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

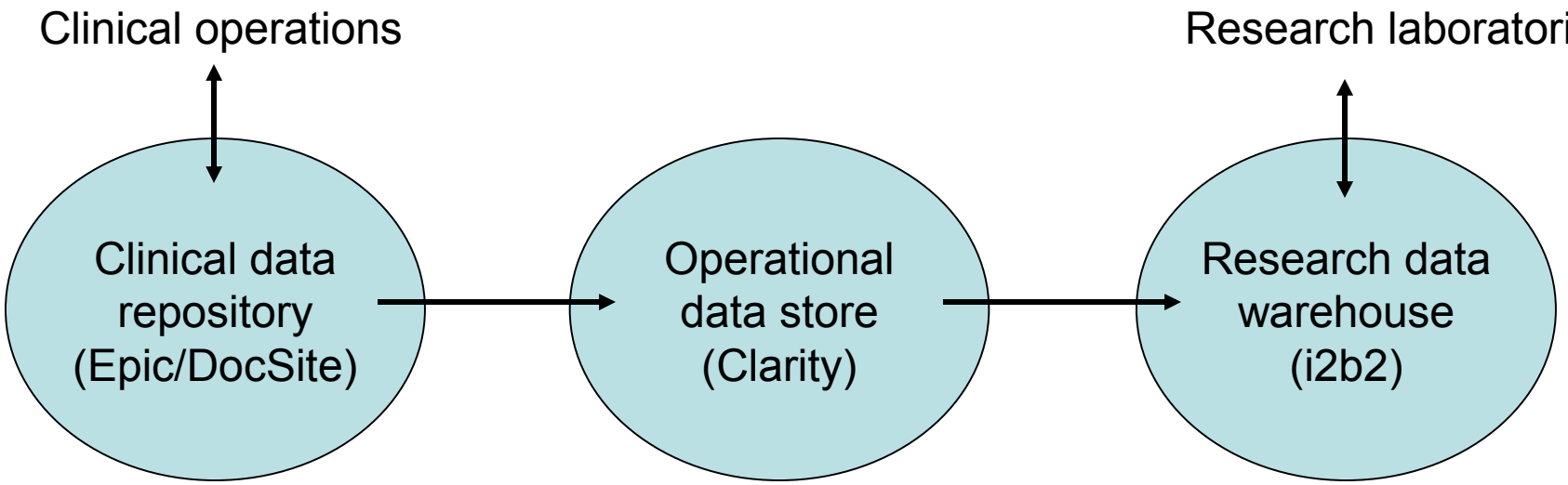
CCHMC

- 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



Transactional
Real-time operation
Focus on visit
Clinical care
- History and Physical
- Physician orders
- Progress notes

Transactional
Near real-time and archival
Focus on both patient & operation of clinic
Effectiveness & outcomes

Integrate multiple data sources
Combine research & clinical
Focus on cohorts & patient populations
Potential for de-identified queries

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

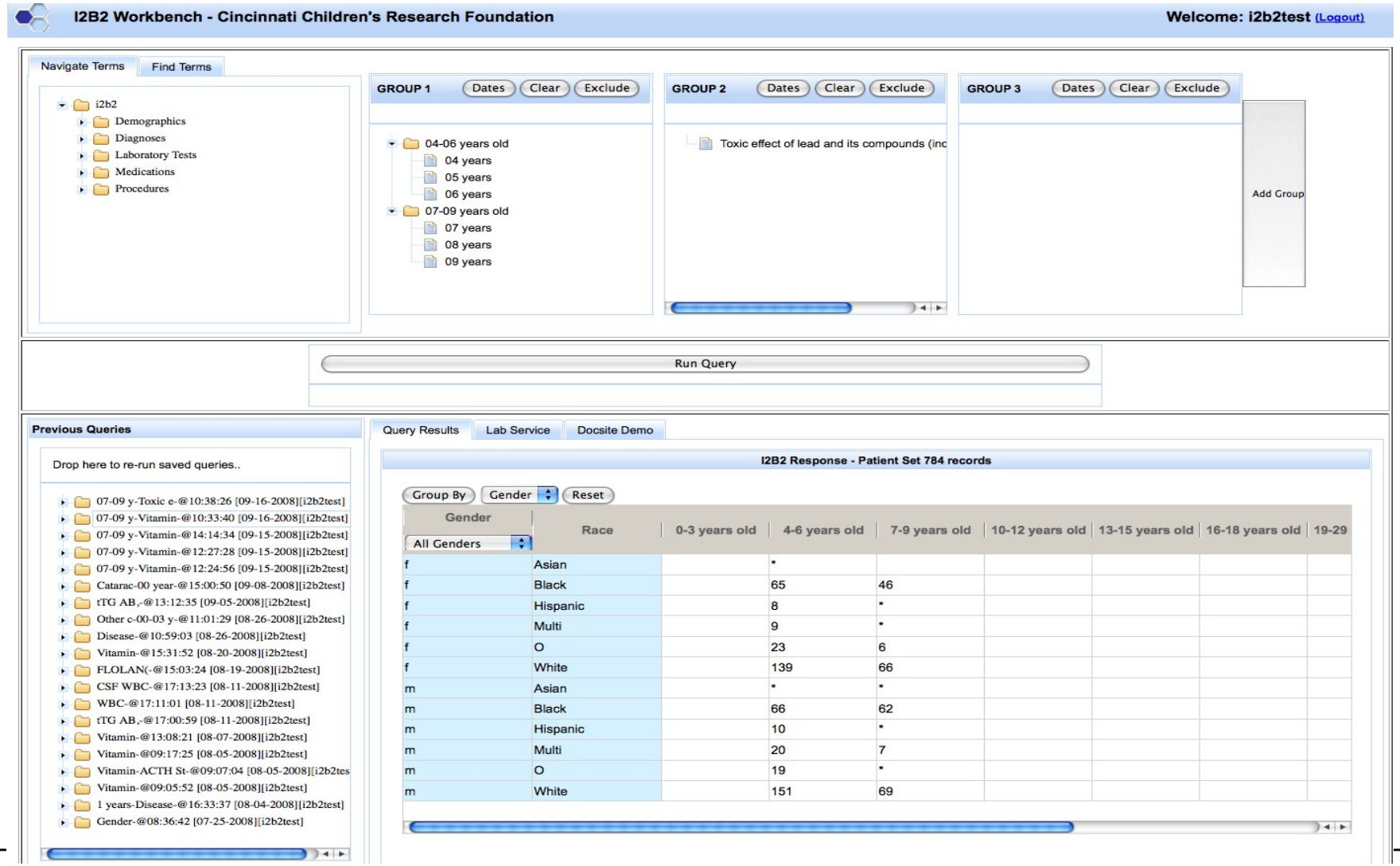
Data-related challenges

- 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

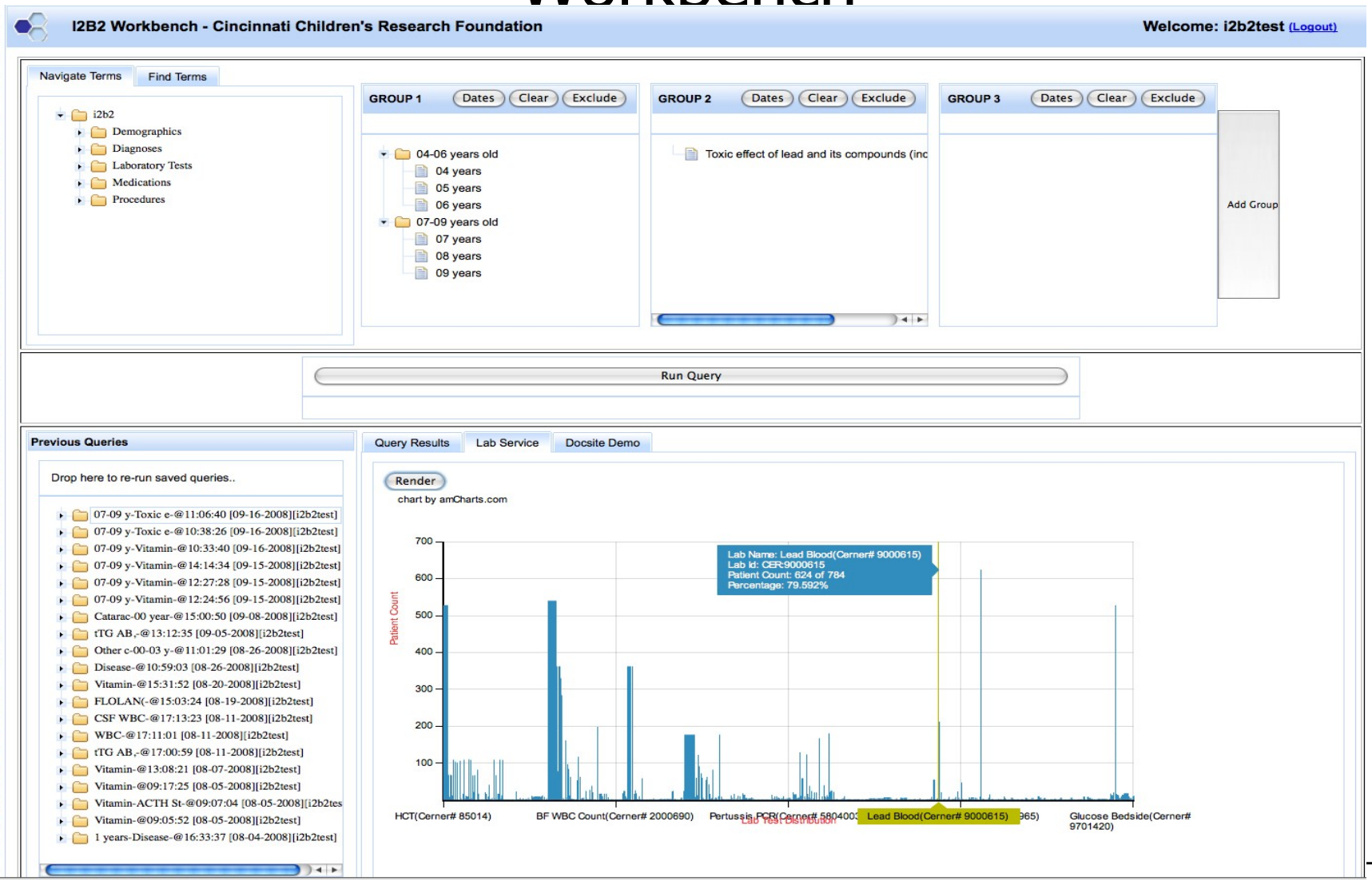
Development @ CCHMC

- 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


I2B2 Workbench - Cincinnati Children's Research Foundation
Welcome: i2b2test (Logout)
Navigate Terms
Find Terms
i2b2
Demographics
Diagnoses
Laboratory Tests
Medications
Procedures
GROUP 1
Dates
Clear
Exclude
04-06 years old
04 years
05 years
06 years
07-09 years old
07 years
08 years
09 years
GROUP 2
Dates
Clear
Exclude
Toxic effect of lead and its compounds (inc
GROUP 3
Dates
Clear
Exclude
Add Group
Run Query
Previous Queries
Drop here to re-run saved queries..
07-09 y-Toxic e-@10:38:26 [09-16-2008][i2b2test]
07-09 y-Vitamin-@10:33:40 [09-16-2008][i2b2test]
07-09 y-Vitamin-@14:14:34 [09-15-2008][i2b2test]
07-09 y-Vitamin-@12:27:28 [09-15-2008][i2b2test]
07-09 y-Vitamin-@12:24:56 [09-15-2008][i2b2test]
Catarac-00 year-@15:00:50 [09-08-2008][i2b2test]
tTG AB,-@13:12:35 [09-05-2008][i2b2test]
Other c-00-03 y-@11:01:29 [08-26-2008][i2b2test]
Disease-@10:59:03 [08-26-2008][i2b2test]
Vitamin-@15:31:52 [08-20-2008][i2b2test]
FLOLAN(-@15:03:24 [08-19-2008][i2b2test]
CSF WBC-@17:13:23 [08-11-2008][i2b2test]
WBC-@17:11:01 [08-11-2008][i2b2test]
tTG AB,-@17:00:59 [08-11-2008][i2b2test]
Vitamin-@13:08:21 [08-07-2008][i2b2test]
Vitamin-@09:17:25 [08-05-2008][i2b2test]
Vitamin-ACTH St-@09:07:04 [08-05-2008][i2b2tes
Vitamin-@09:05:52 [08-05-2008][i2b2test]
1 years-Disease-@16:33:37 [08-04-2008][i2b2test]
Gender-@08:36:42 [07-25-2008][i2b2test]
Query Results
Lab Service
Docsite Demo
I2B2 Response - Patient Set 784 records
Group By
Gender
Reset
Gender
All Genders
Race
0-3 years old
4-6 years old
7-9 years old
10-12 years old
13-15 years old
16-18 years old
19-29
f Asian *
f Black 65 46
f Hispanic 8 *
f Multi 9 *
f O 23 6
f White 139 66
m Asian * *
m Black 66 62
m Hispanic 10 *
m Multi 20 7
m O 19 *
m White 151 69

Workbench


I2B2 Workbench - Cincinnati Children's Research Foundation
Welcome: i2b2test (Logout)
Navigate Terms
Find Terms
i2b2
Demographics
Diagnoses
Laboratory Tests
Medications
Procedures
GROUP 1
Dates
Clear
Exclude
04-06 years old
04 years
05 years
06 years
07-09 years old
07 years
08 years
09 years
GROUP 2
Toxic effect of lead and its compounds (inc
GROUP 3
Add Group
Run Query
Previous Queries
Drop here to re-run saved queries..
07-09 y-Toxic e-@11:06:40 [09-16-2008][i2b2test]
07-09 y-Toxic e-@10:38:26 [09-16-2008][i2b2test]
07-09 y-Vitamin-@10:33:40 [09-16-2008][i2b2test]
07-09 y-Vitamin-@14:14:34 [09-15-2008][i2b2test]
07-09 y-Vitamin-@12:27:28 [09-15-2008][i2b2test]
07-09 y-Vitamin-@12:24:56 [09-15-2008][i2b2test]
Catarac-00 year-@15:00:50 [09-08-2008][i2b2test]
tTG AB,-@13:12:35 [09-05-2008][i2b2test]
Other c-00-03 y-@11:01:29 [08-26-2008][i2b2test]
Disease-@10:59:03 [08-26-2008][i2b2test]
Vitamin-@15:31:52 [08-20-2008][i2b2test]
FLOLAN(-@15:03:24 [08-19-2008][i2b2test]
CSF WBC-@17:13:23 [08-11-2008][i2b2test]
WBC-@17:11:01 [08-11-2008][i2b2test]
tTG AB,-@17:00:59 [08-11-2008][i2b2test]
Vitamin-@13:08:21 [08-07-2008][i2b2test]
Vitamin-@09:17:25 [08-05-2008][i2b2test]
Vitamin-ACTH St-@09:07:04 [08-05-2008][i2b2tes
Vitamin-@09:05:52 [08-05-2008][i2b2test]
1 years-Disease-@16:33:37 [08-04-2008][i2b2test]
Query Results
Lab Service
Docsite Demo
Render
chart by amCharts.com
Lab Name: Lead Blood(Cerner# 9000615)
Lab Id: CER:9000615
Patient Count: 624 of 784
Percentage: 79.592%
Patient Count
700
600
500
400
300
200
100
HCT(Cerner# 85014)
BF WBC Count(Cerner# 2000690)
Lead Blood(Cerner# 9000615)
Glucose Bedside(Cerner# 9701420)

Ontology Browser

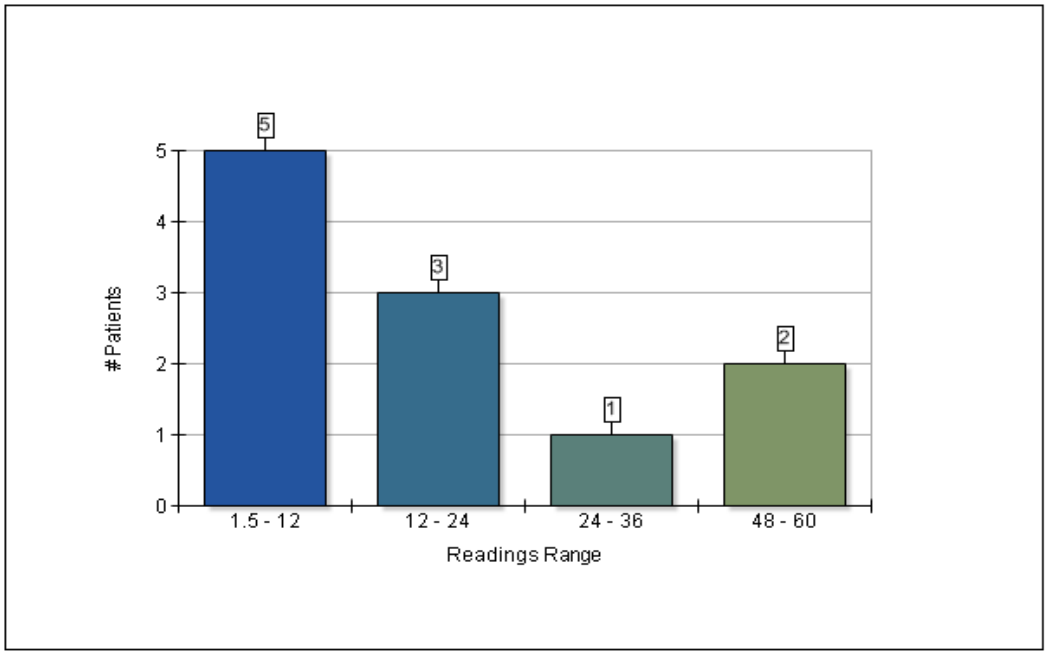
Welcome to the i2b2 Ontology Browser

Find: Search

- Ontology
  - Demographics
  - Diagnoses
  - Laboratory Tests
    - ACTH Stimul
      - Aldosterone
        - 9000055^ALDOSTERON
      - Cortisol Free Serum
    - Allergen Group
    - Amino Acids Level
    - Anaerobic
    - Analgesics
    - Anti Asthmat
    - Anti Convuls
    - Antibiotics
    - Antineoplast
    - Arg/Ins Tol
    - Argi/Clon St
    - BMT
    - Bld Gas/Resp
    - Bld Gas/RespA
    - Blood Bank Group
    - Blood Cult
    - Body Fluid Group
    - CSF

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

9000055^ALDOSTERON



Reading Range: 1.5 - 12

	Age Range	Gender	# Patients
1	0-9	f	2
2	20-29	f	1
3	0-9	m	1
4	20-29	m	1

Home | Help | Logout

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

	CPU	Memory	Storage
Database - Oracle Cluster (2 node Standard Real Application Cluster)	1x Quad Core (each node)	16 GB (each node)	1 TB SAN storage (shared)
ETL - Oracle Server (Oracle Enterprise)	2x Quad Core	32 GB	1 TB SAN storage
i2b2 Middleware - Production (Linux, Apache/Tomcat, JBOSS)	2x Quad Core	16 GB	Local storage
i2b2 Middleware - Development (10x VMware virtual machines)	1-2x Single Core	512 MB - 8 GB (28 GB total)	Local storage
i2b2 Fileserver	1x Quad Core	8 GB	1 TB SAN storage

Staffing

Role	Effort
Project lead	1 FTE (faculty)
Database administrator	1 FTE
Data cleaning, data quality, user reports	2 FTE
Software developers	2 FTE
Customer interface	0.5 FTE

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