Computing our Patient’s Future Using Data from our Healthcare Institutions

Shawn Murphy MD, Ph.D.
NETTAB 2011 Workshop on Clinical Bioinformatics
Example: PPAR\textsubscript{\gamma} Pro12Ala and Diabetes

Overall \( P \) value = \( 2 \times 10^{-7} \)

Odds ratio = 0.79 (0.72-0.86)

Ala is protective

Courtesy J. Hirschhorn
The Power of Numbers: Efficiently Reaching a Large $N$

- High throughput genotyping
- High throughput phenotyping
- High throughput sample acquisition

DHHS Secretary’s Advisory Committee on Genetics, Health, and Society (SACGHS) argues for the health value of a 500,000 to 1M subject study. Estimated cost: $3,000,000,000

Cost of the pediatric 100,000 study recently launched >> $1B + decades.
High Throughput Methods for supporting Research at Partners Healthcare

- 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 Methods for supporting Research at Partners Healthcare

- 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
Research Patient Data Registry exists at Partners Healthcare to find patient cohorts for clinical research

1) Queries for aggregate patient numbers

- Warehouse of in & outpatient clinical data
- 5.0 million Partners Healthcare patients
- 1.3 billion diagnoses, medications, procedures, 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
  Shawna N. Murphy, MD, Ph.D. and Henry C. Chueh, 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.
2009’s usage of RPDR

- 2,227 registered users, 457 new in 2008
- 338 teams gathering data for research studies
- 1286 identified patient data sets returned to these teams, containing data of 7.8 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
106 Total Responses

- Critical: 43%
- Useful: 42%
- Not Useful: 15%

% of Patients Who Fit Required Profile
105 Total Responses

- > 75%: 33%
- 50% - 75%: 26%
- 25% - 50%: 22%
- < 10%: 19%
Organizing data in the Clinical Data Warehouse

Star schema

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

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
- practitioner_key

Encounter DIMENSION
- encounter_key
- encounter_date
- hospital_of_service

Practitioner DIMENSION
- practitioner_key
- name
- service

Binary Tree

Start search

1300 million
FINDING PATIENTS

Query items

Person who is using tool

Query construction

Results - broken down by number distinct of patients
MATCHING PATIENTS

Previous query items

Case set construction

Control set construction

Estimate set size and run program
RPDR Detailed Data Request Wizard -- Web Page Dialog

Using Partners IRB#2002P000381 (Research Patient Data Registry (RPDR)) to obtain data from the RPDR
You are logged in as Duey, Stacey A. 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)

- Allergy Data from PEAR (Partners Enterprise Allergy Repository)
- Demographic Data
- Identifying Patient Information - not available for Limited Data Sets
- LMR (Longitudinal Medical Record)
- Medications, Diagnoses and Procedures
- 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
  - Top three providers for each patient

Step 9
High Throughput Methods for supporting Research at Partners Healthcare

- 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
Set of patients is selected through Enterprise Repository and data is gathered into a data mart.
Data is available through the i2b2 Workbench
Research Investigator Workflow enabled by mi2b2

Query is done
To find patients

Use i2b2

Request Images with Accession #’s

mi2b2

Images Retrieved from Clinical PACS

Derive new data from images

Study Images

BIRN/XNAT

Images

Retrieved from Clinical PACS
Team support for Projects

- Project Manager
- Biostatistician
- Analyst
- Local data extract analyst
- Programmer
- RPDR Support Programmers
NLP Workflow

Communication

I2b2 Project Investigators

Results Delivery

NLP Specialists
NLP (and comedy) is not pretty

SOCIAL HISTORY: The patient is a **nonsmoker**. No alcohol.

SOCIAL HISTORY: **Negative for tobacco**.

SOCIAL HISTORY: **Non-Smoker**

SOCIAL HISTORY: The patient is a **nonsmoker**. No alcohol.

SOCIAL HISTORY: The patient lives in rehab, married. **Unclear smoking** history from the admission note…

SOCIAL HISTORY: The patient lives in rehab, married. **Past Smoker**

BRIEF RESUME OF HOSPITAL COURSE: 63 yo woman with COPD, 50 pack-yr tobacco (quit 3 wks ago), spinal stenosis, ...

HOSPITAL COURSE: ... It was recommended that she receive …We also added Lactinax, oral form of *Lactobacillus* acidophilus to attempt a repopulation of her gut.

SH: widow, lives alone, 2 children, no **tob/alcohol**.

**Hard to pick**
Investigator Review
Project data can be added back to Enterprise Repository
Community

United States
- Arizona State University
- Beth Israel Deaconess Hospital, Boston, MA
- Boston University School of Medicine, Boston, MA
- Brigham and Women's Hospital, Boston, MA
- Case Western Reserve Hospital
- Children's Hospital, Boston, MA
- (Denver) Children's Hospital, Denver, CO
- Children's Hospital of Philadelphia, PA
- Children's National Medical Center (GWU)
- Cincinnati Children's Hospital, Cincinnati, OH
- Cleveland Clinic, Cleveland, OH
- (Weil Medical College of) Cornell, NYC, NY
- Duke Medical College
- Group Health Cooperative
- Harvard Pilgrim Healthcare
- Harvard Medical School, Boston, MA
- Health Sciences South Carolina
- Kaiser Permanente Health
- Kimmel Cancer Center (Thomas Jefferson University)
- Massachusetts General Hospital, Boston, MA
- Maine Medical Center, Portland, ME
- Marshfield Clinic, Wisconsin
- Morehouse School of Medicine, Atlanta, GA
- Ohio State University Medical Center, Columbus, OH
- Oregon Health & Science University, Portland, OR
- Renaissance Computing Institute, Chapel Hill, NC
- South Carolina Clinical and Translational Research Institute
- Tufts Medical Center, Boston, MA
- University of Alabama
- University of Arkansas Medical School
- University of California Davis, Davis, CA
- University of California San Francisco, SF, CA
- University of Chicago
- 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 at Houston, Houston, TX
- University of Texas Health Sciences Center at San Antonio, SA, TX
- Utah Health Science Center, Salt Lake City, UT
- University of Washington, Seattle, WA
- University of Wisconsin Madison
- Veterans Administration Boston and Utah

International
- Georges Pompidous Hospital, Paris, France
- Institute for Data Technology and Informatics (IDI), NTNU, Norway
- Karolinska Institute, Sweden
- University of Erlangen-Nuremberg, Germany
- University of Goettingen, Goettingen, Germany
- University of Pavia, Pavia, Italy
- University of Seoul, Seoul, Korea
Aggregating across 4 hospitals, 3 i2b2 instances

SHRINE (Shared Research Informatics Network) = Distributed Queries
Clinical data in SHRINE

- 10 years (2001-2011)
- 4 hospitals
- 6 million total patients
- >1 billion medical observations
  - Demographics
  - Diagnoses (ICD9-CM)
  - Medications (RxNorm)
  - Labs (LOINC)
High Throughput Methods for supporting Research at Partners Healthcare

- 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

- Automated discovery can be created from these projects to support further hypothesis-driven research
Genotype samples and compare to controls

DNA Genotyping

DE-IDENTIFIED I2B2 DATA REPOSITORY

FIREWALL

Lab. Info. System

Narrative Electronic Medical Record

Codified data (e.g. billing)

i2b2 data mart

Asthma

NLP

13101 21030 30121 Match 93110 41030 22031 44310

1000100101110100 1001100111100111 0011101111110011 0101101001010010 1100100010001001
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. <strong>New</strong></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/week¹</td>
</tr>
<tr>
<td>High-throughput sample acquisition through RPDR and BETR/Crimson.</td>
<td>$20</td>
<td>50-200 subjects /week²</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
Accrual Rates

(a) RA (yr. total = 18840)  
Asthma (yr. total = 42238)  
MDD (yr. total = 39357)

(b) ASD (yr. total = 756)  
Huntington’s (yr. total = 102)

# Specimens per week

Week
High Throughput Methods for supporting Research at Partners Healthcare

- 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
Performing Clinical trials “in-silico”

- 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 with large amounts of well organized medical data.

- Fundamental problems complicate this approach:
  - Patients drift in and out of the healthcare system. Sophisticated statistical models using adequate control populations are necessary to compensate for the drift.
  - Confounding variables may not be found in the database. Natural language processing may be needed to extract the confounders from textual reports to allow confounders to be exposed.
  - Unknown missing data disrupts typical statistical approaches.
  - Biases in the data can easily mislead the investigator to false conclusions; data exploration and visualization tools are needed to expose these kinds of potential problems.
Dashboard used to observe high-level signals
Dashboard used to observe high-level signals
Set of patients is selected through Enterprise Repository and data is gathered into a data mart.

Selected patients

Data directly from EDR
Data from other sources
Data collected specifically for project

Daily Automated Queries search for Patients and add Data

Project Specific Phenotypic Data
Builds complex “Custom Study” displays
Builds complex “Custom Study” displays
Seven important factors enabled by i2b2 platform

1) Enables enterprise-wide repurposing of health care data for research
2) Enables extensible software architecture for developers
3) Extends EHR research so that data may be shared among sites
4) Enables natural language processing
5) Provides method for materializing scientific method for EHR-based investigations
6) Extends EHR research so that data may be shared among sites and samples may be obtained
7) Provides platform for Clinical Trials “in silico”
Collaborators

- **RPDR**
  - Eugene Braunwald
  - John Glaser
  - Diane Keogh
  - Henry Chueh

- **i2b2**
  - Isaac Kohane
  - Susanne Churchill
  - Griffin Weber
  - Michael Mendis
  - Vivian Gainer
  - Lori Phillips
  - Rajesh Kuttan
  - Wensong Pan
  - Janice Donahue
  - William Simons (SHRINE)
  - Andy McMurry (SHRINE)
  - Doug McFadden (SHRINE)

- **Medical Imaging (mi2b2)**
  - Christopher Herrick
  - David Wang
  - Bill Wang

- **Sample Acquisition**
  - Lynn Bry
  - Natalie Boutin

- **i2b2 Driving Biology Projects**
  - Vivian Gainer
  - Victor Castro
  - Raul Guzman
  - Robert Plenge
  - Scott Weiss
  - Stan Shaw
  - John Brownstein
  - Qing Zeng
  - Guergana Savova