Use of the Research Patient Data Registry at Partners Healthcare, Boston

Advancing Clinical Research with Hospital Clinical Records

Shawn Murphy MD, Ph.D.
Massachusetts General Hospital
Outline of presentation

- Introduction to the Research Patient Data Registry (RPDR)
- How data is organized in the RPDR
- How to query the data in the RPDR
- How data is delivered from the RPDR
- The future of using clinical data for research
Research Patient Data Registry exists at Partners Healthcare to allow exploration of clinical data

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. 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.
# 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 of 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
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
  - practitioner_key
  - encounter_key
  - value_type
  - numeric_value
  - textual_value
  - abnormal_flag

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

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

**Binary Tree**

- start search

- 4.6
- 0.06
- 150
- 1.4
- 1200 million

- 1.14
- 0.06150
- 4.6
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/flow 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

OK  Clear  Cancel
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...

- Same Visit

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.

- Any Visit

Patient | Date      | Groups    | 1 | 2 | 3
--------|-----------|-----------|---|---|---
0000004 | 12/15/1995| MGH       |   | Inpatient | Diabetes

0000004 | 2/23/1999  | BWH       |   | Inpatient | Seizure

0000004 | 10/05/1995 | MGH       |   | Outpatient | Afi

0000004 | 12/15/1995 | BWH       |   | Outpatient | Diabetes

Run the query as background job

Run Query | Cancel Query
Research Patient Database Query Tool

Please complete...

Please supply a name for this query:
AMI and CK-MB>3.5

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.

<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, Inpatient, Diabetes</td>
</tr>
</tbody>
</table>

<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, Inpatient, Seizure</td>
</tr>
<tr>
<td>0000004</td>
<td>10/05/1998</td>
<td>MGH, Outpatient, Alib</td>
</tr>
<tr>
<td>0000004</td>
<td>12/15/1995</td>
<td>BWH, Outpatient, Diabetes</td>
</tr>
</tbody>
</table>
Antonarakis Variant Notation

Wildtype Sequence

5' ..TGAAGTGTATCCGACAT.. 3'
   tgacatAggctgtag
tgacatCggctgtag
tgacatGggctgtag
tgacatTggtgtgtag
3' gacataAggtgtgag
   gacataCggtgtgaga
gacataGggtgtgaga
gacataTggtgtgaga

Variant

5' ..TGAAGTGTATCCGACAT.. 3'
   tgacatAggctgtag
tgacatCggctgtag
tgacatGggtgtgtag
tgacatTggtgtgtag
3' gacataAggtgtgag
   gacataCggtgtgaga
gacataGggtgtgaga
gacataTggtgtgaga

Footprint

16,493
Research Patient Database Query Tool

Query Items

- Molecular Medicine
  - Genomics
    - Hearing Loss
    - Hypertrophic Cardiomyopathy
    - Marfan Syndrome, TYP1
    - Non-Small Cell Lung Cancer
  - EGFR
    - 21 25G-A (Responsive)
    - 21 254T (Unclassified)
    - 21 555G-A (Unclassified)
    - 21 555G-A (Responsive)
    - 21 555G-T (Responsive)
    - 21 555G-C (Responsive)
    - 21 555G-C (Novel Presumed Benign)
    - 223CC-T (Novel Presumed Benign)
    - 2237_2235del (Responsive)
    - 2237_2235del (Responsive)
    - 2237_2235delT (Responsive)
    - 2237_2235delC (Responsive)
    - 2238_2235del (Responsive)
    - 2242_2245del (Responsive)
    - 2242_2245del (Responsive)
    - 2242_2251del (Responsive)
    - 2245_2251del (Responsive)
    - 2245_2251del (Responsive)
    - 2253_2256del (Responsive)
    - 2254_2277del (Responsive)
    - 228CC-T (Presumed Benign)
    - 2294_2306dup (Responsive)
    - 2303G-C (Unclassified)
    - 2307_2315dup (Responsive)

Data Status

Logging: Shawn Murphy, MD
Status:
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)
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 protocol number(s)

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

Newton Wellesley Hospital IRB: N/W 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
Please select if you would like a HIPAA-defined (deidentified) limited data set or an identified data set

What's a limited data set?

**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.
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)
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) (Loinc:2158-4)
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:  

Data is gathered from RPDR and other Partners sources.

Files include a Microsoft Access Database.

Identified data is gathered.

Output files placed in special directory.
Distribution method for images
Distribute very large sets of RPDR data in data marts (used mostly to support hospital operations)

Data Mart is created with random patient numbers and encounter numbers. Mappings database matches random patient numbers to identifiers.
Future of using Clinical Data for Research

- Extending Phenotypic Data Collection
- Anonymous specimen repository
- Bayesian inference engine
- Predictive modeling
- Clinical trials performed in-silico
- Discovering correlations within data (relationship networks)
Extending Phenotypic Data Collection

RPDR

Selected patients

Data directly from RPDR
Data from other hospital sources
Data collected specifically for project

Daily Automated Queries search for Patients and add Data

Project Specific Phenotypic Data
Creating Research Data Sets

- Cohort Identification
- Primary data collection
- Research Cohort
- Research Data Set
Partners current model

Clinical Systems
Financial
Admin

Clinical
Electronic Records

Transform Normalize to metadata De-identify

IRB Protocol

RPDR

Limited Data Set

Anonymous Data Set

Research Data Set

Primary data collection

Research Cohort
Research Silos Despite RPDR
Integration into Shared CRC

[ Enterprise Shared Data ]

Shared data of Project 1

Shared data of Project 2

Shared data of Project 3

CRC DB
Project 1

Ontology
Consent/Tracking
Security

CRC DB
Project 2

CRC DB
Project 3
Creating de-identified tissue repository

De-identified data exists in the RPDR from all of Partners Healthcare, that can be linked to the samples in the BWH/CL.

Discarded Tissue

The discarded tissue may be from any source, but only if collected in the course of clinical care, and only if would be discarded.
Creating de-identified tissue repository

Research
Patient
Data Registry

The Identifiers that came in with the data and the samples are substituted for a random identifier. Any identifiers not conforming to the HIPAA limited data set requirements are removed.

BWH/CL
Creating de-identified tissue repository

De-identified Data Registry

De-identified samples

Links to identified patients are broken

BWH/CL
Clinical trials performed 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 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
Relevance Networks – unsupervised learning techniques

- Associates phenotypes and genotypes to generate novel hypothesis.

<table>
<thead>
<tr>
<th>person</th>
<th>concept</th>
<th>date</th>
<th>raw value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z5937X</td>
<td>Surgery</td>
<td>3/4</td>
<td></td>
</tr>
<tr>
<td>Z5937X</td>
<td>ER visit</td>
<td>3/4</td>
<td></td>
</tr>
<tr>
<td>Z5937X</td>
<td>Trauma</td>
<td>3/4</td>
<td></td>
</tr>
<tr>
<td>Z5937X</td>
<td>Gene-Chips</td>
<td>3/4</td>
<td></td>
</tr>
<tr>
<td>Z5937X</td>
<td>Seizure</td>
<td>4/6</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>Gene-Chips</td>
<td>5/2</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>Seizure</td>
<td>5/2</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>Alzheimer’s</td>
<td>5/2</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>Diabetes</td>
<td>5/2</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>CT Scan</td>
<td>3/9</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>Hemorrhage</td>
<td>3/9</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>Trauma</td>
<td>3/9</td>
<td></td>
</tr>
<tr>
<td>Z5956X</td>
<td>Thalamus</td>
<td>3/9</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- Most of the operational cost of building a system like the RPDR occurs when gathering and classifying data from the medical record systems.

- Use of the analytical database “Star schema” allows concepts from radiology and molecular medicine to be queried in a unified manner with other clinical phenotypic data.

- Besides supporting patient recruitment for clinical studies and for clinical operations, large databases of well organized medical record information offer significant promise towards the support of primary clinical research.
References