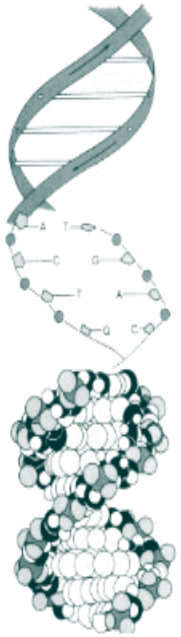
PARTNERS HEALTHCARE

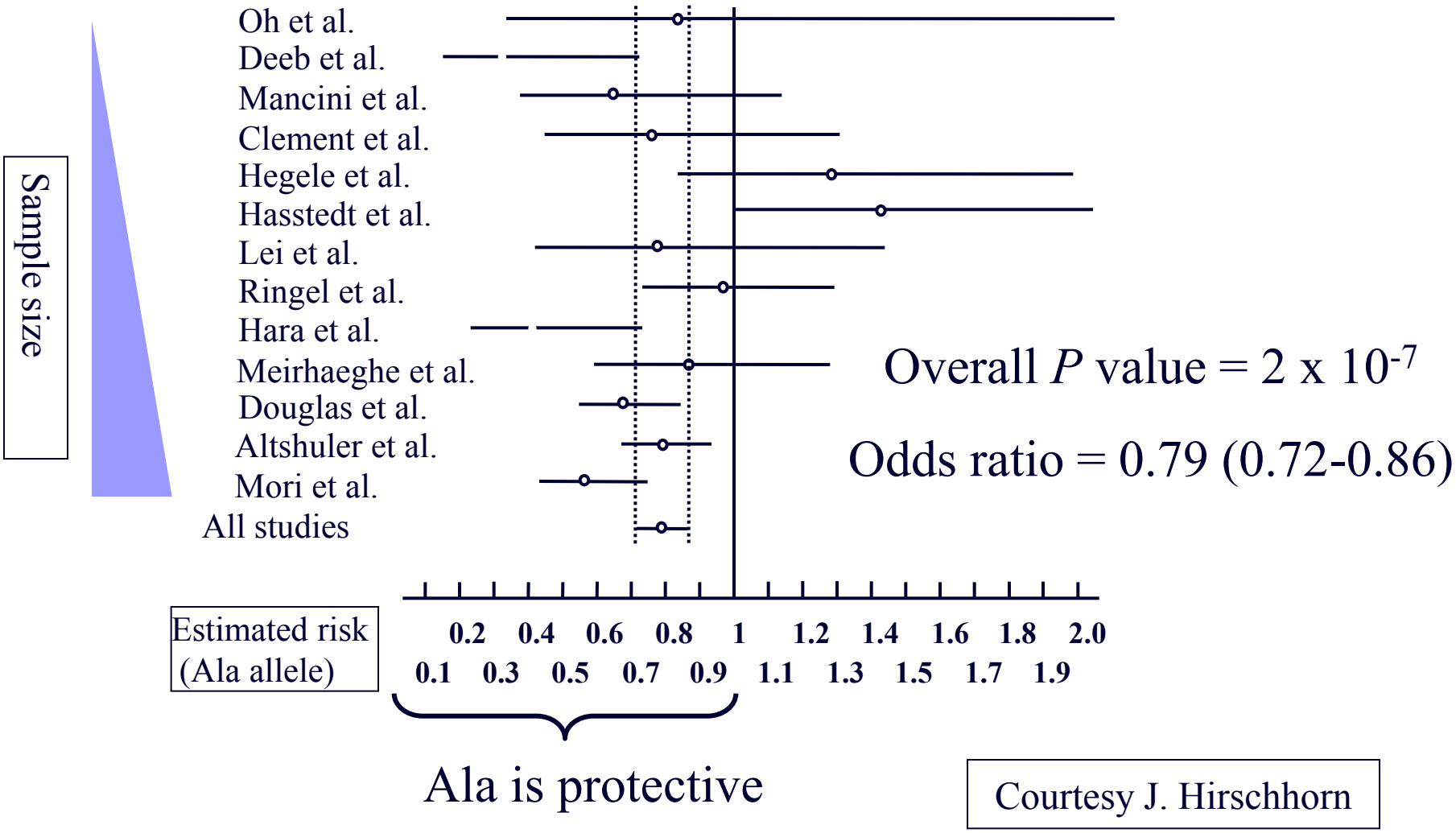


Computing our Patient's Future Using Data from our Healthcare Institutions

Shawn Murphy MD, Ph.D.

NETTAB 2011 Workshop on Clinical Bioinformatics

Example: PPARγ Pro12Ala and Diabetes



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

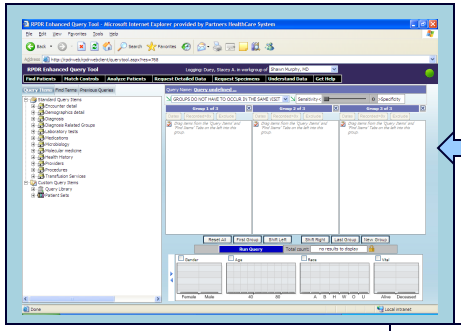
- Set of patients is selected from medical record data in a high throughput fashion
- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data
- Using the Crimson system, tissues of these patients can 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)

Query construction in web tool



De-identified Data Warehouse

Z731984X
Z74902XX
...
...

Encrypted identifiers

2) Returns identified patient data

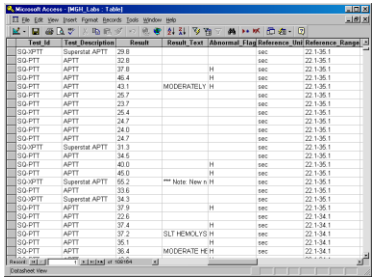
- Start with list of specific patients, usually from (1)
- Authorized use by IRB Protocol
- Returns contact and PCP information, demographics, providers, visits, diagnoses, medications, procedures, laboratories, microbiology, reports (discharge, LMR, operative, radiology, pathology, cardiology, pulmonary,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

0000004
2185793
...
...

OR

0000004
2185793
...
...

Real identifiers



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.

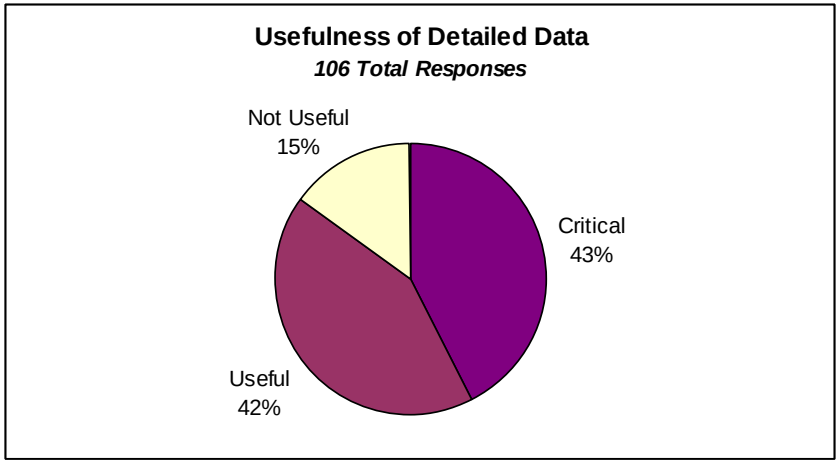
- Queries done in Query Tool available for review by RPDR team, a user lock out will specifically direct a review.
- De-identified data warehouse is a “Limited Data Set” by HIPAA
  - Medical record numbers are encrypted and obvious identifiers are removed from data.
- Concept of “established medical investigator” is promoted by classification as a faculty sponsor.

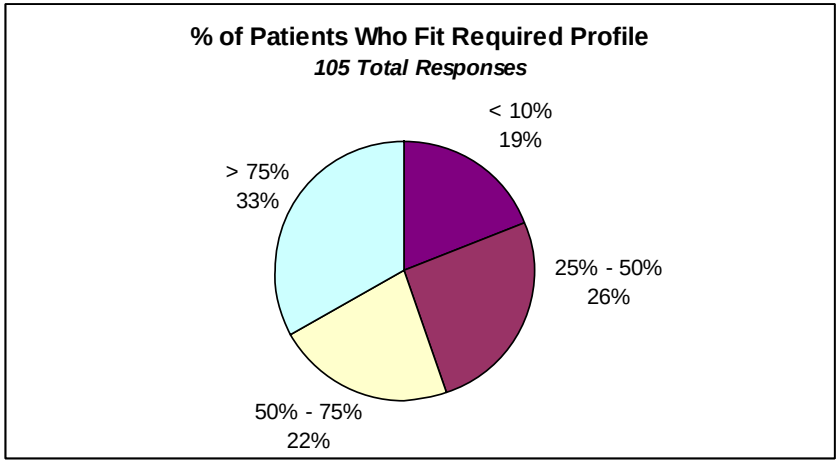
Security and Patient Confidentiality of Step 2

- 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.

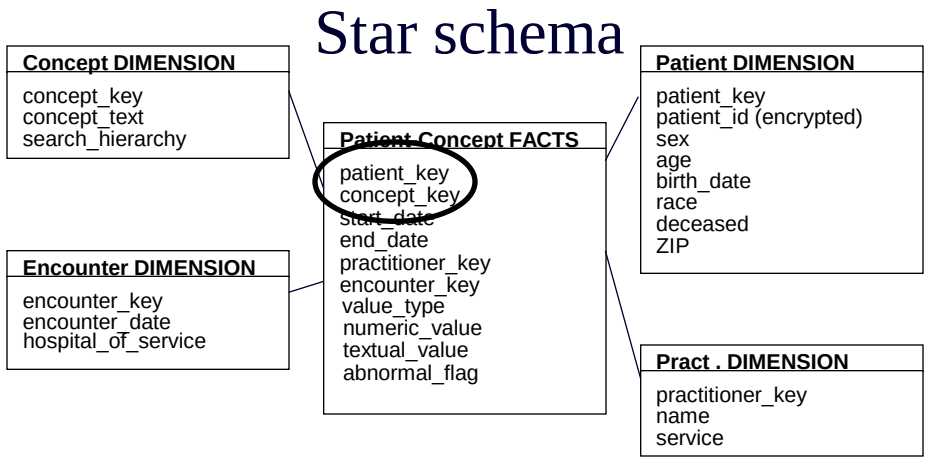
2009's usage of RPDR

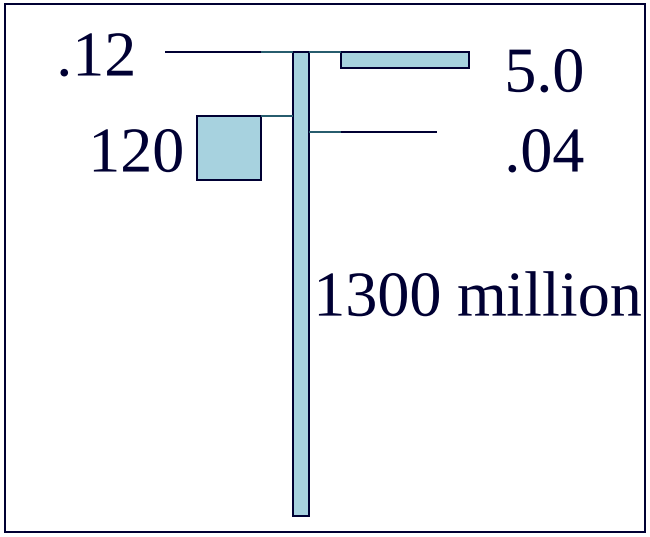
- 2,227 registered users, 457 new in 2008
- 338 teams gathering data for research studies
- 1286 identified patient data sets returned to these teams, containing data of 7.8 million patient records.
- From a survey of 153 teams
  - Importance of the data received from the RPDR was evaluated in relation 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

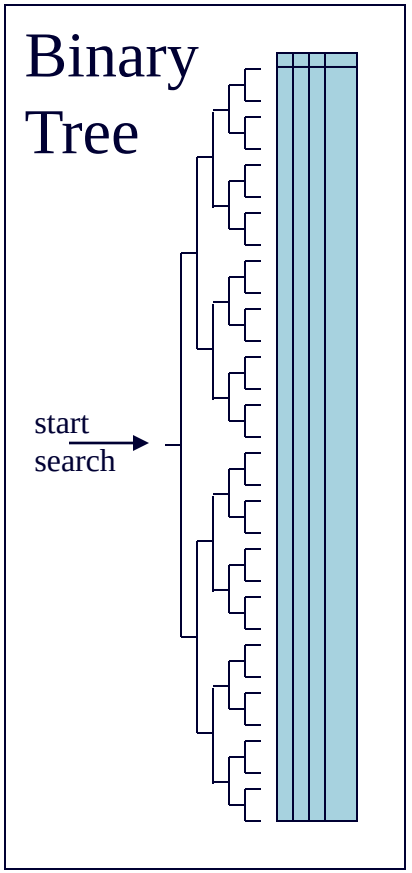




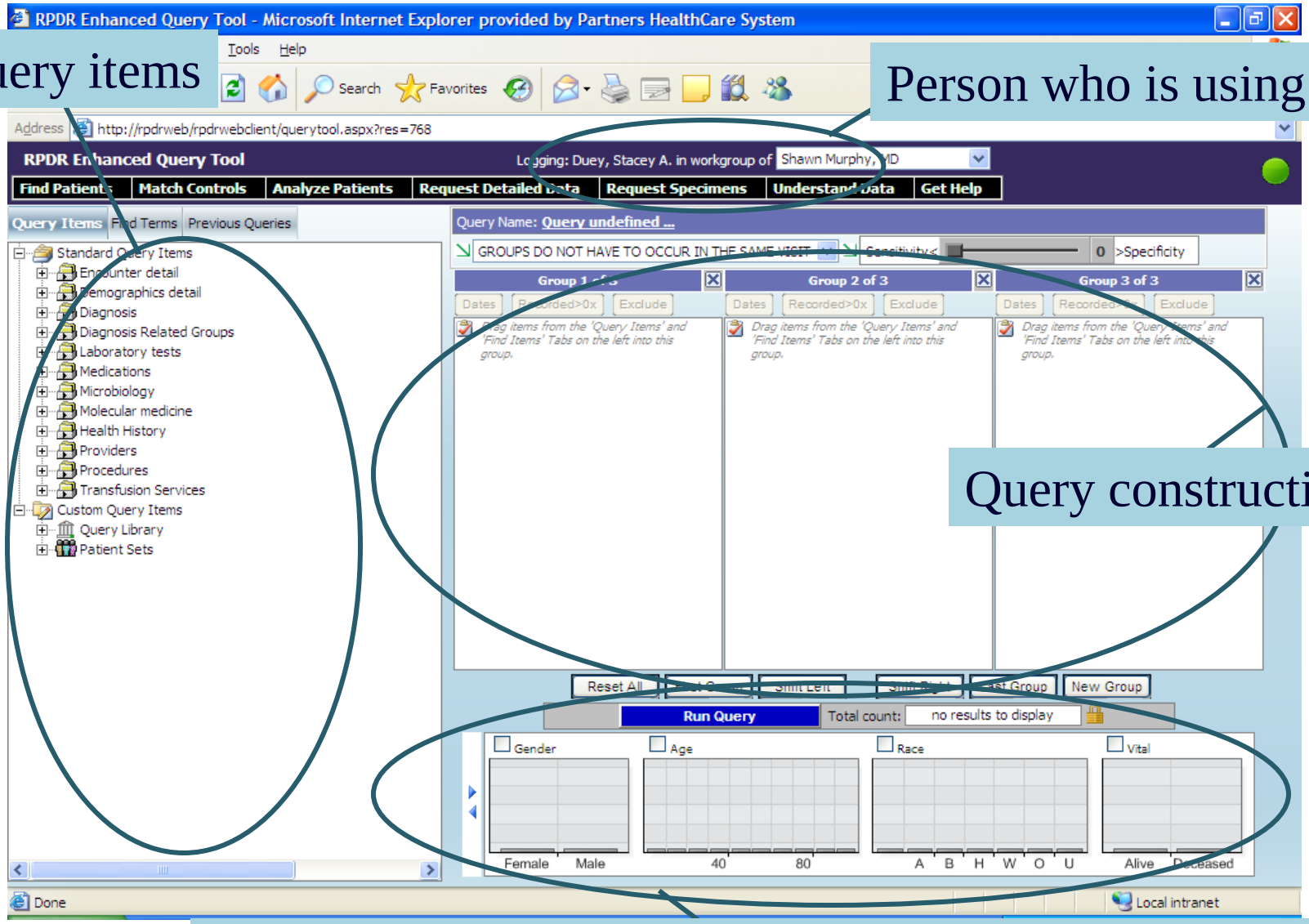
Organizing data in the Clinical Data Warehouse







FINDING PATIENTS

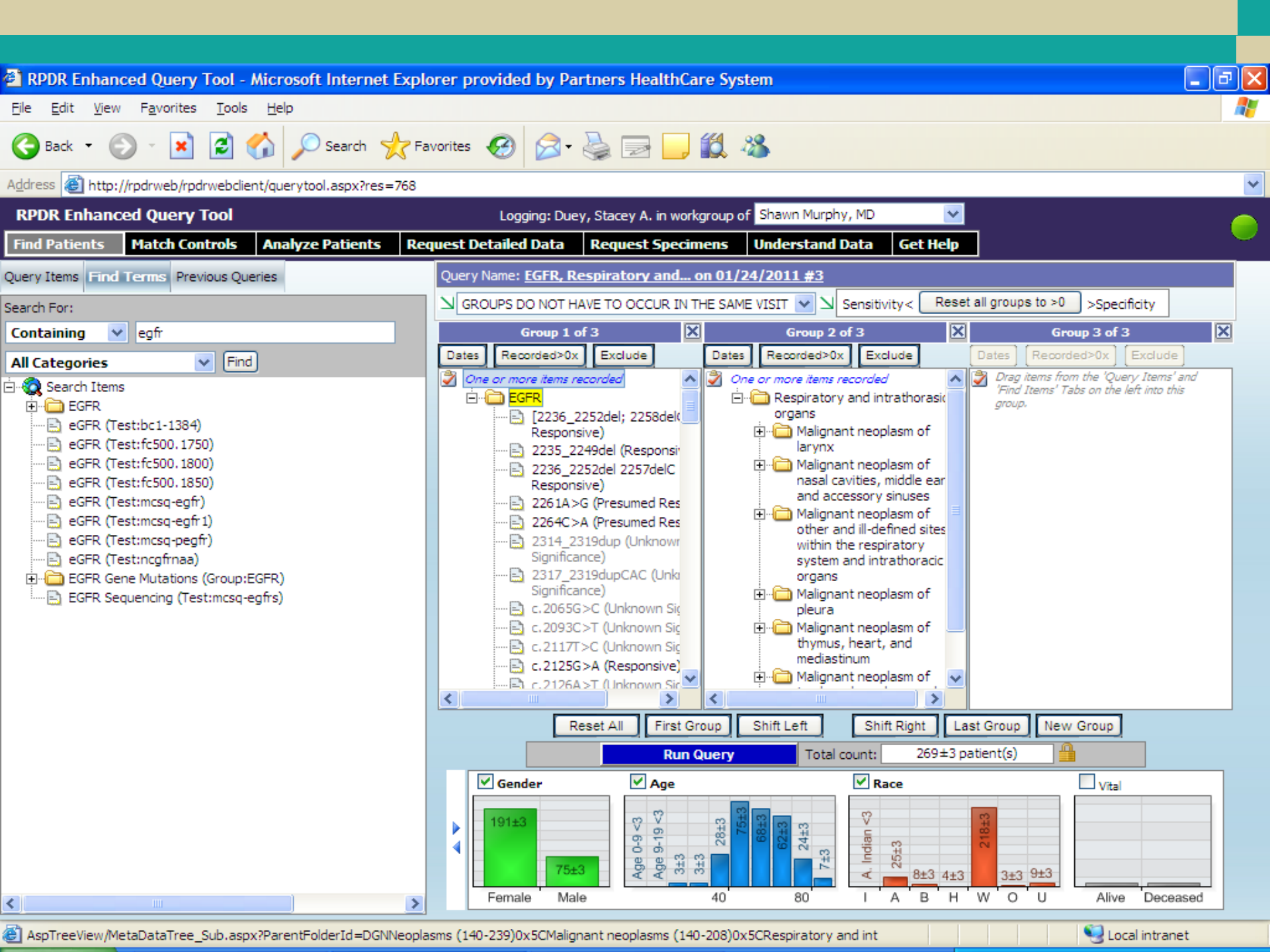


Query items

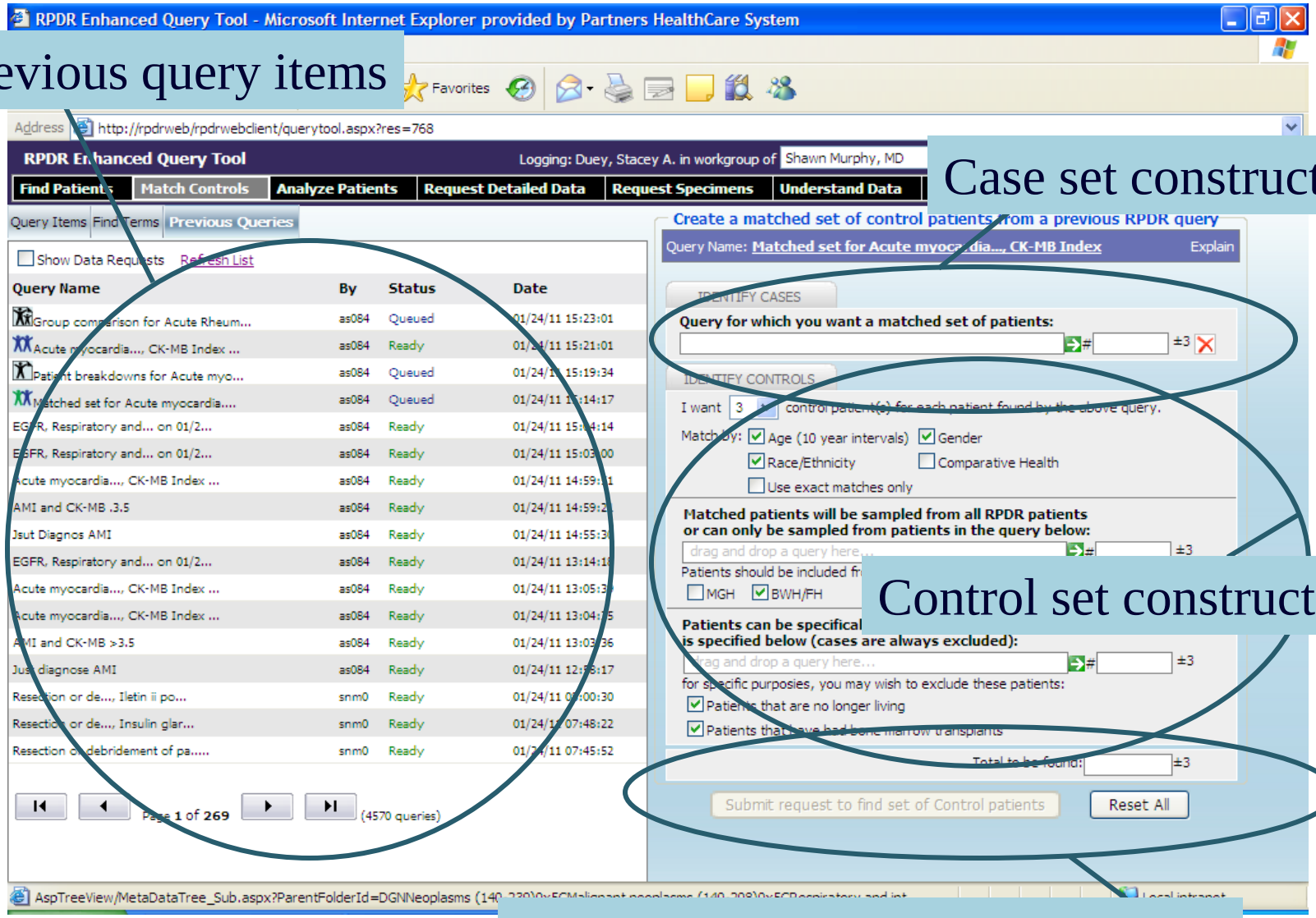
Person who is using tool

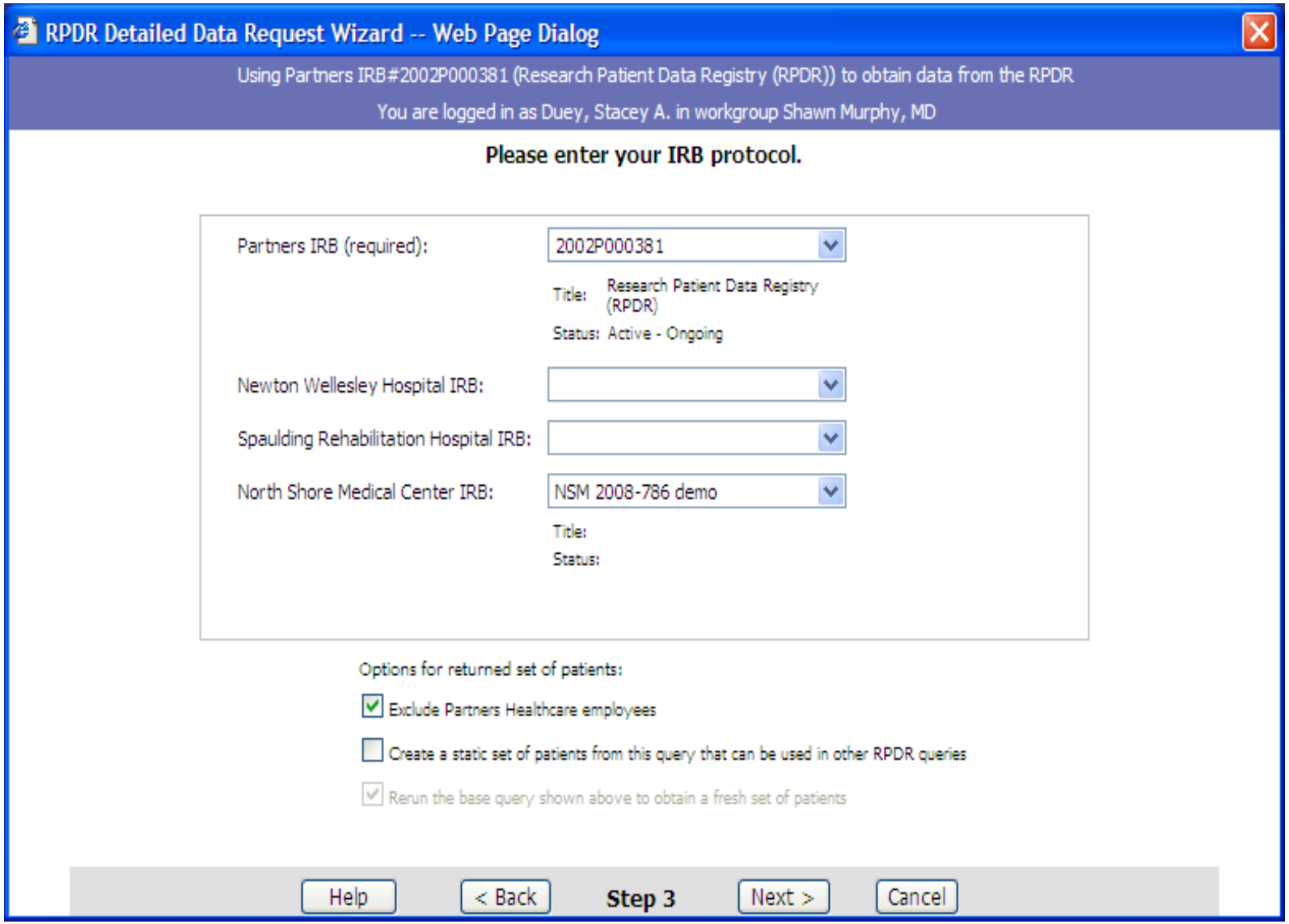
Query construction

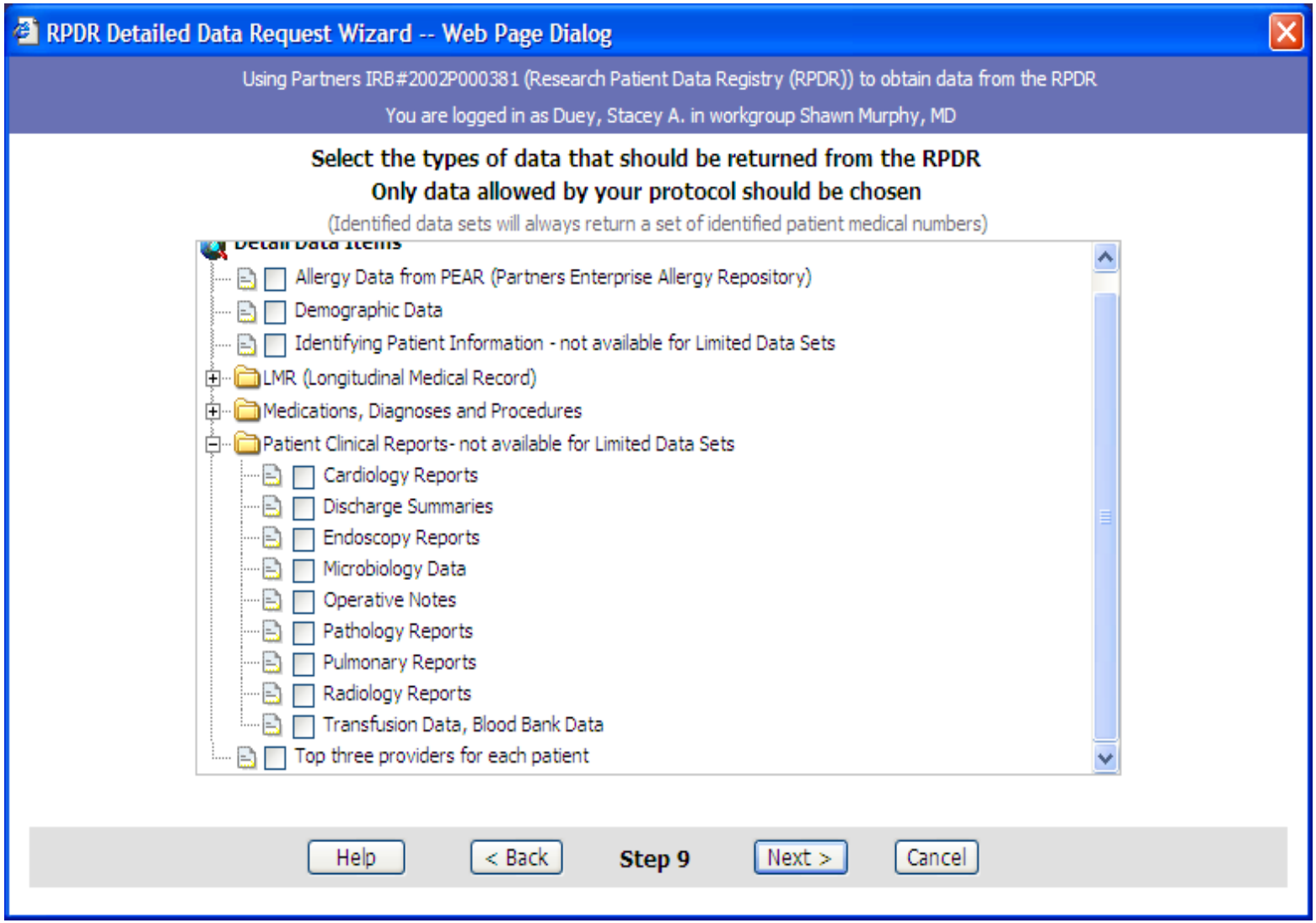
Results - broken down by number distinct of patients


RPDR Enhanced Query Tool - Microsoft Internet Explorer provided by Partners HealthCare System
File Edit View Favorites Tools Help
Back Search Favorites
Address http://rpdrweb/rpdrwebclient/querytool.aspx?res=768
RPDR Enhanced Query Tool
Logging: Duey, Stacey A. in workgroup of Shawn Murphy, MD
Find Patients
Match Controls
Analyze Patients
Request Detailed Data
Request Specimens
Understand Data
Get Help
Query Items
Find Terms
Previous Queries
Search For:
Containing
egfr
All Categories
Find
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
Group 2 of 3
Group 3 of 3
Dates
Recorded>0x
Exclude
One or more items recorded
EGFR
[2236_2252del; 2258delC 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 Sig
c.2093C>T (Unknown Sig
c.2117T>C (Unknown Sig
c.2125G>A (Responsive)
Respiratory and intrathoracic 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
Drag items from the 'Query Items' and 'Find Items' Tabs on the left into this group.
Reset All
First Group
Shift Left
Shift Right
Last Group
New Group
Run Query
Total count:
269±3 patient(s)
Gender
Age
Race
Vital
191±3
75±3
Female
Male
Age 0-9 <3
Age 9-19 <3
3±3
3±3
28±3
75±3
68±3
62±3
24±3
7±3
40
80
A. Indian <3
25±3
8±3
4±3
218±3
3±3
9±3
I
A
B
H
W
O
U
Alive
Deceased
AspTreeView/MetaDataTree_Sub.aspx?ParentFolderId=DGNNeoplasms (140-239)0x5CMalignant neoplasms (140-208)0x5CRespiratory and int
Local intranet

MATCHING PATIENTS



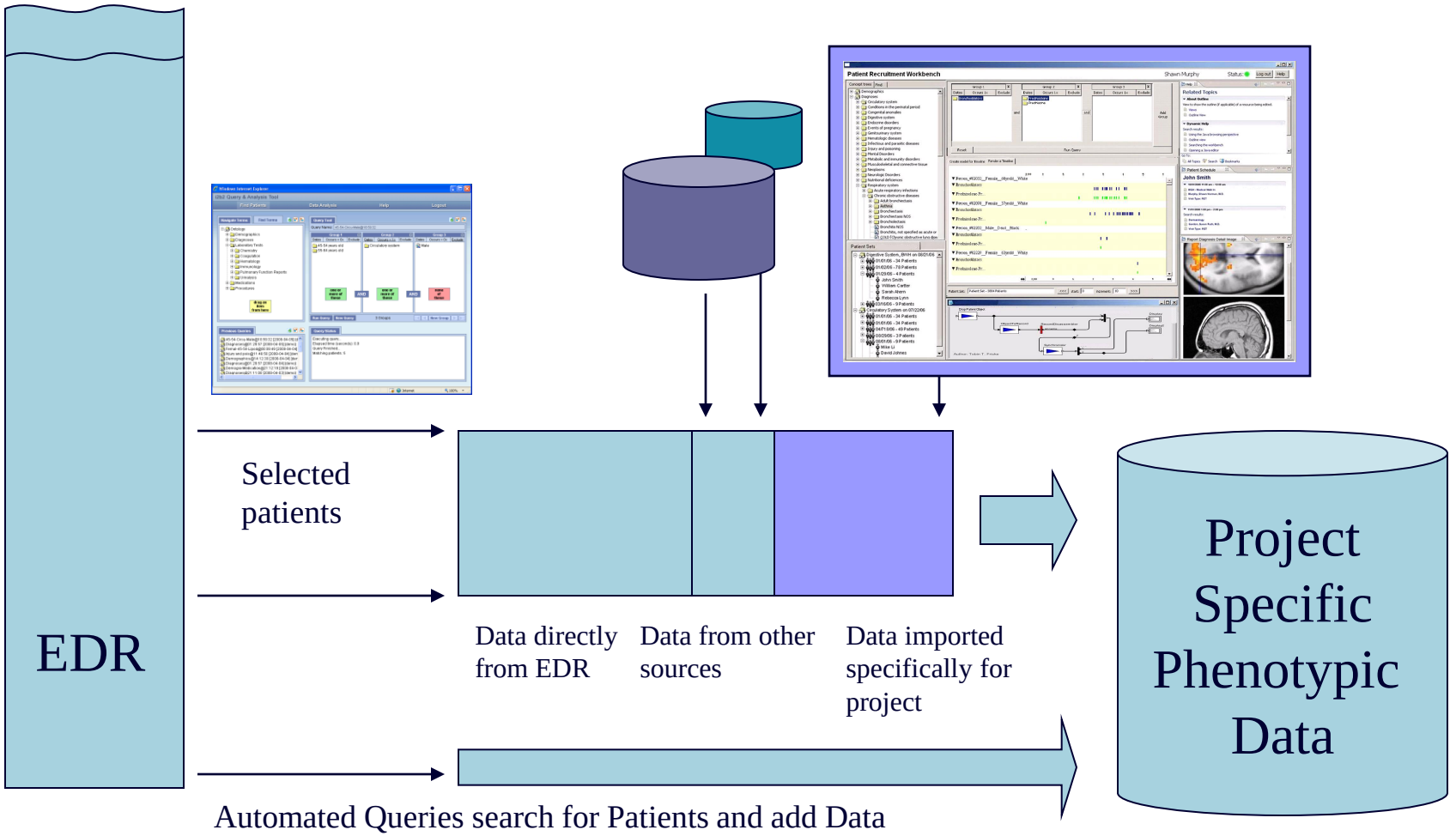

RPDR Detailed Data Request Wizard -- Web Page Dialog
Using Partners IRB#2002P000381 (Research Patient Data Registry (RPDR)) to obtain data from the RPDR
You are logged in as Duey, Stacey A. in workgroup Shawn Murphy, MD
Please enter your IRB protocol.
Partners IRB (required):
2002P000381
Title: Research 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
Help
< Back
Step 3
Next >
Cancel


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
Help
< Back
Step 9
Next >
Cancel

High Throughput Methods for supporting Research at Partners Healthcare

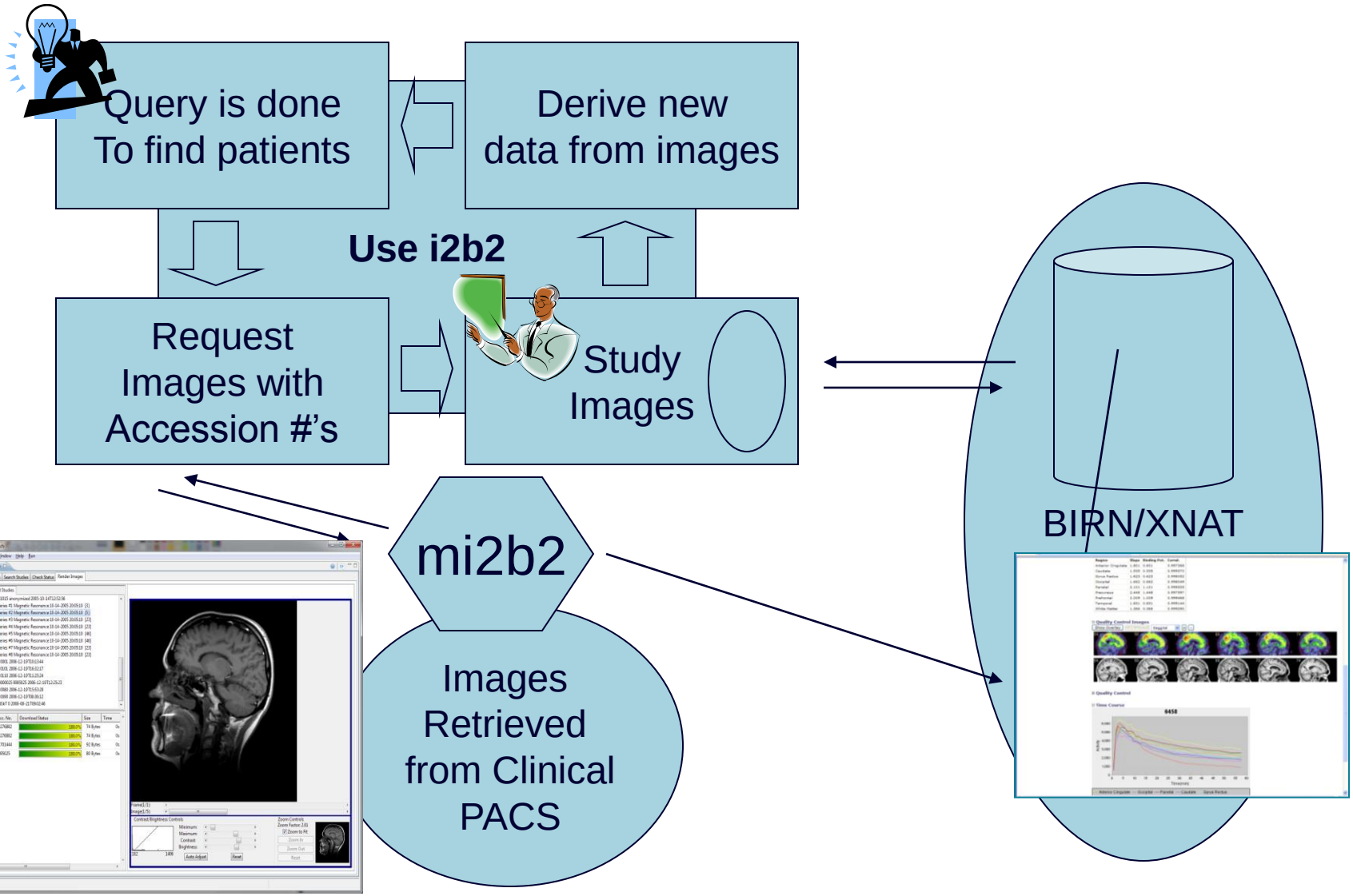
- Set of patients is selected from medical record data in a high throughput fashion
- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data
- Using the Crimson system, tissues of these patients can 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

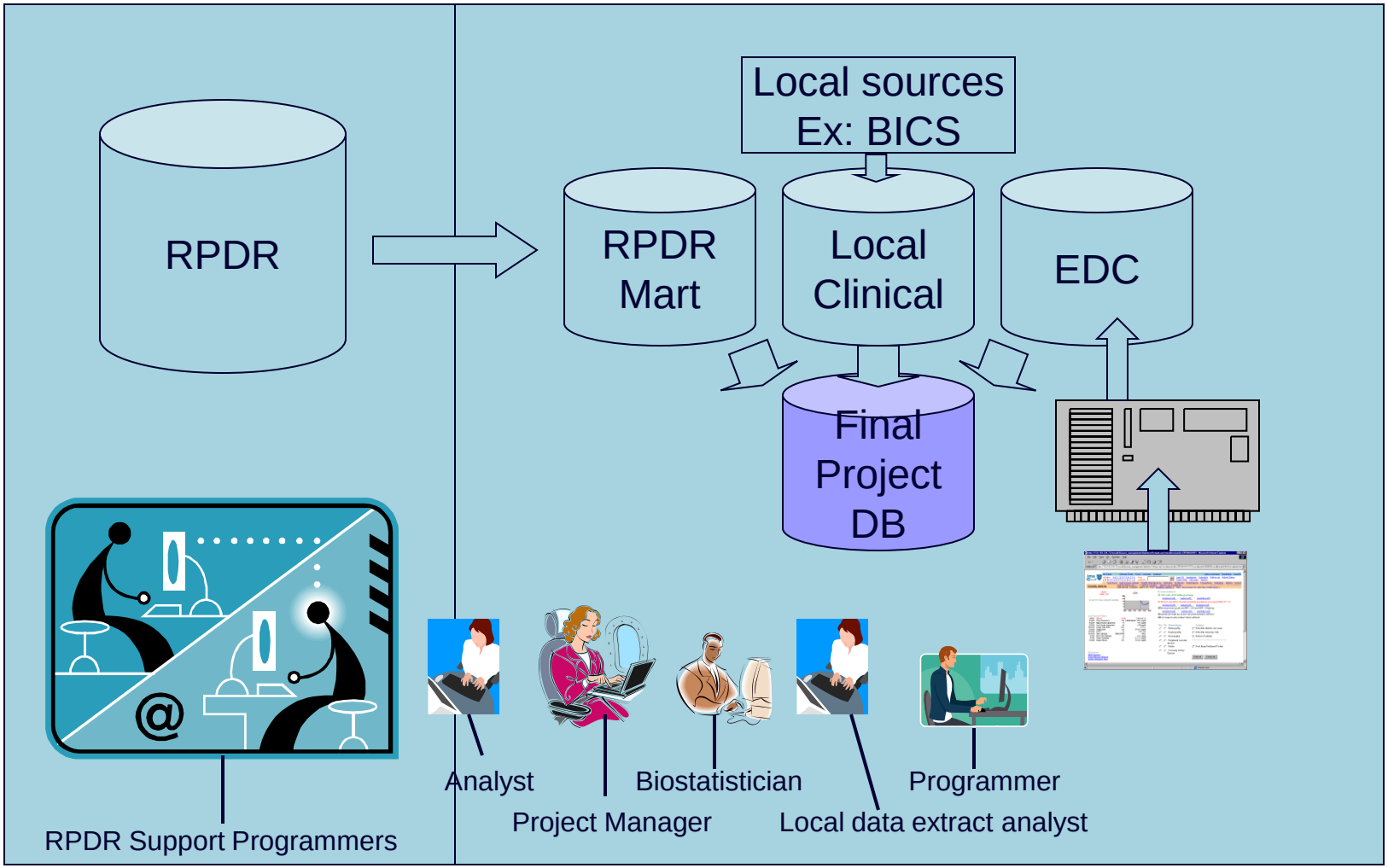


Data is available through the i2b2 Workbench

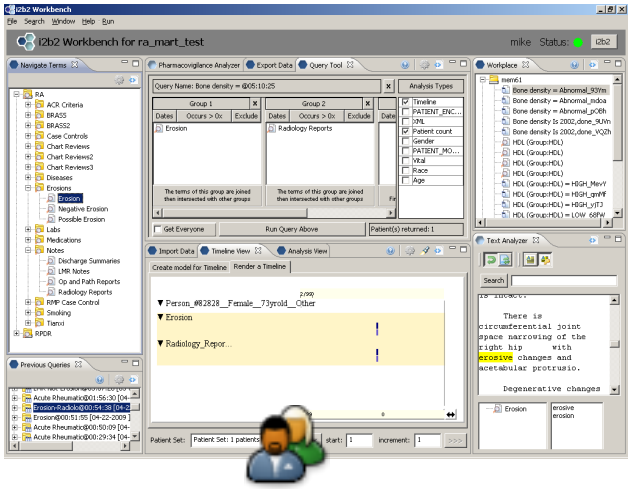
Research Investigator Workflow enabled by mi2b2



Team support for Projects



NLP Workflow



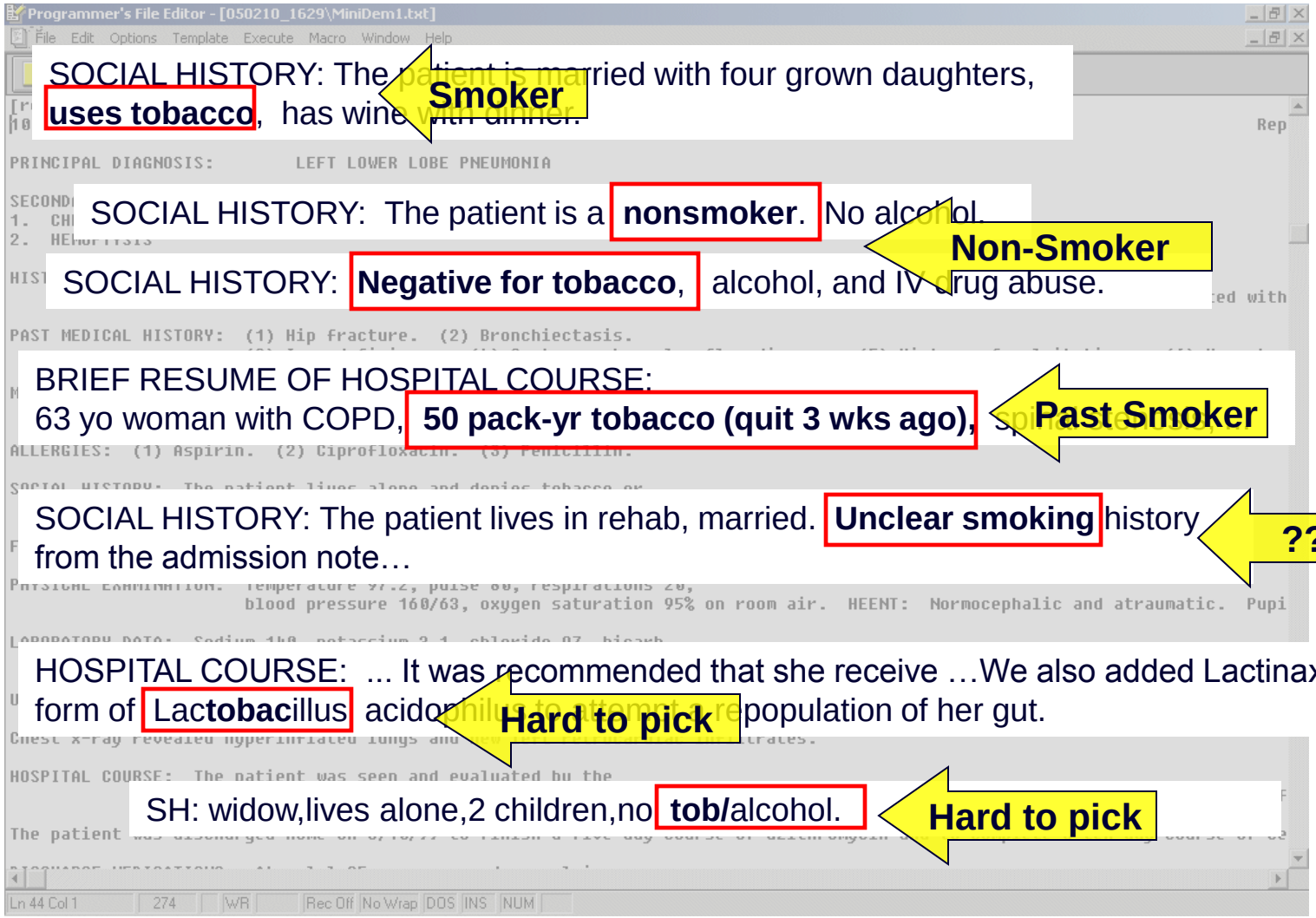
I2b2 Project Investigators

Communication

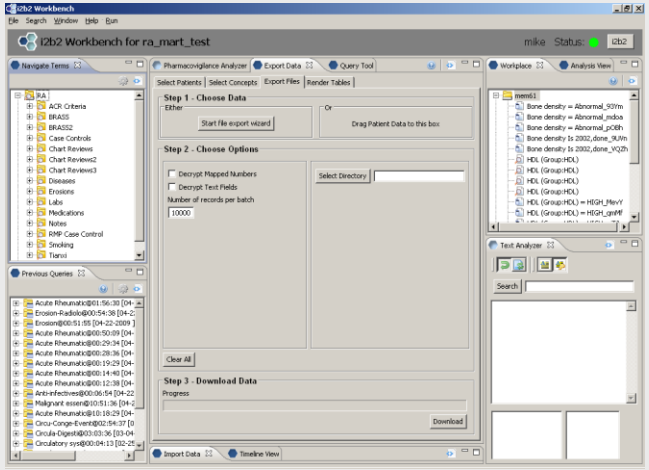
Results Delivery

NLP Specialists

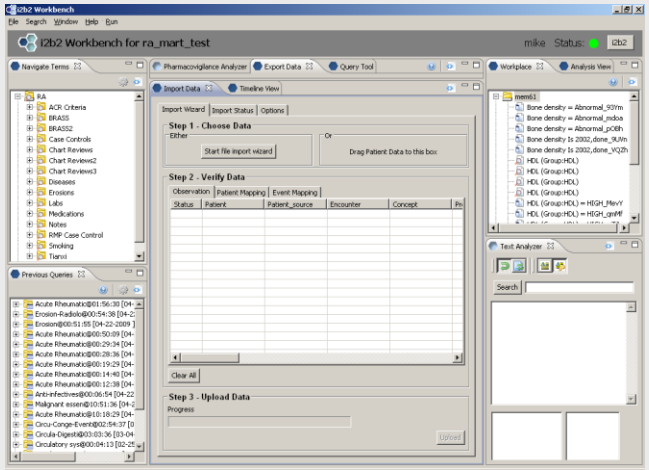
NLP (and comedy) is not pretty



NLP Specialists Workstation



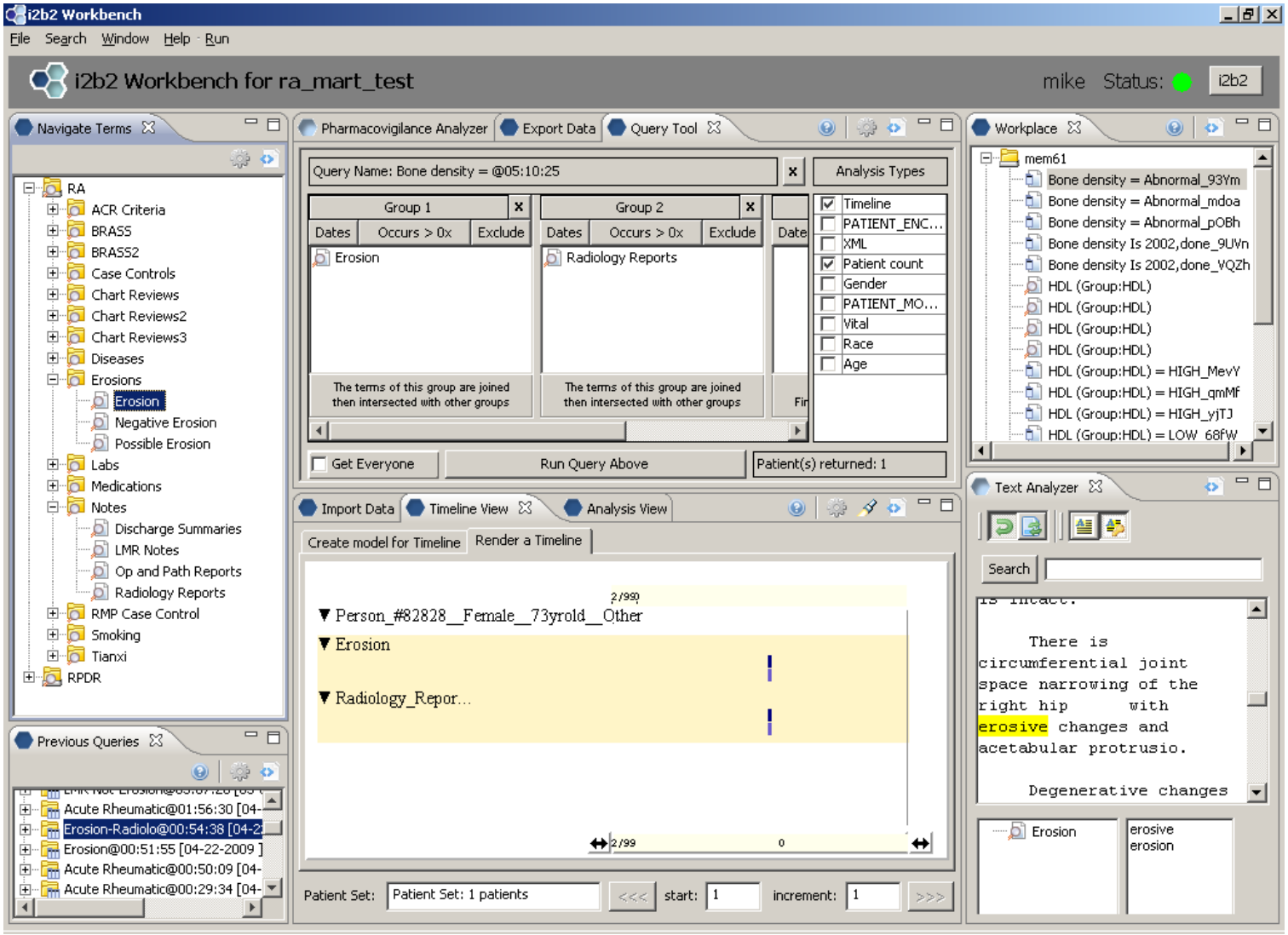
Export Notes



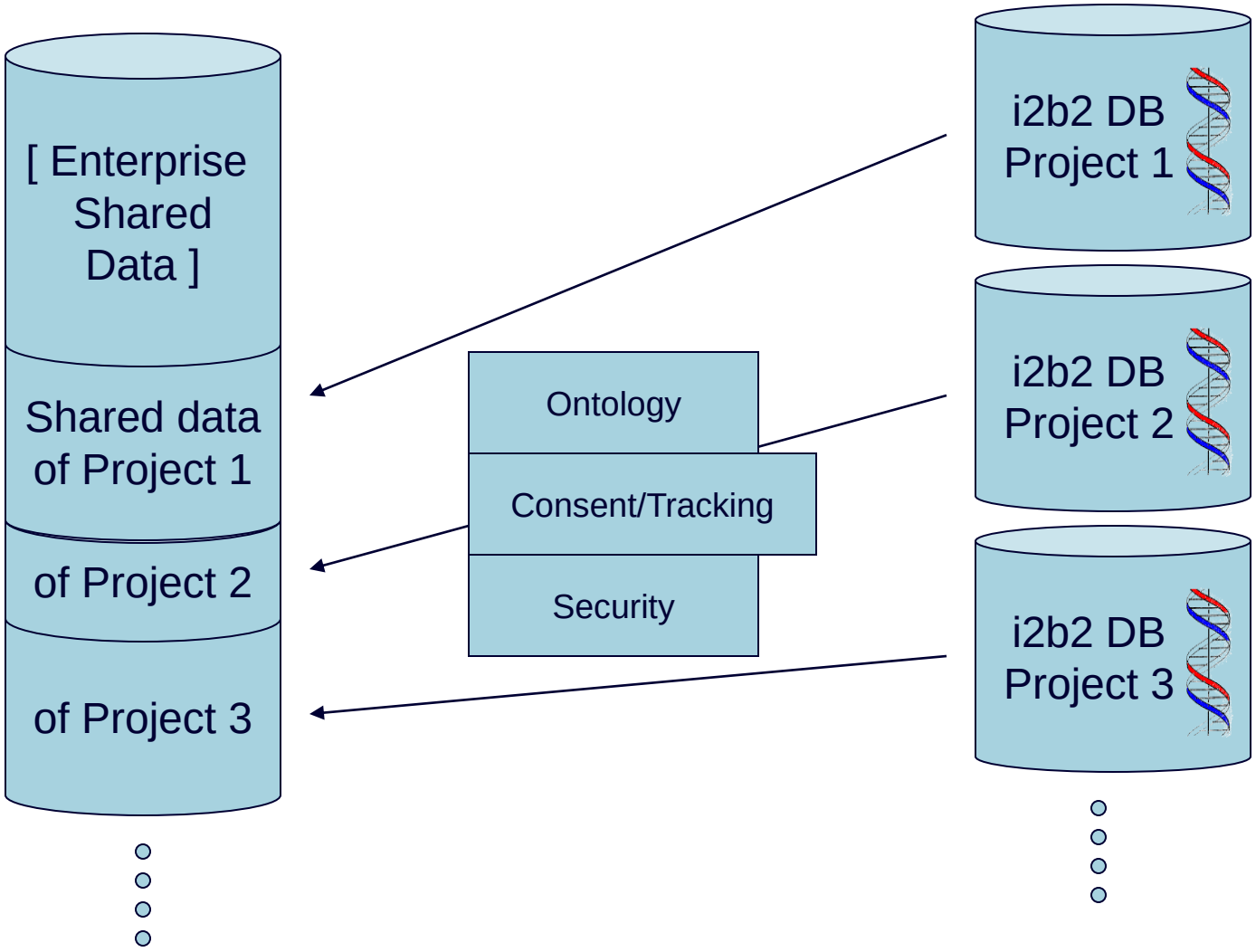
Import Derived Codes

NLP Specialists

Investigator Review



Project data can be added back to Enterprise Repository



Community

United States

- Arizona State University
- Beth Israel Deaconness 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
- Childrens'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
- University of Texas Health Sciences Center Southwestern, Dallas, TX
- Utah Health Science Center, Salt Lake City, UT
- University of Washington, Seattle, WA
- University of Wisconsin Madison
- Veterans Administration Boston and Utah

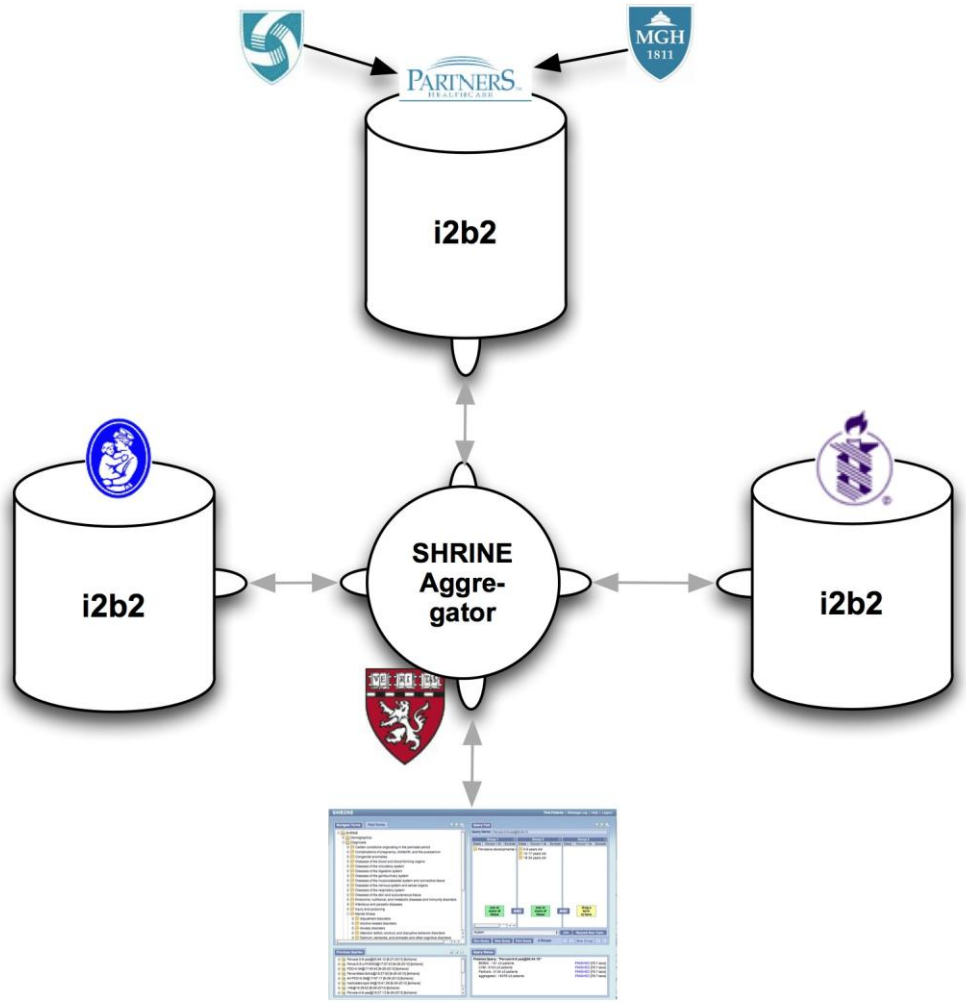
International

- 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 Leicester and Hospitals, England (Biomed. Res. Informatics Ctr. for Clin. Sci)
- University of Pavia, Pavia, Italy
- University of Seoul, Seoul, Korea

Aggregating across 4 hospitals, 3 i2b2 instances

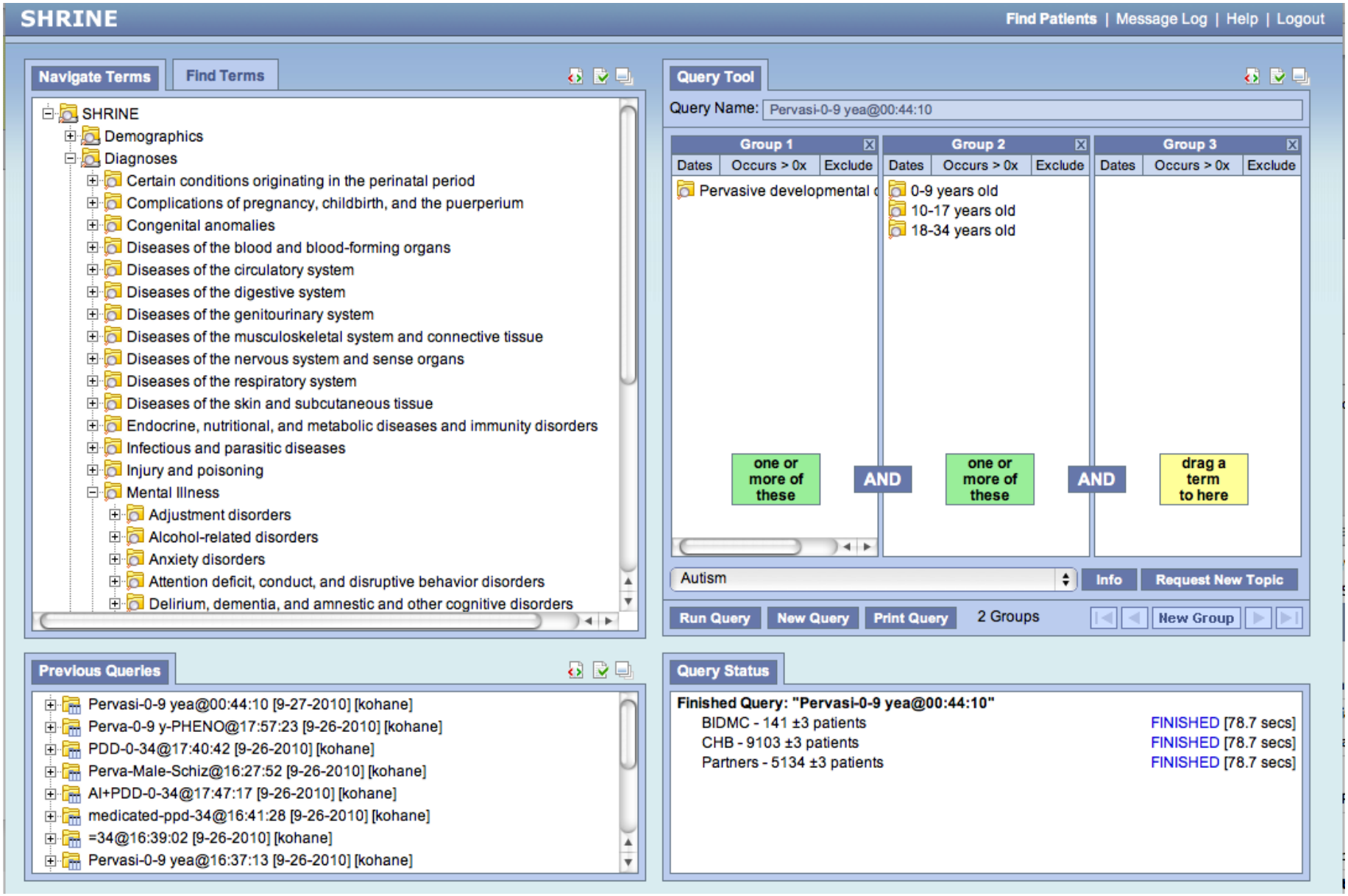
SHRINE (Shared Research Informatics Network)

= Distributed Queries

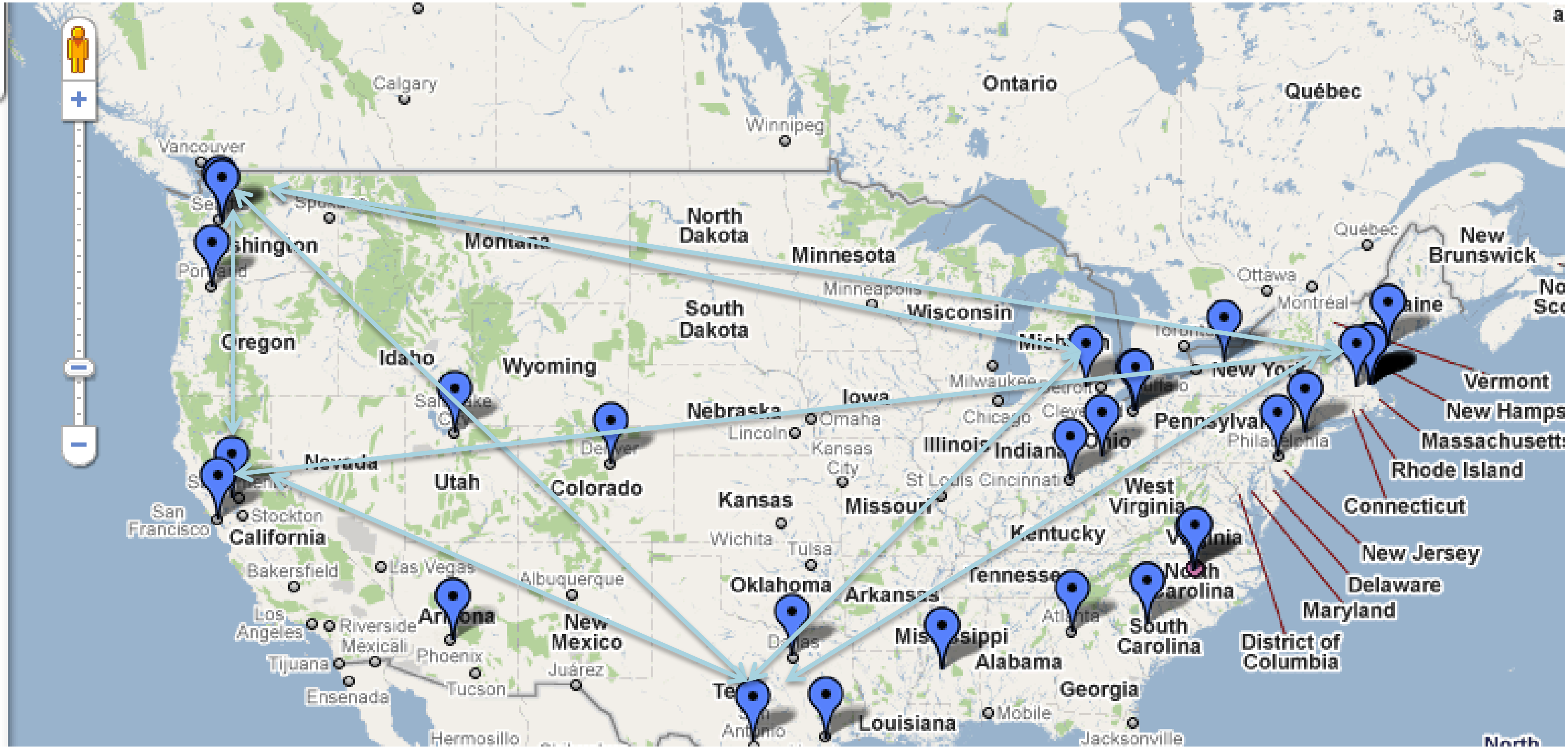


Clinical data in SHRINE

- 10 years (2001-2011)
- 4 hospitals
- 6 million total patients
- >1 billion medical observations
  - Demographics
  - Diagnoses (ICD9-CM)
  - Medications (RxNorm)
  - Labs (LOINC)


SHRINE
Find Patients | Message Log | Help | Logout
Navigate Terms
Find Terms
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 Tool
Query Name: Pervasi-0-9 yea@00:44:10
Group 1
Dates
Occurs > 0x
Exclude
Pervasive developmental
Group 2
Dates
Occurs > 0x
Exclude
0-9 years old
10-17 years old
18-34 years old
Group 3
Dates
Occurs > 0x
Exclude
one or more of these
AND
one or more of these
AND
drag a term to here
Autism
Info
Request New Topic
Run Query
New Query
Print Query
2 Groups
New Group
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]
Query Status
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

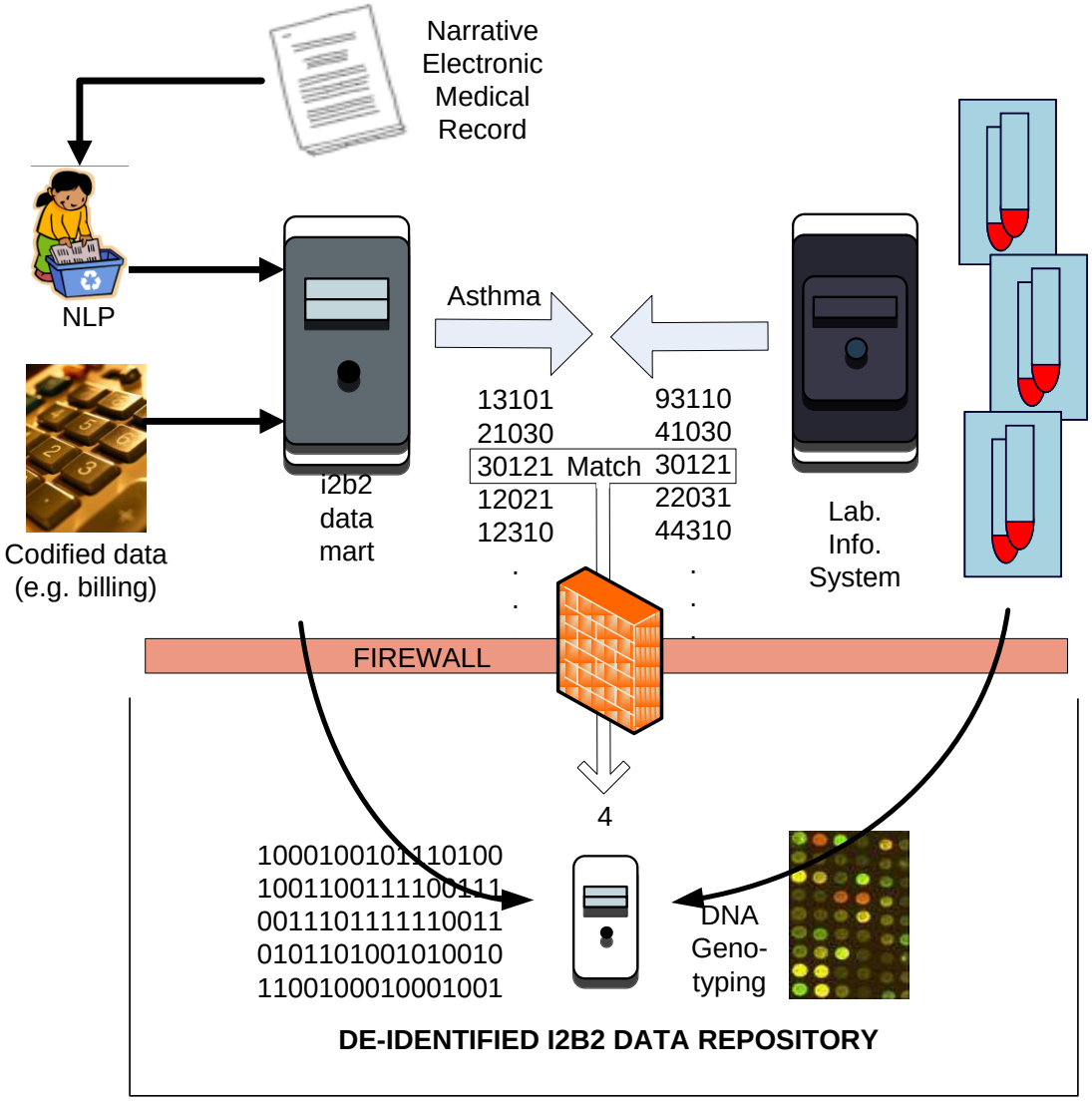


Calgary
Ontario
Québec
Winnipeg
Vancouver
North Dakota
Minnesota
Montana
Washington
Portland
Oregon
Idaho
Wyoming
South Dakota
Wisconsin
Minneapolis
Ottawa
Montréal
New Brunswick
Maine
New York
Milwaukee
Iowa
Nebraska
Omaha
Lincoln
Chicago
Illinois
Indiana
Ohio
Pennsylvania
Philadelphia
Nevada
Utah
Colorado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

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



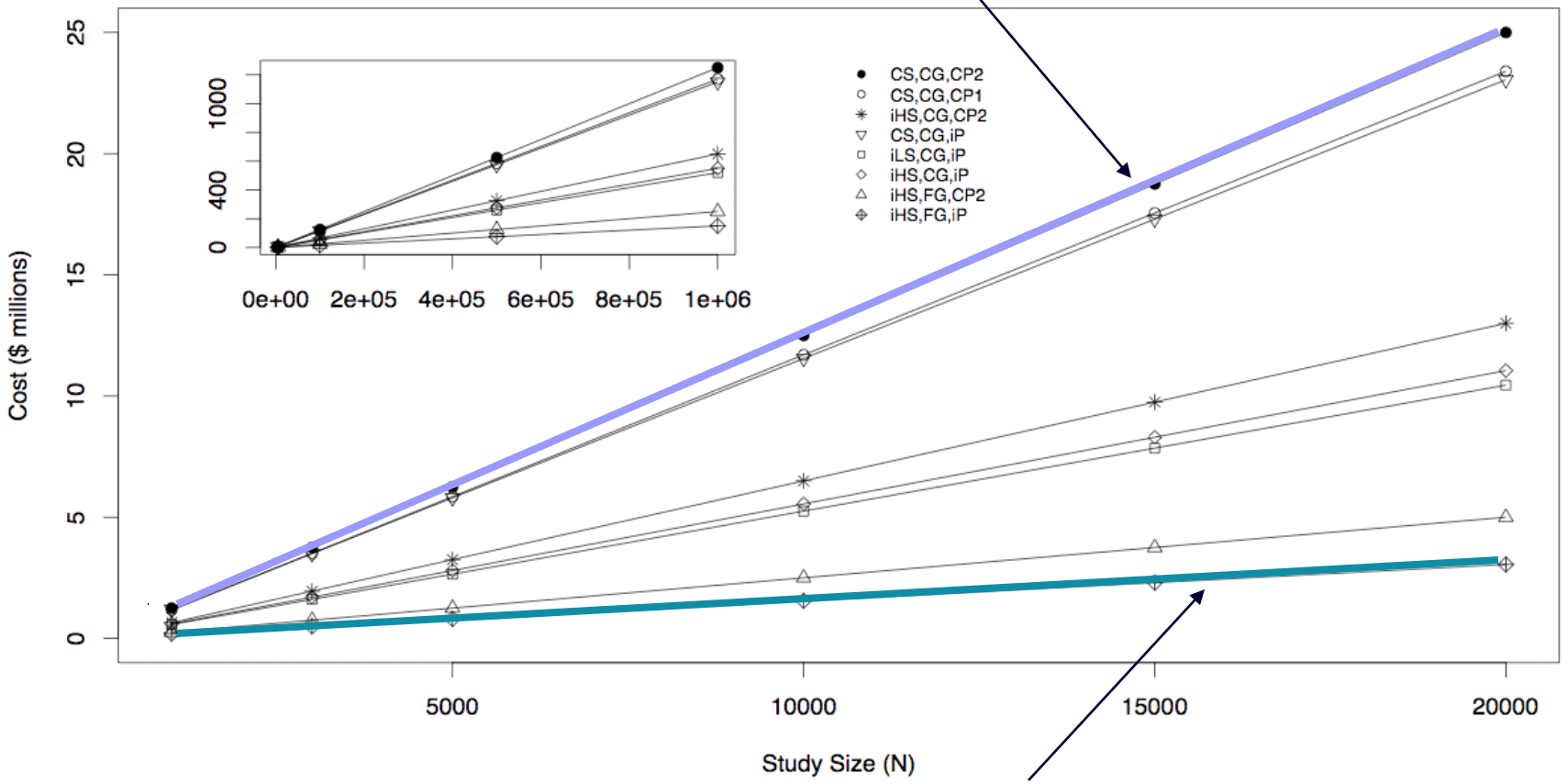
Cost and time benefit of Instrumenting with Sample Collection for Modest-size Study with 10,000 subjects (cases + controls)

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

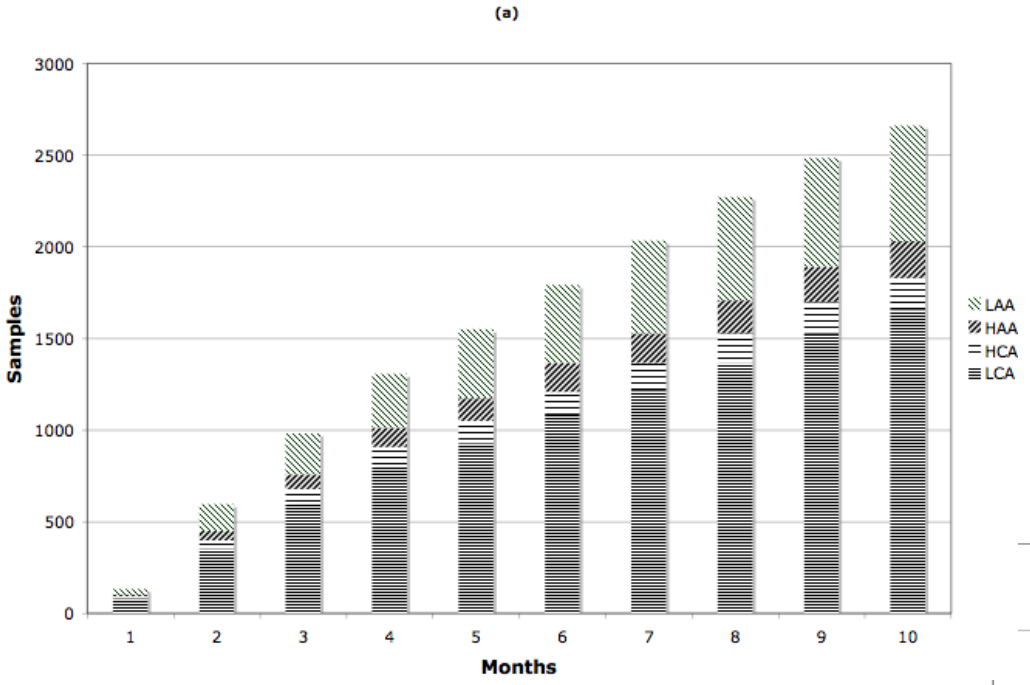
Escalating cost and time benefit of Instrumenting with Sample Collection

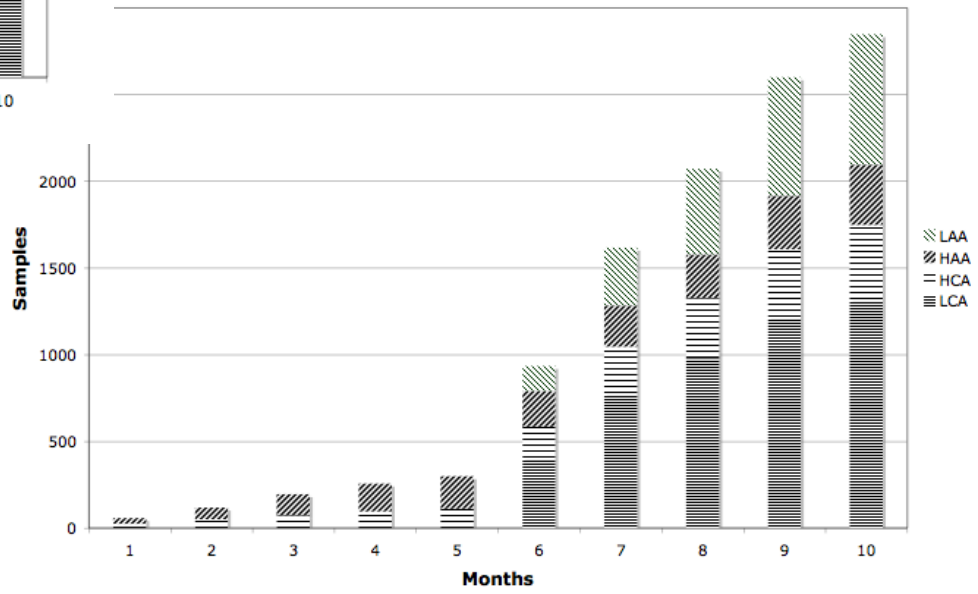
Previous model for collecting specimens



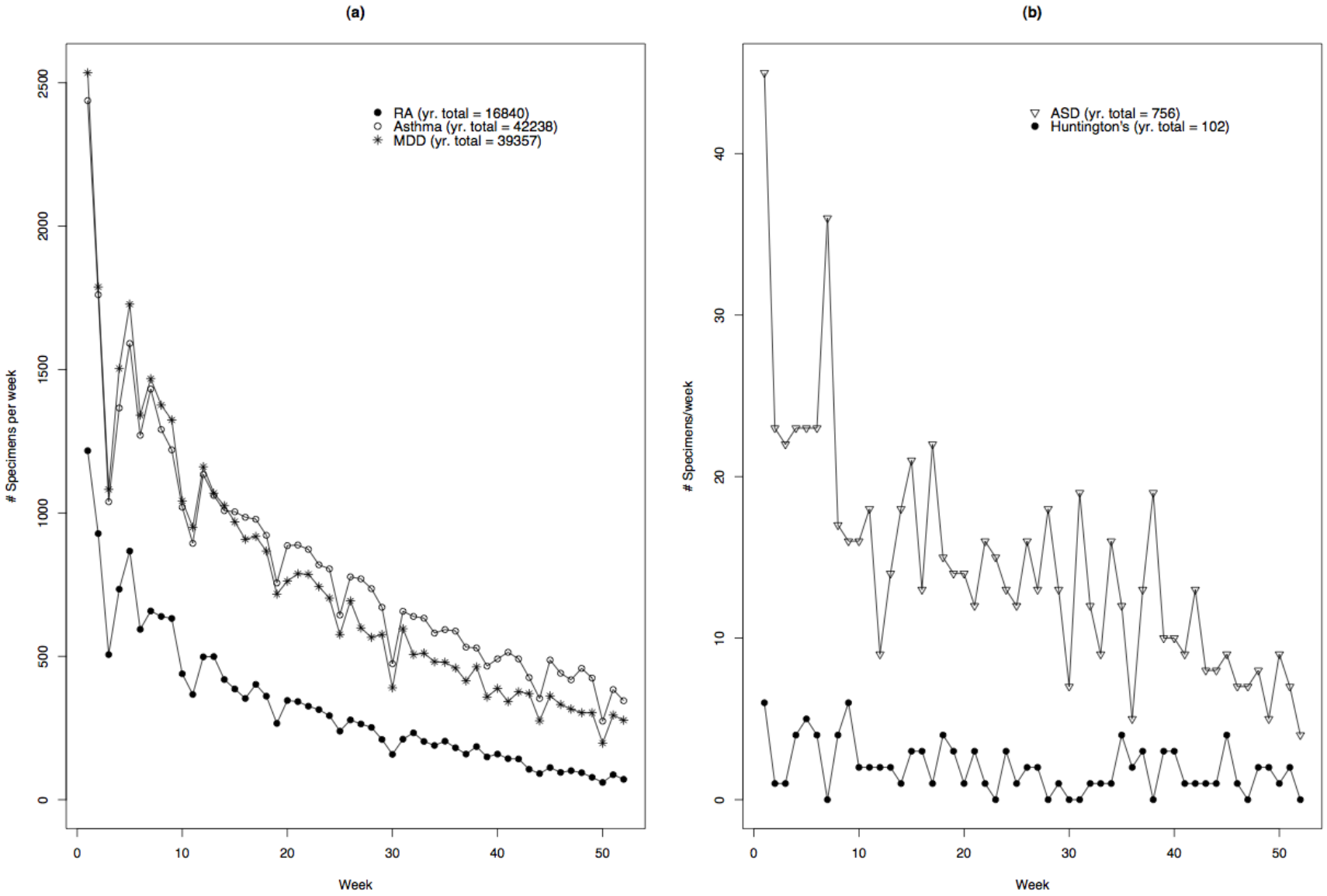
New model for collecting specimens

Meeting Expectations





Accrual Rates



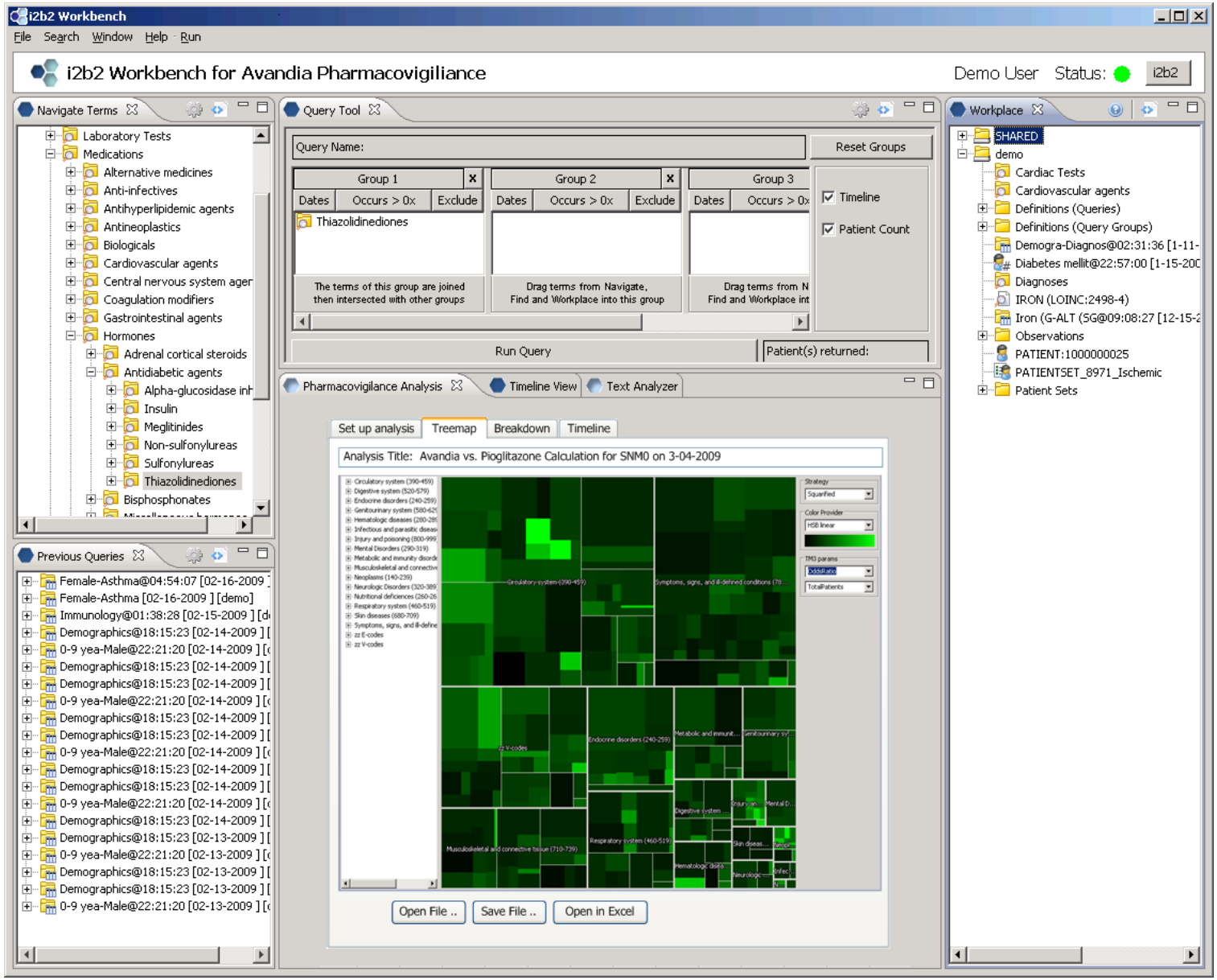
High Throughput Methods for supporting Research at Partners Healthcare

- 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

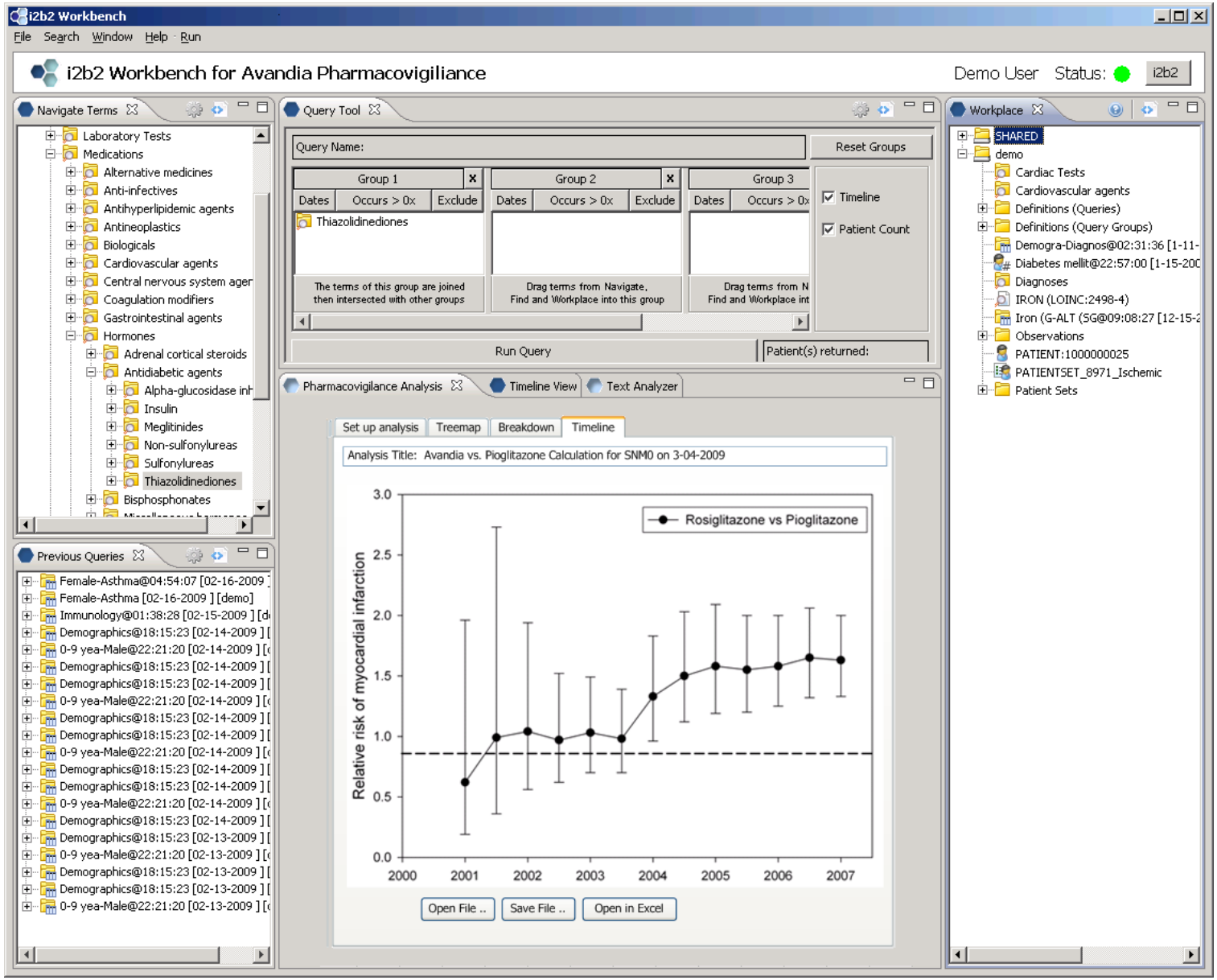
Performing Clinical trials “in-silico”

- 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.

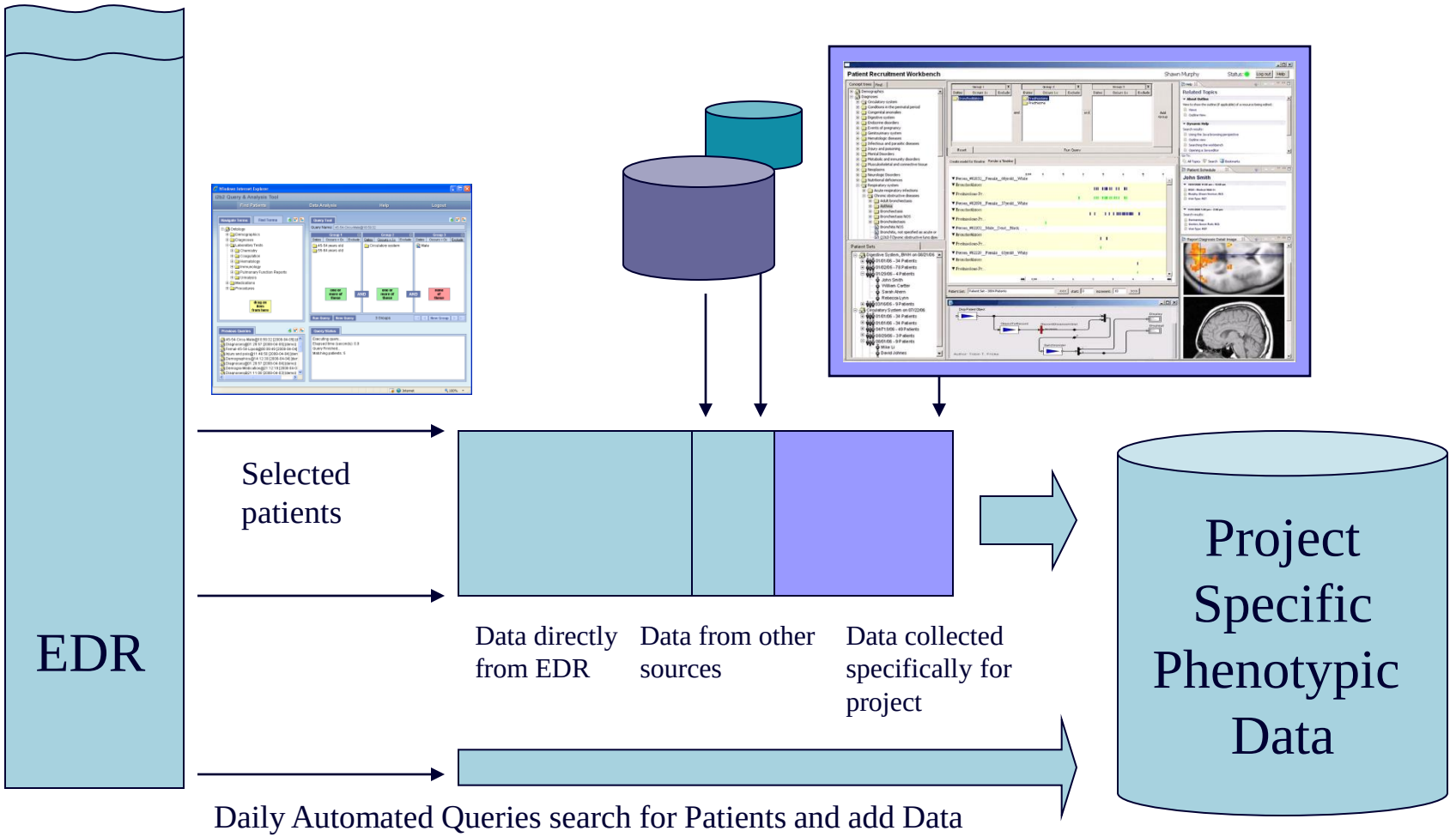
Dashboard used to observe high-level signals



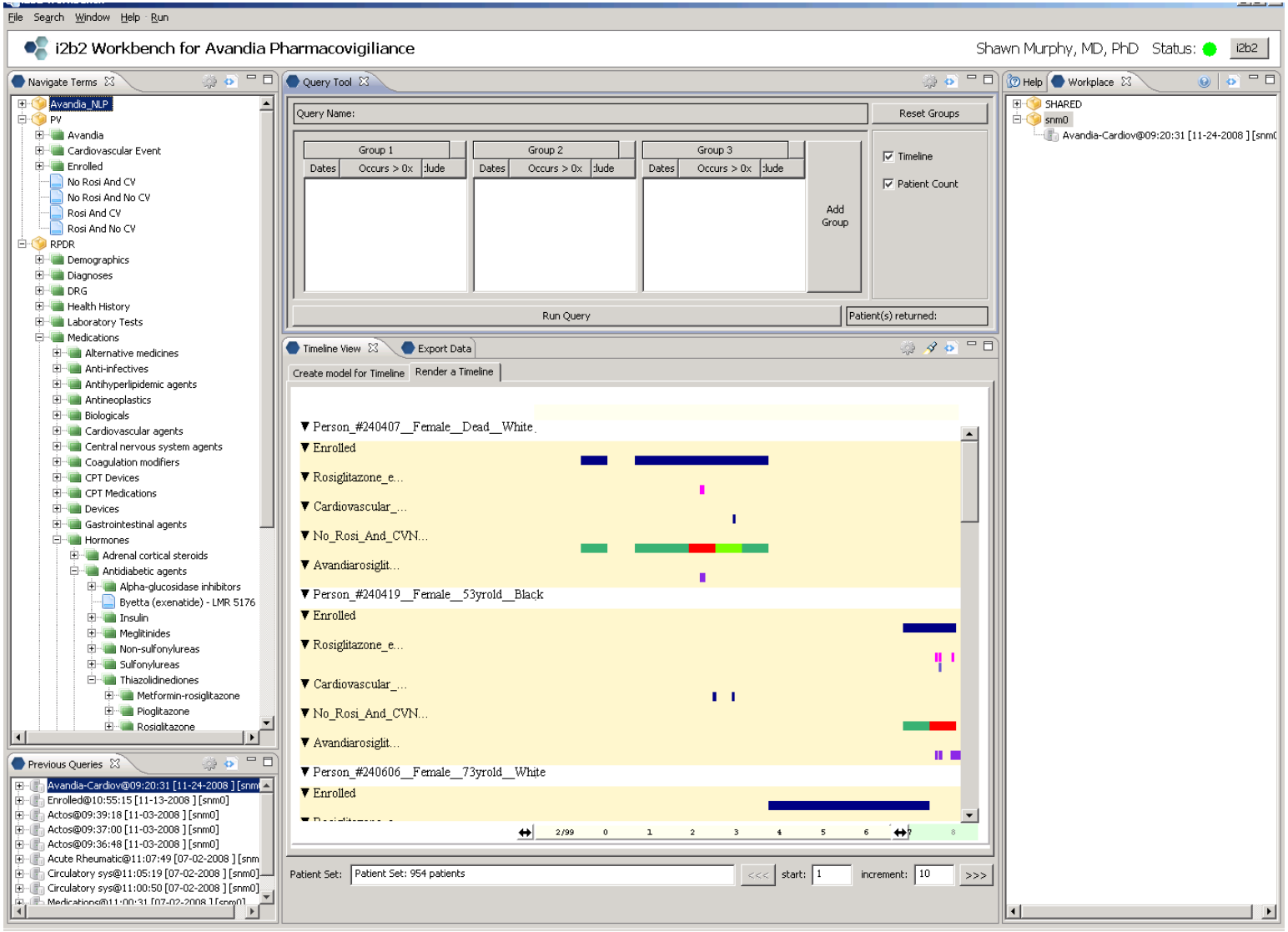
Dashboard used to observe high-level signals



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



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