Cincinnati Children’s Clinical Research Data Warehouse (i2b2)

23 September 2008

Keith Marsolo, PhD
Outline

• Background
  – CCHMC, informatics, RDW

• i2b2 implementation
  – Sources, staffing, hardware

• Extensions & future work
  – Ontologies, modules, bio-repository, CTSA
CCHMC

- Independent, full-service, not-for-profit pediatric academic medical center (affiliated with the University of Cincinnati College of Medicine)
- Major pediatric care provider for southern Ohio, northern Kentucky and eastern Indiana (serve patients from all 50 states, 48 countries)
- 500+ member pediatric clinical and research faculty
- Ranked in the top 5 pediatric institutions in the country
- Second among pediatric centers receiving NIH research grants
Bioinformatics @ CCHMC

• Basic research
  – Proteomics
  – NLP
  – Protein modeling
  – Data and systems integration

• Research IT
  – Hardware and software support
  – Software development
  – Storage and database hosting
  – High-performance computing
Data Management in a Combined Research & Clinical Environment

Clinical operations
- Clinical data repository (Epic/DocSite)
  - Transactional
  - Real-time operation
  - Focus on visit
  - Clinical care
    - History and Physical
    - Physician orders
    - Progress notes
- Operational data store (Clarity)
  - Transactional
  - Near real-time and archival
  - Focus on both patient & operation of clinic
  - Effectiveness & outcomes
- Research data warehouse (i2b2)
  - Integrate multiple data sources
  - Combine research & clinical
  - Focus on cohorts & patient populations
  - Potential for de-identified queries

Research laboratories
Data Warehouse: pre-i2b2

• Custom-built solution
• Primary drawbacks:
  • No coherent data model
  • Lack of data standards/governance
  • Poor data quality
• Resulted in a system with limited functionality
New Warehouse - Existing or Custom?

Custom Solution
• Pros
  – Tailored functionality
  – Control over design
• Cons
  – Expensive
  – Long development time
  – Proprietary data formats or system architecture

Existing Architecture
• Pros
  – Proven success
  – Potential for collaboration
• Cons
  – Missing features
  – Control of source and/or development

Our choice: i2b2
What is i2b2?

• i2b2 = Informatics for Integrating Biology & the Bedside

• National Center for Biomedical Computing (NCBC)
  – Funded by NIH to develop national computational infrastructure for biomedical computing
  – Centered at Partners HealthCare in Boston

• Open-source warehouse architecture
  – Based on Research Patient Data Registry developed at Massachusetts General Hospital (MGH)
  – Geared toward identification and analysis of patient cohorts.
Why i2b2?

• Designed for translational research
• Simple, scalable architecture
  – Supports multiple data types and sources
  – Capable of handling large amounts of data
• Potential for funding/collaboration
  – Share development with other institutions
  – Funding to develop additional functionality
Functionality of i2b2

- Designed around populations and cohorts
- Automated tools for cohort identification and hypothesis generation
- Creation of datamarts for later statistical analysis
- Develop other reporting and analysis tools based on user feedback.
Warehouse status

- 5 years of archive data (~500,000 patients)
- Access to Epic and legacy systems
- Content:
  - Demographics (age, race, gender, marital status)
  - Diagnoses (ICD-9)
  - Laboratory & pulmonary function tests
  - Medications (based on NDC)
  - Procedures (ICD-9 & CPT)
Future data sources

- Epic
  - Gold-standard for demographics
  - Vitals, problem list
  - Research variables
- DocSite (clinical research registries)
- Text-based reports
  - Discharge summaries
  - Pathology, Radiology and Cardiology reports
- Genetics, microarray
The trouble with free-text

• Natural language processing is hard
• Most effective at identifying concepts and keywords
  – Best with structured text and controlled vocabulary
  – What if concept is absent?
• Potential solution:
  – Parse all reports for a set of major concepts
  – Further processing after identification of cohort
Research & i2b2

• Two views:
  – Pull from i2b2 to augment research data
  – Push research data into i2b2
    • Allow others access to new information
    • More data for overlapping patients

• Other services:
  – Use i2b2 tools on project-specific datamart
  – Extracts and reports from Epic
Data-related challenges

• Age
  – Current?
  – At admission? Diagnosis?
  – De-identified: year only (i.e. 0 or 1)

• Overlapping & incomplete terminologies
  – ICD-9 and CPT for procedures
  – ICD-9 for diagnosis

• Medications
  – Ordered meds only
  – Not a complete history
Development @ CCHMC

• Web-based Workbench
  – Cohort identification through browser
  – Tabular breakdown of patient set

• Ontology Browser
  – Basic statistics for each query term
  – Histogram of diagnoses by age, laboratory results by reference range, etc.
Workbench
Ontology Browser

Welcome to the i2b2 Ontology Browser

- Ontology
  - Demographics
  - Diagnoses
  - Laboratory Tests
    - ACTH Stimul
      - Aldosterone
    - Cortisol Free Serum
  - Allergen Group
  - Amino Acids Level
  - Anaerobic
  - Analgesics
  - Anti Asthma
  - Anti Coagula
  - Antibiotics
  - Anticoagulants
  - Arg/Ins Tol
  - Arq/Clen St
  - RMT
  - Bid GasResp
  - Bid GasRespA
  - Blood Bank Group
  - Blood Cult
  - Body Fluid Group
  - CSF

Ontology - Labtests - LAB - ACTH Stimul - Aldosterone - 9000055^ALDOSTERON

9000055^ALDOSTERON

Reading Range: 1.5 - 12

<table>
<thead>
<tr>
<th>Age Range</th>
<th>Gender</th>
<th># Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-9</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>20-29</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0-9</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>20-29</td>
<td>1</td>
</tr>
</tbody>
</table>
Future Development @ CCHMC

• Cohort-based reminders and notifications
  – Adherence to protocol
  – Recruitment for trials
  – Interface with scheduling for near-time alerts
• Search for related terms using UMLS
• Customized ontologies
  – Pediatric-specific (joint efforts with Denver & Boston Children’s)
  – Registry-based (i.e. DocSite)
Other Development

• CTSA-related
  – Federated queries
  – Multi-institution ontologies
  – Identity management

• Biorepository
  – Use cohort criteria to identify samples from discarded specimens

• i2b2
  – Query by value
  – Data import/export
  – File repository & Image annotation

• Potential integration with caBIG
## Hardware

<table>
<thead>
<tr>
<th>System</th>
<th>CPU</th>
<th>Memory</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Database</strong> - Oracle Cluster (2 node Standard Real Application Cluster)</td>
<td>1x Quad Core (each node)</td>
<td>16 GB (each node)</td>
<td>1 TB SAN storage (shared)</td>
</tr>
<tr>
<td><strong>ETL</strong> - Oracle Server (Oracle Enterprise)</td>
<td>2x Quad Core</td>
<td>32 GB</td>
<td>1 TB SAN storage</td>
</tr>
<tr>
<td><strong>i2b2 Middleware</strong> - Production (Linux, Apache/Tomcat, JBOSS)</td>
<td>2x Quad Core</td>
<td>16 GB</td>
<td>Local storage</td>
</tr>
<tr>
<td><strong>i2b2 Middleware</strong> - Development (10x VMware virtual machines)</td>
<td>1-2x Single Core</td>
<td>512 MB - 8 GB (28 GB total)</td>
<td>Local storage</td>
</tr>
<tr>
<td><strong>i2b2 Fileserver</strong></td>
<td>1x Quad Core</td>
<td>8 GB</td>
<td>1 TB SAN storage</td>
</tr>
<tr>
<td>Role</td>
<td>Effort</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project lead</td>
<td>1 FTE (faculty)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database administrator</td>
<td>1 FTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data cleaning, data quality, user reports</td>
<td>2 FTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software developers</td>
<td>2 FTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer interface</td>
<td>0.5 FTE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Participants

• **Implementation Team (Biomedical Informatics - BMI):**
  Keith Marsolo - Project Leader
  Parth Divekar, Pranay Shyam, Hai Ge, Adil Khan

• **Information Services - IS (Data Sources):**
  Frank Menke, Jacquie Keebaugh, Lee Rich, Ron Robinson

• **BMI (Hardware and Database Support):**
  Michal Kouril, Mihir Mishra

• **Special Thanks (Other Assistance):**
  Jason Napora, Marianne James, John Hutton, Paul Steele, Andy Spooner
Questions?

• For further information:
  – E-mail: keith.marsolo@cchmc.org
  – Web: http://i2b2.cchmc.org