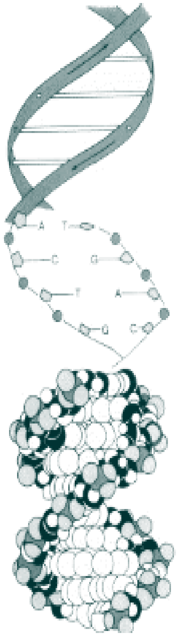
PARTNERS
HEALTHCARE



# Instrumenting the Health Care Enterprise for Discovery Research

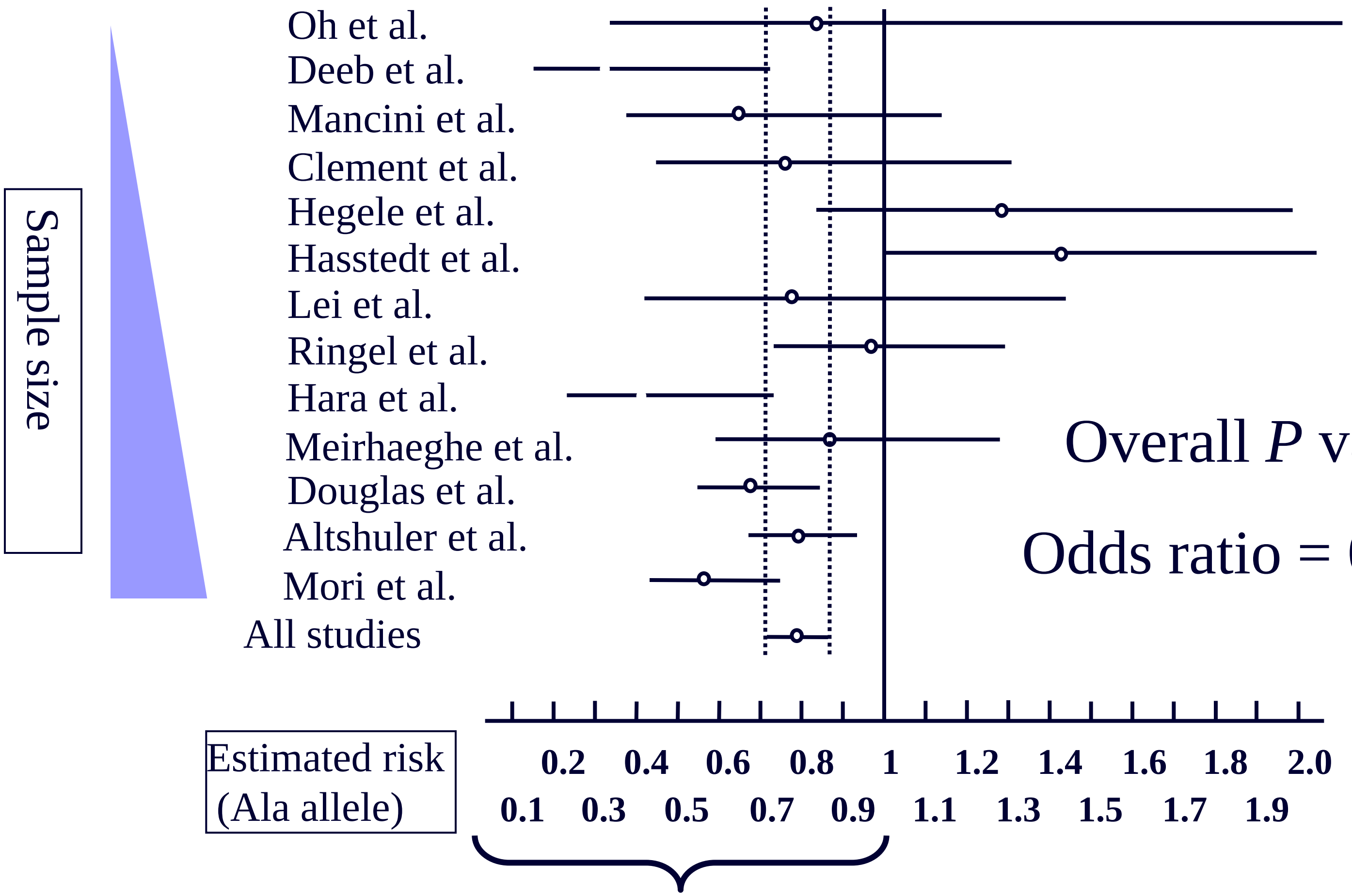
***Shawn Murphy MD, Ph.D.***
***CrEDIBLE  Project Meeting***
***October 15, 2012***

# Conflict of Interest Disclosure (Nothing to Disclose)

Shawn Murphy MD, Ph.D.

Neither I nor members of my immediate family have any financial relationships with commercial entities that may be relevant to this presentation.

# Example: PPARγ Pro12Ala and Diabetes



Overall *P* value = 2 x $10^{-7}$

Odds ratio = 0.79 (0.72-0.86)

Ala is protective

Courtesy J. Hirschhorn

# **The Power of Numbers: Efficiently Reaching a** Large ***N* *for clinical studies***

- High throughput genotyping
- High throughput phenotyping + 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 Translational Research

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

- Investigators explore phenotypes of these patients using i2b2 tools and a translational team developed to work specifically with medical record data

- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing

- Tissues of these patients can be made available for genomic and biochemical analysis

## High Throughput Methods for supporting Translational Research

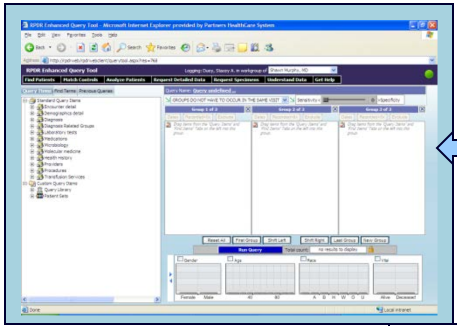
- Set of patients is selected from medical record data in a high throughput fashion
- Investigators explore phenotypes of these patients using i2b2 tools and a translational team developed to work specifically with medical record data
- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing
- Tissues of these patients can be made available for genomic and biochemical analysis

# 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
- 6.0 million Partners Healthcare patients
- 1.5 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)

**Query construction in web tool**



De-identified Data Warehouse

Z731984X
Z74902XX
...
...

Encrypted identifiers

## 2) Returns identified patient data

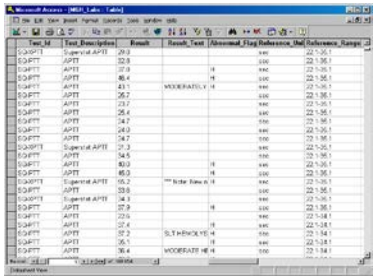
- Start with list of specific patients, usually from (1)
- Authorized use by separate IRB Protocols
- Returns contact and PCP information, demographics, providers, visits, diagnoses, medications, procedures, laboratories, microbiology, reports (discharge, LMR, operative, radiology, pathology, cardiology,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

0000004
2185793
...
...

OR

0000004
2185793
...
...

Real identifiers

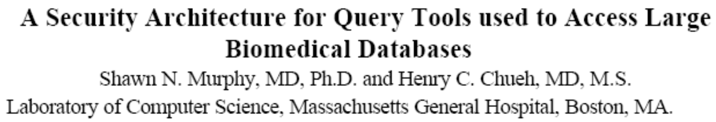

Microsoft Access - [MGH_Labs : Table]
File Edit View Insert Format Records Tools Window Help
Datasheet View

| Test_Id | Test_Description | Result | Result_Text | Abnormal_Flag | Reference_Uni | Reference_Range |
|---|---|---|---|---|---|---|
| SQXPTT | Superstat APTT | 29.0 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 32.8 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 37.0 | | H | sec | 22.1-35.1 |
| SQPTT | APTT | 46.4 | | H | sec | 22.1-35.1 |
| SQPTT | APTT | 43.1 | MODERATELY | H | sec | 22.1-35.1 |
| SQPTT | APTT | 26.7 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 23.7 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 25.4 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 24.7 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 24.0 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 24.7 | | | sec | 22.1-35.1 |
| SQXPTT | Superstat APTT | 31.3 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 34.5 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 40.0 | | H | sec | 22.1-35.1 |
| SQPTT | APTT | 46.0 | | H | sec | 22.1-35.1 |
| SQXPTT | Superstat APTT | 55.2 | *** Note: New n | H | sec | 22.1-35.1 |
| SQPTT | APTT | 33.8 | | | sec | 22.1-35.1 |
| SQXPTT | Superstat APTT | 34.1 | | | sec | 22.1-35.1 |
| SQPTT | APTT | 37.9 | | H | sec | 22.1-35.1 |
| SQPTT | APTT | 22.6 | | | sec | 22.1-38.1 |
| SQPTT | APTT | 37.4 | | H | sec | 22.1-38.1 |
| SQPTT | APTT | 37.2 | SLT HEMOLYE | H | sec | 22.1-34.1 |
| SQPTT | APTT | 36.1 | | H | sec | 22.1-38.1 |
| SQPTT | APTT | 36.4 | MODERATE HE | H | sec | 22.1-34.1 |

# 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.

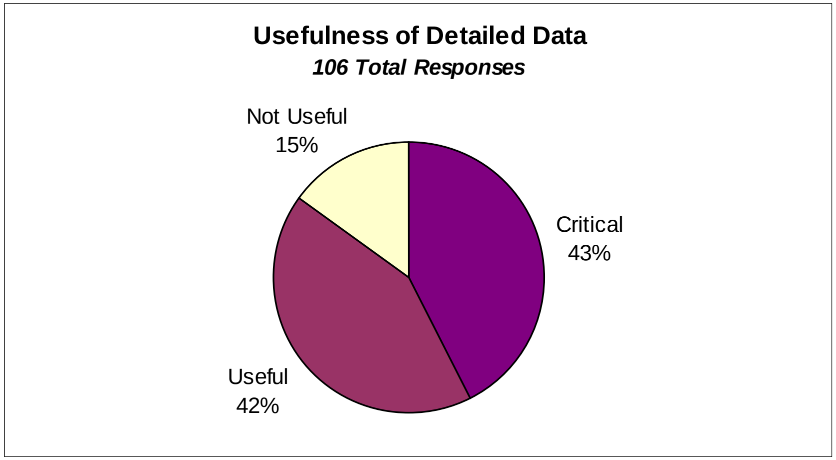
- Queries may be set up by workgroup member, but faculty sponsor on IRB protocol must directly approve all queries that return identified data.

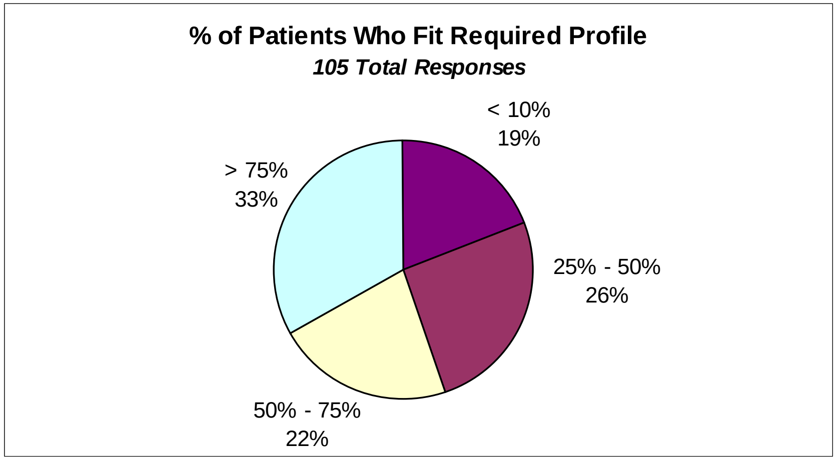
- Special controls exist when distributing data regarding HIV antibody and antigen test results, substance abuse rehab programs, and genetic data, due to specific state and federal laws.

- Queries that return identified data are reviewed (retrospectively) by the IRB.

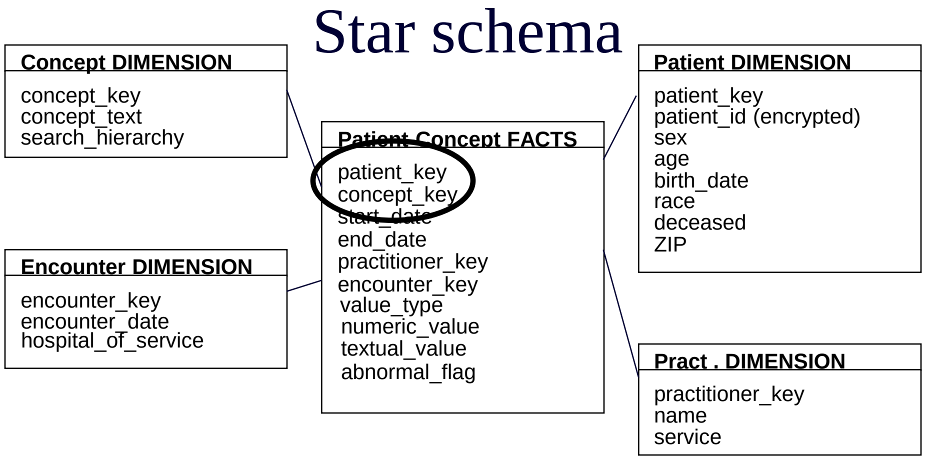
# 2011’s usage of RPDR

- 2,733 registered users, 457 new in 2011
- 462 teams gathering data for research studies
- 1852 detailed 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





# Organizing data in the Clinical Data Warehouse



Star schema

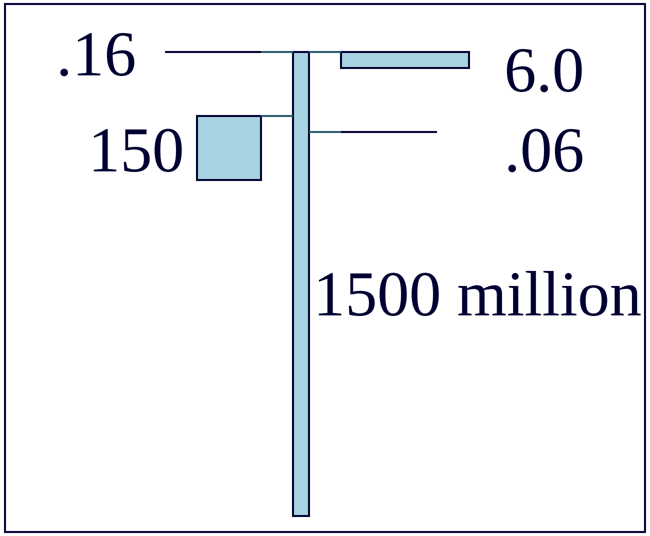
**Concept DIMENSION**
concept_key
concept_text
search_hierarchy

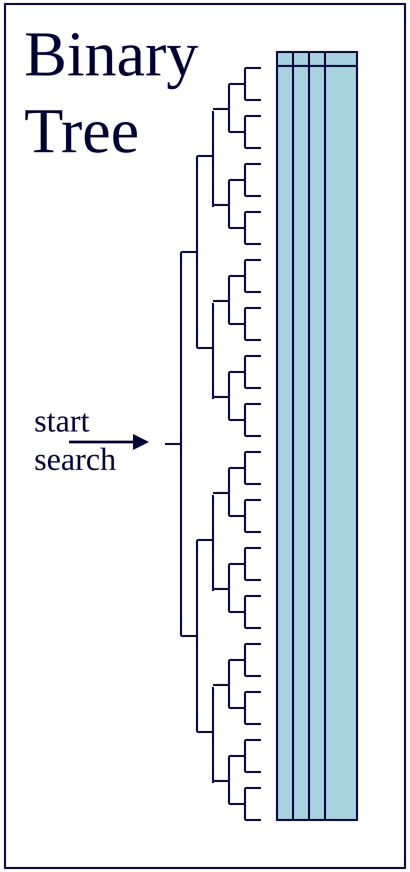
**Encounter DIMENSION**
encounter_key
encounter_date
hospital_of_service

**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

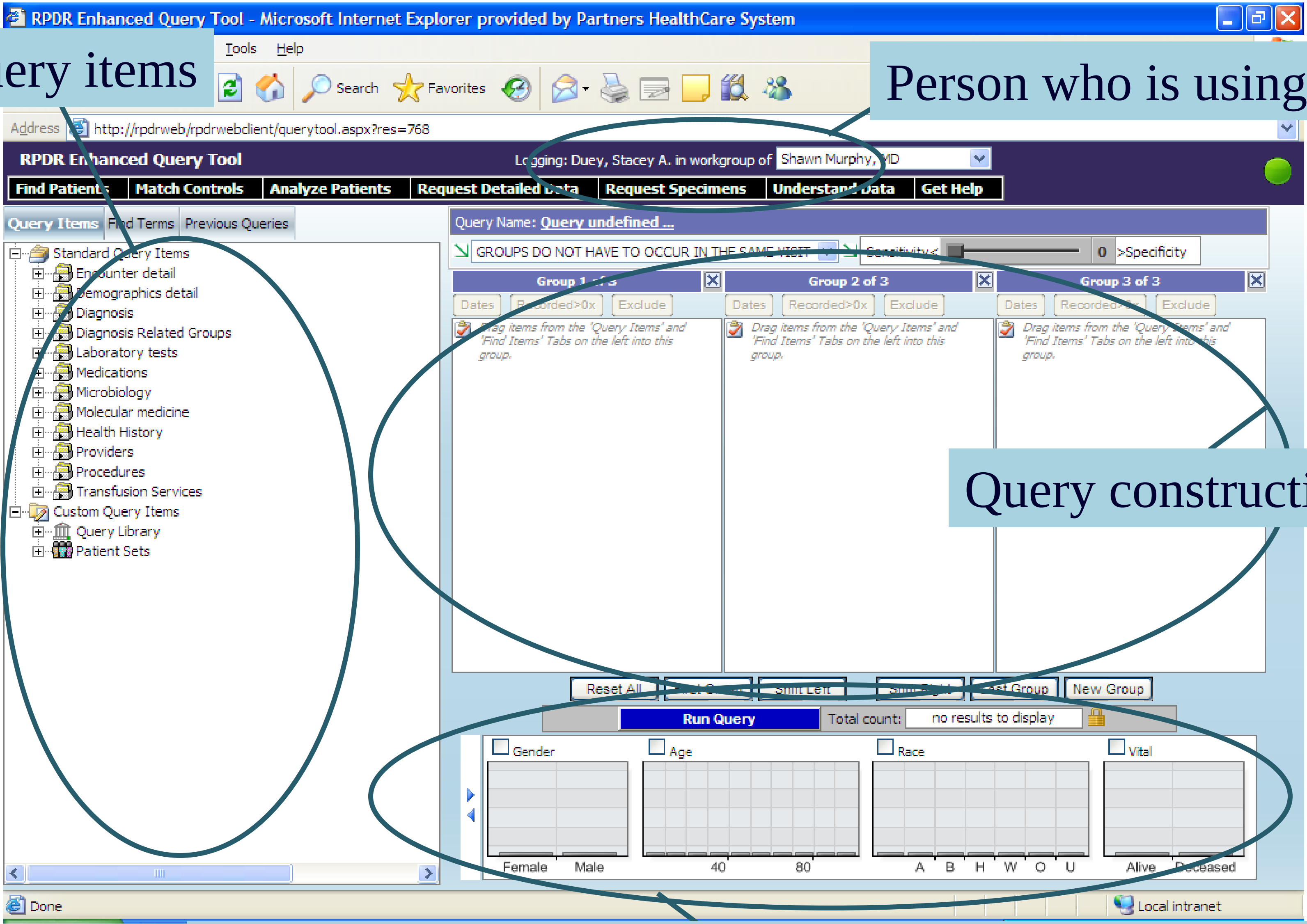
**Pract . DIMENSION**
practitioner_key
name
service





Binary Tree

# FINDING PATIENTS



RPDR Enhanced Query Tool - Microsoft Internet Explorer provided by Partners HealthCare System

Address http://rpdrweb/rpdrwebclient/querytool.aspx?res=768

Logging: Duey, Stacey A. in workgroup of Shawn Murphy, MD

Query Name: Query undefined ...

GROUPS DO NOT HAVE TO OCCUR IN THE SAME VISIT

Sensitivity< 0 >Specificity

Drag items from the 'Query Items' and 'Find Items' Tabs on the left into this group.

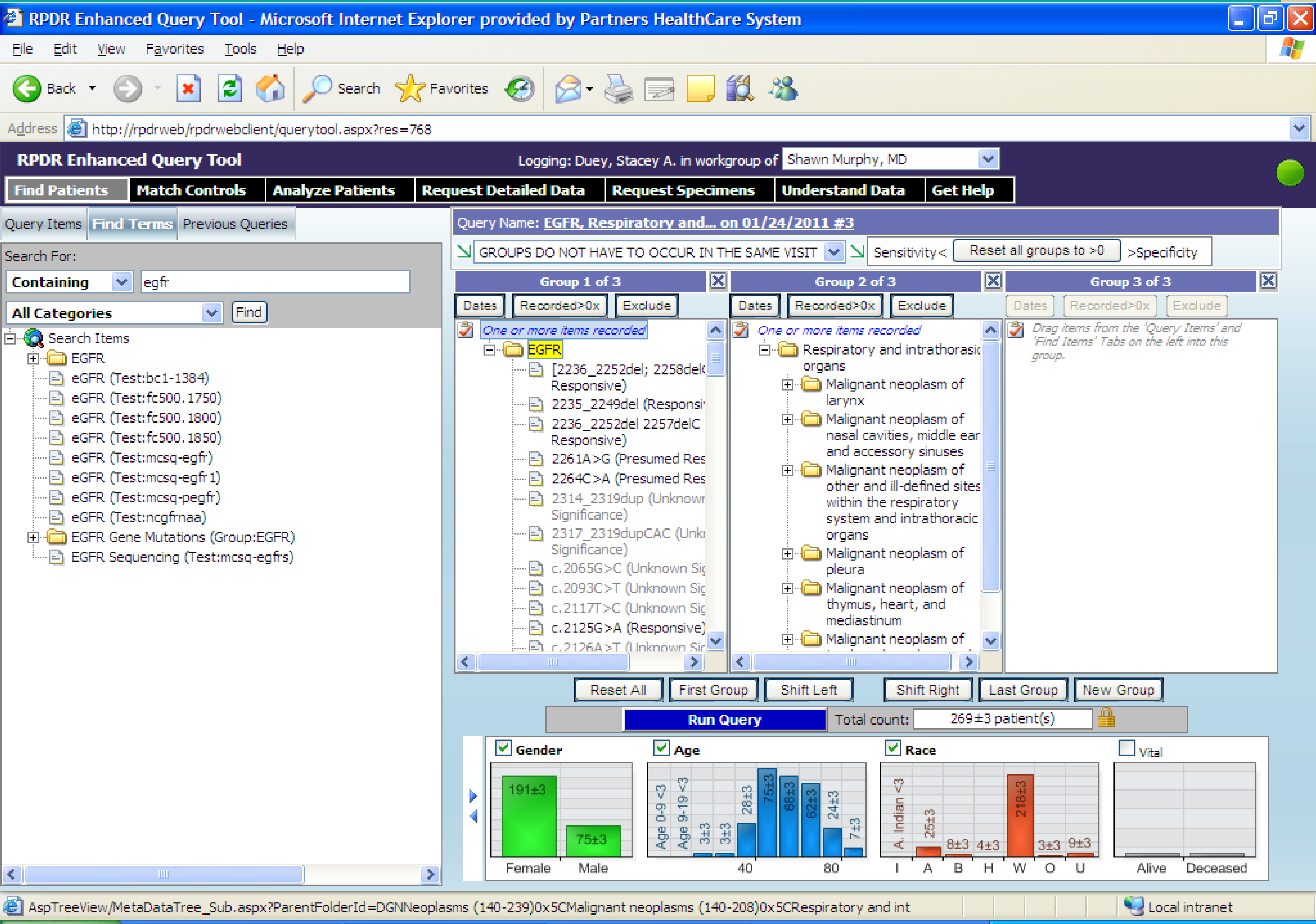
Total count: no results to display

Query items

Person who is using tool

Query construction

Results - broken down by number distinct of patients



Address: http://rpdrweb/rpdrwebclient/querytool.aspx?res=768

Logging: Duey, Stacey A. in workgroup of Shawn Murphy, MD

Search For: Containing egfr

All Categories

- Search Items
  - EGFR
  - eGFR (Test:bc1-1384)
  - eGFR (Test:fc500. 1750)
  - eGFR (Test:fc500. 1800)
  - eGFR (Test:fc500. 1850)
  - eGFR (Test:mcsq-egfr)
  - eGFR (Test:mcsq-egfr1)
  - eGFR (Test:mcsq-pegfr)
  - eGFR (Test:ncgfrnaa)
  - EGFR Gene Mutations (Group:EGFR)
  - EGFR Sequencing (Test:mcsq-egfrs)

Query Name: EGFR, Respiratory and... on 01/24/2011 #3

GROUPS DO NOT HAVE TO OCCUR IN THE SAME VISIT

Sensitivity< Reset all groups to >0 >Specificity

**Group 1 of 3**

One or more items recorded

- EGFR
  - [2236_2252del; 2258del( Responsive)
  - 2235_2249del (Responsi
  - 2236_2252del 2257delC Responsive)
  - 2261A>G (Presumed Res
  - 2264C>A (Presumed Res
  - 2314_2319dup (Unknown Significance)
  - 2317_2319dupCAC (Unk Significance)
  - c.2065G>C (Unknown Si
  - c.2093C>T (Unknown Si
  - c.2117T>C (Unknown Si
  - c.2125G>A (Responsive)
  - c.2126A>T (Unknown Sic

**Group 2 of 3**

One or more items recorded

- Respiratory and intrathorasic organs
  - Malignant neoplasm of larynx
  - Malignant neoplasm of nasal cavities, middle ear and accessory sinuses
  - Malignant neoplasm of other and ill-defined sites within the respiratory system and intrathoracic organs
  - Malignant neoplasm of pleura
  - Malignant neoplasm of thymus, heart, and mediastinum
  - Malignant neoplasm of

**Group 3 of 3**

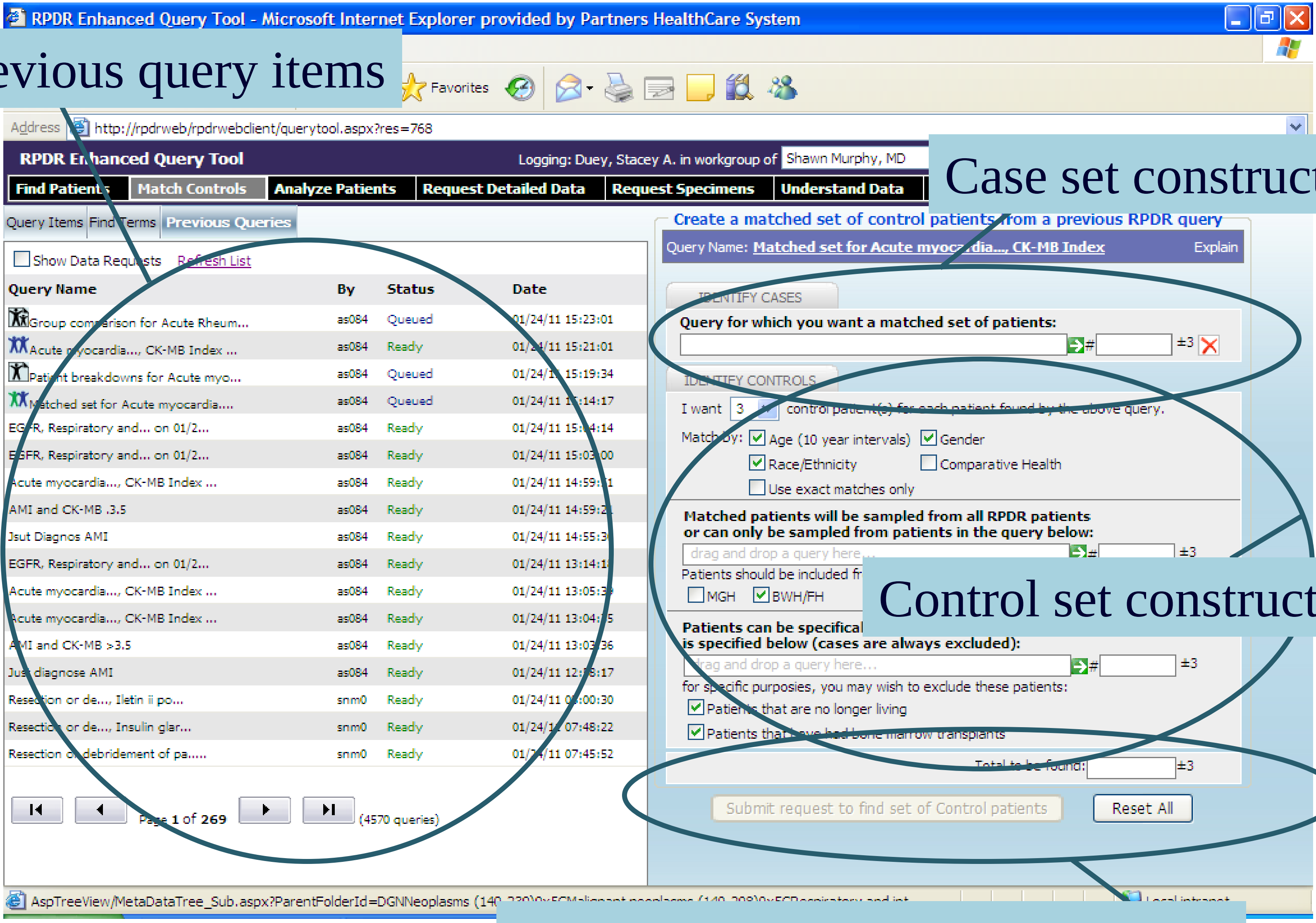
Drag items from the 'Query Items' and 'Find Items' Tabs on the left into this group.

Total count: 269±3 patient(s)

☑ Gender ☑ Age ☑ Race □ Vital

AspTreeView/MetaDataTree_Sub.aspx?ParentFolderId=DGNNeoplasms (140-239)0x5CMalignant neoplasms (140-208)0x5CRespiratory and int

# MATCHING PATIENTS



Previous query items

Case set construction

Control set construction

Estimate set size and run program

Logging: Duey, Stacey A. in workgroup of Shawn Murphy, MD

☐ Show Data Requests

| Query Name | By | Status | Date |
|---|---|---|---|
| Group comparison for Acute Rheum... | as084 | Queued | 01/24/11 15:23:01 |
| Acute myocardia..., CK-MB Index ... | as084 | Ready | 01/24/11 15:21:01 |
| Patient breakdowns for Acute myo... | as084 | Queued | 01/24/11 15:19:34 |
| Matched set for Acute myocardia.... | as084 | Queued | 01/24/11 15:14:17 |
| EGFR, Respiratory and... on 01/2... | as084 | Ready | 01/24/11 15:04:14 |
| EGFR, Respiratory and... on 01/2... | as084 | Ready | 01/24/11 15:03:00 |
| Acute myocardia..., CK-MB Index ... | as084 | Ready | 01/24/11 14:59:51 |
| AMI and CK-MB .3.5 | as084 | Ready | 01/24/11 14:59:21 |
| Jsut Diagnos AMI | as084 | Ready | 01/24/11 14:55:30 |
| EGFR, Respiratory and... on 01/2... | as084 | Ready | 01/24/11 13:14:18 |
| Acute myocardia..., CK-MB Index ... | as084 | Ready | 01/24/11 13:05:39 |
| Acute myocardia..., CK-MB Index ... | as084 | Ready | 01/24/11 13:04:15 |
| AMI and CK-MB >3.5 | as084 | Ready | 01/24/11 13:03:36 |
| Just diagnose AMI | as084 | Ready | 01/24/11 12:58:17 |
| Resection or de..., Iletin ii po... | snm0 | Ready | 01/24/11 08:00:30 |
| Resection or de..., Insulin glar... | snm0 | Ready | 01/24/11 07:48:22 |
| Resection or debridement of pa..... | snm0 | Ready | 01/24/11 07:45:52 |

Page 1 of 269 (4570 queries)

Create a matched set of control patients from a previous RPDR query

Query Name: Matched set for Acute myocardia..., CK-MB Index

IDENTIFY CASES

Query for which you want a matched set of patients: # ±3

IDENTIFY CONTROLS

I want 3 control patient(s) for each patient found by the above query.

Match by: ☑ Age (10 year intervals) ☑ Gender ☑ Race/Ethnicity ☐ Comparative Health ☐ Use exact matches only

Matched patients will be sampled from all RPDR patients or can only be sampled from patients in the query below:

drag and drop a query here... # ±3

Patients should be included fr

☐ MGH ☑ BWH/FH

Patients can be specifical is specified below (cases are always excluded):

drag and drop a query here... # ±3

for specific purposies, you may wish to exclude these patients:

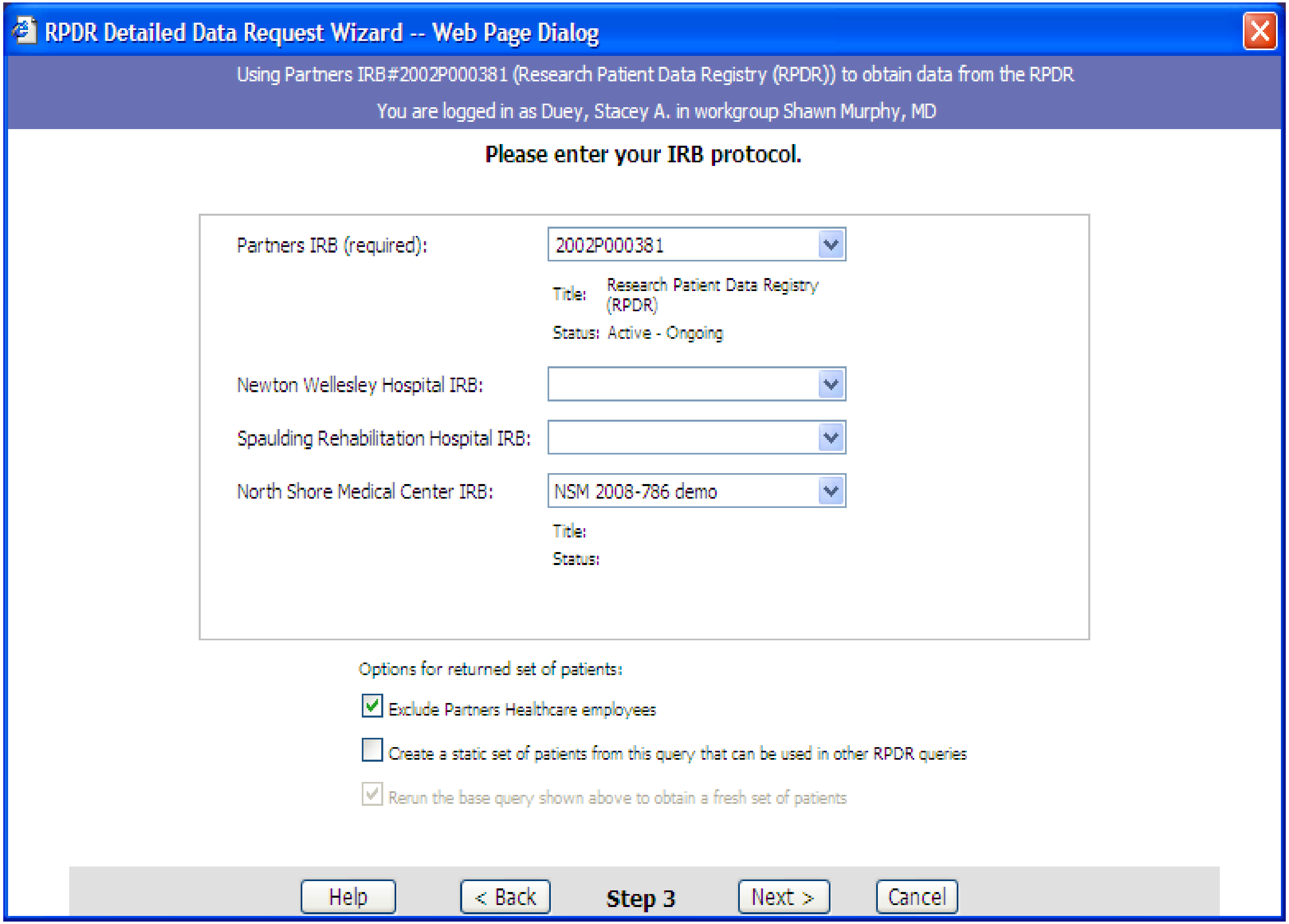
☑ Patients that are no longer living

☑ Patients that have had bone marrow transplants

Total to be found: ±3

AspTreeView/MetaDataTree_Sub.aspx?ParentFolderId=DGNNeoplasms (140-239)0x5CMalignant neoplasms (140-208)0x5CRespiratory and int...

# Obtaining Data Extracts



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

**Please enter your IRB protocol.**

Partners IRB (required): 2002P000381

Title: Research Patient Data Registry (RPDR)

Status: Active - Ongoing

Newton Wellesley Hospital IRB:

Spaulding Rehabilitation Hospital IRB:

North Shore Medical Center IRB: NSM 2008-786 demo

Title:

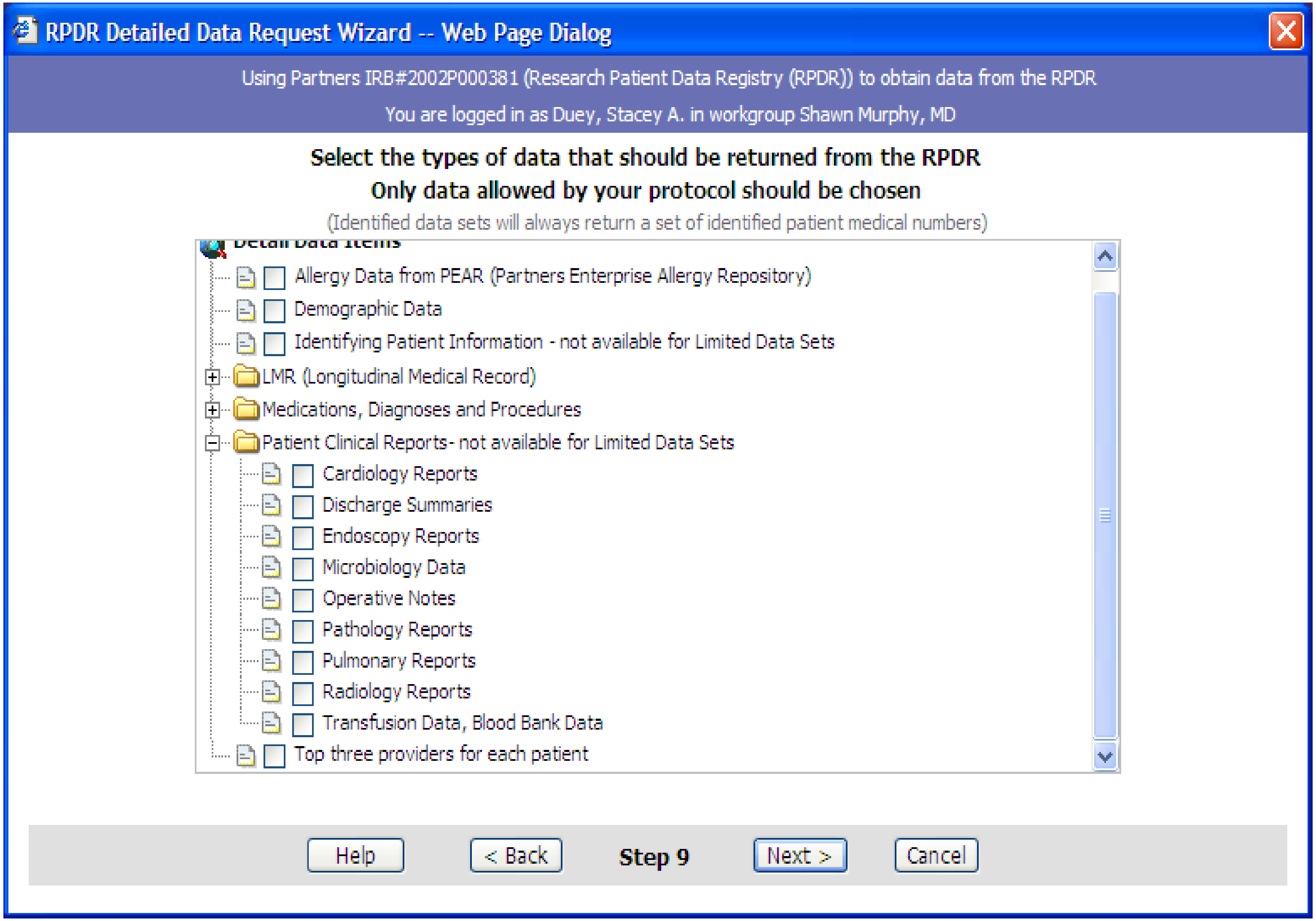
Status:

Options for returned set of patients:

☑ Exclude Partners Healthcare employees

☐ 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 -- 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)

Detail Data Items

- □ 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

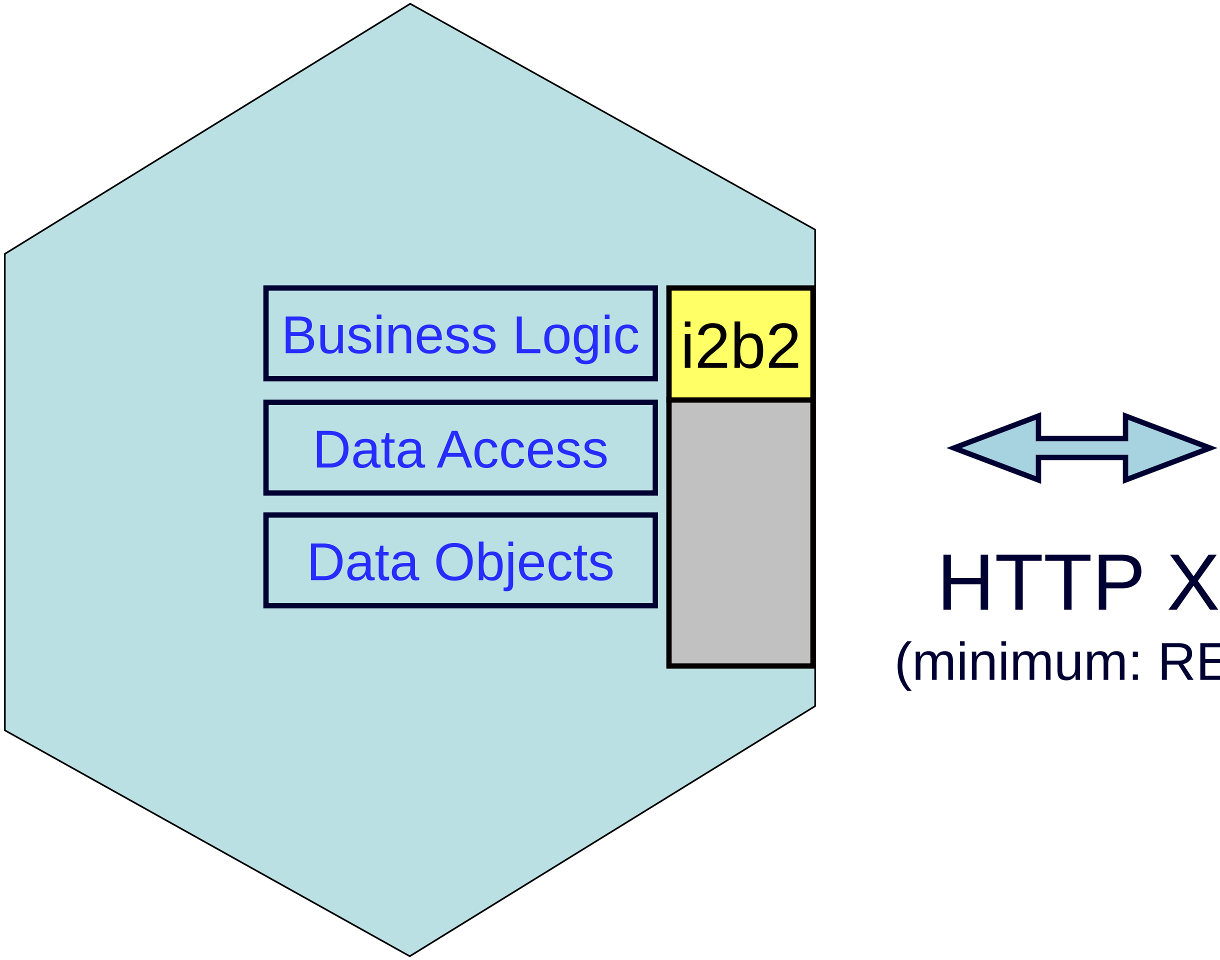
## High Throughput Methods for supporting Translational Research

- Set of patients is selected from medical record data in a high throughput fashion
- Investigators explore phenotypes of these patients using i2b2 tools and a translational team developed to work specifically with medical record data
- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing
- Tissues of these patients can be made available for genomic and biochemical analysis

## The National Center for Biomedical Computing entitled Informatics for Integrating Biology and the Bedside (i2b2), what is it?

- Software for explicitly organizing and transforming person-oriented clinical data to a way that is optimized for clinical genomics research
  - Allows integration of clinical data, trials data, and genotypic data
- A portable and extensible application framework
  - Software is built in a modular pattern that allows additions without disturbing core parts
  - Available as open source at https://www.i2b2.org

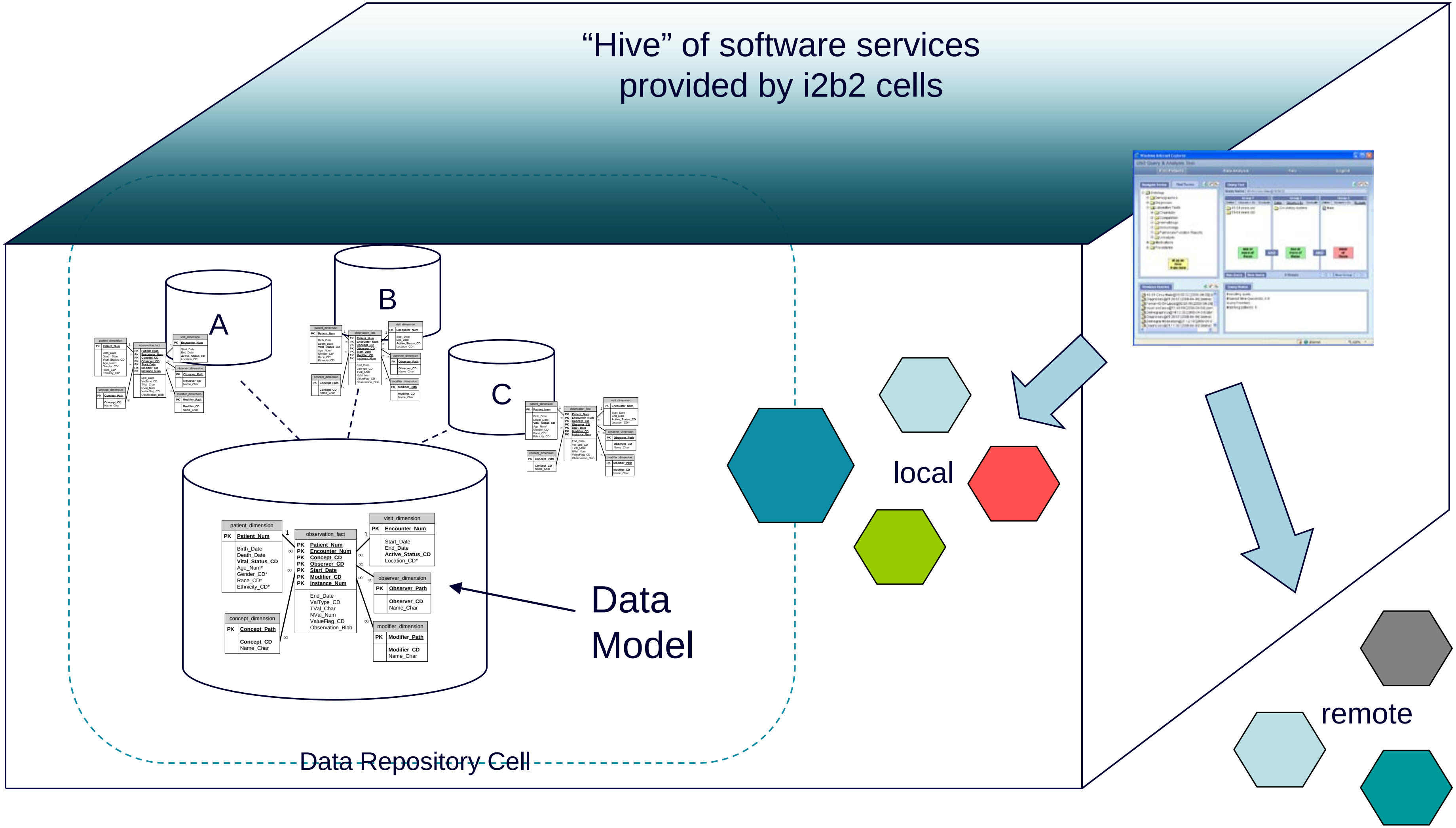
# i2b2 Cell: The Canonical Software Module



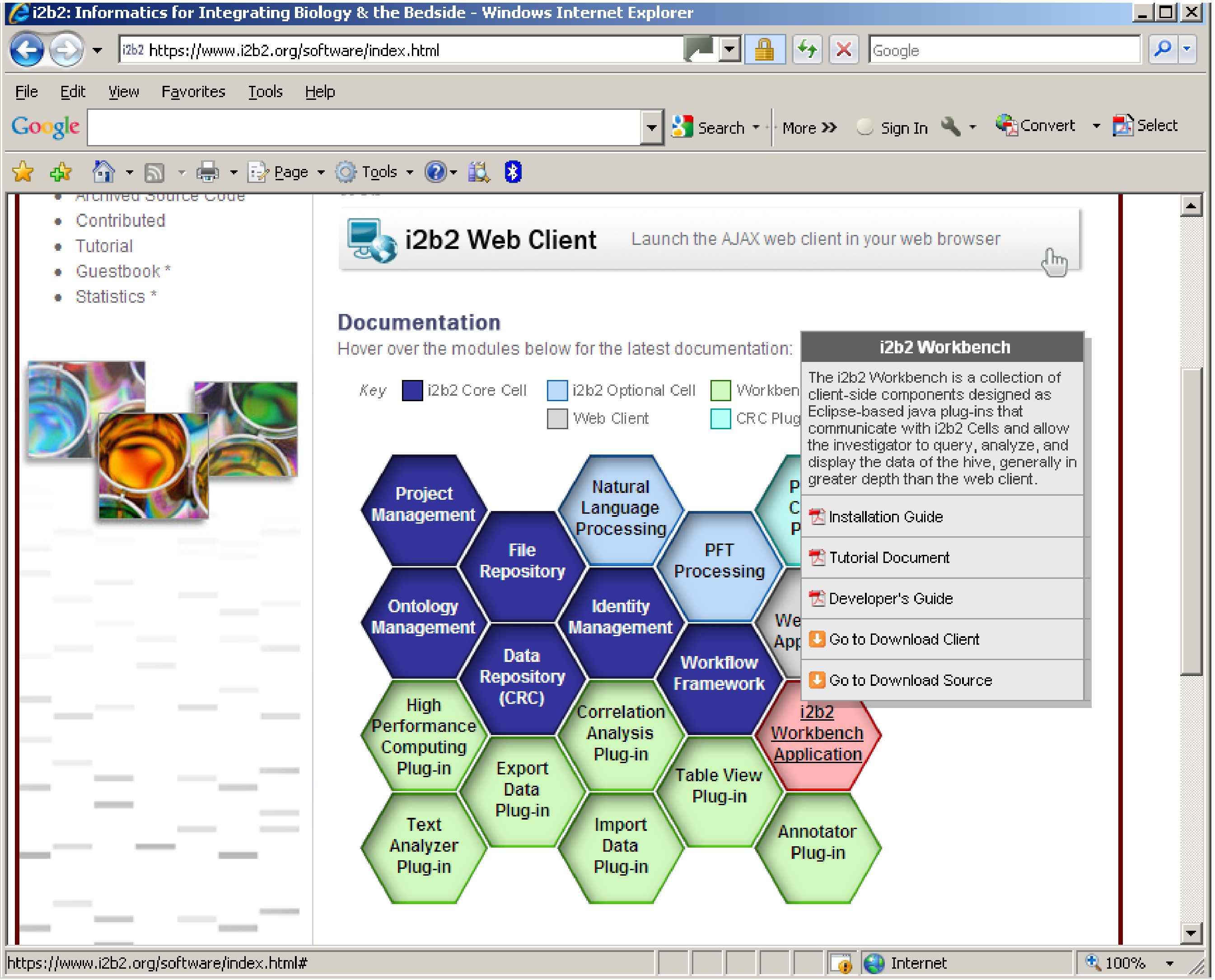
HTTP XML

(minimum: RESTful)

# An i2b2 Environment (the Hive) is built from i2b2 Cells



# I2b2 Software components are distributed as open source

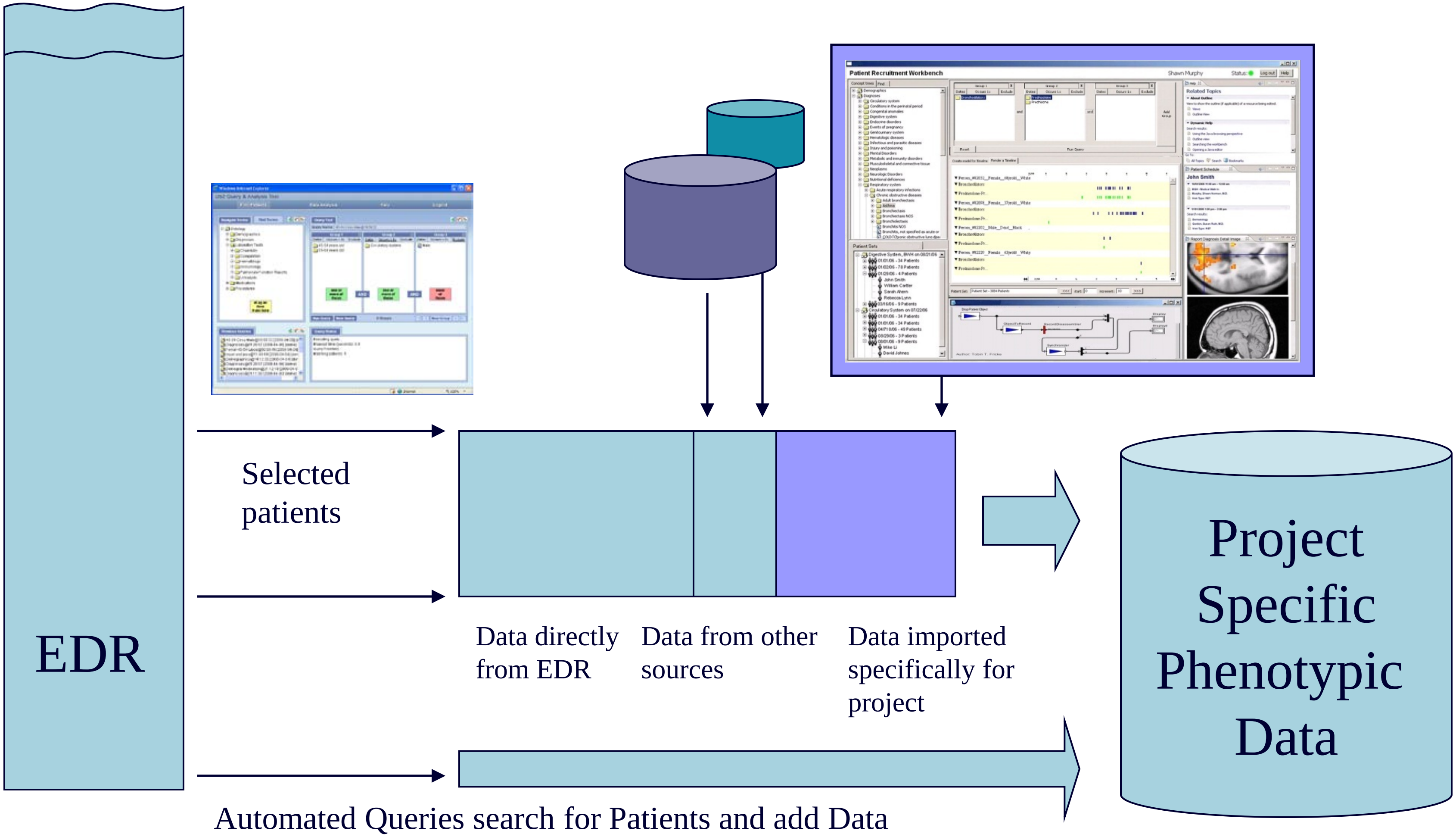


**i2b2 Workbench**

The i2b2 Workbench is a collection of client-side components designed as Eclipse-based java plug-ins that communicate with i2b2 Cells and allow the investigator to query, analyze, and display the data of the hive, generally in greater depth than the web client.

- Installation Guide
- Tutorial Document
- Developer's Guide
- Go to Download Client
- Go to Download Source

# Set of patients is selected through Enterprise Repository and data is gathered into a data mart



EDR

Selected patients

Data directly from EDR

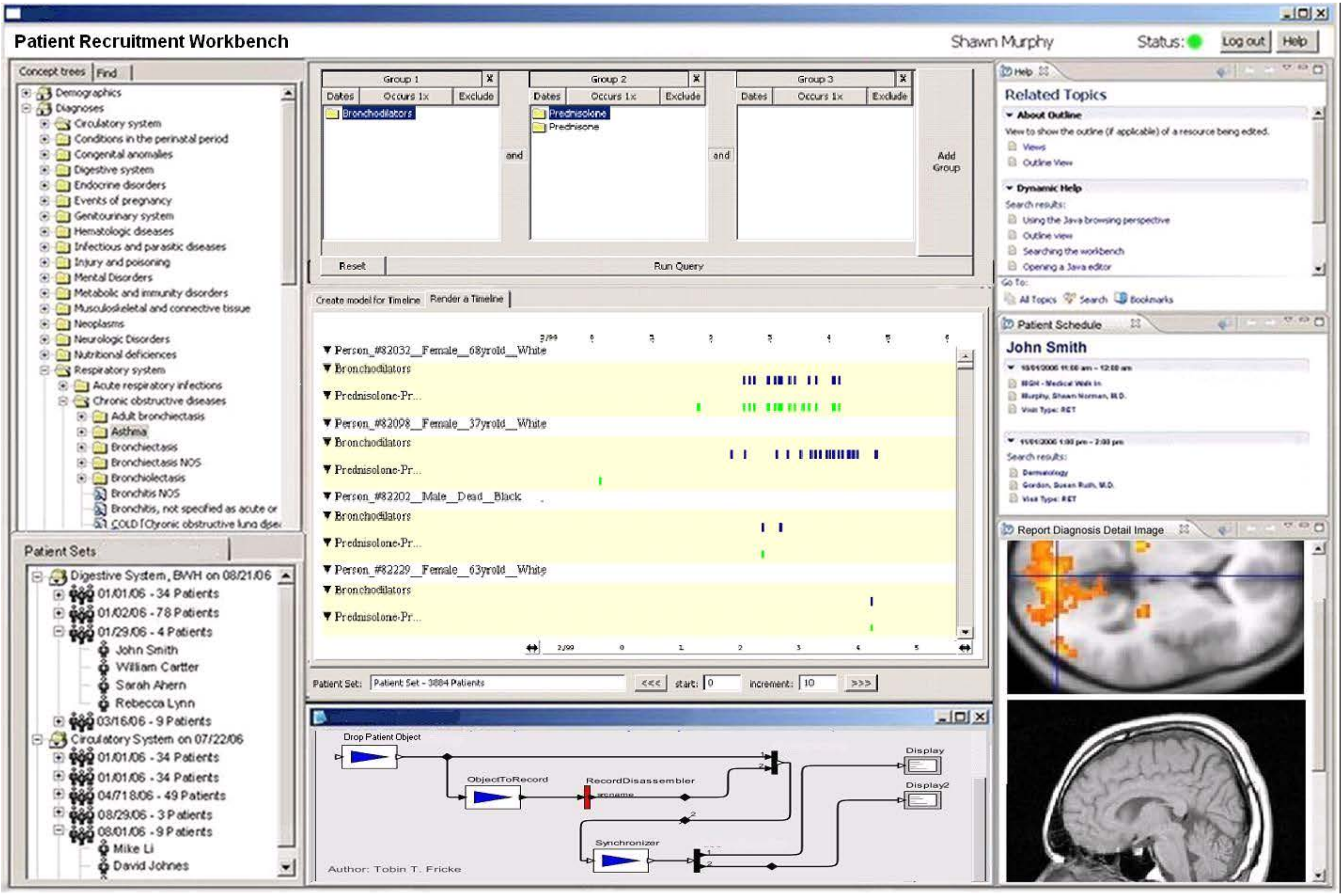
Data from other sources

Data imported specifically for project

Project Specific Phenotypic Data

Automated Queries search for Patients and add Data

# Data is available through the i2b2 Workbench



Patient Recruitment Workbench

Shawn Murphy

Status:

- Demographics
- Diagnoses
  - Circulatory system
  - Conditions in the perinatal period
  - Congenital anomalies
  - Digestive system
  - Endocrine disorders
  - Events of pregnancy
  - Genitourinary system
  - Hematologic diseases
  - Infectious and parasitic diseases
  - Injury and poisoning
  - Mental Disorders
  - Metabolic and immunity disorders
  - Musculoskeletal and connective tissue
  - Neoplasms
  - Neurologic Disorders
  - Nutritional deficiences
  - Respiratory system
    - Acute respiratory infections
    - Chronic obstructive diseases
      - Adult bronchiectasis
      - Asthma
      - Bronchiectasis
      - Bronchiectasis NOS
      - Bronchiolectasis
      - Bronchitis NOS
      - Bronchitis, not specified as acute or
      - COLD [Chronic obstructive lung dise

Patient Sets

- Digestive System, BWH on 08/21/06
  - 01/01/06 - 34 Patients
  - 01/02/06 - 78 Patients
  - 01/29/06 - 4 Patients
    - John Smith
    - William Cartter
    - Sarah Ahern
    - Rebecca Lynn
  - 03/16/06 - 9 Patients
- Circulatory System on 07/22/06
  - 01/01/06 - 34 Patients
  - 01/01/06 - 34 Patients
  - 04/718/06 - 49 Patients
  - 08/29/06 - 3 Patients
  - 08/01/06 - 9 Patients
    - Mike Li
    - David Johnes

| Group 1 | | | | Group 2 | | | | Group 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dates | Occurs 1x | Exclude | | Dates | Occurs 1x | Exclude | | Dates | Occurs 1x | Exclude | |
| Bronchodilators | | | and | Prednisolone<br>Prednisone | | | and | | | | Add Group |

▼ Person_#82032__Female__68yrold__White
▼ Bronchodilators
▼ Prednisolone-Pr...
▼ Person_#82098__Female__37yrold__White
▼ Bronchodilators
▼ Prednisolone-Pr...
▼ Person_#82202__Male__Dead__Black
▼ Bronchodilators
▼ Prednisolone-Pr...
▼ Person_#82229__Female__63yrold__White
▼ Bronchodilators
▼ Prednisolone-Pr...

Patient Set: Patient Set - 3884 Patients

start: 0

increment: 10

Author: Tobin T. Fricke

Related Topics

▾ About Outline

View to show the outline (if applicable) of a resource being edited.

- Views
- Outline View

▾ Dynamic Help

Search results:

- Using the Java browsing perspective
- Outline view
- Searching the workbench
- Opening a Java editor

Go To:

John Smith

▾ 10/01/2006 11:00 am - 12:00 am

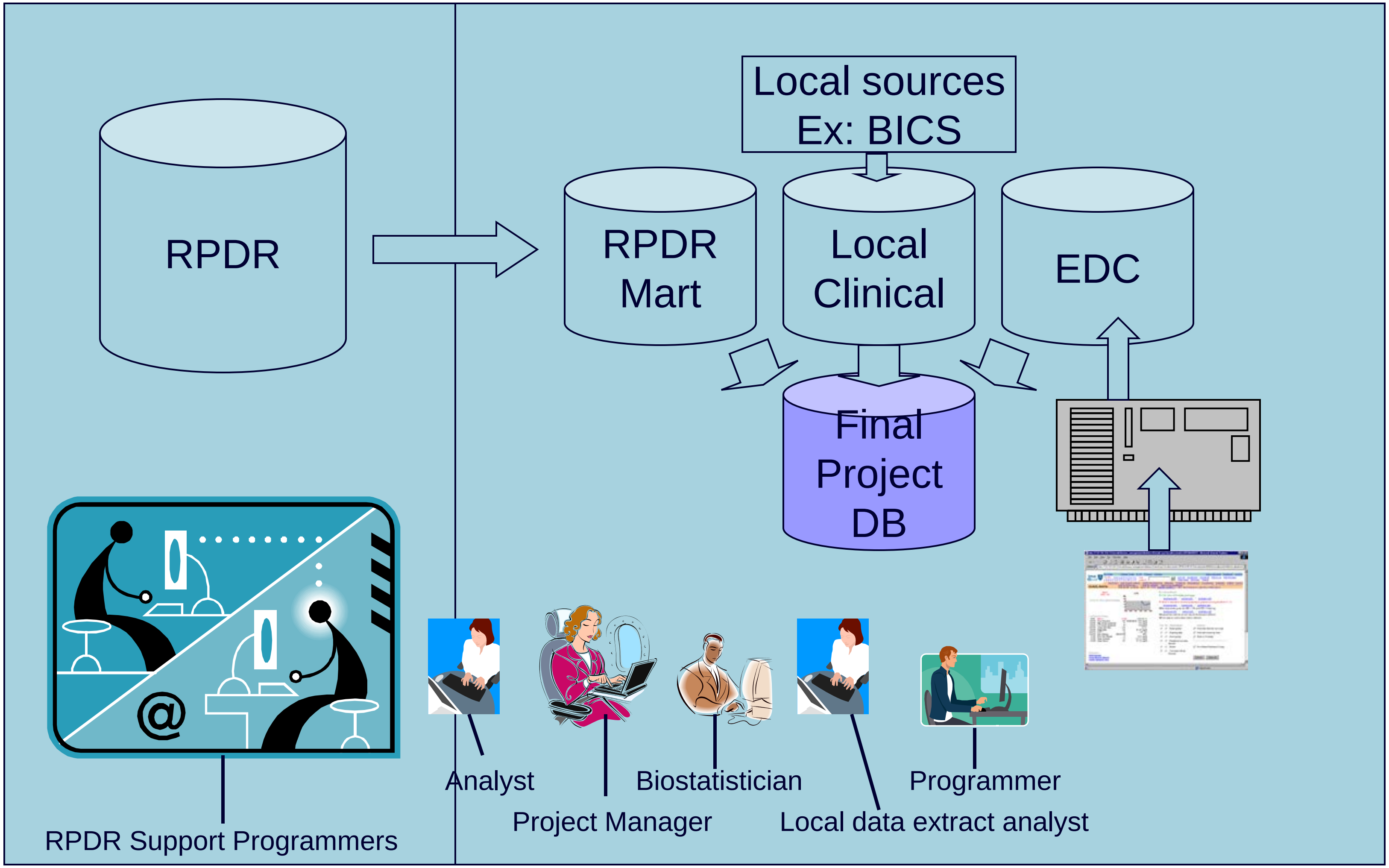
- MGH - Medical Walk In
- Murphy, Shawn Norman, M.D.
- Visit Type: RET

▾ 11/01/2006 1:00 pm - 2:00 pm

Search results:

- Dermatology
- Gordon, Susan Ruth, M.D.
- Visit Type: RET

# Team support for Projects



RPDR

Local sources
Ex: BICS

RPDR
Mart

Local
Clinical

EDC

Final
Project
DB

RPDR Support Programmers

Analyst

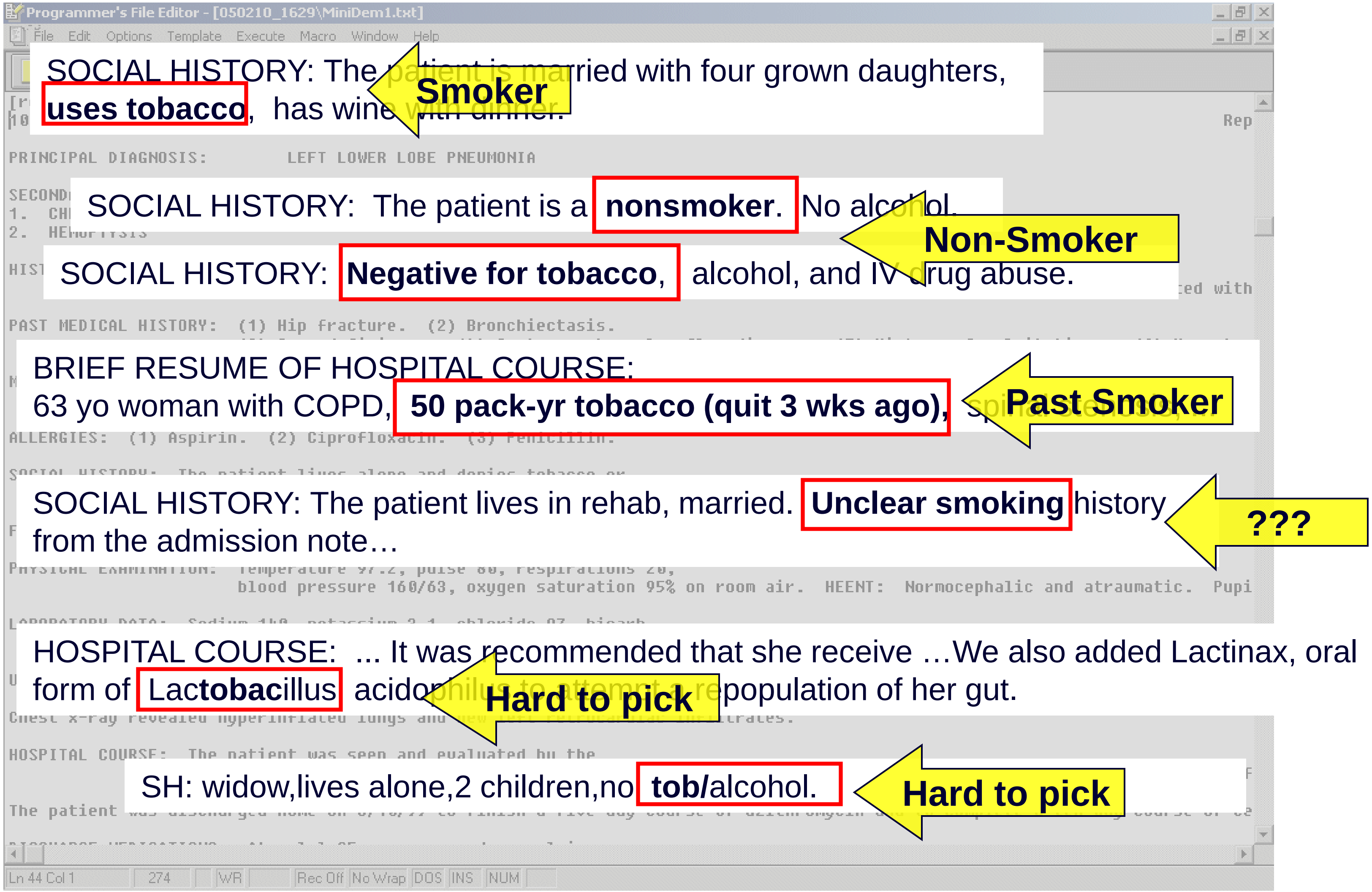
Project Manager

Biostatistician

Local data extract analyst

Programmer

# NLP (and comedy) is not pretty


Programmer's File Editor - [050210_1629\MiniDem1.txt]
File Edit Options Template Execute Macro Window Help
Ln 44 Col 1 274 WR Rec Off No Wrap DOS INS NUM

SOCIAL HISTORY: The patient is married with four grown daughters, **uses tobacco**, has wine with dinner.

**Smoker**

```
PRINCIPAL DIAGNOSIS:        LEFT LOWER LOBE PNEUMONIA
```

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

SOCIAL HISTORY: **Negative for tobacco,** alcohol, and IV drug abuse.

**Non-Smoker**

```
PAST MEDICAL HISTORY:  (1) Hip fracture.  (2) Bronchiectasis.
```

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

**Past Smoker**

```
ALLERGIES:  (1) Aspirin.  (2) Ciprofloxacin.  (3) Penicillin.
```

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

**???**

```
PHYSICAL EXAMINATION:  Temperature 97.2, pulse 80, respirations 20,
                       blood pressure 160/63, oxygen saturation 95% on room air.  HEENT:  Normocephalic and atraumatic.  Pupi
```

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

**Hard to pick**

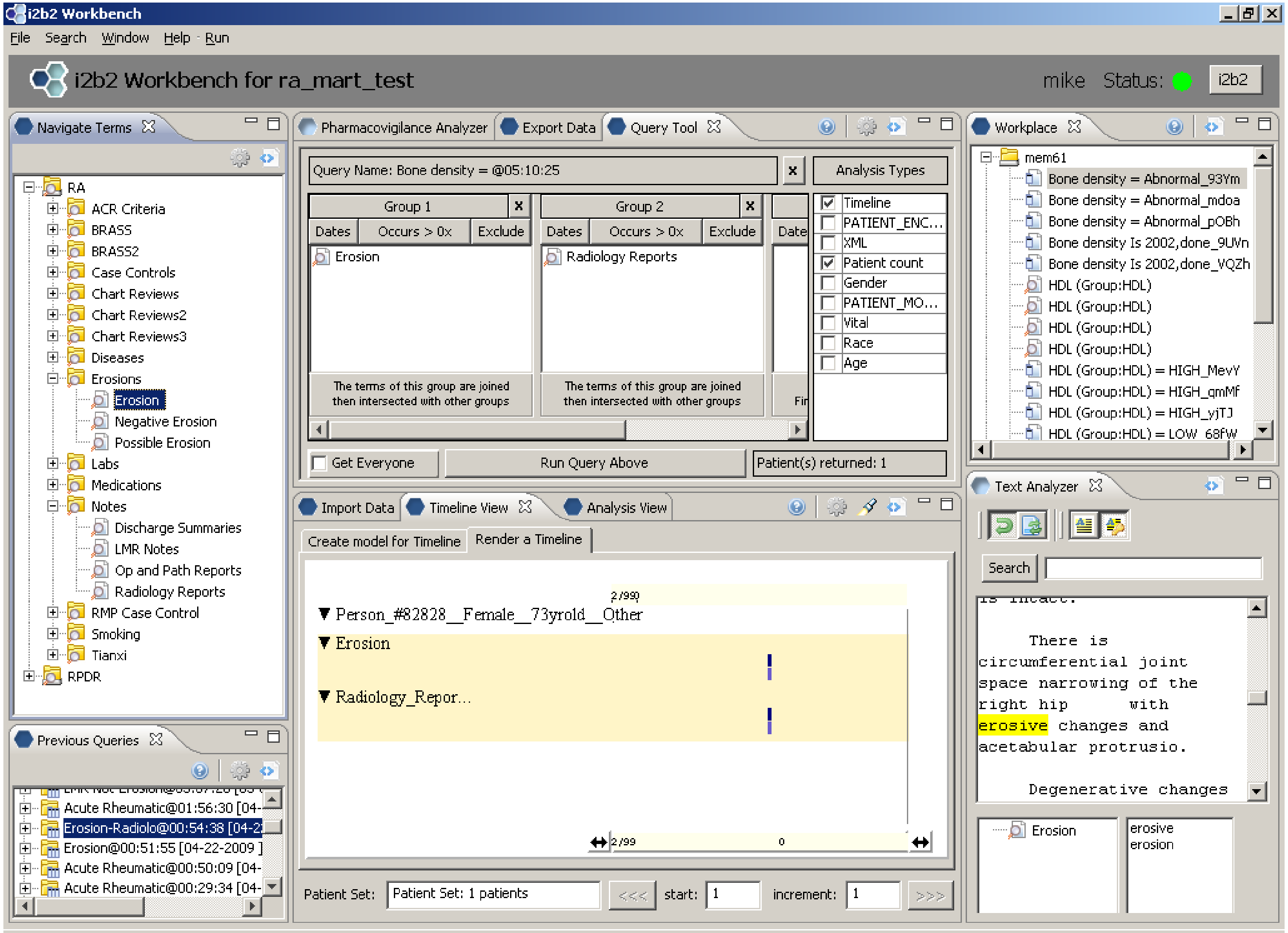
```
Chest x-ray revealed hyperinflated lungs and new left retrocardiac infiltrates.

HOSPITAL COURSE:  The patient was seen and evaluated by the
```

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

**Hard to pick**

# Investigator Review



Query Name: Bone density = @05:10:25

☑ Timeline
☐ PATIENT_ENC...
☐ XML
☑ Patient count
☐ Gender
☐ PATIENT_MO...
☐ Vital
☐ Race
☐ Age

The terms of this group are joined then intersected with other groups

The terms of this group are joined then intersected with other groups

Patient(s) returned: 1

Patient Set: Patient Set: 1 patients <<< start: 1 increment: 1 >>>

```
is intact.

    There is
circumferential joint
space narrowing of the
right hip      with
erosive changes and
acetabular protrusio.

    Degenerative changes
```

# *Can We Trust the Phenotypes?*

## Validation Study (N = 185)

- Evaluate case and control algorithms compared to gold standard of diagnostic interview by expert clinician
- Recruit cases and controls as defined by informatics algorithm
- Interview by clinicians blinded to ascertainment group
- Recruited patients with depression or schizophrenia to enhance blinding

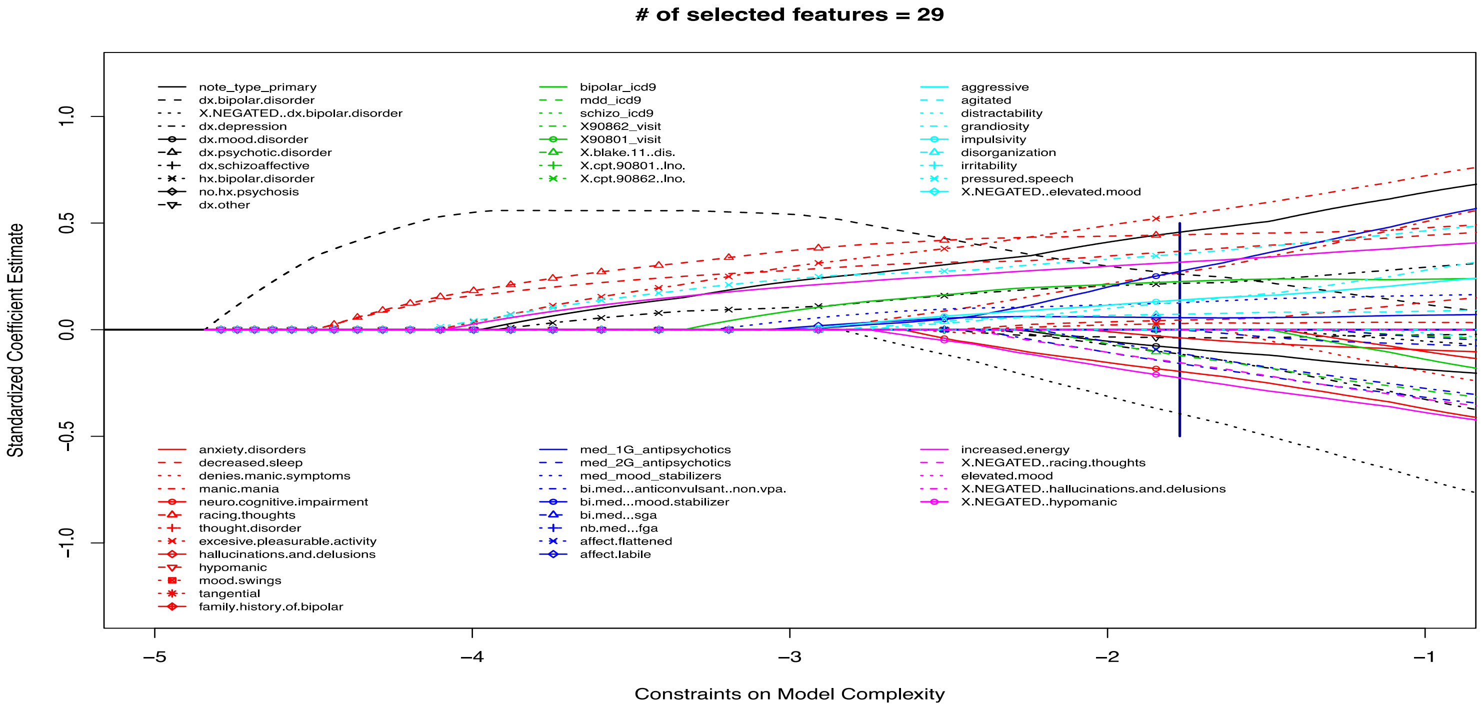
Jordan Smoller MD, ScD and team

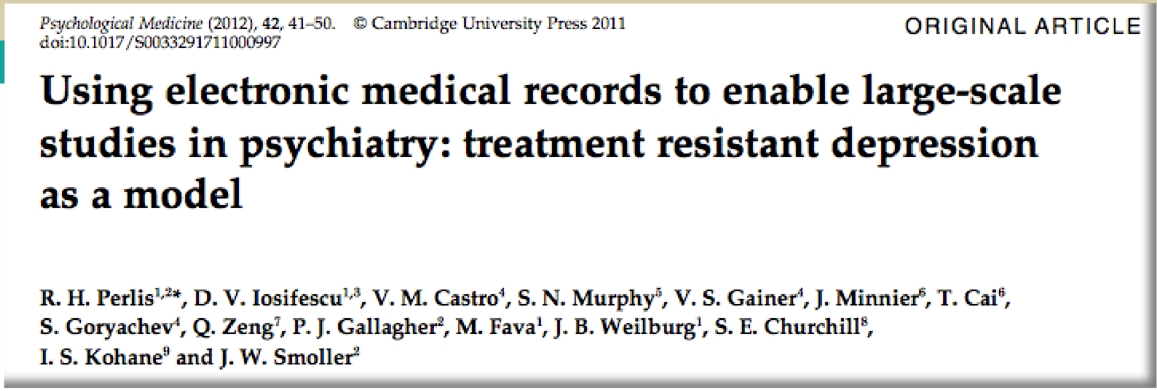
# Train classification algorithms

1. Over 300 words/phrases (features) were identified using chart review

2. Important features were selected for model using adaptive LASSO shrinkage

Tianxi Cai PhD and team





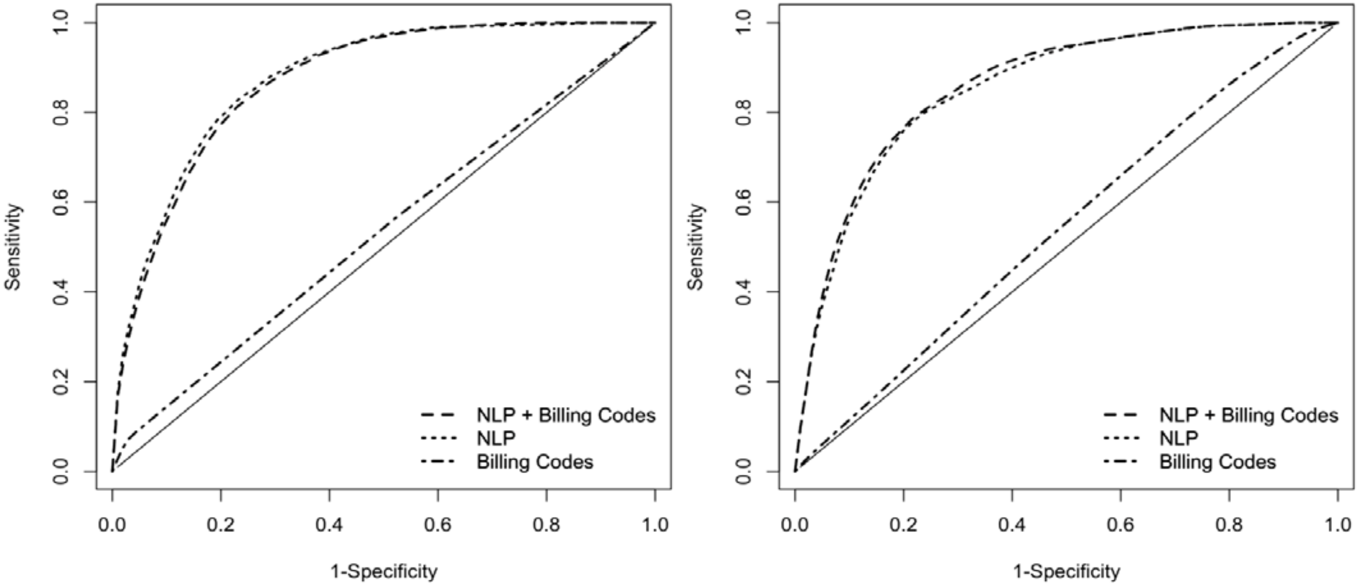
*Psychological Medicine* (2012), **42**, 41–50. © Cambridge University Press 2011
doi:10.1017/S0033291711000997

ORIGINAL ARTICLE

# Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model

R. H. Perlis[1,2]*, D. V. Iosifescu[1,3], V. M. Castro[4], S. N. Murphy[5], V. S. Gainer[4], J. Minnier[6], T. Cai[6], S. Goryachev[4], Q. Zeng[7], P. J. Gallagher[2], M. Fava[1], J. B. Weilburg[1], S. E. Churchill[8], I. S. Kohane[9] and J. W. Smoller[2]

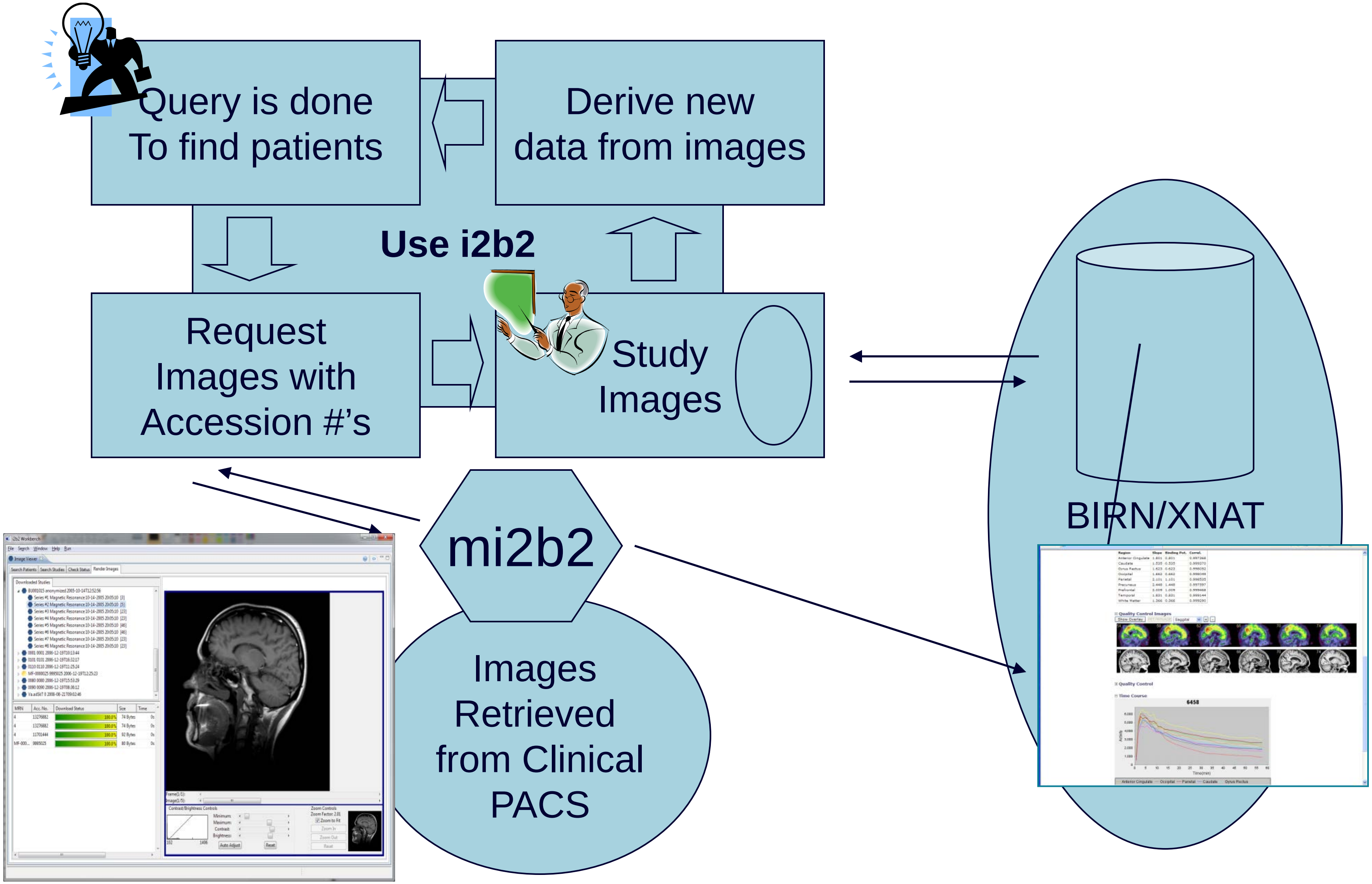
## Use NLP to define cohorts of treatment-resistant and treatment-responsive depression



Specificity: 95%
AUC > 85%

| Clinical Status | Model | Specificity | Sensitivity | Precision | AUC |
|---|---|---|---|---|---|
| Depressed | Billing Codes | 0.95 | 0.09 (0.03) | 0.57 (0.14) | 0.54 (0.02) |
| Depressed | NLP | 0.95 | 0.42 (0.05) | 0.78 (0.02) | 0.88 (0.02) |
| Depressed | NLP + Billing Codes | 0.95 | 0.39 (0.06) | 0.78 (0.02) | 0.87 (0.02) |
| | | | | | |
| Well | Billing Codes | 0.95 | 0.06 (0.02) | 0.26 (0.27) | 0.55 (0.03) |
| Well | NLP | 0.95 | 0.37 (0.06) | 0.86 (0.02) | 0.85 (0.02) |
| Well | NLP + Billing Codes | 0.95 | 0.39 (0.07) | 0.85 (0.02) | 0.86 (0.02) |

# Research Investigator Workflow enabled by mi2b2



Query is done To find patients

Derive new data from images

Use i2b2

Request Images with Accession #'s

Study Images

BIRN/XNAT

mi2b2

Images Retrieved from Clinical PACS

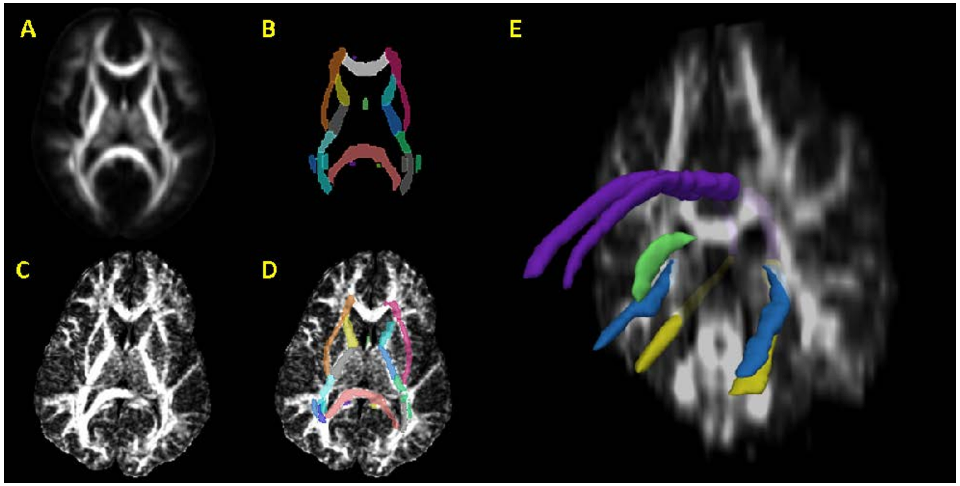
- BU001015 anonymized 2005-10-14T12:52:56
  - Series #1 Magnetic Resonance 10-14-2005 20:05:10 [3]
  - Series #2 Magnetic Resonance 10-14-2005 20:05:10 [5]
  - Series #3 Magnetic Resonance 10-14-2005 20:05:10 [23]
  - Series #4 Magnetic Resonance 10-14-2005 20:05:10 [23]
  - Series #5 Magnetic Resonance 10-14-2005 20:05:10 [46]
  - Series #6 Magnetic Resonance 10-14-2005 20:05:10 [46]
  - Series #7 Magnetic Resonance 10-14-2005 20:05:10 [23]
  - Series #8 Magnetic Resonance 10-14-2005 20:05:10 [23]
- 0001 0001 2006-12-19T10:13:44
- 0101 0101 2006-12-19T16:32:17
- 0110 0110 2006-12-19T11:25:24
- MF-0000025 9995025 2006-12-19T12:25:23
- 0080 0080 2006-12-19T15:53:29
- 0090 0090 2006-12-19T08:36:12
- Va.adSkT 0 2008-08-21T09:02:46

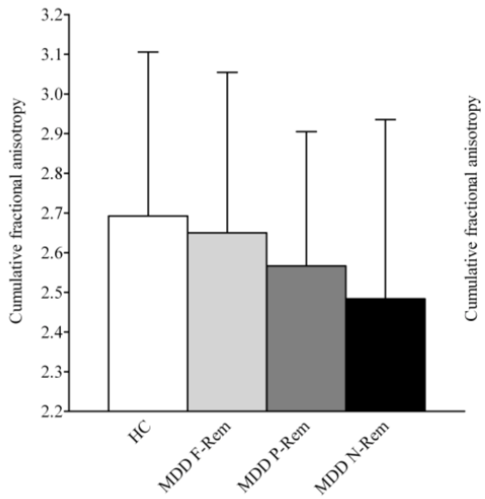
| MRN | Acc. No. | Download Status | Size | Time |
|---|---|---|---|---|
| 4 | 13276882 | 100.0% | 74 Bytes | 0s |
| 4 | 13276882 | 100.0% | 74 Bytes | 0s |
| 4 | 11701444 | 100.0% | 92 Bytes | 0s |
| MF-000... | 9995025 | 100.0% | 80 Bytes | 0s |

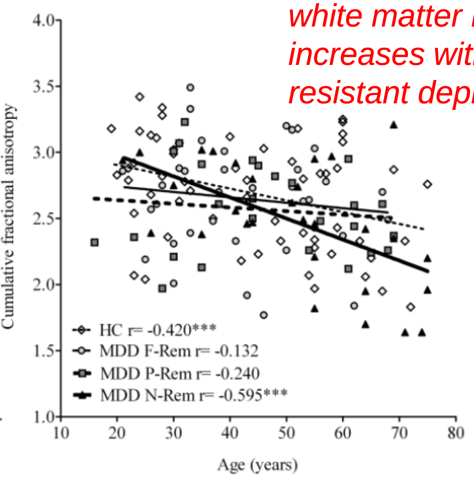
| Region | Slope | Binding Pot. | Correl. |
|---|---|---|---|
| Anterior Cingulate | 1.801 | 0.801 | 0.997268 |
| Caudate | 1.535 | 0.535 | 0.999370 |
| Gyrus Rectus | 1.623 | 0.623 | 0.996092 |
| Occipital | 1.662 | 0.662 | 0.998049 |
| Parietal | 2.101 | 1.101 | 0.998535 |
| Precuneus | 2.448 | 1.448 | 0.997397 |
| Prefrontal | 2.009 | 1.009 | 0.999468 |
| Temporal | 1.831 | 0.831 | 0.999144 |
| White Matter | 1.366 | 0.366 | 0.999290 |

# *White matter abnormalities associated with treatment-resistant depression*

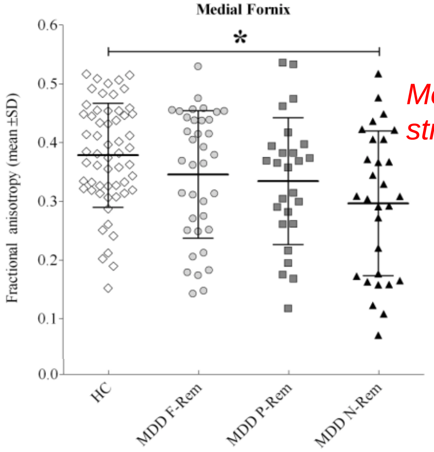
- Scans collected as part of routine clinical care
- NLP identified cohort with treatment outcomes and lack of diagnosed brain pathology on MRI
- Diffusion tensor imaging in 150 pts with best data







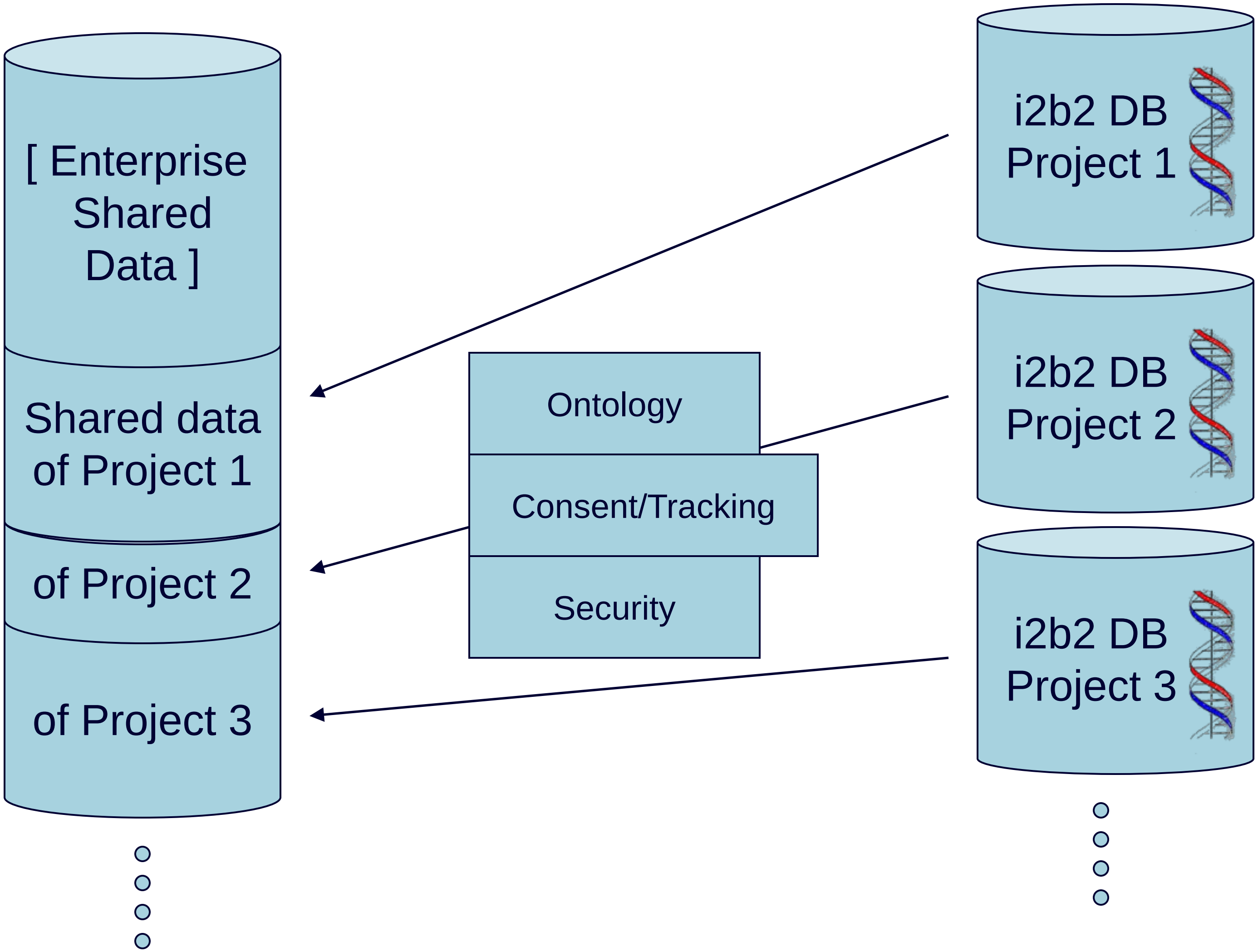
*Age-related decline in white matter integrity increases with treatment resistant depression*



*Medial fornix shows strongest effect*

Hoogenboom et al. World J Biol Psychiatry, 2012

# Ontology-driven data organization allows simplistic data models that paste together



## High Throughput Methods for supporting Translational Research

- Set of patients is selected from medical record data in a high throughput fashion
- Investigators explore phenotypes of these patients using i2b2 tools and a translational team developed to work specifically with medical record data
- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing
- Tissues of these patients can be made available for genomic and biochemical analysis

# i2b2 Implementations

CTSA's

- Boston University
- Case Western Reserve University (*including Cleveland Clinic*)
- Children's National Medical Center (GWU), Washington D.C.
- Duke University
- Emory University (*including Morehouse School of Medicine and Georgia Tech* )
- Harvard University (*including Beth Israel Deaconness Medical Center, Brigham and Women's Hospital, Children's Hospital Boston, Dana Farber Cancer Center, Joslin Diabetes Center, Massachusetts General Hospital*)
- Medical University of South Carolina
- Medical College of Wisconsin
- Oregon Health & Science University
- Penn State MIlton S. Hershey Medical Center
- Tufts University
- University of Alabama at Birmingham
- University of Arkansas for Medical Sciences
- University of California Davis
- University of California, Irvine
- University of California, Los Angeles*
- University of California, San Diego*
- University of California San Francisco
- University of Chicago
- University of Cincinnati (including *Cinncinati Children's Hospital Medical Center*)
- University of Colorado Denver (including *Children's Hospital Colorado*)
- University of Florida
- University of Kansas Medical Center
- University of Kentucky Research Foundation
- University of Massachusetts Medical School, Worcester
- University of Michigan
- University of Pennsylvania (including *Children's Hospital of Philadelphia*)
- University of Pittsburgh (including their *Cancer Institute*)
- University of Rochester School of Medicine and Dentistry
- University of Texas Health Sciences Center at Houston
- University of Texas Health Sciences Center at San Antonio
- University of Texas Medical Branch (Galveston)
- University of Texas Southwestern Medical Center at Dallas
- University of Utah
- University of Washington
- University of Wisconsin - Madison (*including Marshfield Clinic*)
- Virginia Commonwealth University
- Weill Cornell Medical College

Academic Health Centers (does not include AHCs that are part of a CTSA):

- Arizona State University
- City of Hope, Los Angeles
- Georgia Health Sciences University, Augusta
- Hartford Hospital, CN
- HealthShare Montana
- Massachusetts Veterans Epidemiology Research and Information Center (MAVERICK), Boston
- Nemours
- Phoenix Children's Hospital
- Regenstrief Institute
- Thomas Jefferson University
- University of Connecticut Health Center
- University of Missouri School of Medicine
- University of Tennessee Health Sciences Center
- Wake Forest University Baptist Medical Center

HMOs:

- Group Health Cooperative
- Kaiser Permanente

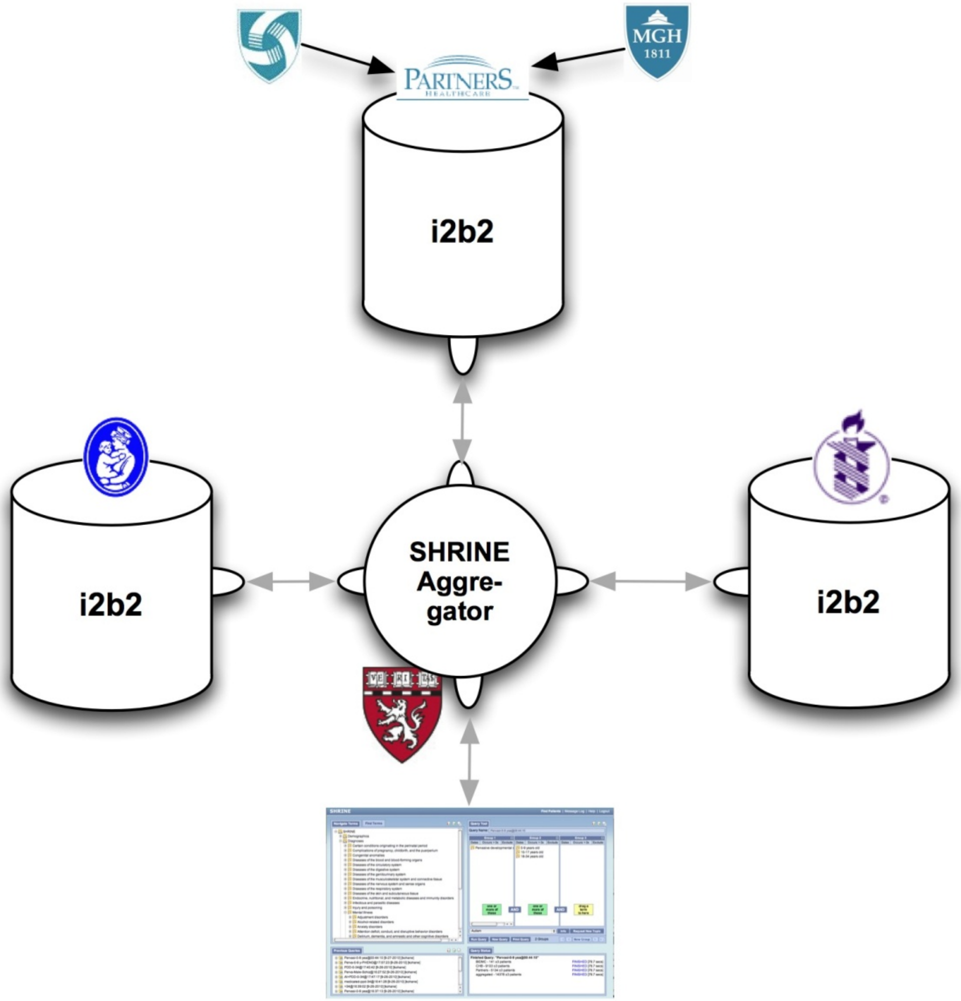
*International:*

- Georges Pompidou Hospital, Paris, France
- Hospital of the Free University of Brussels, Belgium
- Inserm U936, Rennes, France
- Institute for Data Technology and Informatics (IDI), NTNU, Norway
- Institute for Molecular Medicine Finland (FIMM)
- Karolinska Institute, Sweden
- Landspitali University Hospital, Reykjavik, Iceland
- Tokyo Medical and Dental University, Japan
- University of Bordeau Segalen, France
- University of Erlangen-Nuremberg, Germany
- University of Goettingen, Goettingen, Germany
- University of Leicester and Hospitals, England (Biomed. Res. Informatics Ctr. for Clin. Sci)
- University of Pavia, Pavia, Italy
- University of Seoul, Seoul, Korea

Companies:

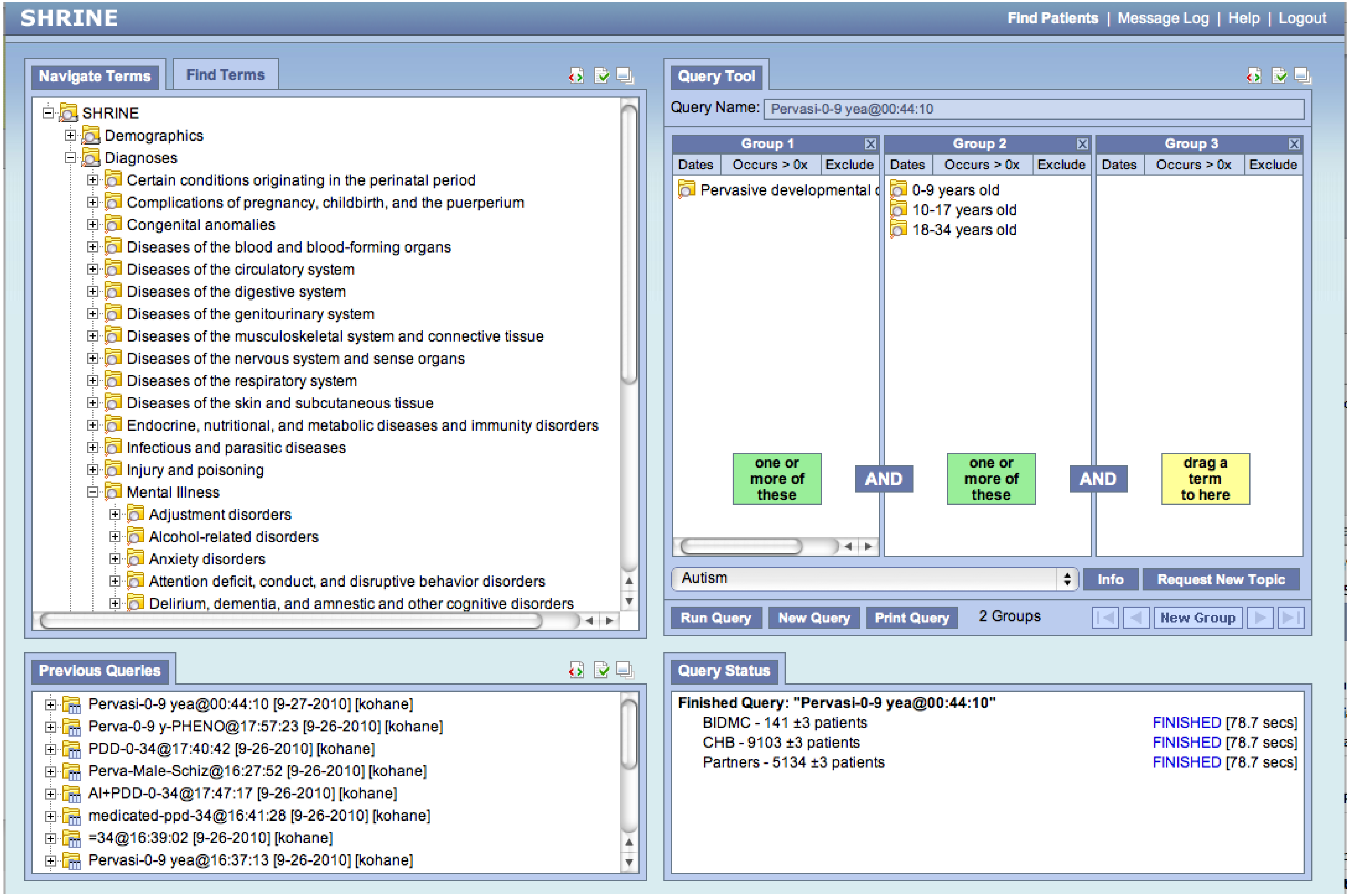
- Johnson and Johnson (TransMART)
- GE Healthcare Clinical Data Services

# 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)



SHRINE
- Demographics
- Diagnoses
  - Certain conditions originating in the perinatal period
  - Complications of pregnancy, childbirth, and the puerperium
  - Congenital anomalies
  - Diseases of the blood and blood-forming organs
  - Diseases of the circulatory system
  - Diseases of the digestive system
  - Diseases of the genitourinary system
  - Diseases of the musculoskeletal system and connective tissue
  - Diseases of the nervous system and sense organs
  - Diseases of the respiratory system
  - Diseases of the skin and subcutaneous tissue
  - Endocrine, nutritional, and metabolic diseases and immunity disorders
  - Infectious and parasitic diseases
  - Injury and poisoning
  - Mental Illness
    - Adjustment disorders
    - Alcohol-related disorders
    - Anxiety disorders
    - Attention deficit, conduct, and disruptive behavior disorders
    - Delirium, dementia, and amnestic and other cognitive disorders

Query Name: Pervasi-0-9 yea@00:44:10

Group 1: Pervasive developmental d

Group 2: 0-9 years old; 10-17 years old; 18-34 years old

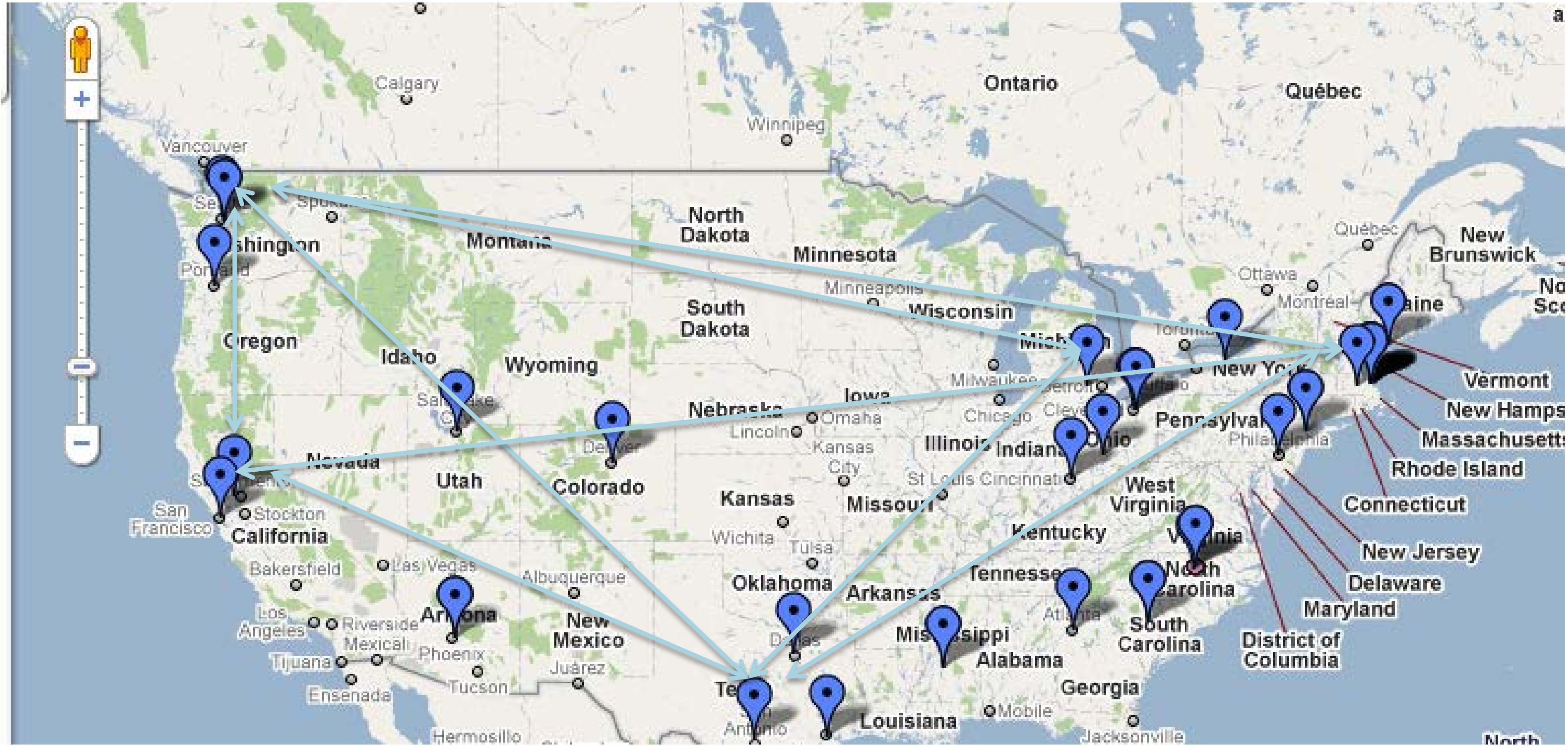
Autism

2 Groups

Previous Queries:
- Pervasi-0-9 yea@00:44:10 [9-27-2010] [kohane]
- Perva-0-9 y-PHENO@17:57:23 [9-26-2010] [kohane]
- PDD-0-34@17:40:42 [9-26-2010] [kohane]
- Perva-Male-Schiz@16:27:52 [9-26-2010] [kohane]
- AI+PDD-0-34@17:47:17 [9-26-2010] [kohane]
- medicated-ppd-34@16:41:28 [9-26-2010] [kohane]
- =34@16:39:02 [9-26-2010] [kohane]
- Pervasi-0-9 yea@16:37:13 [9-26-2010] [kohane]

```
Finished Query: "Pervasi-0-9 yea@00:44:10"
    BIDMC - 141 ±3 patients                    FINISHED [78.7 secs]
    CHB - 9103 ±3 patients                     FINISHED [78.7 secs]
    Partners - 5134 ±3 patients                FINISHED [78.7 secs]
```

# 2012


Calgary
Ontario
Québec
Winnipeg
Vancouver
Spokane
Washington
Portland
Montana
North Dakota
Minnesota
South Dakota
Minneapolis
Wisconsin
Michigan
Toronto
New York
Ottawa
Montreal
Québec
New Brunswick
Maine
Nova Scotia
Oregon
Idaho
Wyoming
Salt Lake City
Nebraska
Iowa
Omaha
Lincoln
Milwaukee
Detroit
Buffalo
Chicago
Cleveland
Pennsylvania
Philadelphia
Illinois
Indiana
Ohio
Nevada
Utah
Colorado
Denver
Kansas
Kansas City
St Louis
Cincinnati
Missouri
West Virginia
Kentucky
Virginia
Vermont
New Hampshire
Massachusetts
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
San Francisco
Stockton
California
Bakersfield
Las Vegas
Los Angeles
Riverside
Mexicali
Tijuana
Ensenada
Arizona
Phoenix
Tucson
Albuquerque
New Mexico
Juarez
Wichita
Tulsa
Oklahoma
Arkansas
Tennessee
North Carolina
South Carolina
Atlanta
Mississippi
Alabama
Georgia
Dallas
Texas
San Antonio
Louisiana
Mobile
Jacksonville
Hermosillo

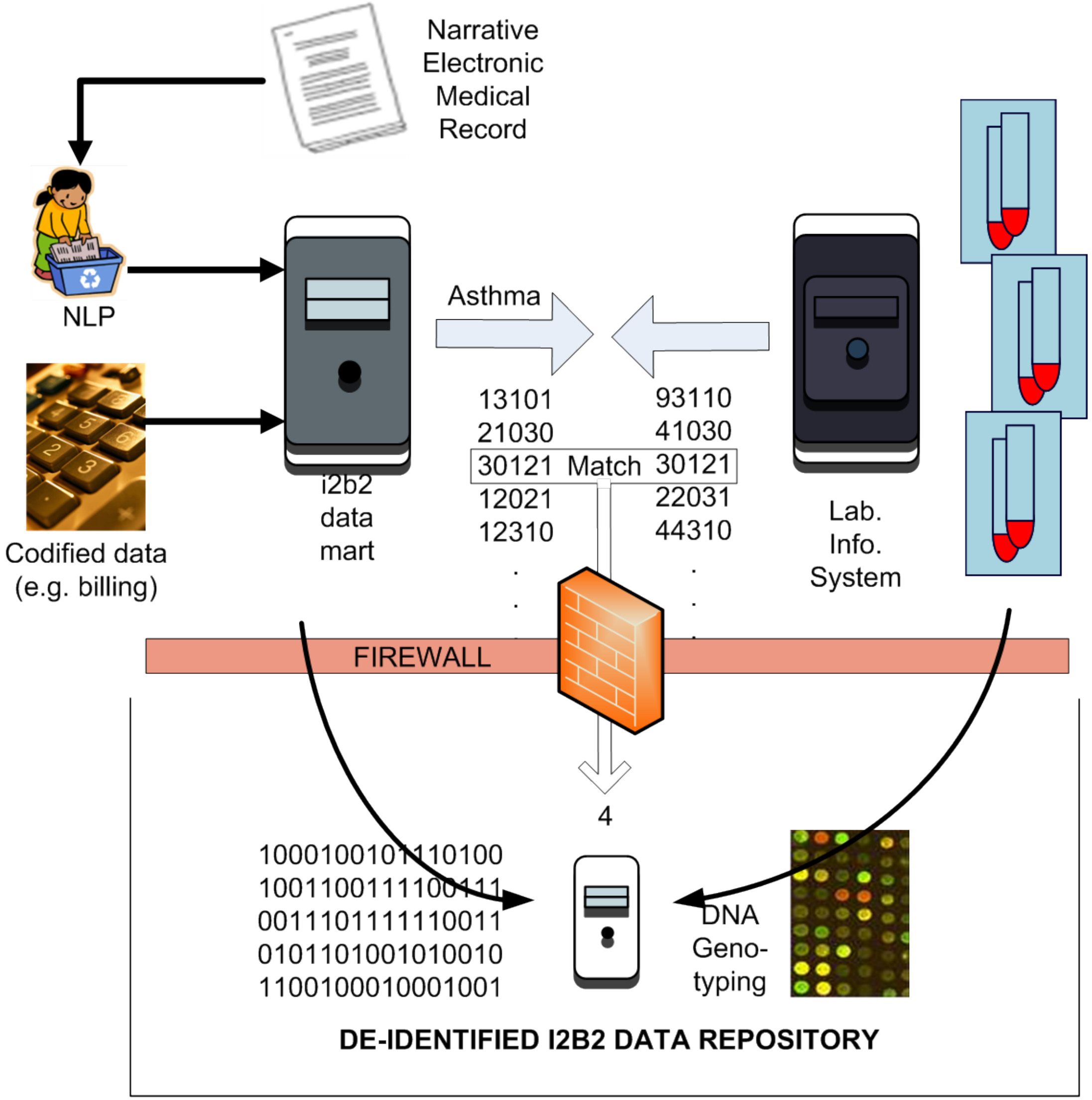
## Performing Clinical trials “in-silico”

- 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.

# High Throughput Methods for supporting Translational Research

- Set of patients is selected from medical record data in a high throughput fashion
- Investigators explore phenotypes of these patients using i2b2 tools and a translational team developed to work specifically with medical record data
- Distributed networks cross institutional boundaries for phenotype selection, public health, and hypothesis testing
- Tissues of these patients can be made available for genomic and biochemical analysis

# Genotype samples and compare to controls



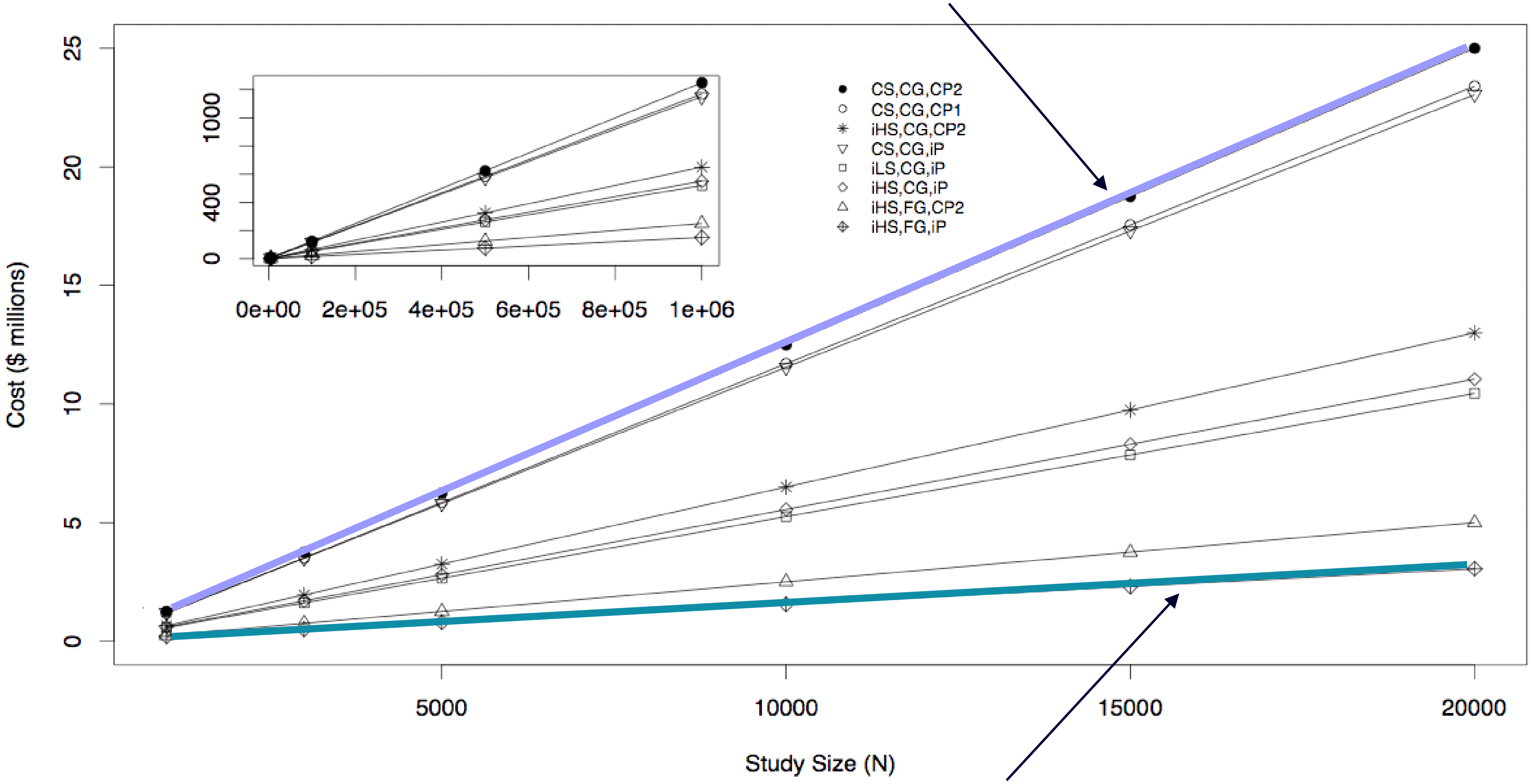
# Cost and time benefit of Instrumenting with Sample Collection for Modest-size Study with 10,000 subjects (cases + controls)

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

= $6.7 million/study vs. $250 thousand/study

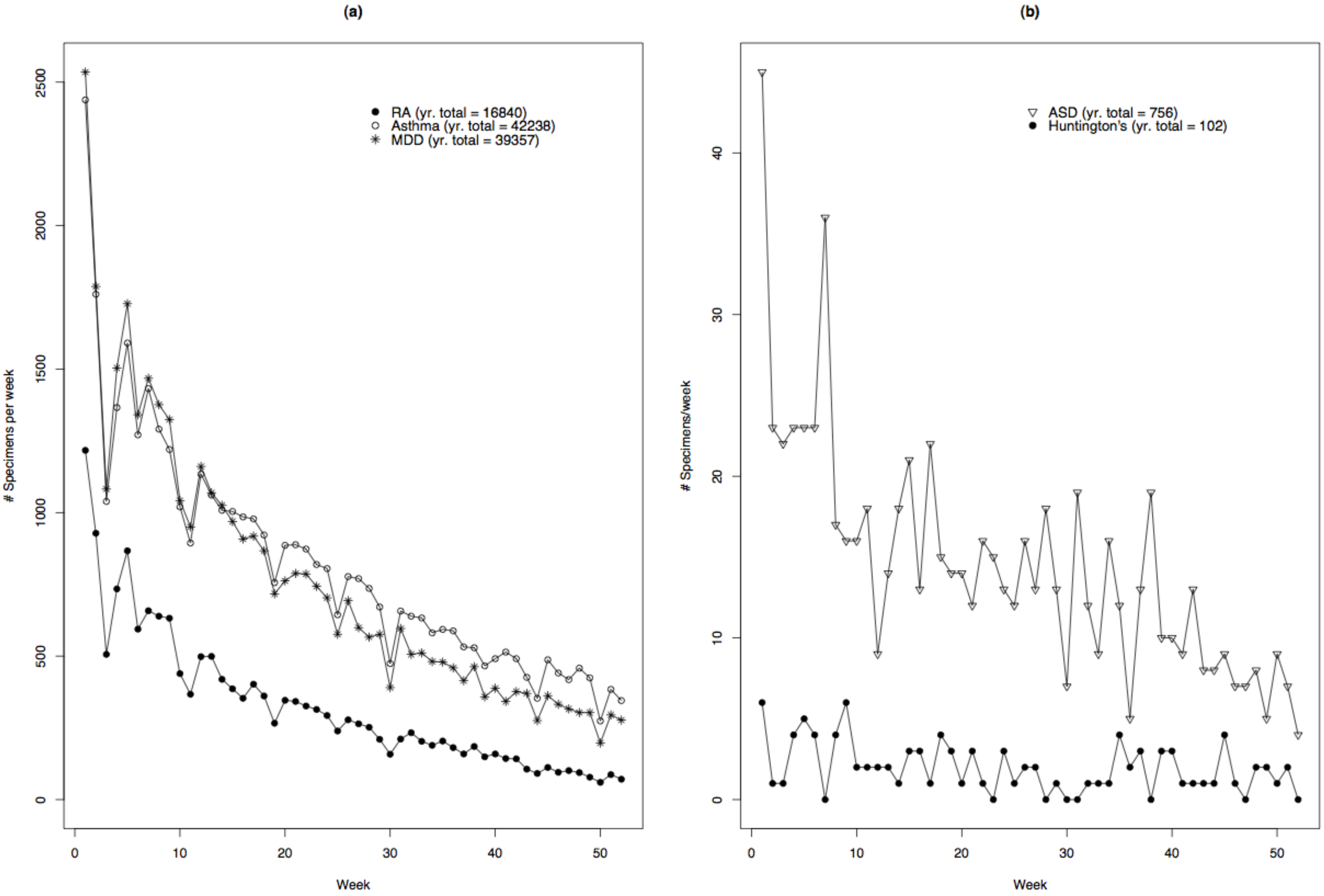
# Escalating cost and time benefit of Instrumenting with Sample Collection

Previous model for collecting specimens

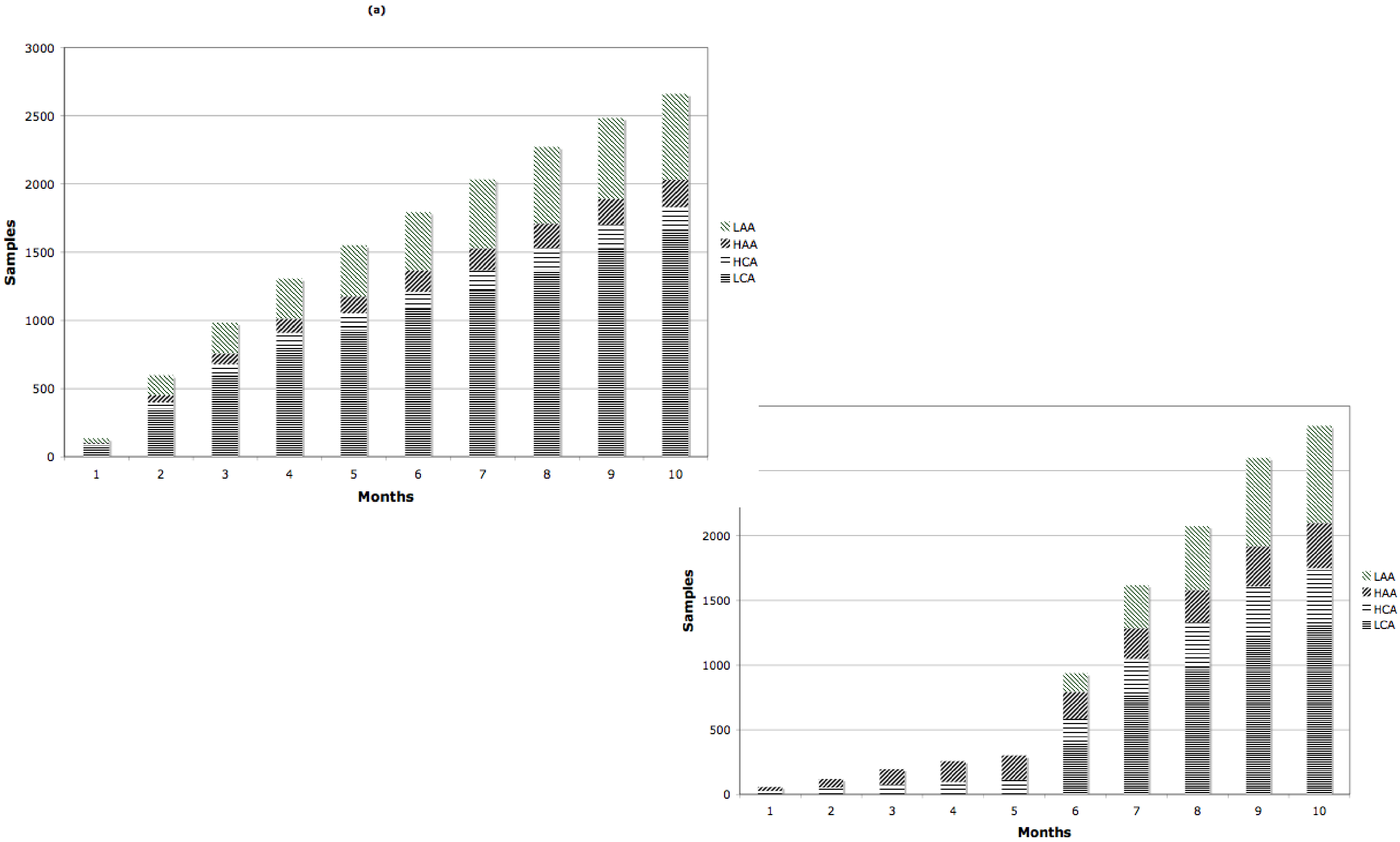


New model for collecting specimens

# Accrual Rates



# Meeting Expectations



# 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
  - Michael Mendis
  - Nich Wattanasin
  - Vivian Gainer
  - Lori Phillips
  - Wensong Pan
  - Janice Donahue
  - William Simons (SHRINE)
  - Doug McFadden (SHRINE)
  - Christopher Herrick (mi2b2)
  - David Wang (mi2b2)
  - Bill Wang (mi2b2)

- Sample Acquisition
  - Lynn Bry
  - Natalie Boutin

- **Depression Driving Biology and Pharmacovigilance Projects:**
- Roy Perlis/Jordan Smoller/Dan Iosifescu (PIs)
  - Victor Castro
  - Caitlin Clements
  - Wouter Hoogenboom,
  - Martha Shenton
  - Patience Gallagher
  - Stefanie Block
  - Alison Hoffnagle

- **International Cohort Collection for Bipolar Disorder:**
- Jordan Smoller**/**Pamela Sklar (PIs)
  - Roy Perlis
  - Victor Castro
  - Alison Hoffnagle
  - Sydney Weill
  - Mireya Nadal-Vicens
  - Niels Rosenquist
  - April Hirschberg
  - Alisha Pollastri
  - Jane Erb
  - Shaun Purcell
  - Nadia Solovieff

**Funding**

NIH/NCBC; National Institute of Mental Health; NCBI; NIH/NCATS

Stanley Center, Broad Institute