an organization, and most recently, a Science Map visualization, which allows users to visually explore the scientific strengths of a university, school, department, or person in the VIVO instance. VIVO enables high quality data to be revealed about researchers, their collaborators, their funding sources, and more. These data can serve as the foundation for further network analyses and elegant visualizations of the research enterprise on the individual, local, and global levels.

By storing data in VIVO in RDF and using standard ontologies, the information in VIVO can either be displayed in a human-readable web page or delivered to other systems as RDF. This allows the open researcher data in VIVO to be harvested, aggregated, and integrated into the Linked Open Data cloud, making discovery across VIVOs and other compatible platforms targeted and efficient.

VIVO enjoys a robust open source, open community space on SourceForge. The VIVO software and ontology are publicly available at http://vivo.sourceforge.net along with content that supports implementation, adoption, and development efforts around the world. VIVO and other compatible applications produce a rich network of information that can be searched to foster collaboration and enable open research discovery at all stages of the translational cycle.

VIVO: Enabling National Network of Scientists is supported by NIH Award U24 RR029822.
Abstract

Using byproducts of the health care system in the form of electronic medical records, SHRINE provides the ability to aggregate large populations of human subjects while maintaining patient privacy. Prior to this work it was necessary to establish policy agreements and technical systems, both of which are now approved for use at Harvard affiliated teaching hospitals and other locations across the US. Now that these systems are built and deployed, it is now possible to address a wide range of T1 research goals. This presentation emphasizes meaningful scientific utility for multi-site research of both rare and common diseases.

System Description

SHRINE enables authorized investigators the ability to query millions of patient observations to aggregate large numbers of human subjects matching specific research criteria. SHRINE has been IRB approved for use at Harvard and other CTSAs across the country. Over 60 institutions, many of which are CTSAs, are evaluating or using the SHRINE software to enable translational research for a wide range of research interests. The software is available for free with a growing open source community.

Goals of presenting at IKFC

Scientific utility and meaningful research use is the principle focus of this presentation:

1) present results of completed medical studies enabled by SHRINE;
2) pose research questions that can now be addressed via SHRINE enabled networks; and
3) show how SHRINE is being evaluated for other T1 goals such as clinical trials.

In this context, we will show how SHRINE has been used to conduct very large studies, such as the largest study of co-morbidity in Autism Spectrum Disorders, and the construction of the largest US registry of pediatric rare diseases (rheumatoid arthritis).

Impact on CTSA Consortium Goals

SHRINE meets several needs of CTSA consortium goals 1, 3, and 5. SHRINE has already made an impact towards establishing policies for national data sharing and federated clinical data search across independent CTSA locations. In June of this year, 51 institutions were assembled for a conference on using SHRINE to “guide the emergence of the national translational research network”. SHRINE investigations address T1 goals such as selecting patients for clinical trials, analyzing EMR outcomes to study medication safety, and measuring population scale health outcomes.
Omics Data Standards Working Group: update on activities

Jessie Tenenbaum

The Omics Data Standards Working Group was created to address common issues encountered in the management and use of “omics” data in the context of clinical and translational research. Our goal is to identify the appropriate standards (and gaps) for use with omics data in integrated data repositories, cross-laboratory data sharing, and federated query approaches. Activities to date have included:

- Formulate high level use cases based on real world projects
- Identify existing standards through investigation and presentations from outside experts
- Deploy REDCap survey to core lab facilities to gain a better understanding of how omics data is currently managed and exchanged.

The survey yielded 58 responses which will be used to flesh out use case “vignettes” for genomic and proteomic data. We will then use these vignettes to generate criteria by which to evaluate existing standards. Future work will include application of findings to real-world integrative data projects, documenting and publishing results.

This work is of significant relevance to the CTSA consortium as more and more clinical and translational studies include the incorporation of these types of high throughput omics datasets. This work is relevant to SG3 to facilitate data sharing, and SG5 to enable T1 translational research.
DIRECT2Experts – Distributed Interoperable Research Experts Collaboration Tool

Griffin Weber, Harvard, Boston, MA; William Barnett, Indiana U, Bloomington, IN; Mike Conlon, U Florida, Gainesville, FL; David Eichmann, U of Iowa, Iowa City, IA; Warren Kibbe, Northwestern, Chicago, IL; Holly Falk-Krzesinski, Northwestern, Chicago, IL; Michael Halaas, Stanford, Menlo Park, CA; Layne Johnson, U of Minnesota, Minneapolis, MN; Eric Meeks, UCSF, San Francisco, CA; Donald Mitchell, Stanford, Stanford, CA; Titus Schleyer, U Pittsburgh, Pittsburgh, PA; Sarah Stallings, U Colorado, Denver, Aurora, CO; Michael Warden, Elsevier, Ann Arbor, MI; Maninder Kahlon, PhD, UCSF, San Francisco, CA; Members of the Clinical and Translational Science Awards (CTSA) Research Networking Group

Introduction. Research networking tools use data-mining, social networking and semantic web approaches to bring scientists together. Several commercial and open source platforms have been built, and numerous institutions across the country are adopting these tools. In August, 2010, the Research Networking Group of the Clinical and Translational Science Awards (CTSA) consortium launched an initiative to design a national federated network, and in less than six months, we built a prototype site at http://direct2experts.org, which searches 29 institutions and 8 products representing over 50,000 researchers.

Methods. 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 compared to 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. Although several of the platforms, including VIVO, Harvard Catalyst Profiles, and Elsevier’s SciVal Experts had existing networks connecting instances of their own products, we sought to solve a complementary challenge, which is combining these networks to reach the broadest set of institutions. The resulting network is defined by two components: 1) a federated technical architecture and common RESTful interface and 2) an agreement between participating institutions that required only sharing counts of the number of people who match a search phrase.

Next Steps. We recognize that much work lies ahead in the next phase of Direct2Experts, which will address governance, scalability, user experience, data quality, and other issues. However, the pilot marks several significant milestones: 1) The major research networking platforms used across the country agreed to a common federated architecture. 2) 28 institutions agreed to share information about their investigators through a public website. 3) Momentum was generated that will encourage additional institutions to participate and will facilitate discussions about the future of Direct2Experts. Our vision for the next phase of Direct2Experts includes using Linked Open Data to discover cross-institutional connections among people, expand beyond people to other types of data, create metrics of impact beyond publications, perform social network analysis across institutions, connect academic faculty with industry/community partners, access expertise/resources not available at one’s home institution, and extend beyond biomedicine to other disciplines. However, we are aware of numerous concerns that institutions might have, such as competitive intelligence, orphaned investigators, data currency, privacy/policy/security, and the costs of local implementation of research networking tools or converting existing tools to use Linked Open Data.

Importance to the IKFC and CTSA. Research Networking is a top priority for CTSA, and Direct2Experts is a CTSA-wide informatics initiative related to this goal that involves two dozen CTSA institutions and is open to others who want to participate.
Use-case Driven Development of a Federated Research Data Warehouse
Monika Ahuja and David Eichmann
Institute for Clinical and Translational Science (ICTS), University of Iowa

Objectives: To design and deploy an enterprise-wide federated research data warehouse (RDW) and use it to build cohorts and registries, and to help researchers analyze patient health trends, discover and focus on these trends through statistical analysis and data mining activities.

Background: The Institute for Clinical and Translational Science (ICTS) at the University of Iowa initiated construction of the data warehouse to support on-demand and comprehensive views of patient information assets. ICTS is building a dynamic, federated data model, based on multiple biomedical, clinical and imaging data systems – ultimately providing a consolidated and standardized metadata layer at the enterprise level. This project will enable investigators to see the “big data picture” across widespread systems.

Project Aims:
1.) Provide enterprise level integrated patient data view to researchers.
2.) Integrate patient medication data with vendor provided drug information systems.
3.) Create cohorts and virtualized environments for researchers to access IRB-approved data
4.) Support various types of data use cases including data extracts, real-time data federated views, and materialized views (as a snapshot on a particular date), etc.
5.) Aid researcher data analysis and integration of study generated data
6.) Provide a Service Oriented Architecture for research collaborations at the consortium level

Example use case 1: (Impact on subject recruiting):
An investigator team needed to approach asthma patients without identifying the patients, this platform brokered an on-demand cohort (undisclosed to the researchers) and based on IRB approval to subject identification scheme, supported subject recruitment processes through mail to patients in a targeted fashion without disclosure of subject entity to investigators.

Example use case 2: (Impact on drug discovery): Researchers working on combining different existing obesity drugs will benefit through definition of a cohort of obese patients with specific symptoms and prescribed specific medications. Based upon race, gender, and prescribed drugs, data mining can highlight changing trends in the patient weight. This process can aid the researcher in observing potential additive-versus-synergistic effects as well as predictors of response (gender, age, non-study, and study medications, drug-drug impact, drug-allergy impact, drug-disease interactions, side effects, renal function, psychological factors, etc.). During clinical trials, this platform will enable researchers to assess drug impact by observing trends regarding patient health – supporting tuning of drug discovery process.

Importance to the IKFC: All CTSAs engage in the activity of integrating patient data. Our approach to Research Data Warehouse development supports substantial flexibility in the modeling and integration of data types and sources. The spectrum of clinical data needs requires just such a broad perspective and leads to the federation of these resources into true knowledge bases. This includes investigator analytical requirements, both in the context of a single CTSA and in an aggregated perspective of all such resources as a single consortium data federation.

Impact on CTSA Consortium Strategic Goals: This platform addresses Goal #3 of the CTSA Consortium by providing a SOA enabled platform supporting research collaborations for data sharing and the development of a pool of standard analytic programs that can be shared and contributed towards an “Inventory of Resources.”
Cross CTSA Institutions Cohort Discovery Approach

Adil Alaoui, Chris Piepenbring, Dongkyu Kim, Andrew Shinohara, Chris St. Clair, Baris Suzek, David M. Tanenbaum, Allan Cunningham, Deepak Nalli, Bernie Galla, Subha Madhavan.

Abstract:

The Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS) is a collaborative research center that includes two major universities and three affiliated hospital and research systems. GHUCCTS institutions include the Georgetown University Medical Center, Howard University, MedStar Health Research Institute, the Washington DC Veterans Affairs Medical Center (VAMC) and Oak Ridge National Laboratory.

We initiated a project to integrate electronic health records from several research and clinical databases at different GHUCCTS institutions and develop a system that enables access to aggregated data from these disparate data sources to facilitate cohort discovery and information sharing across our CTSA. Our aim is to establish a secure exchange architecture that enables intra- and inter-institutional clinical and translational research information sharing across boundaries as well as to develop methodologies for querying electronic health records in support of cohort discovery across our participating institutions. The end result of this effort will facilitate clinical trial recruitment and identify patients for research studies. We selected the Integrating Biology and the Bedside (i2b2) software, an open source analytical tool, and implemented and configured it for the purpose of cohort discovery at GHUCCTS. The use of i2b2 web-based tools will insure that researchers at each of the GHUCCTS institutions will be able to easily identify potential clinical trial participants across institutions, while protecting the confidentiality and security of their health information. i2b2 also supports the analysis and visualization of research and clinical data and enables investigators easy access to available populations across multiple institutions.

We established an Information Architecture team to identify and document available data sources (data dictionaries, metadata and ETL processes) at each institution, to consistently map data from these sources into a common repository, and to create a process and implement tools for facilitating secure information sharing. In our setting, due to different institutional policies and goals, it is challenging to implement an integrated research data repository that includes all participating institutions’ data. Our final approach was to establish a searchable GHUCCTS research data repository complemented by a process to handle query submission to institutions and data sources which cannot be readily integrated into the GHUCCTS i2b2 repository.

The goal of this paper is to present our approach in designing the informational model and integrating various data sources into the architectural framework. We also discuss the rationale behind the selection of our approach and suggested methodologies used to address our institutions’ unique requirements and achieve the project’s goal, including some challenges we encountered and how these were addressed through technical and data governance committees.
**Exploration of Two Approaches for Determination of Recipients of Targeted Communication in a Scientific Collaborative Network (WorkWeb)**

Suzanne Bakken, Yalini Senathirajah, Daniel Dine, Stephen B. Johnson, Chunhua Weng
The Irving Institute for Clinical and Translational Research, Columbia University, New York, NY

**Objective:** The Biomedical Informatics Resource of the Irving Institute for Clinical and Translational Research (IICTR) has developed a collaborative platform called WorkWeb that represents scientist profiles and their publications and grants and enables researchers to search for collaborators, submit requests for IICTR services, and collaborate online through a wiki. Motivated by principles from social marketing to improve the signal-to-noise ratio in our communication, we conducted two studies to explore potential approaches for providing communication related to three use cases: targeted informational messages for CTSA events; targeted informational messages for CTSA services or tools; and targeted CTSA survey administration. Social marketing is defined as “the application of commercial marketing technologies to the analysis, planning, execution, and evaluation of programs designed to influence voluntary behavior of target audiences in order to improve their personal welfare and that of society.”(1)

**Methods:** In Study 1, we invited 15 members of the IICTR leadership to participate in a 30 minute interview. During the interview, participants answered three open-ended questions and were asked to rate the usefulness of their index terms and term frequencies extracted from Medline and the NIH CRISP database to support four use cases (a use case focused on find/be found for collaboration is not addressed in this analysis). Answers to open-ended questions were thematically analyzed. We calculated a discrimination index for each term. In Study 2, we developed and implemented a survey to characterize our target audience of clinical and translational investigators and applied a clustering algorithm to discover audience segments.

**Results:** In Study 1, the 11 respondents found a variable proportion of automated terms useful – only one investigator found 100% of the terms useful. Half the investigators discriminated among use cases in choice of index terms. Qualitative data suggested the need to allow investigators to deselect or add index terms associated with their profiles. One respondent noted that this was critically important because publications and, to a less extent, grants reflect where the investigator has been, not where he or she wants to go with his research. In Study 2, application of the k-means algorithm categorized 88% of 257 survey respondents into five categories (i.e., target audience segments). The most discriminating variable was the % time spent on various work activities; other important discriminators were education, length of time as independent investigator, and confidence to carry out specific research tasks with and without assistance.

**Discussion:** The findings of Study 1 support an indexing approach that is automated, but that allows human review for both deletions and additions. The clusters identified in Study 2 provide the foundation for targeted messages about opportunities and resources as well as targeted views of the WorkWeb portal. The methods and measures that we applied may be useful to other CTSA institutions wishing to characterize and segment the target audience for informational messages about their informatics innovations and resources as well as for survey administration. Because of the role of these approaches in facilitating communication that is “attended to” by the message recipient, they provide a component of methods for determining the impact of informatics’ influence on translational research. Moreover, the research findings are relevant for two CTSA consortium strategic goals: provide training and career development of clinical and translational scientists) and enhance consortium-wide collaborations.

**Acknowledgment:** UL1 RR024156 from the National Center for Research Resources (NCRR)

Improving Robustness and Automating Workflow of the Resource Discovery System with iBIOFind
Michael B. Baldonieri, Nancy Whelan, Charles Borromeo, Harpreet Singh, Michael J. Becich, MD, PhD, Jeff B. Cromwell, PhD

Finding and assembling the right set of resources (e.g. databases, expertise, laboratory equipment, etc.) directly impacts the overall success of a research project. The right software package or knockout animal saves time and money, and allows researchers to better focus on research endeavors. Often these resources are found at a researcher’s home institution, but their existence is obscured within a warren of websites and documents. While many academic health centers (AHCs) are engaging in projects to compile these resources, a significant hurdle in this pursuit is the discovery of resources outside of departmental curation and contact networks. Due to the large size of AHCs and the large number of research faculty, an unrealistically high level of inter-departmental cooperation and a large amount of time is necessary to develop an accurate registry of institutions’ resources. This curation process becomes increasingly achievable by implementing an automated tool to search an institution’s affiliated websites. One specific resource discovery tool, iBIOFind, shows great promise and functionality in resource curation.

This analysis of iBIOFind exposes great utility to the world of Biomedical Informatics. This project explores means of maintaining current information in databases and staying abreast of the evolution of individual resources. By rapid discovery of new resources, iBIOFind increases the power and robustness of resource databases, in this case, the Resource Discovery System (RDS). The method by which this application operates represents the best practice in resource discovery. By navigating an academic health center’s websites and affiliated pages automatically utilizing the Biomedical Resource Ontology (BRO) and the user-entered website root, it returns results which may be considered true resources. After a quality assurance test requiring human oversight performed in Microsoft Excel on a percentage of randomly selected returned URLs, the resources may be added to the RDS and made available to researchers over the internet.

The evaluation of iBIOFind used the application to harvest potential resources from 8 CTSA collaborating sites (6 Sharing Partnership for Innovative Research in Translation members and 2 potential SPIRiT members). The results were then classified as resources (true positives) or non-resources (false positives) via manual examination of each webpage and comparison to the definition of a resource developed for the RDS. After validation of true positives, the data were used to numerically and graphically analyze iBIOFind’s potential impact on the robustness of RDS. The results were also helpful in refining the definition of an RDS-appropriate resource. Presented here is the analysis and human/RDS resource determination methodology. Overall, 34,058 hyperlinks were returned and sorted and 10% were considered true positives. Of the total returned hyperlinks, 13,747 were unique and accessible. Given 3,382 true positives, 25% of unique accessible links are appropriate for entry into RDS.

This resource discovery tool presents a unique opportunity for broad adoption across all CTSA sites and CTSA-related fields of research because of its ease of use and clear results, as well as low cost of implementation. iBIOFind illustrates economic advantages to collaboration over individual projects as it has already undergone multiple stages of research and development, therefore making the CTSA IKFC the perfect venue to promote the use and robust functionality of the application. New availability of these resources will catalyze increased collaborations between both individual researchers and AHCs. Once the new resources are incorporated into databases, the scientific community will experience the utility and functionality of automated resource collection. More efficient use of currently available facilities and research resources can increase member institutions’ focus on contribution to the initiatives of the CTSA by making previously unknown resources available intra- and inter-institutionally for use by researchers of all fields and seniority.
An Informatics Course Based on the CTSA Informatics Competencies

**Background.** The CTSA Strategic Goal 2 Committee has endorsed a set of competencies for clinical and translational researchers, including nine relating to informatics. In 2009, the IKFC Education Affinity Group surveyed CTSA sites to assess the extent to which these competencies were being taught. Results showed that many sites were not including them at all in the educational programs for clinical and translational researchers and many were not addressing them well. Some sites did have informatics included in the training programs, but these programs were not necessarily based on the identified competencies. Other sites did not fully engage the IKFC members in developing courses, but even if they were willing, funding constraints limiting the roles of informatics experts may also have discouraged extensive involvement in the education programs. We developed an educational program based on the nine informatics competencies. The program provides basic information and hands on experience for clinical and translational researchers that can be taken as a course in the master’s program for clinical and translational science, but also provides information for translational researchers on the services of the biomedical informatics component of the UAB Center for Clinical and Translational Science (CCTS).

**Methods.** A one-credit hour graduate course was developed that was available as an elective for clinical and translational science masters students, but was also open to the university community. The course was implemented on a pilot basis in the summer of 2011 as a seminar series. The course consisted of seven sessions addressing that addressed the competencies. Each session included either an in-person or an online presentation and selected reading materials and other exercises. Students evaluated each session as well as the series as a whole. The topics covered are listed below.

<table>
<thead>
<tr>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is informatics?</strong> Topics include: Overview of biomedical and health informatics, key issues in the field, impact of informatics on medical research, education and patient care, UAB informatics resources</td>
</tr>
<tr>
<td><strong>Electronic Health Records for Healthcare and Research.</strong> Topics include: description, functionality, use for research, and challenges for both healthcare and research.</td>
</tr>
<tr>
<td><strong>Research using Clinical Data in Electronic Health Records.</strong> Topics include: Demonstration of tools and resources for accessing UAB’s electronic health record systems, strengths and limitations of EHRs for research. Students get to try their own queries.</td>
</tr>
<tr>
<td><strong>Health Information Technology Standards and Interoperability of Clinical Systems.</strong> Topics include: Importance of standards. What standards exist in healthcare? How are standards determined?</td>
</tr>
<tr>
<td><strong>Using Informatics Methods and Tools for Clinical Research.</strong> Topics include: Demonstration/discussion of tools for clinical data capture and data management.</td>
</tr>
<tr>
<td><strong>Access to Research Literature and Online Resources.</strong> Topics include: Tools, strategies and resources for effective literature searching for researchers in clinical and translational science.</td>
</tr>
<tr>
<td><strong>Bioinformatics and the Analysis of Genomic Data.</strong> Topics include: Overview of bioinformatics, including the tools and resources at UAB for genomic and other “--omic” analyses.</td>
</tr>
</tbody>
</table>

**Results and Discussion.** Student evaluations to the individual sessions and the overall series were very positive. On a 10 point rating scale measuring attendee satisfaction, with 10 being the highest, the sessions averaged between 7.5 and 10, with 6 of the 7 sessions over 8.0. Although a short course such as this was not sufficient for mastery of the competencies, it was an introduction targeted to both the CTSA competencies and the resources available at UAB where participants could get additional experience when they used those resources. The sessions also provided needed background for anyone using the resources that the Biomedical Informatics Component could provide. Since the student response to both the in-person and the online sessions was positive, more sessions may be able to be converted to an online format, which would lead to more flexibility for both students and faculty. The online materials could also be shared with other CTSA sites and the annual IKFC meeting will provide a forum to discuss sharing these and other educational materials.
Finding Collaborators: Towards Interactive Tools for Research Social Networks

Charles D. Borromeo, Titus K. Schleyer, DMD, PhD, Michael J. Becich, MD, PhD, Harry S. Hochheiser, PhD
University of Pittsburgh, Pittsburgh, PA

Project Description
Although helping researchers find collaborators has been identified as one of the primary use cases for research social networks, little is known about the functionality and data that would best support this goal¹. We are interested in developing interactive tools that will leverage the data in research social networks to help scientists build multi-disciplinary, collaborative teams. To identify user needs and preferences for such tools, we are conducting semi-structured interviews with researchers, including the evaluation of prototype interface screens. These interviews will inform the design of a prototype that will be evaluated in a usability study with translational researchers. The results of the project will be applicable to collaborative systems within the CTSA (e.g. VIVO, Digital Vita, Profiles, CAP, etc.).

Importance of Work to the 2011 IKFC Meeting
The IKFC meeting raises awareness about this project and provides additional recruiting opportunities. Attendees can participate in the project contributing their opinions and insights into the requirements gathering process. The research team will present preliminary results at the IKFC meeting. Other collaborative tool developers can use these results to inform their own application design. The meeting allows our team to interface with other scientific networking development teams. These discussions should accelerate collaborator tool development thereby benefiting the CTSA as a whole.

Impact on Strategic Goals
Our project focuses on CTSA’s Strategic Goal 3. Several social networking projects have made great progress establishing a comprehensive database of profiles of the researchers at an institution. Our project explores the next question: what tools can be built on top of these databases to establish multi-disciplinary, collaborative teams? Our research will improve these databases and promote collaborative projects within the CTSA.

References
**REDCap Goes to the Rodeo**  
Linda Carlin, PhD and Jessie Robinson, MS  
University of Colorado Denver

**Objective:** The Colorado Area Health Education Center (AHEC) recruits medical students from the University of Colorado every year to work at health screening events so that the students can get first-hand experience in community health. This year, at the National Western Stock Show in January, for the first time, data were collected from participants electronically, instead of with paper and pencil, using REDCap (Research Electronic Data Capture) on iPads. In the 16 days of the event, 200 medical students screened almost 3,000 people.

In addition to the health screening, a study protocol was implemented to contact screening participants at a later date to see whether they had made recommended behavior changes or discussed health issues with their doctor. This follow-up study would not have been feasible with paper-based data collection. However, use of REDCap enabled the researchers to link a participant’s health screening data with their contact information after consenting for the study. In addition, branching logic used in the electronic screening forms determined automatically whether a person met the criteria for the follow-up study, reducing time and error in recruitment efforts.

Use of REDCap to collect screening data was well-received by both screeners and participants. The project coordinators described several improvements over traditional data collection methods. The main advantages were: 1) the ability to guide screeners through the screening protocol, using online scripts and branching logic, which improved quality and consistency of data collection; 2) data validation capability in a field research setting; and 3) the ability to make modifications on the fly to address issues or to improve workflow.

Productivity was greatly improved by removing the need for data entry at the end of the project. And, for the first time, project coordinators were able to evaluate the screening process itself by measuring factors such as start-to-finish screening time and counselor effectiveness.

One of the surprising findings was how engaging screening participants found the electronic data collection. In addition to their interest in their health information, they were very enthusiastic about the process itself.

**Why Important to Present at 2011 IKFC Meeting**  
The increasing use of mobile devices for data collection presents both opportunities and challenges for investigators. We hope that demonstrating our success in a high-risk, high-throughput data collection environment may reduce barriers to acceptance of mobile technology in research. The ability to evaluate and modify the process in real time follows the best practices of the Learning Healthcare System model best practices.

**Impact on CTSA Consortium Strategic Goals**  
**Goal #1** Enhancing national clinical and translational research capability: Our project is an example of implementing REDCap, which is widely-used across the CTSA consortium, in a novel community health setting using mobile technology.

**Goal #4** Enhancing the health of our communities and the nation - community engagement: Health screening is an effective way for health researchers and practitioners to engage with a community. Our use of iPads for electronic collection of health screening data not only improved the quality of data collected, but had a positive impact on the screening process for both the screeners and the participants.
The TIES-SPIRiT Collaborative Tissue Network will transform valuable but isolated archives of data and research materials into a network for sharing data and research materials. In the pilot phase we will leverage existing CTSA informatics infrastructure to establish a national test-bed for collaborative tissue research among three CTSA-SPIRiT sites – University of Pittsburgh (lead institution), University of Pennsylvania and University of Chicago. Eventually the network will be expanded to include all six SPIRiT sites. In the pilot year, we propose to complete four activities: 1. Customize the TIES system to meet the TIES-SPIRiT CTN security and regulatory requirements, 2. Develop a governance structure for the network and establish all necessary policies, 3. Deploy the TIES software at the three participating institutions, and 4. Develop a set of 3-5 pilot projects that will utilize this new network. The Text Information Extraction System (TIES) is a CTSI and NCI funded project developed at the University of Pittsburgh, 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.

The TIES-SPIRiT Collaborative Tissue Network system will directly support CTSA Consortium Strategic Goals 1 and 3 by providing a mature, HIPAA compliant, collaborative research system for institutions that is based on a mature production tested software architecture. IKFC attendees are the targeted end users of the TIES-SPIRiT system and will benefit by learning more about its capabilities.
Progress in our understanding of brain disorders increasingly relies on the costly collection of large standardized brain magnetic resonance imaging (MRI) data sets. Moreover, clinical interpretations of brain scans benefit from compare and contrast analyses of scans from patients with similar, and sometimes rare, demographic, diagnostic, and treatment status. A solution to both needs is to acquire standardized, research-ready clinical brain scans and to build the information technology infrastructure to share such scans, along with other pertinent information, across hospitals. The resulting research-ready brain imaging resource would provide a wealth of accessible standardized brain imaging data relevant to patient care and research.

In 2010, CTSIs at the University of California Irvine (UCI) and the University of Southern California (USC) began collaborating on a pilot project that is creating such a brain imaging resource. The goals of our pilot study are to acquire standardized, research-ready clinical brain scans across the USC and UCI hospitals and to build the infrastructure to share such scans, along with other pertinent information, across both institutions. The outcome of this pilot project will be a research-ready clinical brain imaging resource accessible to both clinicians and researchers. The specific goals of the pilot project are: (1) to collect standardized brain imaging data for every patient who receives a clinical brain MRI scan in the radiology departments at the UCI and USC hospitals and to securely store and manage that imaging data at each institution, (2) to develop general, open source, end-to-end software infrastructure to support data federation from image capture to distributed image query and download, (3) to deploy a pilot federation across the two institutions that supports cohort identification and retrieval of images based on subject characteristics, and (4) to explore models for patient consent with medical ethics teams from the UCI and USC CTSIs. Our long term goal is to expand this pilot, first to a small number of additional institutions, and eventually to regional and national levels. We anticipate that our federated image repository will eventually be linked with other data resources containing patient information, including genomic and phenotypic data, subject to Institutional Review Board restrictions on patient privacy and data use.

To date, the pilot project has successfully deployed a multi-site federated repository of image data populated by clinical information. The system supports a distributed query method that allows us to make full SQL queries on the imaging metadata at each site to identify images with specified attributes.

The chief neuroradiologists and brain imaging researchers at UCI and USC agreed on a standard Short Imaging Protocol (SIP) for use in the pilot project. This SIP includes a high-resolution structural scan and a Diffusion Tensor Imaging (DTI) scan, which are acquired as part of most clinical and research protocols.

The software architecture of the pilot project supports federation of data and metadata repositories that are populated from clinical sources at each hospital. The system leverages existing security mechanisms, such as Shibboleth identity providers and authentication infrastructure. In each hospital, an MR Technician pushes Short Imaging Protocol scans for each patient to a node inside the hospital network that runs a Forwarder component that de-identifies the images. The Forwarder maintains a database that maps patient identifiers to de-identified, universally unique identifiers; this mapping information never leaves the secure hospital network. After de-identification, the Forwarder pushes the DICOM images outside the hospital network to a node running an Image Gateway Service. This service extracts metadata describing the images from the DICOM headers and stores that metadata in a database; the images themselves are stored in a file server. Finally, the system includes a Federated Query Engine that receives client SQL queries for images with certain attributes and distributes those queries to the metadata databases in each Image Gateway service to identify matching images. The client then interacts directly with the Gateway service to download the desired image files.
Clinical translational research laboratories make technologies accessible to professors, researchers, nurses, and technical support staff to perform research protocols for both patient safety and research quality. CTSA institutions are designed to provide technical expertise to researchers for both innovation and conformity to study requirements. Because of the number of biomedical specialties available as well as the number of CTSAs, metrics to assess the impact of informatics on translational research along with resource utilization for consortium wide collaborations is needed. Thus, the unit of analysis presented here is CTSA lab information which needs to be compared across institutions for effective resource identification, utilization and impact assessment. These comparisons require a list of indicators that can be weighted to generate an overall CTSA lab impact factor (LIF). The following four indicators are proposed for each CTSA lab as a basis for the ranking and assessment:

1. **Scale (Sc)** represents the number of web pages devoted to the CTSA lab.
2. **Lab Visibility (Vi)** measures the total number of unique external links received (inlinks) by a CTSA lab. The number of external inlinks received by a CTSA lab measures the visibility of their published material. Self-archiving and other repositories initiatives can be approximated by both relevant documents and citation analysis.
3. **Relevant Docs (RDocs)** measures the relevance to both academic and publication activities using the following file formats: Adobe Acrobat (.pdf), Adobe PostScript (.ps), Microsoft Word (.doc) and Microsoft Powerpoint (.ppt).
4. **Citation Analysis (Scholar)** measures the output from Google Scholar and provides a citation analysis for each CTSA lab based on papers, reports and other academic artifacts.

The overall Lab Impact Factor (LIF) is computed based on a weighting of (1)-(4) to rank each CTSA lab. The Biomedical Resource Ontology (BRO) is used as a filtering mechanism to obtain web information along with the computational search engine iBIOFind and four main search engines of Google, Yahoo, Live Search and Exalead. The analysis reveals the importance of developing a dialogue in the CTSA community regarding effective resource identification and utilization as well as relative rankings within specific biomedical communities for impact assessment. This unique study examines and ranks labs in a small number of biomedical specialties from a number of current 60 medical resource institutions of the CTSA program based on the overall CTSA lab impact factor.

A detailed web presence for a CTSA lab provides an opportunity for the effective and efficient transmission of the activities of a CTSA which can possibly attract new students and scholars. The result of this analysis and ranking is to motivate both CTSA institutions and scholars to generate a web presence that reflects their activities along with an understanding of their relative position within biomedical specialty communities in the CTSA program. A result considered important for presentation and discussion at this 2011 IKFC meeting.
**Modeling Electronic Tools to Improve Source Document Verification Audits**

Stephany Duda, PhD; Firas Wehbe, MD, PhD  
Department of Biomedical Informatics, Vanderbilt University

Source verification audits are the gold standard for assessing data quality in clinical research. During such audits, external auditors compare research data to the original documentation of patient care, which may include paper clinical charts, laboratory reports, or the contents of electronic medical record and laboratory systems at the study sites. Protocol-driven studies such as clinical trials often engage teams of clinicians and data managers to perform such audits, to ensure the study generates accurate results and follows best practice guidelines required by the U.S. Food and Drug Administration.

Unfortunately, most verification audits of clinical data use paper forms, which have been shown in general to be less effective and efficient than electronic tools. Available audit software packages focus on analyzing an existing, electronic dataset for errors and unusual patterns, rather than facilitating the comparison between the dataset and a physical source document. We believe a computer-assisted audit tool specifically designed for research data auditing can improve the process of source document verification.

We collaborated with the clinical sites of the Caribbean, Central, and South America network for HIV epidemiology, an HIV observational research network, and the network’s data coordinating center (DCC) at Vanderbilt University. Through a review of experiences documented during 14 paper-based data audits, post-audit debriefings, and audit-related discussions with the DCC and sites, we identified five desirable attributes for a computer-assisted audit tool, including multi-user support through networking, simple file formats for importing and exporting audit data, standard error assessment protocols, audit decision support, and tools for calculating and displaying audit results.

We developed a prototype system to demonstrate the significance of these key attributes as well as the feasibility of implementing them in a simple web-based application. We tested this application during audits of twelve cancer and tuberculosis datasets in nine locations and found the system drastically improved the audit process by decreasing preparation time from over six hours, on average, to less than one hour; enforcing consistent, on-site error coding for all records; and eliminating difficulties reading other auditors’ completed forms, as well as the manual error tallying required for a post-audit report.

Data quality is a relevant issue to ongoing research studies at all CTSA institutions. We believe computer-supported auditing is particularly suited to multi-site studies, as audit software installed on a laptop can serve as a mobile audit tool, and to investigators wishing to implement a simple, internal audit program (in contrast to supervised clinical trials required to contract with external regulatory agencies). We hope to further explore the data quality assurance needs of small studies, expand our prototype software, and collaborate with other CTSA sites interested in low-cost approaches to data quality control.
**Description:** The Duke Biobank Informatics Working Group (IWG) was formed in 2010 and consists of operational and informatics representatives from Duke’s major biobanking groups. The members are working to address the disadvantages stemming from Duke’s separate and distinct biobanking informatics systems. The objectives of the IWG are to determine the best solution for the biobanks to exchange information; for samples and their associated data to move seamlessly between biobanks; and for investigators to be able to turn to a single system for reports and queries about their specimen collections regardless of where they are banked at Duke.

The IWG is considering three possible solutions: 1) the adoption of a single enterprise-wide biobanking informatics system by all, 2) a federation of systems with each biobank maintaining their current system but sending their data to a warehouse for reporting and querying, or 3) a ‘hybrid’ system in which the majority of biobanks adopt a central biobanking system, while established biobanks maintain their current system, and send their data to a central system which acts as a replication repository for reporting and querying.

**Activity:** In selecting a solution, the IWG must weigh the cost vs. benefit for adopting a new system and migrating substantial legacy data. In addition, it is vital to provide a system for individual, independent researchers who may not currently have access to a commercial biobanking system and would benefit substantially if an affordable solution was available. Most importantly, the chosen system must satisfy the individual, specific workflows and requirements of each biobank.

The IWG is preparing a formal Request for Applications (RFA) and will invite several commercial biobanking vendors to apply. Respondents will include vendors of commercial biobanking software systems as well as service providers for open source software. Each biobanking software product must meet an initial set of prerequisites in order to be invited to participate in the RFA. The most critical component of the RFA is a list of system requirements developed with input from each member of the IWG. The requirements document is over 30 pages, and contains over 300 individual requirements. The vendor will be required to respond in a specific manner to each requirement. For example, choices include: 1) Fully supported in system, 2) Planned enhancement, 3) Not supported, and 4) Requires a third party product. A vendor’s response will subsequently be incorporated into the final contract. At the end of the review process a vendor will be chosen that best meets the needs of Duke’s biobanks.

**Importance for presentation at IKFC Meeting:** Although the initial focus is within Duke, it is our intention to share the documents and the process as well as the lessons learned with the CTSA community. It is clear that a majority of research institutions engaged in translational research struggle with disparate biobanking activities and find it challenging and time consuming to assess the number and types of biospecimens in their own institution. Duke will reach out to sister CTSA institutions at the IKFC meeting to enable others to build from our work. **Impact on Strategic Goals:** This project will have significant impact on multiple strategic goals within the CTSA consortium. For strategic goal #3, the process, valuable experience, and lessons learned of the IWG will help inform the other CTSA sites, if they undertake a similar endeavor. For strategic goal #5, a central view of biospecimen collections would facilitate collaboration between clinical researchers who have access to patients and their biospecimens and bench researchers who do not, within Duke, as well as between Duke and other CTSA sites.
Secondary Data Reuse in Comparative Effectiveness Research: A Model Framework and Proof of Concept
Daniel Fort, MPH1, Adam Wilcox, PhD1
1Columbia University, New York, NY

We propose a model of data use to articulate the influence of various informatics interventions in Comparative Effectiveness Research (CER), a significant component of translational research. Informatics interventions are partitioned into overlapping conceptual spaces as described below. Difference in relative sizes of conceptual overlaps as well as changes in size over time allow the assessment of the effect of various informatics interventions on current research.

The model shows the potential overlaps between previously recorded information, medically relevant measurements, and a target population. Region 1 is medically relevant information which defines the population of interest. This region does not simply define the boundaries of the study instrument, but the cohort definition, study endpoints, and patient status. Region 1 represents the conceptual space of primary research.

Region 2 represents recorded information which already exists for the population of interest in any form. This information need not be limited to that regarding health status. Region 3 is medically relevant information collected in the course of ordinary treatment, as opposed to particular diagnostic tests used to characterize the study population. The Useful Fraction is the overlap of all three. This fraction represents useful, medically relevant information which has already been recorded for the population of interest and useful to research using secondary data analysis.

The model also suggests ways in which these regions can be increased. One way to increase potential region size is to increase the size of the target population, for example. Increasing recorded information could occur by actively surveying the target population. Methods for extraction and analysis of new information from previously opaque data, like Natural Language Processing, also essentially increase the value of previously recorded information.

As proof of concept of this model, we performed a literature review and meta-analysis to determine where and how secondary data is used in Comparative Effectiveness Research. Secondary data use in this context implies an informatics solution such as EHR and electronic clinical and lab data, aggregate databases, and document indexing. Secondary data use was considered in both cohort identification and study data realms. In the final sample, 60% of studies used secondary data and there were significant differences in proportion of secondary data use between prospective and retrospective study groups. This proof of concept corresponds to a baseline exploration of Region 1 in the model. Research into parameterizing other regions is ongoing.

Impact on CTSA Strategic Goals:
This model, as demonstrated by the proof of concept, creates a framework for assessing and quantifying the influence of various informatics interventions on current CER, a significant component of translational research. That process will foster an understanding of how various informatics interventions best support current research, which will, in turn, lead to better-supported future research.

Acknowledgment: This research was supported by AHRQ grant R01 HS019853-01, Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER).
Objective: The aim of this presentation is to report on a terminology content management process being deployed at UC Davis which will improve the usability and data quality of the UCD Cohort Discovery Tool, based on the i2b2 software application.

Background: The UC Davis Biomedical Informatics Program is using the Language Engine Access Portal or “LEAP” suite of products and services supplied by Health Language, Inc (HLI) to maintain the currency of “reference content” or coding systems and terminologies utilized in Cohort Discovery.

LEAP is a web-based terminology management suite of applications and services which allow healthcare organizations to access, browse, and customize content sets and maps; create or upload additional content; and download the content for use in other applications. UC Davis is using LEAP to build and manage a more comprehensive, accurately-coded ontology cell for the UC Davis Cohort Discovery tool than was previously possible with available tools and resources. The most visible effect of this improvement to the ontology cell is reflected in a more robust, complete i2b2 query build user-interface. Additionally, utilization of LEAP will improve the quality of the data in and provided by Cohort Discovery through maintenance of maps from concepts in the ontology cell to the most recent, or a specific historical version of a terminology or coding system. LEAP provides the UC Davis technical team access to versioned reference coding systems content, a user-interface that permits mapping to desired codes and creates a file that can be directly loaded into Cohort Discovery. Initially, the implementation of LEAP at UC Davis has three use cases:

1) Annual / periodic terminology updates. This requires changes in the i2b2 ontology cell schema and review by the UC Davis technical team and potential changes to codes mapped to database elements. This may also result in a change to the i2b2 query build hierarchy. LEAP creates a file from the UC Davis information stored in the web portal which uses an API to extract the coded reference information stored a terminology server hosted by Health Language. This creates a file can be uploaded into the UC Davis i2b2 ontology cell.

2) Adding new fields to the i2b2 database and / or nodes on the ontology cell. A new data field requested or made available by feeder systems to the i2b2 database, requiring a change to the ontology cell to reflect the availability of that new data within the ontology cell navigation hierarchy. The LEAP portal allows the new data element (node) to be mapped to a reference terminology with a code associated with that data element and produces a file inclusive of the appropriate change(s) that can be uploaded to the i2b2 ontology cell.

3) New implementation(s) of i2b2. Each data element in the i2b2 database gets assigned to one or more code(s) and the output of the LEAP application is a file that can uploaded to the i2b2 ontology cell at installation. LEAP improves the accuracy of, and should accelerate, the ontology cell build in new implementations.

Why is this important to the IKFC? The i2b2 platform is in widespread use across the consortium. Novel, improved and alternative tools, processes and methodologies that enhance the quality of the i2b2 data and the overall use experience of i2b2 is valuable to any site with or contemplating an i2b2 installation. The UC Davis experience with LEAP is the only application of HLI services & products for this purpose.

Impact on CTSA strategic goals

Goal #3 Enhancing consortium-wide collaborations Maintaining the data in the i2b2 research repository that is consistent with and reflects the latest versions of coded reference content supports semantic interoperability necessary to achieve data sharing across computing platforms. Currently, systems that lack interoperability in site-to-site implementation of the same platform or across divergent computing platforms are expensive to maintain and persist as a major barrier to use of systems and data for research.
i2b2 at the University of Michigan: Cohort discovery and beyond . . .

Stephen Gendler, Ph.D., University of Michigan, Ann Arbor, Michigan 48109
Director, Medical Center Information Technology, Business Intelligence, Interfaces, and Research
Kurt Richardson, University of Michigan, Ann Arbor, Michigan 48109
Manager, Medical Center Information Technology, Business Intelligence, Interfaces, and Research
Brian D. Athey, Ph.D., University of Michigan, Ann Arbor, Michigan 48109
Director, Academic Informatics (Medical School)

In 2010, several University of Michigan Health System (UMHS) groups explored the 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 (MICHR) and the UMHS Hospitalist Program. NCIBI completed a “proof of concept” integration of its automated annotation tool (Gene2MedSH). MICHR completed an in-depth evaluation of i2b2 in the context of the NIH-funded Physio-MIMI project. Medical Center Information Technology (MCIT) completed a demonstration project that connected UMHS with three other institutions together via the i2b2 and SHRINE (i.e. the multi-institutional networked version of i2b2).

Current State of i2b2 at UM

i2b2 Cohort: In early 2011, the 1.5 version of the i2b2 web client and data mart (3.2 million patients) was made generally available to all UMHS faculty and staff for cohort discovery. The UMi2b2 data mart includes patient demographics, diagnoses, procedures, medications and lab results data. To date, 20+ faculty and staff have been trained in cohort discovery using UMi2b2. The next phase of UMi2b2 Cohort will be the addition of Intelligent Medical Objects (IMO) coded diagnosis data (with SNOMED-CT codes) and our transition to the Epic EMR as the primary UMHS clinical data and documentation source system beginning in 2012.

i2b2 HOMERUN: Design and development work continues on the Hospital Medicine Reengineering Network (HOMERUN) project using the SHRINE capability of i2b2. This project group (made up of a number of University Hospital Consortium members) is now focusing its attention on VTE prophylaxis in the inpatient setting. We have partnered with Harvard for the CTSA SHRINE demonstration project as well.

i2b2 Cancer Registry: An i2b2 data mart loaded our NAACCR-based Cancer Registry data (6,400+ patients) is being piloted as a cohort discovery tool. An ICD-O ontology was developed to query this data in a meaningful way.

i2b2 TextMiner: In our current document repository, there are ~ 29M unstructured text-based documents. A project is underway to make that data available to researchers, using smart query and high speed indexing technology, via the i2b2 framework. The strong institutional commitment to i2b2 (based on preliminary activities) as a tool for significantly improving the UMHS cohort discovery process and study enrollment metrics, as well as the institutional commitment to Epic as our primary clinical system vendor, has spawned a larger institutional research architecture effort that hopes to design and implement a research data and integration platform that will serve the institution for years to come.

Value Proposition: An institutionally-supported i2b2 initiative has met 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.
Withdrawn
Identifying duplicate clinical records using administrative and clinical data
Michael J. Byrne, Susan C. Guerrero, Phillip S. Reeder, Elmer V. Bernstam
UTHealth School of Biomedical Informatics, Houston, TX

Abstract
Duplicate records are relatively common in clinical data warehouses. Duplicate records complicate clinical care and the re-use of clinical data for research. To re-use clinical data for research, these duplicate records must be identified and removed. Traditional record-linking algorithms (RLAs) are able to identify the vast majority of duplicate records using only demographic data. However, a small number of record-pairs, such as those belonging to infant (identical) twins, are difficult to classify since they share most (and possibly all) demographic variables. Adding clinical and administrative data along to demographic data can help classify the majority of “difficult” record pairs.

Introduction
Clinical data warehouses are prone to duplicate records where a single patient may have multiple Medical Record Numbers present in the database. Thus, a patient’s medical history may be spread across multiple records which are problematic for clinicians and for data mining. This problem has been recognized for decades and multiple record-linking algorithms (RLAs) have been developed to detect duplicate records. RLAs typically use demographic data (e.g., name, address, phone number, etc.) to determine whether a pair of records represents the same patient or two different patients. Thus, record-pairs are sorted into three categories: non-matches, potential duplicates, and matches.

In previous work, we evaluated three types of RLAs (deterministic, probabilistic and fuzzy inference) and found that most record-pairs were correctly classified by all RLAs. However, several types of “difficult” record-pairs limited the ability of the algorithms to reduce the number of potential duplicates. For example, identical twins share almost all demographic data and occasionally have very similar first names. The best-performing RLA could not reduce the potential duplicates to less than 3% of all record pairs based solely on demographic data.

Due to difficult records-pairs and the requirement that no records can be falsely matched, there was a limit to the performance achieved with demographic data alone. To classify the remaining potential duplicates, we included administrative (e.g., billing codes) and clinical data (e.g., labs, medications, notes, etc.). Clinical data are relatively patient-specific and can differentiate record-pairs. In the case of twins, (text) notes often include details that indicate a patient was a twin or triplet. However, twins or siblings also tend to see the same doctor, and have the same immunizations and tests performed, often at the same time. Thus distinguishing what clinical data are useful was not trivial and an optimized combination of clinical data must be found which can be used to identify record pairs without false matches.

Introducing clinical data into de-duplication can be used to identify at least 75% of record pairs as matches or non-matches. As with demographic data, clinical data are not always available and may be ambiguous. In some cases, lab results may be on one record and medications on another, making comparison impossible. In general, the addition of the clinical data to the de-duplication effort does allow us to significantly reduce the number of potential duplicates. Improving the quality of eMPI data is a necessary step in reusing clinical data for research.
Achieving Clinical Research Management System (CRMS) Adoption Through Coordinated Changes to Clinical Research and Billing Practice

Daniel E. Ford MD MPH, Michael B. Amey MAS, Harold P. Lehmann MD PhD, Diana Gumas MS
Johns Hopkins University, Baltimore MD

Background: Federal Medicare regulations require that clinical research medical charges (Hospital and Physician Practice), beyond accepted standard of care, must not be billed to Medicare and other Federal payers. Additionally, Johns Hopkins Medicine (JHM) need such excluded research charges to be paid by research study funds. At the same time, JHM collaborated with Medical Decision Logic Inc. to develop and implement the Clinical Research Management System (CRMS) to assist research teams in enrolling and tracking participants, resulting in registry of studies and research patient participants for research team tracking as well as clinical research billing compliance. CRMS is built with a Ruby on Rails front end, Microsoft SQL Server database and the XSB rules-based engine. Knowing that adoption of any enterprise technology in a distributed system like the typical academic medical center (and like most CTSA institutions) requires explicit attention, we attempted to combine these two initiatives. Having faculty engaged in enterprise-level clinical-research environments is the first step to satisfying Consortium Strategic Goals of building a national research capability and of enhancing Consortium-wide collaboration.

Intervention: Effective July 1, 2008, all protocols submitted to the Johns Hopkins School of Medicine IRBs are evaluated to determine whether the study could result in a clinical bill and thus require a Prospective Reimbursement Analysis (PRA). The PRA documents which procedures of each study meet the Medicare National Coverage Decision criteria for billing and which do not. The study protocol is reviewed against local and national CMS coverage decisions and CMS approved journals documenting current standard of care to indicate whether an event is billable to insurance and self-pay, or is to be paid for by the study. Studies which require a PRA must use the Clinical Research Management System (CRMS) to register the study and participants. A study is not approved by the IRB until study team members have taken required web training for both billing compliance and the use of the CRMS system, and the PRA has been finalized. Auditing is conducted to assure that study teams are registering new study participants into CRMS in a timely fashion. In addition to assisting study staff with the management of their study, CRMS assists backend office staff to determine whether a candidate study participant has insurance which will cover standard of care study events. An interface was created between the CRMS and our electronic IRB system which is the system of record for protocols. CRMS has also been interfaced with our Electronic Patient Record in order to assure that the participants' demographic data are correct and that links can be made with the clinical data repository if desired.

Results: Begun in July 2008, as of August 1st, 2011 the CRMS usage is up to 4,635 clinical research studies and 58,881 study participants. CRMS is used for all protocols at JHM, including Oncology and Pediatric protocols. Study PIs and research coordinators have been trained, and enrolled subjects patient care bills are diverted for manual review to confirm compliance with the PRA. CRMS standardized reports provide data for study teams, billing review, Sponsors, and research administration. Study teams also use the electronic protocol template tool which generates study and patient calendars, and the electronic case report form tool. Study team satisfaction with CRMS has been very high with only 10% of respondents dissatisfied, despite having CRMS mandated for use. As of June 2011, 750 PRAs have been created and more than $120 million have been adjudicated through the clinical research pre-bill review process, using CRMS.

Conclusion: The need to comply with Federal Medicare billing regulations can be used to foster adoption of an enterprise CRMS that also provides Principal Investigators and their study managers with a study and participant registry and electronic calendars and case report forms.
Withdrawn
An Open-source Informatics Management System and Community for the National Children’s Study

William R. Hogan\textsuperscript{1,2}, Umit Topaloglu\textsuperscript{1,2}, Anthony McGuire\textsuperscript{2,3}, Topeka Stacey\textsuperscript{2}, Shariq Tariq\textsuperscript{1,2}, Akheel Ahmed\textsuperscript{2}, Melissa Casteel\textsuperscript{2,3}

\textsuperscript{1}Division of Biomedical Informatics - University of Arkansas for Medical Sciences, \textsuperscript{2}Arkansas Study Center of the National Children’s Study, \textsuperscript{3}Arkansas Children’s Hospital Research Institute

We have adapted our Comprehensive Research Informatics Suite (CRIS) to the needs of the Arkansas Study Center of the National Children’s Study (NCS) and created the infrastructure for an open-source community of Study Centers to collaborate on its development and use for the NCS. CRIS is composed of several, entirely open-source software applications, and we share all our software modifications and other resources, including instrument implementations, with the CTSA and NCS community. Links to download software, documentation, instrument implementations, and other resources specific to the NCS are at \url{http://code.google.com/p/ncs-open}. All resources are freely available without obligation or registering at the site. The components of the IMS include: (1) caBIG Central Clinical Participant Registry (C3PR) for participant registration, collection of address and demographic data, and management of consents, consent versions, and consent signing data; (2) LimeSurvey for conducting NCS instruments and collecting numerous operational data elements; (3) a data warehouse for collating data from other applications and submitting all data to the Vanguard Data Repository (VDR) of the NCS; (4) Talend for extract, transform, and load (ETL) processes in the warehouse including formatting data into the VDR schema; (5) Open sYSTem for Entity Resolution (OYSTER) for resolving duplicate address records; and (6) our own portal based on LifeRay open-source software that gives users a single point of access to all applications as well as implements single sign-on. We plan to use the open-source Patient Study Calendar for tracking study activities and the open-source caTissue Suite for managing biological and environmental specimens. To date, we have registered 100 women, 50 babies, and * (rounds to zero) fathers in C3PR. We have conducted 500 instruments in LimeSurvey. Note that we have applied official NCS rounding rules to these statistics per the mandate of the NCS Program Office. We use custom-developed Talend ETL scripts to (1) extract data from C3PR and LimeSurvey and load them into staging tables in the data warehouse, (2) transform data into tables structured according the VDR XML schema, and (3) export data in XML format from those tables. In addition, the scripts contain logic to remove protected health information, check data integrity, and apply other business logic for data cleaning. We have submitted all study data to the NCS VDR in this manner. Overall, despite the fact that substantial effort was required to adapt C3PR and LimeSurvey to the needs of the NCS and to create the research data warehouse de novo, CRIS served as highly functional starting point for the informatics requirements of the NCS relative to beginning from scratch. In addition, CRIS functionality for other purposes, including cancer clinical research, has increased.
Identifying and Assessing the Impact of Biomedical Research

Kristi L. Holmes, Ph.D.¹, Cathy C. Sarli, MLS, AHIP¹, and Mae O. Gordon, PhD³

¹Translational Research Support Division, Bernard Becker Medical Library, Washington University in St. Louis School of Medicine, St. Louis, MO; ²Washington University Institute of Clinical and Translational Sciences; ³Department of Ophthalmology and Visual Sciences, Washington University in St. Louis School of Medicine, St. Louis, MO

Abstract:
The recent emphasis in demonstrating translational outcomes of clinical and biomedical research findings into clinical practice and community benefit has spurred a need for new methods beyond traditional citation metrics to document the impact of research. To understand the true impact of research, it is necessary to identify tangible indicators of impact that are not readily discernible using citation analysis. These indicators serve as a foundation for documenting the dissemination of research outputs and resulting diffusion of knowledge into meaningful outcomes such as: development of a research tool or database; contribution to the knowledge base; change in understanding of a disease, disorder, or condition; change in practice; change in community health; or change in public law or policy.

The objective of this effort was to determine a model to assess the diffusion of knowledge and resulting research impact of individuals, groups, institutes and institutions engaged in biomedical research. The research group initially used for this study was an NIH-funded randomized clinical trial, the Ocular Hypertension Treatment Study (OHTS). The focus of OHTS was to study the safety and efficacy of ocular hypotensive medication in preventing glaucoma. Bibliographic analysis was performed on 26 peer-reviewed journal articles resulting from OHTS. Additional research outputs and activities were also examined to locate evidence of research impact.

Traditional bibliometric analysis was not a sufficiently robust means for assessing the full impact of OHTS findings, nor did it reveal subsequent clinical implementation and community benefit. More in-depth analysis was required in order to provide a meaningful narrative of outcomes. Given the depth of evidence of research impact for OHTS that was revealed, the authors refined the framework into a standardized model: the Becker Medical Library Model for Assessment of Research Impact.

The Becker Model¹,² provides indicators of evidence of impact based on resulting diffusion of research outputs and activities, resources for locating evidence of impact, and also includes strategies that can be utilized by biomedical scientists to enhance their research impact. This model can be applied across disciplines and can be used to assess the application and diffusion of various informatics efforts. This presentation will discuss the Becker Model and give a variety of ways that it is currently being applied. Strategies for enhancing research impact will also be discussed.

Informatics Education: Partnerships and Opportunities

Kristi L. Holmes, PhD¹; Leslie McIntosh, PhD²; Lili Wang, MD, MS¹; and Rakesh Nagarajan, MD, PhD²

¹Translational Research Support Division, Bernard Becker Medical Library, Washington University in St. Louis School of Medicine, St. Louis, MO; ²Washington University Institute of Clinical and Translational Sciences; ³Center for Biomedical Informatics, Washington University in St. Louis School of Medicine, St. Louis, MO

OBJECTIVES
The Washington University (WU) Center for Biomedical Informatics (CBMI) and the Bernard Becker Medical Library work in cooperation to support a translational informatics initiative, which offers access to relevant information resources, and effective instructional approaches to better support information dissemination, facilitate a fluid exchange of data, and support the application and integration of software and databases.

METHODS
The WU CBMI aims to integrate informatics into all aspects of clinical and translational research functions through the development and application of novel software tools and also through hands-on support of these tools and approaches. The CBMI provides interdisciplinary teams composed of members with expertise in informatics and computer science, biostatistics and computational biology, regulatory knowledge and study coordination, biobanking, and genomics. The Bernard Becker Medical Library has on staff two bioinformaticists who bring diverse backgrounds (biochemistry, medicine, and biostatistics) have developed and implemented an extensive range of training resources and classes which reflect themes ranging from general to more specific topics focusing on specialized software analysis tools, databases, and biostatistics. The CBMI and Becker Medical Library have worked hand-in-hand for the last 5 years to offer training and consultation support services to everyone from novice users to experts.

RESULTS AND CONCLUSIONS
Effective support of translational informatics endeavors on campus can be challenging. Sometimes it can be difficult to identify immediate needs and also respond to those needs in an appropriate manner. These translational informatics education efforts have been received positively by the research community at WU, due in large part to the formation of a successful partnership between the CBMI and Becker Library. This approach may offer a novel way to promote learning opportunities that connect the clinical and research communities and offers the library an integral role in the support of the educational and research goals of translational science research centers.

This presentation will describe the programs in detail and will present our future plans for this partnership.
The Center for Clinical and Translational Science (CCTS) Biomedical Informatics (BI) Core at the University of Illinois-Chicago (UIC) would like to highlight our recent efforts to better utilize informatics for university-community partnerships. Our work emphasizes CTSA strategic goals in the areas of enhancing the health of our communities and our nation through community engagement and enhancing consortium-wide collaborations through social networking and data sharing. Presenting our three major projects in this area, the UICollaboratory, Community Partner Profiles in VIVO, and the UIC Community Mapping Project, has the potential to contribute to a national conversation about utilizing informatics to leverage research partnerships within the broader community.

The UICollaboratory:
We have developed a research social networking tool called the UICollaboratory through services provided by SciVal/Elsevier. The UICollaboratory includes faculty members at UIC who have an affiliation with one of the health science colleges and/or the CCTS. Each profile is built based on basic human resources information, publication data, and grant information. Human resources information includes college and department affiliations and contact information. Publication data includes National Library of Medicine PubMed citation data. Grants information includes automated pulls from NIH Exporter, the VA and the UIeRA, an internal grants database at the university. All data are linked to individuals in the faculty list, disambiguated, and presented in a web-based interface as a researcher profile. Researchers are mapped to their departments within the university, allowing users to browse researchers by department, search by name, or search by topic to find experts in a particular area. Version 1.1 launched in September 2010 and currently includes 2070 researchers, 1865 grants and 45,651 publications (as of August 1, 2011). Working with SciVal, we are developing database integration to connect the UICollaboratory with Community Partner Profiles in VIVO.

Community Partner Profiles in VIVO:
A key strength of UIC is the broad range of community-based organizations partnering in research endeavors. Leveraging these partnerships we are expanding research social networking beyond the traditional faculty research profiling tools to link with profiles highlighting community partners. We used the VIVO platform to create a new web-based new community organization profile template. The result of our efforts is a functional research networking site that includes profiles of community partners. Community sites may create a profile page and update as needed. Using VIVO, we created RDF triples that capture aspects of community partner organizations, including a short description, target population, community areas, services and programs offered. Community Partner Profiles also creates links on each profile page to relevant researcher profiles in the UICollaboratory and makes it easier for researchers and community-based organizations interested in partnering with the university to find each other. Testing and evaluation is still underway, but the Community Partner Profiles site will soon be available publically as linked open data on the web. As we further develop these VIVO enhancements to our research, we will share the tools through the VIVO user community.

The UIC Community Mapping Project:
The UIC Community Mapping Project creates networking infrastructure that will allow all interested parties-community leaders, faculty researchers, funders, community residents and others - to have access to real-time information about research partnerships active within the 77 Chicago community areas. This project enhances the functionality and expands the scope of an already-successful prototype interactive map of active, partnered research in Chicago community areas developed by our Community Engagement Core. We are also developing strategies to incorporate data from the other Chicago-area CTSAs (Northwestern University and the University of Chicago) and examining possible linkages to other research databases (such as Community Partner Profiles) that could further enrich the mapping program and create a more complete picture of active community/university research partnerships in Chicago. The processes, databases and programs developed in this effort will serve as a model for similar efforts for other CTSAs in urban areas, particularly those with multiple CTSAs.
Registar

Registar is a registry management system developed at Northwestern University Biomedical Informatics Center with the aim of assisting researchers with study recruitment by allowing them effectively recruit and manage their own personal pool of potential research participants. Registar has a patient-facing portal that allows participants to self-register for any one or combination of the available registries. The system has specific steps to facilitate the collection of informed consent, collection of general demographic information as well as any additional customizable questions that the researcher may want to ask participants. Registar has several useful features that help with the management of potential research participants:

- Ability to automatically send out emails for participants to renew/update their information after a configurable amount of time
- Ability to track and send personalized emails to participants reminding them to complete the registration process or thank them for doing so
- Ability to search through specific participant responses while trying to determine eligibility for a given study
- Ability to track participants who have been recently contacted and are currently on a study.
- Ability to download a data dump of all information on all participants

Registar is available to all researchers both within and outside the Northwestern family of Institutions, as long as they have IRB approval for their registry.

eNOTIS

The Enterprise NOrthwestern Trial Information System was designed and built to help all researchers at the Northwestern family of institutions track and manage research subjects. In addition to providing researchers with the ability to track basic subject information and demographics as well as various clinical events, eNOTIS also provides for the electronic data capture of both case report forms as well as patient recorded outcomes. eNOTIS has been integrated into our Enterprise Data Warehouse for the efficient building of patient cohorts, participant information is merged into our enterprise master patient index, and is used to accurately identify research participants with our clinical partners.

Both applications were written in Ruby on Rails using agile methodologies including continuous integration, are fully tested and source code is available at github.com/nubic/
Title: NCS Navigator MDES Warehouse

Overview

NCS Navigator is a combination of time-tested tools and utilities custom-made for the needs of the National Children’s Study. NCS Navigator MDES Warehouse is a specialized tool for extracting data from the NCS Navigator suite. It transforms the data according to the semantics defined by the study’s Master Data Element Specification and makes it ready to load into the Vanguard Data Repository.

The warehouse automatically refreshes and validates itself on a daily basis, allowing study analysts to identify data problems in advance of the biweekly VDR submissions to the program office. And while the system was developed for performing VDR submissions, as a data warehouse it supports other forms of reporting.

Technical Details

The warehouse is built using the Ruby programming language on top of the PostgreSQL database management system. While it was designed with the needs of the NCS Navigator suite in mind, its transform modules are replaceable. This means that it could provide the same services to other NCS-supporting applications or suites.

The warehouse derives its understanding of the Master Data Elements automatically from the VDR documentation where possible. This means that its data model can largely be automatically updated as new versions of the MDES are published.
Faculty Connect

Faculty Connect is a web-based application that matches FSM faculty to non-federal biomedical research funding opportunities. Faculty are identified for specific grants through keyword curating and a sophisticated matching algorithm that uses semantic logic to search demographics, research interests and MeSH terms from publications to select the best fit for each opportunity. Faculty Connect enables navigators, who have extensive experience with non-federal funding, to verify the match fidelity, supplement the opportunity announcement with additional sources of guidance, and assist faculty who are interested in pursuing an opportunity as needed. The application will present matches for approximately 400 non-federal funding opportunities during the next year to faculty whose research interests closely match the eligibility and funding criteria.

Faculty Connect is an informatics resource for non-federal funding that addresses several important issues:

1. Faculty Connect performs an automatic import of complex information about FSM faculty from the FSM faculty Database.
2. Faculty Connect imports funding opportunities from public and privately licensed sources and provides navigators with a way to review and update.
3. Faculty Connect has a high fidelity algorithm that matches faculty to funding.
4. Faculty Connect facilitates collaboration among faculty using LatticeGrid.
5. Faculty Connect augments communication between research navigators and faculty.

The goal of Faculty Connect is to enhance faculty efforts to secure non-federal funding.

Key figures since launch on April 28, 2011

- Matches generated: 2602
- Matches presented: 674
- Matches pursued: 30
- Value of matches pursued: $29M

Faculty Team

Michael Abecassis MD.................................................................Co-Principal Investigator
Philip Greenland MD...............................................................Co-Principal Investigator
Warren A. Kibbe PhD.................................................................Investigator

Project Team

Cynthia Csernansky PhD’ ...............................................................Project Lead
Michael Gurley ............................................................................Lead Developer
Paul Friedman .............................................................................Developer
Cara Clifford .............................................................................Data Administrator
Abby Cosentino-Boehm .............................................................Project Manager
Lorraine Padour ........................................................................FSM Database Administrator
Laura Wimbiscus Yoon .................................................................Usability Analyst

This project was supported by a grant from the state of Illinois Excellence in Academic Medicine fund through Northwestern Memorial Hospital, and the NIH CTSA grant award UL1RR025741 to the Northwestern Clinical and Translational Sciences Institute from the National Center for Research Resources.
LatticeGrid

LatticeGrid is a modular, flexible approach to assessing collaboration patterns using institutional and public data, such as PubMed. LatticeGrid can incorporate data from a variety of sources, integrate it around organizational constructs such as centers and departments, and build models of collaboration patterns. Importantly, LatticeGrid provides flexible mechanisms for acquiring and processing these data and can consume ontological representations of annotations, enabling an automated, knowledge-based analysis of semantic content of connections between investigators and projects. The ability of LatticeGrid to define collaboration teams “on the fly” is a unique feature not found in other collaboration tools. Another unique component of LatticeGrid is that it provides multiple ways to represent physical and virtual relationships between investigators and map those organizations onto collaboration patterns between investigators. This makes LatticeGrid an ideal companion to existing social networking and knowledge management platforms.

LatticeGrid models enable evidence-based decisions about how best to manage and direct organizational change to maximize the effectiveness of translational science initiatives. LatticeGrid enables biomedical research organizations to monitor and measure more effectively collaboration and funding patterns. By studying how these patterns change with time, organizations can use this information to evaluate the effectiveness of changes in organizational structure and policies. Understanding how shifts in institutional structure and policy affect translational science patterns of collaboration and funding is of fundamental importance to the biomedical research community. This knowledge provides a mechanism to evaluate the “return on investment” for team science efforts.

The core of LatticeGrid can be viewed on the Robert H. Lurie Comprehensive Cancer Center of Northwestern University website at [http://latticegrid.cancer.northwestern.edu](http://latticegrid.cancer.northwestern.edu) and at the Northwestern Feinberg School of Medicine website at [http://latticegrid.feinberg.northwestern.edu](http://latticegrid.feinberg.northwestern.edu). The UCSF Helen Diller Family Comprehensive Cancer Center website is at [https://latticegrid.cancer.ucsf.edu](https://latticegrid.cancer.ucsf.edu). The Fox Chase website is at [http://staffpubs.fccc.edu](http://staffpubs.fccc.edu). Other groups that have adopted LatticeGrid include the Cleveland Clinic Cancer Center, the Cancer Institute of New Jersey and the University of Michigan Comprehensive Cancer Center, although these instances are behind institutional firewalls.

LatticeGrid is designed to:

1. Identify the network of collaborations that currently exist in a biomedical research organization.

2. Highlight the expertise and existing collaboration patterns for an individual in that organization.

3. Analyze changes in the pattern of collaboration for a given individual, virtual organizational unit, academic unit, or other organizational structure over time.

4. Facilitate the application of this information to the promotion and support of intra- and inter-institutional collaboration.
A Paradigm Shift: Electronic Health Records Data in Clinical Practice

Ketan K. Mane\textsuperscript{1}, Chris Bizon\textsuperscript{1}, Phil Owen\textsuperscript{1}, Javed Mostafa\textsuperscript{2}, Kenneth Gersing\textsuperscript{3}, and Charles Schmitt\textsuperscript{1}

\textsuperscript{1}Rennaisance Computing Institute, Univ. of North Carolina, Chapel Hill, NC; \textsuperscript{2}School of Information and Library Science, Univ. of North Carolina, Chapel Hill, NC; \textsuperscript{3}Psychiatry Dept., Duke University Medical School, Durham, NC

In clinical practice, the longitudinal data in electronic health record systems (EHRs) for patients with chronic medical conditions remain an untapped resource. Retrospective data analysis of patient data in EHRs carries tremendous potential to provide clinical evidence to aid the clinician in their decision making process. From a clinical decision support (CDS) perspective, we consider two fundamental ways EHR data can assist the clinician. First, EHR data can be used to gain a better understanding of an individual patient’s treatment profile to support clinical decisions. Second, by examining historical EHR data from comparative patient populations, the clinician can understand the potential impacts of treatments options on a presenting patient. However, a significant bottleneck to use EHR data at point of care is its large volume. In a clinical setting, clinicians need an enhanced capability to make sense of large volume of patient data in the limited amount of time they spend during a patient visit or in one-on-one interactive sessions.

The visual analytics (VA) approach offers a potential solution to effectively present evidence from EHRs in clinical practice. As part of the collaboration between Renaissance Computing Institute (RENCI) and Duke University, we explored the use of VA techniques to aggregate and visually present EHR level data for patients with Major Depressive Disorder. Compared to a tabular view of patient’s treatment outcome, a visual representation of EHRs patient-level data helps to better understand the trends in a patient’s response to past treatments. Visual data layering techniques helps build context for other patient-level information (medications, visit type, side-effects, overall wellness, etc.), and to understand treatment impacts. The VA approach helps to define, aggregate, and build summarized views of clinical evidence from a comparative population. Flexibility to apply ad hoc data filters in VA helps to customize presented evidence at different levels – demographics, comorbidity, etc. Figure 1 shows an interactive CDS dashboard prototype that was build using the VA approach. Interactive dashboard user interface presents patient-level and comparative patient-level evidence for clinical decisions, and supports a patient-centric clinical care paradigm. This initial work is extended in collaboration with UNC CTSA to directly target the meaningful use of EHRs with a particular focus on improving the human computer interface of CDS systems.

Our long term aim is to take advantage of the CTSA Network to disseminate the findings of our work to a broader community. Further, we intend to extend the work to health IT research training and as part of educational consortium established between UNC and Duke focused on health IT graduate training programs.

Figure 1: Dashboard view of the visual analytics based clinical decision support tool: (1) Patient demographics, (2) Outcome response of similar patient, (3) Comorbid conditions of patient and comparative population, (4) Guideline-patient visit view, (5) Temporal view of patient outcome response, (6) Projected response to selected medication, (7) Duration of prescribed medication, and (8) Visit type information.
Docking simulations and virtual screening have become an integral part of drug design pipelines, facilitating experimental and translational efforts in this regard. However, performing virtual screening and analyzing the results with the goal of identifying lead compounds requires resources and expertise that may not be available within individual academic centers, or could be shared across multiple centers to optimize the use of such resources. We have developed a number of software solutions that streamline setting up docking simulations for virtual screening, integration with cheminformatic packages, parsing of small molecule libraries, as well as analysis and visualization of the results. In particular, tailored extensions of the recently developed Polyview-MM portal can be used for automated (web-based) visualization of docking poses for large compound libraries, and re-ranking of hits by clustering of poses coupled with cheminformatic analyses. Owing to the ease of use and higher level of integrations, both basic researchers and medicinal chemists can directly work together to analyze the results and streamline the selection of hits for further validation and lead development. Successful collaborations of this type within the University of Cincinnati College of Medicine and Cincinnati Children’s Hospital Research Foundation, spurred by translational efforts within the local CTSA funded center, indicate that these solutions can now be extended to other CTSA centers. Demonstration of these capabilities, as well as discussion of future extensions, will be the focus of this presentation falling under Theme 2 of the conference.

PROJECT SUMMARY:
Performing research in community settings requires both a network of practices and community partners committed to research and a system to facilitate researchers’ access to these resources. The Indiana CTSI has established primary and specialty care practice-based research networks (PBRNs) and community partner organizations, and it is fortunate to have one of the nation’s most sophisticated health information networks that includes most Indiana hospitals, laboratory and imaging facilities, pharmacies, and more than 10,000 physicians’ practices. In addition, several major hospital systems and health care groups in Central Indiana (representing >1500 physicians and 500,000 patients) have merged into a single community-based clinical entity, Indiana University (IU) Health, that has a strong research mission. To take full advantage of these resources, researchers need to know how to navigate an unfamiliar array of networks, practices, and information systems. The Indiana CTSI created the Central Indiana Innovation Network (CI-Net) and its Research Planning Team (RPT) established (1) linkages among existing PBRNs; (2) enhanced subject recruitment into a wide variety of studies by importing registration and contact data and upcoming appointments from registration and scheduling systems in CI-Net practices; (3) a registry of patients who are interested in participating in randomized trials of new drugs and devices, linking enrolled patients to their electronic medical records to enhance condition-focused recruiting; (4) enhanced subject recruitment and management in collaboration with the IU Clinical Trials Program; and (5) training for researchers to access the RPT and its information and management resources to plan and implement translational research projects. By the end of the 2-year supplement and a 1-year extension, CI-Net will be the primary connection between the community and researchers at Indiana, Purdue, and Notre Dame Universities. CI-Net will be sustained through ongoing research and the increased volume and efficiency of recruiting subjects into clinical trials of drugs, devices, and biologics, helping Central Indiana become a national resource for testing of new therapeutics and translating practice-based evidence into evidence-based practice.

Providing the best care requires knowing how well it works in everyday community practices and providing access to community-based research resources requires a knowledgeable team of facilitators who can help investigators identify which community practices or individuals in the community are appropriate subjects for their studies and how to recruit them efficiently and effectively. CI-Net and the RPT will be that resource for central Indiana and the Indiana CTSI.
ResearchIQ: Semantic Search for the Research Community
Omkar Lele, MS1; Rakesh Dhaval, MS1; Tara Borlawsky, MA1; Peter J. Embi, MD, MS1,2; Philip R.O. Payne, PhD1,2

1The Ohio State University, Department of Biomedical Informatics, Columbus OH
2The Ohio State University, Center for Clinical and Translational Science, Columbus, OH

Introduction
The development and execution of integrative clinical or translational research is significantly limited by the propagation of data and knowledge “silos”. As such, there is increasing interest in building systems that facilitate interaction with distributed biomedical knowledge collections1, and a growing need for tools that will allow for the semantically anchored discovery of both local and public heterogeneous and multi-dimensional biomedical linked data spanning organizational boundaries2. To address these challenges we have developed ResearchIQ, a query platform that employs Semantic Web technologies that harnesses the knowledge embedded within ontologies in order to mine and discovery a wide variety of resources like websites, databases, expertise profiles, and grid-based resources.

System Design
Metadata Annotation Engine (Meta-Engine): The Meta-Engine is a semantic annotation engine that generates semantic metadata as RDF triples (http://www.w3.org/RDF) corresponding to the various input resources (e.g., websites, databases, grid data services). It uses lightweight natural language processing (NLP) tools to generate semantic annotations and relationships.

Knowledge Repository and Query Engine: The semantic metadata obtained from the Meta-Engine is stored in a centralized knowledge repository, or RDF triple store. RDF provides a common framework for expressing this metadata with minimal loss of meaning and context. Knowledge that has been mined from a subset of collections within UMLS (SNOMED-CT, MSH, NCI-T, and GO), and associated metadata about resources, are stored as named graphs in the triple store.

Discussion
The beta-version of ResearchIQ has been deployed at The Ohio State University in the context of motivating clinical and translational research use cases, and allows researchers to intuitively query for resources across a set of 20 shared resource websites, 2 publicly available datasets (OAI3 and StudySearch4) and a subset of the OSU expertise profile system (OSUPro5). The pilot system has allowed us explore the technical features described above and to seek feedback from end-users. A production implementation and deployment of ResearchIQ is planned for Fall 2011, to be followed by usability and human factors evaluations to determine factors that will influence the adoption of the system.

Acknowledgement
This project was supported by NIH/NCRR UL1-RR025755.

References
5. OSUPro, http://pro.osu.edu/
StudySearch: A Tool For Connecting Potential Participants with Locally Recruiting Studies

Tara Borlawsky, MA¹; Daniel Carpenter¹; Bart Kelsey¹; Robert Rice, PhD¹; Carson Reider, PhD²; Peter J. Embi, MD, MS¹²; Philip R. O. Payne, PhD¹²

¹The Ohio State University, Department of Biomedical Informatics, Columbus OH
²The Ohio State University, Center for Clinical and Translational Science, Columbus, OH

Introduction & Background
Sung et al.¹ identified a paucity of individuals willing to participate in clinical studies as one of the central challenges facing the clinical research enterprise. ClinicalTrials.gov provides a centralized listing of federally and privately supported clinical trials. However, not all non-intervention research studies are included. Additionally, some reports suggest that the available search methods are not particularly user-friendly, and the information presented is difficult to digest by the general public. StudySearch² provides a web application for easily searching a comprehensive and layperson friendly listing of research studies that are approved for recruitment at The Ohio State University (OSU), as well as allows regulatory personnel to iteratively communicate with investigators to refine the study listings (Figure 1).

System Description
All studies actively recruiting at OSU and listed on ClinicalTrials.gov are imported monthly into StudySeach. Additionally, the OSU Center for Clinical and Translational Science (CCTS) staff can manually add studies not otherwise listed in ClinicalTrials.gov to the StudySearch database. All materials are reviewed by the CCTS Regulatory Core, which also adds the corresponding Institutional Review Board (IRB) protocol and approval information for all studies. The respective research team is then provided with the opportunity to review and edit details (e.g., contact information) and verbiage (e.g., purpose, eligibility criteria) of the study. After both the study investigator(s) and CCTS staff approve all details, it is made available to a public search on the CCTS web site (http://ccts.osu.edu/studysearch). End users may search for studies based on participant age, gender and/or diagnosis or intervention. MeSH diagnosis and/or intervention terms are either parsed from the ClinicalTrials.gov entry or added by the CCTS staff. StudySearch has been developed as an open-source Drupal module and is available to be adopted by the clinical and translational community (https://wiki.citih.osumc.edu/display/studysearch).

Discussion
There are currently 132 studies recruiting at OSU that are listed on the CCTS website, and 28 under review. Future work includes evaluating the impact of StudySearch on the recruitment process, integrating it with the eIRB, and enhancing the user experience through the use of semantically anchored contextual and definitional information. In conclusion, StudySearch allows individual sites to closely monitor the regulatory compliance and web-based advertising of locally recruiting studies, as well present the general public with an easily digestible listing of recruiting studies.

Acknowledgement
The authors acknowledge the input provided by Michael Para, MD and Rose Hallarn. This project was supported by NIH/NCRR UL1-RR025755.

References
It has become clear in recent years that to better understand diseases we cannot continue to focus
only on single genes, proteins or metabolites operating in single linear-ordered pathways. The integration
of data derived from large scale high throughput technologies enables a more comprehensive view of the
complex interactions occurring within body fluids or tissues at any one time. To be able to support these
large translational studies we have established an informatics infrastructure that is capable of supporting
high throughput bio-specimen banking workflows. Our development is based on the open source
biorepository software tool caTissue Suite 1.2.

Analysis of high quality molecular data requires an informatics infrastructure that is not only
capable of capturing all of the events related to a biospecimen’s lifecycle, from the bedside to the
laboratory, but is also capable of reducing of the number of errors that result in biospecimen
misidentification. To ensure the quality of biospecimens and to document their history, including the
chain of custody, an automated data collection (ADC) system utilizing barcodes was implemented. This
ADC system, caTrack, includes a portable digital assistant (PDA) with software that recognizes
sequences of barcode scan events as actions in the real-world in the form of associations between objects.
For example, associations of patients with collection tubes and collection tubes with aliquots are made by
scanning sequentially the patient’s identification bracelet, the specimen container, and the aliquot tubes.
To ensure that this implementation is scalable nationwide, we implemented a global unique identifier
scheme based on ISO object identifiers that consist of a worldwide prefix, type-of-object code, and an
unique ID that consists of the parent ID and the biospecimen serial number. Encoding the type of object
into the barcode enables our ADC software to recognize the types of objects scanned and to infer the kind
of associations to be made.

In addition to tracking the events on all biospecimens, large translational studies also require a
robust infrastructure to track consents and all metadata associated with a subject’s consent, such as the
version signed, the person obtaining consent or the date the consent was signed. Currently this
information is typically tracked manually by a study coordinator, which is time- and resource consuming.
To capture the data from the consents and to store the digital image of the consent within the database, we
have implemented an automated processing method using a forms optical character recognition (OCR)
product, Cardiff / Teleforms, that captures hand-written user input from the consent forms, routes the
consent forms, and integrates the data from the consent into caTissue Suite 1.2. In our implementation, all
scanned forms are dropped off in a common inbox. The system not only recognizes the form’s barcode
and routes it to the study-specific archive folder but it also recognizes check-marks, hand printed digits
and the pasted subject labels. We have implemented an enhanced interface to a common relational
database, which holds data from all projects and forms. In a final step the data is moved into caTissue
Suite 1.2 by direct database operations controlled by metadata.

Translational studies also require the collection of clinical data at the point in time the specimen
is collected. In our experience, paper-based data collection remains the preferred method of physicians
and clinical research coordinators. The drawback is that the data has to be manually entered into the
database. To allow physicians to use their preferred mode for collecting data in the clinic but eliminate the
error prone manual data entry step, we have implemented the same automated forms-processing systems
that we used for capturing the consent for the clinical data forms and questionnaires. This enables us to
capture structured clinical data and allows for storing the digital image of the questionnaires within the
database. To allow the acquisition of data through telephone interviews using specialized software we
integrated the software tool WinCATI, which was specifically developed for conducting telephone
interviews, with catissue Suite 1.2.

In summary we have developed and integrated several software tools to support high throughput
bio-specimen banking workflows at the Indiana Clinical and Translational Sciences Institute.
Translational and clinical studies in neuro-oncology would benefit from a refined molecular classification of diffuse gliomas that allows therapies to be directed at those signaling networks that drive tumor progression. It is becoming increasingly clear that diffuse gliomas should be recognized as a number of molecularly discrete diseases, each with its own tumor-promoting signaling networks, which can be exploited therapeutically once master regulators have been elucidated. Signaling network pathogenesis is best interpreted with a quantitative understanding of the interplay between histopathology, genetics, and genomic expression. An integrated analysis of these multi-scale data (micro-anatomic morphology, genomics, molecular networks, and clinical outcome) can lead to a better classification of glioma subtypes with the ultimate goal of therapeutic targeting of underlying mechanisms of progression for each subtype. The computational and data management requirements associated with this type of analyses are extremely substantial. Analyses can involve hundreds to thousands of images, executions of many interrelated algorithms, and query and exploration of hundreds of billions of morphological features and integration of these features with genomic and outcome data.

This project, supported in part by the NCI/cabIG® In Silico Centers of Excellence program and by the Atlanta Clinical and Translational Science Institute, develops information models, analytic pipelines, high performance computing and data management to support systematic, large-scale comparative analyses of brain tumors using high resolution whole slide images and high throughput genomic data. We will present our work on the development of techniques for analysis of high resolution pathology images, high-performance computing support for executing data and computation intensive analysis pipelines on large image datasets, and data models and databases to handle very large volumes of analysis results. The initial scientific work using image analysis supported by high performance computing and data management middleware has led to results that demonstrate morphological subtypes of glioblastoma not previously recognized by pathologists. The survival characteristics of this morphology-driven stratification are significant when compared to the survival of molecular subtypes, suggesting that morphology is a significant predictor of prognosis.

Why it is important to be presented at the 2011 IKFC meeting: High-resolution, high-throughput data offers enormous opportunities for examining characteristics of complex disease mechanisms and driving new treatment strategies. With improving cost-effectiveness of scanners, it is rapidly becoming feasible for even medium-scale studies to routinely generate thousands of whole slide images. Nevertheless, image-based data has been underutilized in biomedical research, because of the formidable information synthesis and management challenges that it presents. Our work addresses these challenges that are increasingly common in translational and clinical studies.

Impact on the CTSA Consortium Strategic Goals and on local CTSA activities: Our effort is driven by the vision of delivering interoperable, integrative information technologies that facilitate quantitative data-directed healthcare. These technologies are designed to efficiently interrogate and explore data at multiple scales and multiple modalities, integrate it across scales and dimensions, and extract and compare information across individuals and cohorts. They have potential to significantly enhance the study of common and discriminating biological characteristics across cohorts of subjects creating categorizations of patient subpopulations and determine targeted prevention and treatment strategies tailored to individual patients.

Acknowledgement. This research is supported in part by SAIC/NCI Contract No. HHSN261200800001E from the National Cancer Institute, PHS Grant UL1RR025008 from the CTSA program, NCI Contract No. N01-CO-12400, 85983CBS43, and 94995NBS23, by R24HL085343 from the NHLBI, by Grant R01LM009239 from the NLM, and P20 EB000591 by the BISTI program.
Abstract 65

Integration of drug classification information with EMR data in the utCRIS Data Warehouse for adverse drug event association analysis

Richard H. Scheuermann1,2*, John M. Armstrong3, Teresa Bosler4, Christen Buseman2, Jennifer Cai4, Ramona R. Leibnitz3, Dipti Ranganathan2,4, Donglu Xie4

1Department of Pathology, 2Division of Biomedical Informatics, 4Academic Information Systems, U.T. Southwestern Medical Center, Dallas, TX, USA, 3Lead Horse Technologies, Junction City, KS, USA

As part of the CTSA program, the UT Southwestern Bioinformatics Key Function began the development of an integrated UT Southwestern Clinical Research Information System (utCRIS) to support the data management and data analysis needs of clinical and translational investigators. One of the major components of this system is the utCRIS Data Warehouse (utCRIS-DW), built upon the i2b2 infrastructure. utCRIS-DW implementation involved the development of extraction, transformation, and loading scripts to pull, clean, and normalize data from five different UT Southwestern transactional systems – the University Hospital and Ambulatory Clinics EMR systems, the enterprise Master Person Index (eMPI), and the Ambulatory billing systems (IDX and EPIC Resolute) The current utCRIS-DW is focused on structured data fields with records from >3.8 million patients, including over 67 million diagnoses, 115 million procedures, and 7.1 million ordered medications. One of the uses of the utCRIS-DW has been to mine this large dataset for evidence of adverse drug events.

Although clinical trials of new therapeutic formulations are designed to assess safety and efficacy, in many cases the number of participants enrolled may not be sufficient to identify relatively rare adverse events. The CTSA Consortium provides a potential mechanism to assess adverse drug reactions in post-market studies by increasing the number of treated subjects by developing data sharing mechanisms and common analysis strategies as highlighted by the Data-Sharing Sub-goal B of the Enhancing Consortium-wide Collaborations Strategic Goal Committee. In order to establish the relevant data standards and analysis methods, we performed a pilot analysis using the utCRIS-DW. To generate a baseline dataset we applied artificial intelligence (AI) algorithms to identify 3896 different drug-adverse event association rules from the MedWatch database (2002 – 2010), which contains information about possible drug-related adverse events observed by the medical community post marketing approval from the FDA’s Adverse Event Reporting System (AERS). In order to determine if we find evidence for drug-adverse event associations in our patient populations, we queried the utCRIS-DW to determine the number of subjects treated with the identified drug or drug class, the number of subjects with a diagnosis equivalent to the putative adverse event, and the number of subjects treated with the identified drug AND with a diagnosis equivalent to the putative adverse event for each rule. These data were assembled into 2 x 2 contingency tables of drug treatment and no drug treatment versus diagnosis and no diagnosis, followed by odds ratio statistical tests. Of the adverse events classified as serious, we found that 82% of the adverse event AI rules identified from the FDA AERS/MedWatch database could be observed as significant overrepresentation of drug/diagnosis combinations in our utCRIS-DW patient population. Through this pilot we have defined the procedures necessary to interrogate local clinical research data warehouses for evidence of adverse drug reactions that will be shared with the CTSA Consortium, and are now designing strategies to perform cross-consortium meta-analysis.

Supported by UL1 RR023468
* Richard Scheuermann is a member of the Science & Medicine Advisory Board of Lead Horse Technologies
**DV Docs: Generating CVs and NIH biosketches from VIVO data**

William Shirey, MS¹, Shiyi Shen, MS¹, Stella Mitchell², Brian Lowe², Jon Corson-Rikert², Linda M Schmandt, MS¹, Titus Schleyer, DMD, PhD²

1 University of Pittsburgh    2 Cornell University

**Description of the project:** Research networking systems such as VIVO and Digital Vita (DV) facilitate the discovery of researcher interests and the establishment of high-impact, productive collaborations. While these functions are important, they are not always sufficient to fully motivate adoption of the systems among their target audiences. At the University of Pittsburgh, the Digital Vita development team has adopted the strategy of attracting users by providing them with the ability to generate various documents relevant to their work, such as a Curriculum Vitae (CV), NIH biosketches, evaluation forms and online profile pages from their profile data. The DV Docs software will extend those benefits to the entire VIVO community.

The VIVO application provides the capability to manage most of the same data as DV, but to date has had no tools for generating documents populated with the data. To fill that gap our teams developed a way to combine DV’s document generation features with VIVO data. First we mapped the DV data elements to VIVO, identifying the VIVO data elements needed to populate a CV and a biosketch. We then obtained data in VIVO RDF-XML format and used XSPARQL (http://xsparql.deri.org/) to transform that format into the XML that DV accepts as input to its document generation module. DV’s document generation features were made available through an Application Programming Interface written in Java and provided as a .jar file; a web-service version is also available. We are now using the API to generate CVs from VIVO profiles. Generating other types of documents – such as performance evaluation forms, departmental reports, and others – from the same data will require only providing different XSLT instructions.

Digital Vita allows users to maintain a library of personal statements for NIH biosketches, specify which personal statement, positions, publications and grants should appear on a biosketch, and identify the publications that are most relevant to the current application. Since those features are essential to creating a complete NIH biosketch and are currently absent from VIVO, the next step will be to create a standalone utility for biosketch generation and management that is based on current DV functionality and uses the web service in order to provide VIVO users with these capabilities. This will require implementing a persistent data store for non-VIVO NIH biosketch data and a means for users to authenticate themselves.

**Importance to the 2011 IKFC meeting:** The two-year VIVO project, funded as part of the American Recovery and Reinvestment Act, is coming to an end, and it is important to its success that as many users as possible from all CTSA institutions create profiles as the basis for research collaboration. Many participants at this meeting will be aware of some of the results of the VIVO project, but they may not be aware of the additional features available as a result of the DV Docs project. Our poster will provide additional incentive to use VIVO.

**Impact on the CTSA Consortium Strategic Goals:** This work is central to CTSA Strategic Goal 3, enhancing consortium-wide collaborations. By providing researchers with the immediate, tangible benefit of CV and biosketch creation, it encourages use of VIVO, CTSA’s primary research networking data repository. When this application is complete, any researcher with a well-populated VIVO profile will be able to generate not only a CV but also targeted NIH biosketches and other useful summaries of scholarly interests and achievements.
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.
Integration of Clinical Data for an ICU Registry

Theme: Integrated Data Repositories

Description:
A prospective trial of a biomarker for medical intensive care unit (MICU) patients was proposed with a shift to a long term research registry after 300 patients were enrolled. Our hospital has 3 MICUs with a total capacity of 30 patients. The majority of the data was available electronically but some manual data collection would be required at the bedside.

The informatics plan was to collect and store all data into one database from all sources. Oracle 10g was chosen as the platform. Real-time data from the ADT system (Admission, Discharge and Transfer) was required to determine which patients had been admitted and discharged from the units. This was accomplished using a commercial HL7 parser which loaded the ADT information into the database. Based on these medical record numbers and dates, data was acquired from the electronic medical record (EMR) and the APACHE© system. In addition, lab results from the new biomarker were loaded into the research database. Data entry forms were also created in Oracle Apex to complete the data collection at the bedside.

300 patients were accrued in 6 weeks at which point the manual data collection was discontinued and the automated data feeds continued in the registry mode. This novel approach to utilizing multiple data sources for both prospective and retrospective data collection provides the investigators with rapid access to a wealth of clinical data for research on a patient population traditionally difficult to study because of the volume of data to be collected.

Why it is important to present at the IKFC:
This project represents a novel approach to a data repository in a challenging population. It provides a roadmap for others who are targeting ICU patient populations. This also lays groundwork for incorporating genomic or proteomic data with clinical for future studies.

Impact on Strategic Goals
Enhancing Clinical and translational research, specifically, this project demonstrates an approach to combining biomarker data with data from multiple clinical sources. While studying ICU patients is a major challenge using manual data collection, integrating data from clinical sources using available data transfer methods can enable both prospective studies and retrospective registries in this population.

Background: Like many academic institutions, Duke’s systems for research support have traditionally been fragmented and driven by funding. Researchers working in better-funded departments have enjoyed access to a rich infrastructure supporting multi-center clinical trials including statistical support, data management support and a robust set of clinical research applications. However, researchers in poorly funded areas may have lacked the necessary tools to do efficient, compliant and effective research. This inequity may have resulted not only from differences in funding, but also from a lack of knowledge that support resources existed.

Our Approach: In 2009, the Duke Bioinformatics Core (BIC) began to bridge these gaps by providing consulting services and supporting technology. As this service evolved, it became apparent that success required connecting the researcher directly with a service that both meets their need and was appropriate for the project funding level. The BIC identified the Duke School of Nursing’s Research Management Team (RMT) and extended their model of research service support across the Duke Medicine enterprise. Coupling RMT with Vanderbilt’s REDCap created a much improved service but not complete. First, Duke recognized the need for testing, validation, and compliance for all collected research data regardless of the researchers’ financial means. The validation process was considered critical not only regarding local decisions about the use of REDCap within Duke, but we also believed we could provide a validation framework, consisting of documentation and templates that would benefit the entire REDCap consortium. Second, Duke has offered enterprise level support systems with great success because they have provided accompanying support. It was necessary to reduce the support gap created by differing levels of funding by emulating this support on a small-scale for a reasonable cost offered to a much wider and disparate customer base.

Observations and Findings: As a HIPAA covered entity, Duke required formal testing, validation and change control for systems carrying PHI. This had not been done by Vanderbilt or the consortium. Duke spearheaded this effort using its existing expertise with validation and change control offering SOPs, templates and documentation. Additionally, Duke influenced Vanderbilt to modify its process for introducing new releases separating fix releases from enhancement releases. Other institutions have successfully adopted these processes and templates. Adding REDCap into the toolbox of RMT’s existing research support service model with extension across the Duke Medicine enterprise has been overwhelmingly successful. RMT has grown with the addition of REDCap, from servicing 30 individual research projects with 6.5 FTEs of staff in 2009, to servicing over 90 projects with 9 FTEs of staff in 2011. In a customer satisfaction survey sent to 117 RMT users (52% response rate) this fiscal year, 98 percent of responders were somewhat or very satisfied with the cost of contracting RMT staff and 96 percent of responders were somewhat or very satisfied with data integrity, compliance and security.

Conclusions: The significant benefits of using REDCap as a validated, low cost data capture tool with built-in compliance components within the existing RMT research support infrastructure has facilitated a reduction in the research support gap across funding levels. The model has the potential for adoption at other CTSA institutions with potential impact on the quality of translational medicine data.
A Web-based Tool for Cataloging Primary Care Electronic Medical Record Federated Data: FInDiT
Kari A. Stephens, Ching-Ping Lin, Laura-Mae Baldwin, Abigail Echo-Hawk, & Gina Keppel

Introduction
Data sharing across disparate ambulatory care based electronic medical records is necessary to facilitate comparative effectiveness research (CER) in primary care. Adoption of electronic medical records in primary care, from single to large group practices, has reached critical mass in the last decade. Practices, in response to national Health Information Technology incentives that promote efforts such as meaningful use and patient-centered medical homes, are engaging architectural solutions to conduct data sharing. This engagement is building the data sharing capacity critical to promote research. The CTSAs are well positioned to promote these architectural solutions and are engaged in developing data sharing capacity across large population bases. Our CTSA efforts at the University of Washington's Institute of Translational Health Sciences include our LC Data QUEST pilot project that is creating data sharing capacity within our regional Practice Based Research Network efforts and across our tribal communities, in collaboration with the DARTNet group, a consortium of national practice-based data sharing networks. As data sharing capacity matures, understanding the types, quantity, and quality of the data available for CER efforts is crucial to promoting access and engagement between the community partners, research teams, and the data sharing architectures. The Federated Information Dictionary Tool (FInDiT) is a web-based tool that catalogues type, quantity, and quality of the data that are available across the LC Data QUEST data sharing architecture.

Methods (Tool Development)
We will extract data in aggregate form, containing counts and unique listings of codes within the core data elements across domains (i.e., problems, demographics, encounters, laboratory, medications, etc.). From these extractions we will derive content to augment basic data dictionary (schema) information, including listings of data elements, types, and definitions. These aggregations and data dictionary metadata will be imported and stored in an SQL database backend. The web tool GUI interface will display the information in meaningful categorizations. To drive design efforts, we have evaluated five research use cases across our LC Data QUEST environment and collaborate with our CORT Core partners and community/practice partners. We will use and track future research projects that use LC Data QUEST to improve content within FInDiT continuously as study specific data retrievals and aggregations result in significant data quality and ontology based discoveries.

Results (System Functionality)
Our design allows for easy addition of future sites by: 1) defining a set extract format for aggregated data content and metadata needed from any additional federated repository wishing to be added; 2) allowing for simple upload of this extract into a SQL database; and 3) dynamic access to the data via the web-based front-end GUI. As we have iterated on the FInDiT design, we have engaged our colleagues within the DARTNet consortium to promote scalability to and compatibility with outside networks with existing compatible data sharing architectures. As a result, we have begun to expand our ontology definition to consider the Observational Outcomes Medical Ontology (OMOP) that will likely be adopted by the DARTNet consortium.

Discussion and Conclusions
FInDiT addresses several gaps in existing data sharing architecture efforts by: 1) giving web-based access to meaningful detail about what data are available in the network; 2) offering detailing of semantic alignment across the federated repositories; and 3) providing a tool that can be used to facilitate on-going and future research collaborations. In the immediate term, FInDiT will help us understand and address semantic alignment and data quality issues across our existing architecture. Our goal is to attract health researchers with specific research interests in understanding the types and volume of available data to promote engagement in CER with our partners. Furthermore, this tool can help community partners target semantic alignment gaps and increase research capacity.
Warehousing of Trials Data for Submission and Reporting

Shariq Tariq\textsuperscript{1,2}, Akheel Ahmed\textsuperscript{2}, Umit Topaloglu\textsuperscript{1,2}, William Hogan\textsuperscript{1,2}

\textsuperscript{1}Division of Biomedical Informatics - University of Arkansas for Medical Sciences,  
\textsuperscript{2}Arkansas Study Center of the National Children’s Study

We are creating ETL (Extract, Transform, and Load) processes over an integrated collection of biomedical research data from heterogeneous sources and multiple relational databases, and subsequently populating them into a research data warehouse/repository. We are using Talend Open Studio, an industry standard open source solution for ETL and data profiling, to build these ETL processes. This was chosen not because of its ease of process development and data integration but because it is easily adoptable. The reason is twofold. First, the community that has developed around Talend enables collaboration. Furthermore, its ability to integrate seamlessly with other Talend open source products allow data quality, data management, and application integration to be all handled under one platform.

The high level management of these ETL processes will allow users to provide input into design as well as enable flexible re-use over common/shared tasks, source-code optimization, data quality/integrity/security, and rapid development of new processes. The ETL process building will be incremental to keep up with the expanding scope of the data warehouse/repository to include additional sources.

Our current process extracts data from multiple, heterogeneous sources that use different relational database systems (MySQL, PostgreSQL, MSSQL). Instrument and operational data are collected using LimeSurvey. Participant data such as demographics, address, and consent data are collected in the caBIG Central Clinical Participant Registry (C3PR). The ETL process extracts and transforms data from these processes, and loads them into the warehouse. For the National Children’s Study (NCS), we then created other Talend processes to create XML files from the warehouse for submission to the NCS Vanguard Data Repository (VDR), where all Study Centers are required to submit their data. The processes ensure consistency of the files with the VDR XML schema and execute other functions like removing personally-identifying information. Also, the warehouse supports reports that staff at our Study Center can use to track recruitment and retention.
Budgeting Issues and solutions for a Clinical Research in an Academic Health Center

Umit Topaloglu PhD, Jiang Bian PhD, William R. Hogan MD MS, Cheryl Lane BS, Thomas Wells MD MBA, Laura Hutchins MD
University of Arkansas for Medical Sciences, Little Rock, AR

Introduction:
Clinical Research (CR) is the driving force of advancements in healthcare for new drug and device discoveries. The funding of a new trial, however, is an expensive and a big challenge. With the June 7, 2000 Presidential Order and the resulting 2007 guideline from Center for Medicare and Medicaid Services (CMS), charges may be made for routine patient care and/or complications associated with a clinical trial. Although these developments created financial incentives for CR, compliance with appropriate charging creates many legal and administrative burdens in a typical Academic Health Center (AHC). In this paper, we present challenges related to research billing and describe an informatics solution under development at the University of Arkansas for Medical Sciences in which our Research Support Center, Clinics, and Compliance and Billing groups are the stakeholders and are well represented.

Methodology
As widely accepted, the issues related clinical research billing cannot be addressed with the help of tool development only. Recently, we have developed a model that has the potential to address the problem. This model includes a) Developing a budgeting tool, b) Customizing Clinical Trial Management System (CTMS) solutions, and c) Implementing e-billing with electronic encounter forms.

a) Budgeting Tool: CLinical Research Administrator (CLARA) is a web-based portal developed at UAMS for study planning, budget creation, submission, review, approval, and tracking of human subject research. Compared to various existing electronic review systems, CLARA’s modular design and utilization of web 2.0 technologies (e.g. Ajax, JSON etc.) offer streamlined data entry, provide an easy to use and intuitive user interface for study budget matrix creation, eliminate unnecessary steps and errors in the review process, and provide a comprehensive reporting system to monitor key metrics in the approval process.

b) CTMS: UAMS has successfully implemented the National Cancer Institute (NCI) supported cancer Biomedical Informatics Grid (caBIG®) CTMS suite (i.e. Patient Study Calendar (PSC)). For budgeting and billing purposes we implemented several procedures and customizations in PSC. To identify study activities, SNOMED CT curation has been completed by the clinic for each study. Each activity is also annotated with one or more CPT codes and charge types (i.e. R, C etc.) from CLARA, as well as an ICD-9-CM code as justification for each activity and safety code (if applicable). Creating studies as mentioned is time consuming, on the other hand, for a registered subject’s visit, we know in detail what are the procedures to be done, what are the charge types and CPT codes for charging along with the diagnosis codes.

c) E-Billing: Another key component is the development of an electronic encounter form system, e-billing, which is designed to improve coding efficiency and positively impact physicians’ practice. The e-billing system not only eliminated the need of the paper forms, but it also automated the sharing of clinical and billing information. The ICD-9 and CPT codes are entered into electronic encounter forms by clinicians and transferred to the billing system. The billing clerk can quickly review the charges and diagnosis codes to ensure the correctness and compliance of the bill. The e-billing system can be tailored to make the commonly used diagnoses, billing codes, modifiers and action codes are readily available for each clinic or even for each physician, streamlining the process.

Conclusion
In UAMS, after many lengthy discussions and tireless efforts from several people, we are implementing new tools and workflows to address many challenges come with the CR budgeting and billing. As the software solutions can provide interoperability and data consistency for certain number of functions, the human operations are indispensable and still the key of the success for a research enterprise.
Towards an Oncology Database (ONCOD): A Case Study Using Data Warehousing Technology

Xiaoming Wang1, Lili Liu1, James Fackenthal2, Paul Chang3, Peng Liu3, Gilliam Newstead3, Steven Chmura4, Ian Foster1, Olufunmilayo I Olopade2

1Computation Institute, 2Department of Medicine, 3Department of Radiology, 4Department of Radiation Oncology, University of Chicago

Motivation: Through the ONCOD project, we explore potential applications of an alternative data warehousing strategy with the ability to support a broad range of translational data integration demands.

Methods and Results: ONCOD is a pioneer data mart project implemented within our internally developed data warehousing system known as the Translational data Mart (TraM). The TraM system is developed to support a variety of data marts without recreating project-specific datasets (Fig 1). Along the same line efforts, our approach has the following features: 1) The TraM schema is developed upon an entity-relationship (ER) model, which demands column semantic homogeneity. 2) To support ONCOD, we have developed a data transformation workflow to systematically supply unified, standardized, and curated data for integration. 3) By separating domain concept entities from research object entities, the TraM model is able to boost data capture capacity for recruiting ever-involving domain concepts, and can associate these concepts to research objects in any combination. 4) Although the system allows ONCOD users to create their own project specific accounts for data privacy protection, it simultaneously enforces all accounts to share a unified data structure and controlled vocabulary. 5) As a result, individualized ONCOD data can be effectively queried horizontally (across domains) and longitudinally (over time). De-identified and structured data from query returns is directly displayed to the ONCOD web interface, so users can adjust their query strategy according to tangible information for desirable results. Finally, the TraM strategy allows iterative curation during and after ONCOD data integration, which will add significant value to this data source and potentially make it an oncology database.

Importance: Although ONCOD focuses on data in cancer research, the TraM system is designed for a broad range of biomedical information integration tasks. In particular, our model-driven approach is proven effective, sustainable, modifiable, and scalable for our ultimate goal. The mechanism and components shown in Figure 1 have been completely implemented and ONCOD has been in use in our cancer translational research community for four years. This approach may be applied in general informatics practice and will add knowledge in this field.

Fig 1 Overview of the ONCOD/TraM system and data integration mechanism
(http://tram.uchicago.edu, Wang et al, JBI 2009)
A Machine Learning Based Approach for Retrieval and Organization of Translational Research Articles

Firas H. Wehbe, MD, PhD

Department of Biomedical Informatics, Vanderbilt University

Molecular medicine encompasses the application of molecular biology techniques and knowledge to the prevention, diagnosis and treatment of diseases and disorders. Statistical and computational models can predict clinical outcomes, such as prognosis or response to treatment, based on the results of molecular assays. For advances in molecular medicine to translate into clinical results, clinicians and translational researchers need to have up-to-date access to high-quality predictive models. The large number of such models reported in the literature is growing at a pace that overwhelms the human ability to manually assimilate this information. Therefore the important problem of retrieving and organizing the vast amount of published information within this domain needs to be addressed. The inherent complexity of this domain and the fast pace of scientific discovery make this problem particularly challenging.

A semantic analysis of this domain was performed to inform the development of a specialized annotation scheme of published articles that can be used for meaningful organization and for indexing and efficient retrieval. This annotation scheme was codified using an annotation form and accompanying detailed guideline document that were used by 9 human experts to annotate over 1000 articles. “Relevant articles” were operationally defined by the annotation guideline as articles that establish or test statistical relationships between molecular data and clinical outcomes. These articles were annotated along multiple dimensions that include: the biological source and the type of molecular assay used to obtain the molecular data as well as the type of clinical outcome (prognosis, response to treatment, etc). The human experts varied in educational background and training. They include: board certified physician scientists, post doctoral trainees in basic biological sciences, medical students, epidemiologists, and medical librarians. A dataset of over 1000 randomly selected articles from journals in the domains of lung cancer research, breast cancer research, and basic bioinformatics research were manually annotated. This dataset was then used to train and test support vector machine (SVM) machine learning classifiers. Articles were formatted for machine learning by using a variety of text preprocessing and feature extraction techniques that analyzed individual terms in the abstract, title, or medical subject headings (MeSH) within their MEDLINE records.

The classifiers were designed to provide a scalable mechanism to replicate human experts’ ability (1) to retrieve relevant MEDLINE articles and (2) to annotate these articles using the specialized annotation scheme. They were independently validated using the human annotation and were found to have very good predictivity using the AUC metric. The predictivity was minimally affected when different feature extraction techniques were used (e.g. adding/removing MeSH terms and using natural language processing to identify UMLS concepts in the MEDLINE record). Also, the classifiers’ predictivity was found to successfully generalize to articles in another disease domain as well as to articles that were annotated by a different set of experts. The results of these experiments point to the importance of the development of an unambiguous operational definition of the semantic entities that define molecular medicine predictive models. Specifically, it is important that there exists a clear set of annotation instructions that can be interpreted and applied in a consistent manner by human annotators.

This approach for using machine learning classifiers is a promising scalable technique for addressing the problem of large-scale retrieval of relevant articles in the domains of molecular medicine and translational research. The construction and analysis of the dataset used was done using open-source machine learning libraries and readily available public resources (PUBMED and eutils by the NCBI).
A Method for Standardized Patient Reported Outcome Data Collection at the Point of Care

**Introduction:** Collection of outcome data is an important element of comparative effectiveness research, a key focus of CTSA Strategic Goal 4. It is also important in direct clinical care as well as other clinical research. Patient reported outcomes (PROs) based on standardized, validated questionnaires used to capture detailed data on disease status (i.e. depression severity) and other associated domains (i.e. medication adherence, co-morbid substance abuse, etc.) provide more robust data than anecdotal reports. At the University of Alabama (UAB) 1917 HIV/AIDS Clinic (1917 Clinic) we have used PRO data to inform clinical decision-making, facilitate research and enhance care. In order to provide the benefits of PRO data capture in other chronic care settings we partnered with Palliative/Supportive Care, Rheumatology and 1917 Clinic Social Workers to build web-based software supporting PRO questionnaires specific to each practice setting.

**Methods:** A broad group of stakeholders at UAB was assembled including member of the Biomedical Informatics of the UAB Center for Clinical and Translational Science, Health System Information Services, several Divisions of the Department of Medicine (Clinical Immunology and Rheumatology; Gerontology, Geriatrics and Palliative Care) and Centers (Comprehensive Cancer Center, Comprehensive Arthritis Center, Center for AIDS Research and Comprehensive Diabetes Center). After project approval, personnel from the target implementation sites Social Services at 1917 Clinic, Palliative and Supportive Care at the 1917 Clinic and Rheumatology met regularly with the programming team. Each site constructed a specific survey each containing validated instruments covering specific domains relevant to their practice.

**Results:** The application is written in C# .NET and is web based. The web server uses a 128 bit encryption SSL certificate to protect transmitted data and all user passwords are hashed into a table in the backend ORACLE database. We pursued a web based model in order to reach a broad range of devices including desktops, tablets and ultimately smart phones. While surveys are administered, clinic personnel have access to a “mission control” screen that allows providers to see individual users progress in the application providing real time assistance as needed. Software allows for alerts to be sent if pre-specified (i.e. score thresholds or specific answers) are reported by patients, allowing for clinic response. The data collected is normalized in a way that makes cross clinic and cross domain queries easy and reliable. We purposefully left the database slightly open ended so that we might accommodate any future instruments with question types/answers we have not yet encountered. Data are accessible via querying tools built into the web-based software and each partner can access their data for operational or research purposes. The operational reports are custom built to each clinic’s specifications and focus on facilitating workflow, while research reports allow for data retrieval across specified date ranges.

**Discussion:** Integrating PROs at the point of care in a routine clinical setting may represent a transformative health informatics technology for routine clinical care (i.e. point of care decision making) and comparative effectiveness research providing added value to both enterprises and furthering the CTSAs mission to support clinical and translational research. In addition, using mobile technologies, PRO data capture can travel outside of the immediate hospital environment opening up exciting new data capture possibilities.
Using a MeSH-based Index to Demonstrate the Multidisciplinary Propensity of KL2 Scholars
Xiao Dong, Klara Papp, Satya Sahoo, Chime Ogbuji, GQ Zhang
Case Western Reserve University

Objective: In 2005, the Cleveland Clinical Translational Science Collaborative (CTSC) implemented the KL2 Multidisciplinary Clinical Research Training Program to train the next generation of clinical and translational scholars to lead multidisciplinary, team-based research. Research productivity of the KL2 scholars may be assessed by traditional metrics such as number of publications, funding support and impact factor. However assessment of their propensity to engage in multidisciplinary research remains difficult to measure. Yet identifying whether the KL2 training program is achieving its objectives for encouraging multidisciplinary collaborations is of key importance. We propose an informatics approach based on the scope of published research work to estimate multidisciplinary propensity in order to assess this objective. In order to assess whether scholars’ research topics are increasingly multidisciplinary, we use MTI – a MeSH-Tree-based Index – to estimate scholars’ multidisciplinarity and assumed that the MTI values would increase more rapidly across the training period compare to those of a non-KL2 cohort.

Method: We randomly selected 6 scholars at Cleveland CTSC who have completed KL2 training, alongside with a match cohort of 5 additional scholars engaged in other type of K programs during the same period (2005-2010). In order to characterize the overall research portfolio for these scholars, we gathered their publication information on PubMed and retrieved the MeSH terms associated with these publications via NCBI’s Entrez API. The MeSH terms are accumulated during the award period. Using the MeSH terms’ tree number index we further construct a distinctive set of personalized MeSH hierarchies for all scholars. Such a hierarchy is used to calculate the MTI as the sum of the reciprocals of the depth of the occurring MeSH terms. These reciprocals account for the intuition that branching at the levels closer to the root carries more weight than branching near the leaf level.

Result: The total and percentage growths of the MTI at the point of entry to the program and at the end of the program are obtained, for each scholar (see Figure above). The average percentage growth for the KL2 cohort (150%) is higher than that of the match cohort (128%), whereas the match cohort tops the KL2 cohort in average absolute MTI growth (37.7 vs 25.5).

Discussion: The average percentage growth of MTI and the average absolute growth of MTI show opposite comparative signs, although both have increased substantially. A single index alone is unlikely accountable for all the factors affecting an individual scholar’s training performance, although the higher rate of average percentage increase among the KL2 cohort is encouraging. With these limited cohort sizes, we can partially conclude that the overall multidisciplinary profiles of the KL2 scholars have expanded significantly during the training period, and our KL2 training program has successfully fulfilled the goal of promoting multidisciplinary research.