Enabling the Healthcare Enterprise for Discovery Research with i2b2

Shawn Murphy MD, Ph.D.
High Throughput Method for supporting Research

- Set of patients is selected from medical record data in a high throughput fashion

- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data

- Using the Crimson system, tissues of these patients can be made available for genomic and biochemical analysis

- Automated discovery can be created from these projects to support further hypothesis-driven research
High Throughput Method for supporting Research

- Set of patients is selected from medical record data in a high throughput fashion
  - Eugene Braunwald
  - Henry Chueh
  - John Glaser
  - Diane Keogh
Research Patient Data Registry exists at Partners Healthcare

1) Queries for aggregate patient numbers
- Warehouse of in & outpatient clinical data
- 4.6 million Partners Healthcare patients
- 1.2 billion diagnoses, medications, procedures, genomics, laboratories, & physical findings coupled to demographic & visit data
- Authorized use by faculty status
- Clinicians can construct complex queries
- Queries cannot identify individuals, internally can produce identifiers for (2)

2) Returns identified patient data
- Start with list of specific patients, usually from (1)
- Authorized use by IRB Protocol
- Returns contact and PCP information, demographics, providers, visits, diagnoses, medications, procedures, laboratories, microbiology, reports (discharge, LMR, operative, radiology, pathology, cardiology, pulmonary, endoscopy), and images into a Microsoft Access database and text files.
Security and Patient Confidentiality of Step 1

- All patients at Partners are added
  - HIPAA notification that their data may be used for research upon registration.

- RPDR data is anonymized at the Query Tool.
  - Aggregated numbers are obfuscated to prevent identification of individuals; automatic lock out occurs if pattern suggests identification of an individual is being attempted.

A Security Architecture for Query Tools used to Access Large Biomedical Databases
Shawn N. Murphy, MD, Ph.D. and Henry C. Chneh, MD, M.S.
Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA.

- Queries done in Query Tool available for review by RPDR team, a user lock out will specifically direct a review.

- De-identified data warehouse is a “Limited Data Set” by HIPAA
  - Medical record numbers are encrypted and obvious identifiers are removed from data.

- Concept of “established medical investigator” is promoted by classification as a faculty sponsor.
Security and Patient Confidentiality of Step 2

- Only studies approved by the Institutional Review Board (IRB) are allowed to receive identified data.

- Queries may be set up by workgroup member, but faculty sponsor on IRB protocol must directly approve all queries that return identified data.

- Special controls exist when distributing data regarding HIV antibody and antigen test results, substance abuse rehab programs, and genetic data, due to specific state and federal laws.

- Queries that return identified data are reviewed (retrospectively) by the IRB.
Getting data into the RPDR from clinical systems

Source Systems

- BRIGHAM AND WOMEN'S HOSPITAL
- MASSACHUSETTS GENERAL HOSPITAL
- Faulkner Hospital
- NSMC
- Newton-Wellesley Hospital
- HPCGG Partners
- Reference Lab1
- Reference Lab2

Enterprise Integration Engine

PACS Services
MGH Services
BWH Services

User Interfaces

- LMR
- WEB
- Results Review
- ED
- CPOE

Data Services Layer

- VB/COM
- Cache'

Enterprise Integration Engine

RPDR

Cache'

CDR

LMR

EMPI

Loinc

Results Manager

(Cache')
Organizing data in the Clinical Data Warehouse

Star schema

- **Concept DIMENSION**
  - concept_key
  - concept_text
  - search_hierarchy

- **Encounter DIMENSION**
  - encounter_key
  - encounter_date
  - hospital_of_service

- **Patient Concept FACTS**
  - patient_key
  - concept_key
  - start_date
  - end_date

- **Patient DIMENSION**
  - patient_key
  - patient_id (encrypted)
  - sex
  - age
  - birth_date
  - race
  - deceased
  - ZIP

- **Pract. DIMENSION**
  - practitioner_key
  - name
  - service

Binary Tree

1200 million

150

.06

4.6

.14
Query items

Person who is using tool

Query construction

Results - broken down by number distinct of patients
Contains commands for working with the selected items.
Choose value of Heart Rate

Searches using health history can be constrained to be within certain values through the dialog below.

- **No value**
- **By high/low flag**
- **By numeric value**

Please select operator:
- **GREATER THAN (>)**

Please enter value:
- **100**

Click on a bar segment to help you specify a value or range:

- **50**
- **100**

**Units:** beats/minute

**Create Query** | **Request Data** | **Manage Results** | **Run Query**

**Gender:**
- **Male**: patients
- **Female**: patients

**Age**
- **0**
- **40**
- **80**

**Race:**
- **IAB/HWOU**

**Vital status:**
- **Alive**: patients
- **Dead**: patients
Please complete...

Please supply a name for this query:
JustDiagnosesAMI

Create a patient set from this query

Please choose one of these two timing models...

Either (1) Some item from each group must have occurred in the same visit; or (2) Some item from each group may have occurred at any visit. For example, if one wanted to find patients whose diabetes was treated at the MGH while an inpatient, choose the timing model where items occurred at the same visit.

Same Visit

<table>
<thead>
<tr>
<th>Patient</th>
<th>Date</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000004</td>
<td>12/15/1995</td>
<td>MGH</td>
</tr>
</tbody>
</table>

Any Visit

<table>
<thead>
<tr>
<th>Patient</th>
<th>Date</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000004</td>
<td>2/23/1999</td>
<td>BWH</td>
</tr>
<tr>
<td>0000004</td>
<td>10/05/1996</td>
<td>MGH</td>
</tr>
<tr>
<td>0000004</td>
<td>12/15/1995</td>
<td>BWH</td>
</tr>
</tbody>
</table>
Choose value of CK-MB Index

Searches by Lab values can be constrained by the high/low flag set by the performing laboratory, or by the values themselves.

- No value
- By high/low flag
- By numeric value

Please select operator:
- GREATER THAN (>)

Please enter value:
- 3.50

Click on a bar segment to help you specify a value or range:
- 0.00
- 2.50
- 3.50
- 5.00

The items of [ ] are intersected with [ ]

Create Query
- OK
- Clear
- Cancel

Gender: patients
- Male
- Female

Age:
- 0
- 40
- 80

Race:
- IAB|HWOU

Vital: patients
- Alive
- Dead
Please complete...

Please supply a name for this query:

AMI and CK-MB>3.5

Create a patient set from this query

What's this?

Please choose one of these two timing models...

Either (1) Some item from each group must have occurred in the same visit; or
(2) Some item from each group may have occurred at any visit. For example, if
one wanted to find patients whose diabetes was treated at the MGH while an
inpatient, choose the timing model where items occurred at the same visit.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Date</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000004</td>
<td>12/15/1995</td>
<td>MGH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inpatient</td>
</tr>
<tr>
<td>0000004</td>
<td>10/05/1995</td>
<td>MGH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Outpatient</td>
</tr>
<tr>
<td>0000004</td>
<td>12/15/1995</td>
<td>BWH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inpatient</td>
</tr>
<tr>
<td>0000004</td>
<td>2/23/1999</td>
<td>BWH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inpatient</td>
</tr>
<tr>
<td>0000004</td>
<td>2/23/1999</td>
<td>BWH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Seizure</td>
</tr>
<tr>
<td>0000004</td>
<td>2/23/1999</td>
<td>BWH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Outpatient</td>
</tr>
</tbody>
</table>
Welcome to the RPDR Data Request Wizards

The RPDR is a HIPAA compliant system, which returns aggregate patient information via a Query Tool, based on user-defined criteria. With proper IRB approval, RPDR users can:

- use their previously queried patient set
- or import their own approved set of Medical Record Numbers

...to request detailed or identified patient clinical data. These wizards are included in the RPDR for human research investigators to request identified patient data from their respective Partners sites.

You are now launching a wizard in order to request identified patient data.

Your request must conform and comply with the allowances of your Partners sponsored IRB human studies protocol. This responsibility rests entirely on the faculty sponsor who is requesting the identified data or who is approving the request of identified data from a workgroup member. It is very important that the correct IRB protocol number be chosen for each request of protected health information.

This information is protected under the Partners Privacy and Confidentiality Policy and provided with approval by the Human Research Committee only for the use specified in your protocol number. It may not be used for any other purpose without specific approval by the Human Research Committee. It may not be distributed to any individual not specifically authorized under that approval. The data must be managed in a manner that complies with HIPAA Security Regulations.

I accept responsibility for the data returned by this query.

Accept    Cancel

Partners Healthcare System HIPAA Compliance

Additional HIPAA information for the research community is available from these links, sponsored by Partners and the Human Research Council (PHRC).

HIPAA and the Privacy Rule
HIPAA Central
Select the sites from which you would like to receive data
(Please note that different sites and Institutional Review Boards (IRBs) may have different policies regarding obtaining patient data. Detail of the policies can be read here)

- Massachusetts General Hospital (MGH)
- Brigham and Women's Hospital (BWH)
- Newton Wellesley Hospital (NWH)
- Spaulding Rehabilitation Hospital (SRH)
- Faulkner Hospital (FH)
- North Shore Medical Center (NSMC)
Select protocol number(s)

Partners IRB (required): mgh-demo-1
Title: RPDR protocol - Demonstration IRB number for Dr. Murphy
Status: Active

Newton Wellesley Hospital IRB: NWH Demo 1
Title: test
Status: Active

Spaulding Rehabilitation Hospital IRB:

Options for returned set of patients:
- [ ] Create a static set of patients from this query that can be used in other RPDR queries
- [ ] Rerun the base query shown above to obtain a fresh set of patients
RPDR DETAILED DATA REQUEST WIZARD

Using IRB#mgd-demo-1 (found in the RPDR Identified database) to obtain data from the RPDR.
You are logged in as Murphy, Shawn N. in workgroup Shawn Murphy, MD

Select others who require access to this data

| Selected faculty sponsor(s): | Shawn Murphy, MD |

<table>
<thead>
<tr>
<th>Available</th>
<th>Chosen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peng, Zhaoping Z.</td>
<td>Gainer, Vivian S.</td>
</tr>
<tr>
<td>Nalichowski, Ruth</td>
<td></td>
</tr>
</tbody>
</table>

[Buttons: Add, Add All, Remove, Remove All]
Please select if you would like a HIPAA-defined (deidentified) limited data set or an identified data set

Limits a limited data set
- The files that result from this request will be available in a protected file share with no special encryption.

Identified Data Set
- The text files that result from this request will be encrypted and the Microsoft Access file will be password protected. In order to access the data, a password will be provided.
RPDR Detailed Data Request Wizard -- Web Page Dialog

RPDR DETAILED DATA REQUEST WIZARD
Using IRB#mgh-demo-1 (found in the RPDR Identified database) to obtain data from the RPDR
You are logged in as Murphy, Shawn N. in workgroup Shawn Murphy, MD

Select the types of data that should be returned from the RPDR
Only data allowed by your protocol should be chosen
(Identified data sets will always return a set of identified patient medical numbers)

Detail Data Items
- Demographic Data
- Identifying Patient Information - not available for Limited Data Sets
- LMR (Longitudinal Medical Record)
- Medications, Diagnoses and Procedures
- Medications, Diagnoses and Procedures from Billing Data - only visits where query criteria occur all in the same visit
- Patient Clinical Reports- not available for Limited Data Sets
  - Cardiology Reports
  - Discharge Summaries
  - Endoscopy Reports
  - Microbiology Data
  - Operative Notes
  - Pathology Reports
  - Pulmonary Reports
  - Radiology Reports
  - Transfusion Data, Blood Bank Data

Help  < Back  Step 9  Next >  Cancel
Please select the laboratory tests that should be returned
(Drag the lab item from the left panel and drop into the panel below)

Selected Data Items
- CK-MB Index (Group:CKMBRI) (Loinci:2159-4)

Other Laboratory Data Items:
- Laboratory tests
  - Cell Counts
  - Cell Markers
- Chemistry
  - Anemia Related Studies
  - Blood Gases
  - Blood Gases/Oximetry
- Cardiac Tests
  - BNP (Group:BNP)
  - CK Isos (Group:CK-MM1)
  - CK-BB (CK1) (Group:CK1/BB)
  - CK-BB% (Group:%CK1BB)
  - CK-MB (CK2) (Group:CK2/MB)
  - CK-MB Index (Group:CKMBRI)
  - CK-MB QUANT (Group:CPK-1)
  - CK-MB% (Group:%CKMB)
RPDR DISCLAIMER: All requests, such as this one, are reviewed by the IRB to assure compliance with the written protocol. If the data requested is not authorized by the protocol, those involved with its retrieval face disciplinary action including loss of federal funding, loss of employment, and/or criminal prosecution.

Read Full IRB Policy

Enter IRB Password:
Identified data is gathered

Output files placed in special directory

Data is gathered from RPDR and other Partners sources

Files include a Microsoft Access Database
2007’s usage of RPDR

- 1,580 registered users, 332 new in 2007
- 294 teams gathering data for research studies
- 815 identified patient data sets returned to these teams, containing data for 8.8 million patient records.
- From a survey of 153 teams
  - Importance of the data received from the RPDR was evaluated in relation to the study it was supporting.
  - The adequacy of the match of a patient profile that could be obtained through the RPDR query tool was estimated.
- $94-136$ million total research support critically dependent on RPDR from patient data received throughout life of funding.
- ~300 data marts were created to support hospital operations, representing about 80 million patient records

![Usefulness of Detailed Data](chart1)

![% of Patients Who Fit Required Profile](chart2)
Enterprise-wide repurposing and distribution of medical record data for research

- Enable high performance collection of medical record data for querying and distribution
  - Enterprise web client
  - Create patient cohorts for further investigation
- Enable discovery within data on enterprise wide scale
  - Relevance networks
  - Health Surveillance
Enterprise web client
http://services.i2b2.org/webclient/
Technical Overview

- Formed as a collection of interoperable services provided by i2b2 Cells
- Loosely coupled
- Makes no assumptions about proximity
- Connected by Web services
- Activity can be directed manually or automatically
i2b2 Cell: Canonical Hive Unit

- Business Logic
- Data Access
- Data Objects

HTTP XML
(minimum: RESTful, others like SOAP optional)

Programmatic Access
i2b2 Environment

Clinical Research Chart

shared / central

A

B

C

local

“Hive” of software services and core CRC cells

remote
Community

United States
- Beth Israel Deaconness Hospital, Boston, MA
- Boston University School of Medicine, Boston, MA
- Brigham and Women's Hospital, Boston, MA
- Children's Hospital, Boston, MA
- Denver Children's Hospital, Denver, CO
- Cincinnati Children's Hospital, Cincinnati, OH
- Cleveland Clinic, Cleveland, OH
- Weil Medical College of Cornell, NYC, NY
- Group Health Cooperative
- Harvard Medical School, Boston, MA
- Massachusetts General Hospital, Boston, MA
- Maine Medical Center, Portland, ME
- Marshfield Clinic, Wisconsin
- Morehouse School of Medicine, Atlanta, GA
- Oregon Health & Science University, Portland, OR
- Ohio State University Medical Center, Columbus, OH
- Philadelphia Children's Hospital, Philadelphia, PA
- Renaissance Computing Institute, Chapel Hill, NC
- Tufts New England Medical Center, Boston, MA
- University of California Davis, Davis, CA
- University of California San Francisco, SF, CA
- University of Massachusetts Medical School, Worcester, MA
- University of Michigan Medical Center, Ann Arbor, MI
- University of Pennsylvania School of Medicine, Philadelphia, PA
- University of Rochester Medical Center, Rochester, NY
- University of Texas Health Sciences Center Houston, Houston, TX
- University of Texas Health Sciences Center San Antonio, SA, TX
- University of Texas Health Sciences Center Southwestern,
- Utah Health Science Center, Salt Lake City, UT
- University of Washington, Seattle, WA

International
- Georges Pompidous Hospital, Paris, France
- University of Goettingen, Goettingen, Germany
- University of Pavia, Pavia, Italy
- University of Seoul, Seoul, Korea
SHRINE (Shared Research Informatics Network) = Distributed Queries

Central “aggregator” broadcasts query to local hospital “adaptors”, which return aggregate counts only.
High Throughput Method for supporting Research

- Set of patients is selected from medical record data in a high throughput fashion

- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data
  
  - Isaac Kohane
  - John Glaser
  - Susanne Churchill
  - Henry Chueh
  - Griffin Weber
  - Michael Mendis
  - Andrew McMurry
  - Vivian Gainer
  - Lori Phillips
  - Rajesh Kuttan
  - Wensong Pan
Set of patients is selected through Enterprise Repository and data is gathered into a data mart.
Use of medical record data in clinical studies focused upon teamwork and workflow

- Repurpose medical record information for research studies
  - I2b2 Workbench
  - Natural language processing
- Enable Genomic Studies and Public Heath
  - Tissue/blood selection
  - Health Reporting
Data is available through the i2b2 Workbench
Team support for Projects

- RPDR
- Local sources (Ex: BICS)
- Final Project DB
- EDC
- RPDR Support Programmers
- Analyst
- Project Manager
- Biostatistician
- Local data extract analyst
- Programmer
NLP Workflow

I2b2 Project Investigators

Communication

Results Delivery

NLP Specialists
NLP (and comedy) is not pretty
NLP Specialists Workstation

Export Notes

Import Derived Codes

NLP Specialists
Investigator Review
Select patients for clinical trials
Workbench display of Mutual Inform. Correlations
Project data can be added back to Enterprise Repository

[ Enterprise Shared Data ]

Shared data of Project 1

Shared data of Project 2

Shared data of Project 3

i2b2 DB Project 1

i2b2 DB Project 2

i2b2 DB Project 3

Ontology

Consent/Tracking

Security
High Throughput Method for supporting Research

- Set of patients is selected from medical record data in a high throughput fashion.

- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data.

- Using the BETR/Crimson system, tissues of these patients can be made available for genomic and biochemical analysis.
  - Lynn Bry
  - Natalie Boutin
Genotype samples and compare to controls
Cost and time benefit of Instrumenting with Sample Collection for Modest-size Study with 10,000 subjects (cases + controls)

<table>
<thead>
<tr>
<th>Old vs. New</th>
<th>Cost ($)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 chart review per patient (CP1)</td>
<td>$20</td>
<td>15 minutes/subject</td>
</tr>
<tr>
<td>High-throughput phenotyping (iP) through RPDR and i2b2</td>
<td>$50K Total</td>
<td>1 month total (conservative high estimate)</td>
</tr>
<tr>
<td>Sample acquisition through primary care provider (CP)</td>
<td>$650</td>
<td>3-5 subjects/week(^1)</td>
</tr>
<tr>
<td>High-throughput sample acquisition through RPDR and BETR/Crimson.</td>
<td>$20</td>
<td>50-200 subjects /week(^2)</td>
</tr>
</tbody>
</table>

= $6.7 million/study vs. $250 thousand/study
Escalating cost and time benefit of Instrumenting with Sample Collection

Previous model for collecting specimens

New model for collecting specimens
Meeting Expectations

Graph showing the number of samples over months for different categories:

- LAA
- HAA
- HCA
- LCA

The graphs display the cumulative number of samples for each category across 10 months, with the y-axis representing the number of samples ranging from 0 to 3000.
Accrual Rates

(a) RA (yr. total = 15840)
- Ashima (yr. total = 45238)
- MDD (yr. total = 39357)

(b) ASD (yr. total = 755)
- Huntington’s (yr. total = 102)
High Throughput Method for supporting Research

- Set of patients is selected from medical record data in a high throughput fashion.

- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data.

- Using the Crimson system, tissues of these patients can be made available for genomic and biochemical analysis.

- Automated discovery can be created from these projects to support further hypothesis-driven research:
  - John Brownstein
  - Judy Colecchi
First signal:
- 1 year after Celecoxib
- 8 months after Rofecoxib
Effect on patient age

- Negative association between mean age at MI and prescription volume
- Spearman correlation -0.67, P<0.05
Performing an observational, phase IV study is an expensive and complex process that can be potentially modeled in a retrospective database using groups of patients available in the large amounts of highly organized medical data.

Fundamental problems complicate this approach:
- Patients drift in and out of the system. Sophisticated statistical models using adequate control populations are necessary to compensate.
- Confounding variables are not found in the database. Sophisticated natural language processing is needed to extract the confounders from textual reports to allow these confounders to be controlled.
- Missing data disrupts typical statistical approaches
Odds Ratios for Diseases expressed in comparing Rosiglitazone vs. Pioglitazone
Take-away points from i2b2

1) Enable Enterprise use of Patients for Research
2) Power comes from Numbers of Patients Recruited
3) Extensible Architecture for Developers
4) Enable Scientist Workflow
5) Enable Teamwork between Informaticians and Researchers
6) Communicate through Visualizations
7) Enable Natural Language Processing
8) High throughput Tissue acquisition for Genomic Research
9) Enable Health Surveillance
10) Enable Data Sharing
References


Murphy, SN; Mendis, M, K. Hackett; et al.; Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside; AMIA, Fall Symp. 2007.

Qing T Zeng, Sergey Goryachev, Scott Weiss, et. al., Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system, BMC Medical Informatics and Decision Making, 2006.

Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. BMJ. 2009 Jan 27;338:b81