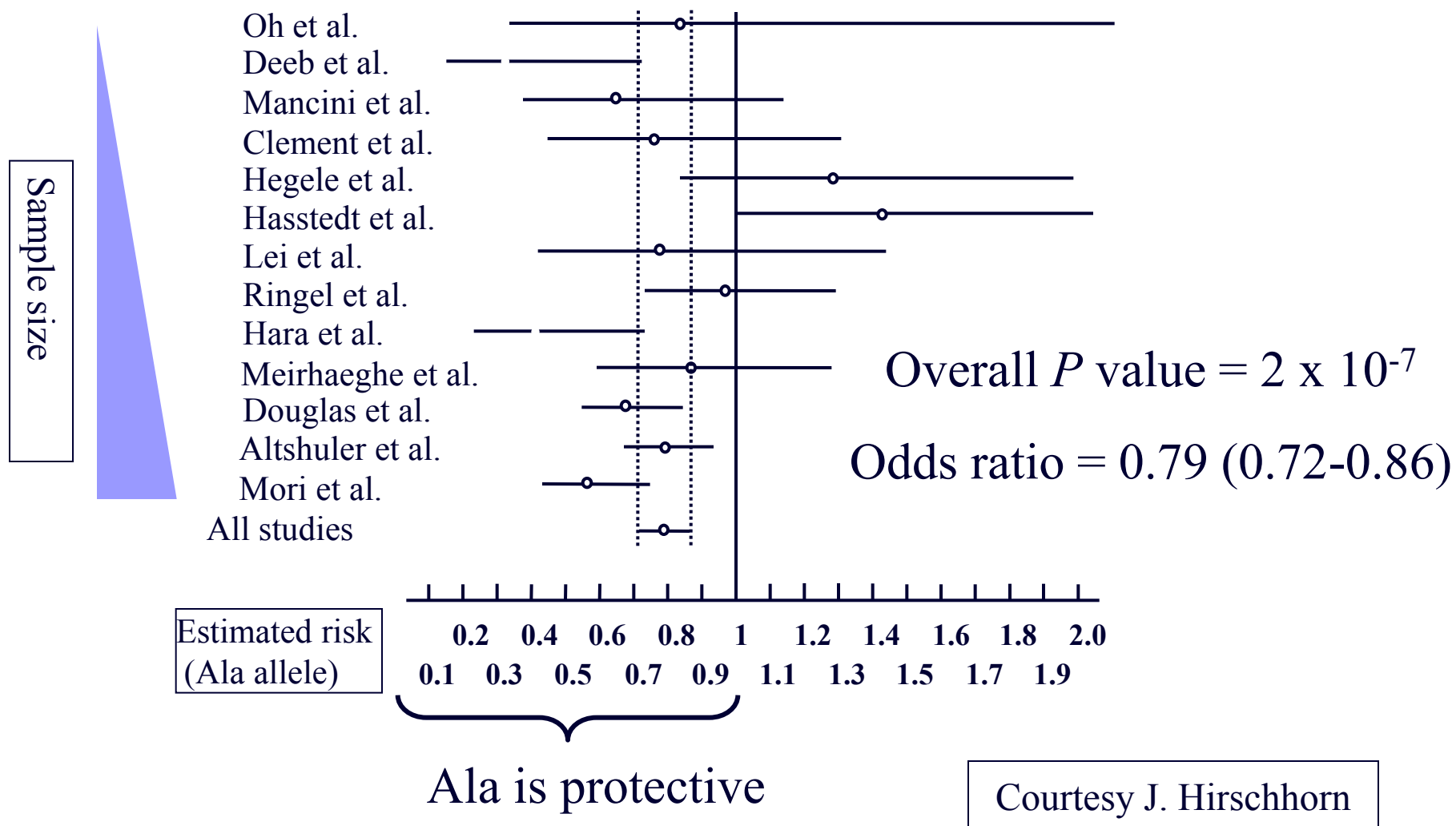


Computing our Patient's Future Using Data from our Healthcare Institutions

Shawn Murphy MD, Ph.D.

NETTAB 2011 Workshop on Clinical Bioinformatics

Example: PPAR γ Pro12Ala and Diabetes



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

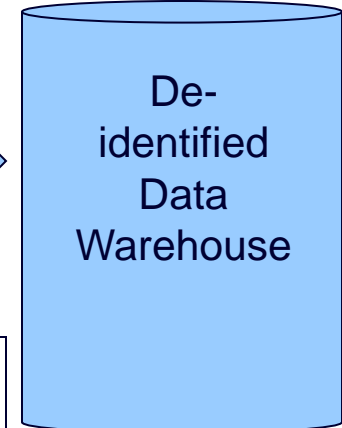
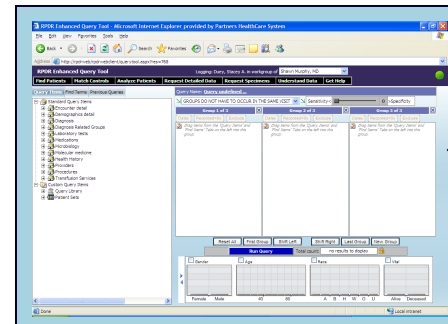
- Set of patients is selected from medical record data in a high throughput fashion
- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data
- Using the Crimson system, tissues of these patients can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

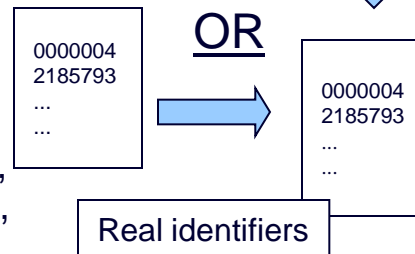


Encrypted identifiers

Z731984X
Z74902XX
...

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.

A screenshot of a Microsoft Access database window. It displays a table with columns: Task ID, Task Description, Result, Result Text, Abnormal Flag, Reference, and Interval Range. The data includes various medical tasks like "Superior APFT" and "SILT HEMOLYSIS" with their corresponding results and flags.

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.

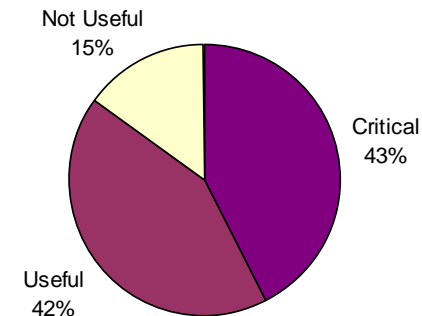
Security and Patient Confidentiality of Step 2

- 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.

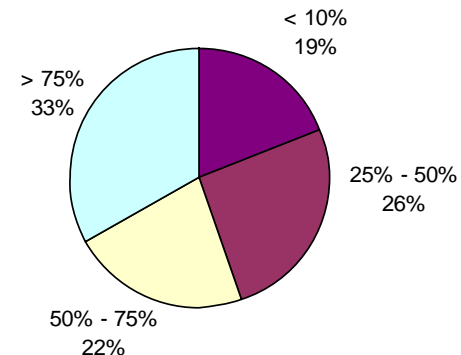
2009's usage of RPDR

- 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

Usefulness of Detailed Data
106 Total Responses

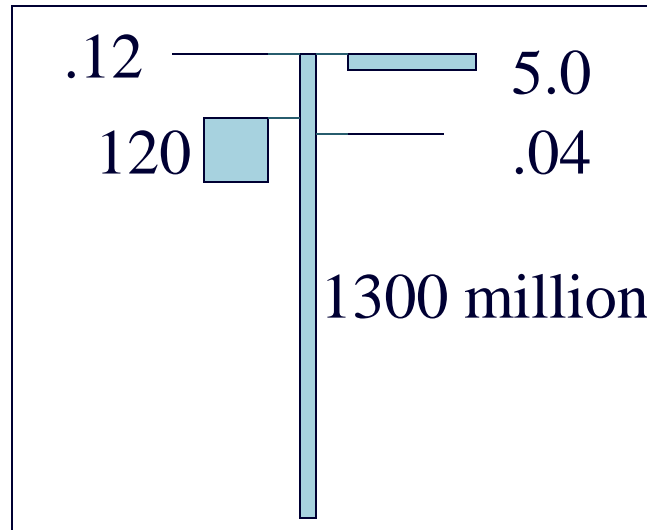
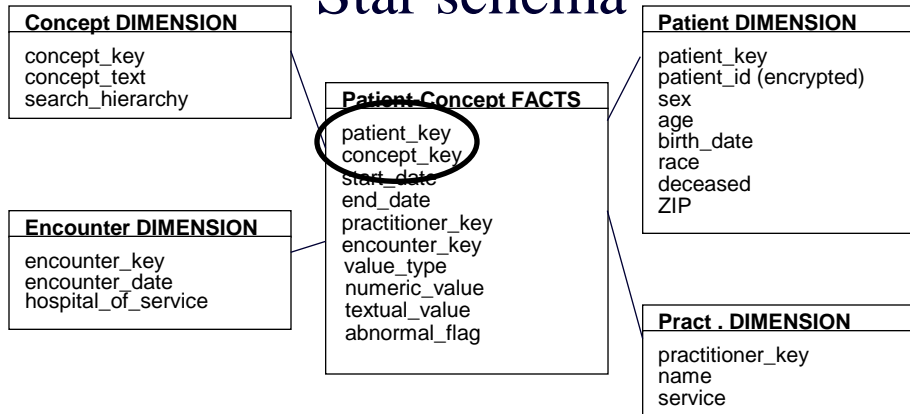


% of Patients Who Fit Required Profile
105 Total Responses



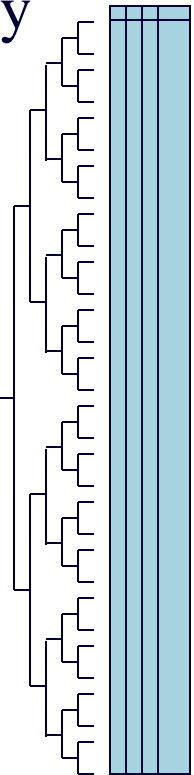
Organizing data in the Clinical Data Warehouse

Star schema



Binary Tree

start
search →



FINDING PATIENTS

Query items

Person who is using tool

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

Tools Help

Address <http://rpdweb/rpdwebclient/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

Standard Query Items

- Encounter detail
- Demographics detail
- Diagnosis
- Diagnosis Related Groups
- Laboratory tests
- Medications
- Microbiology
- Molecular medicine
- Health History
- Providers
- Procedures
- Transfusion Services

Custom Query Items

- Query Library
- Patient Sets

Query Name: Query undefined ...

☒ GROUPS DO NOT HAVE TO OCCUR IN THE SAME VISIT ☐ Sensitivity > Specificity

Group 1 of 3

Dates Recorded > 0x Exclude

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

Group 2 of 3

Dates Recorded > 0x Exclude

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

Group 3 of 3

Dates Recorded > 0x Exclude

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

Reset All Run Query Total count: no results to display

Gender Age Race Vital

Female Male 40 80 A B H W O U Alive Deceased

Done Local intranet

Query construction

Results - broken down by number distinct of patients

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

- 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

Dates Recorded>0x Exclude

One or more items recorded

EGFR

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

Group 2 of 3

Dates Recorded>0x Exclude

One or more items recorded

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
Malignant neoplasm of

Group 3 of 3

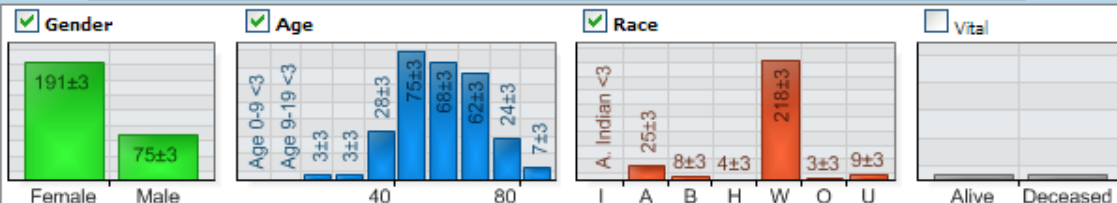
Dates Recorded>0x Exclude

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)



MATCHING PATIENTS

Previous query items

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

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

Query Items | Find Terms | Previous Queries

☐ Show Data Requests [Refresh List](#)

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 15:14:14
EGFR, Respiratory and... on 01/2...	as084	Ready	01/24/11 15:03:00
Acute myocardia..., CK-MB Index ...	as084	Ready	01/24/11 14:59:11
AMI and CK-MB >3.5	as084	Ready	01/24/11 14:59:12
Just Diagnos AMI	as084	Ready	01/24/11 14:55:34
EGFR, Respiratory and... on 01/2...	as084	Ready	01/24/11 13:14:11
Acute myocardia..., CK-MB Index ...	as084	Ready	01/24/11 13:05:33
Acute myocardia..., CK-MB Index ...	as084	Ready	01/24/11 13:04:35
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 debr..., Ilestin il po...	snm0	Ready	01/24/11 07:40:30
Resection or debr..., 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 Explain

IDENTIFY CASES

Query for which you want a matched set of patients:

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...

Patients should be included from: ☐ MGH ☒ BWH/FH

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

drag and drop a query here...

for specific purposes, 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

Submit request to find set of Control patients Reset All

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

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)

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

Help

< Back

Step 9

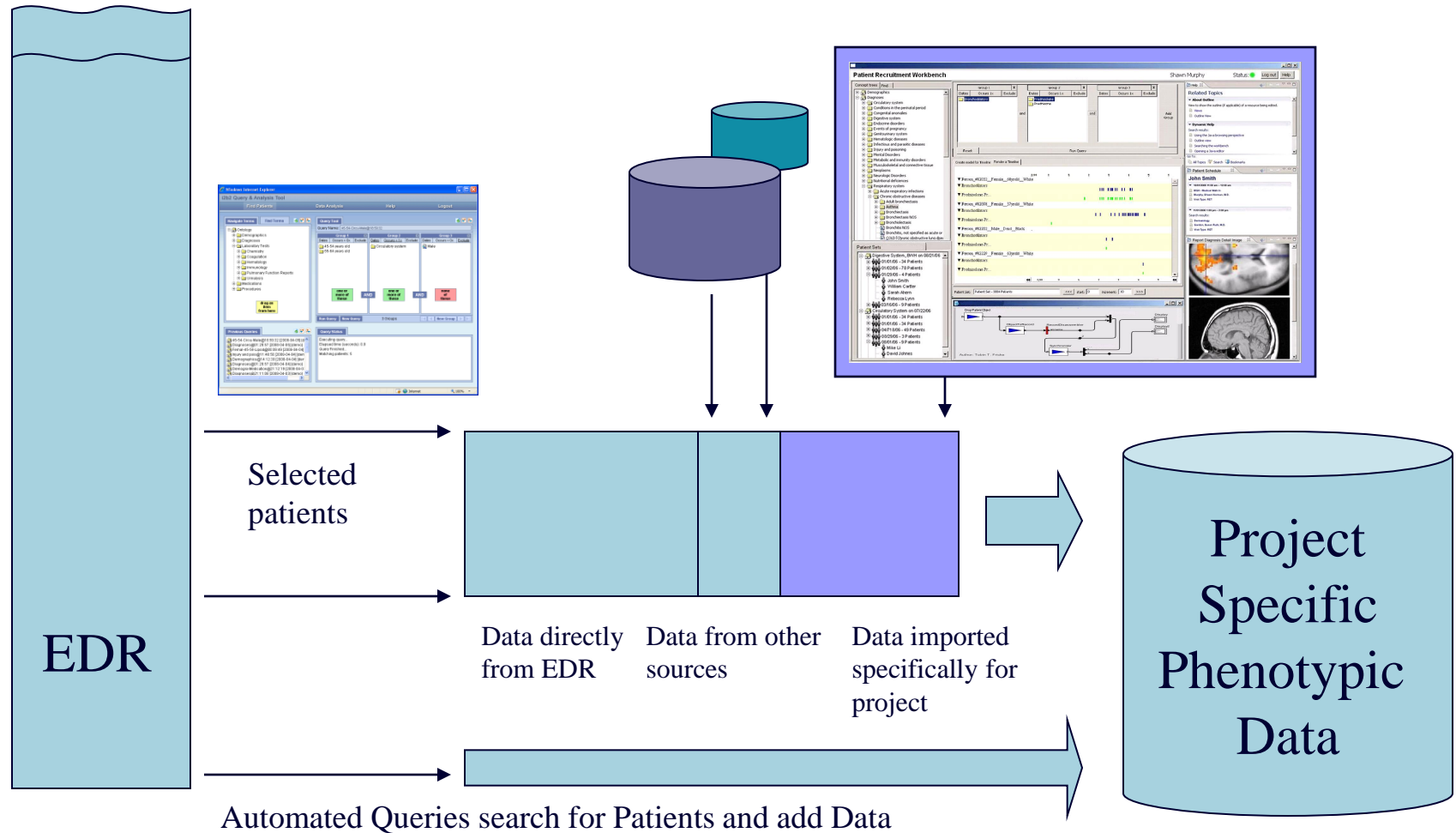
Next >

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

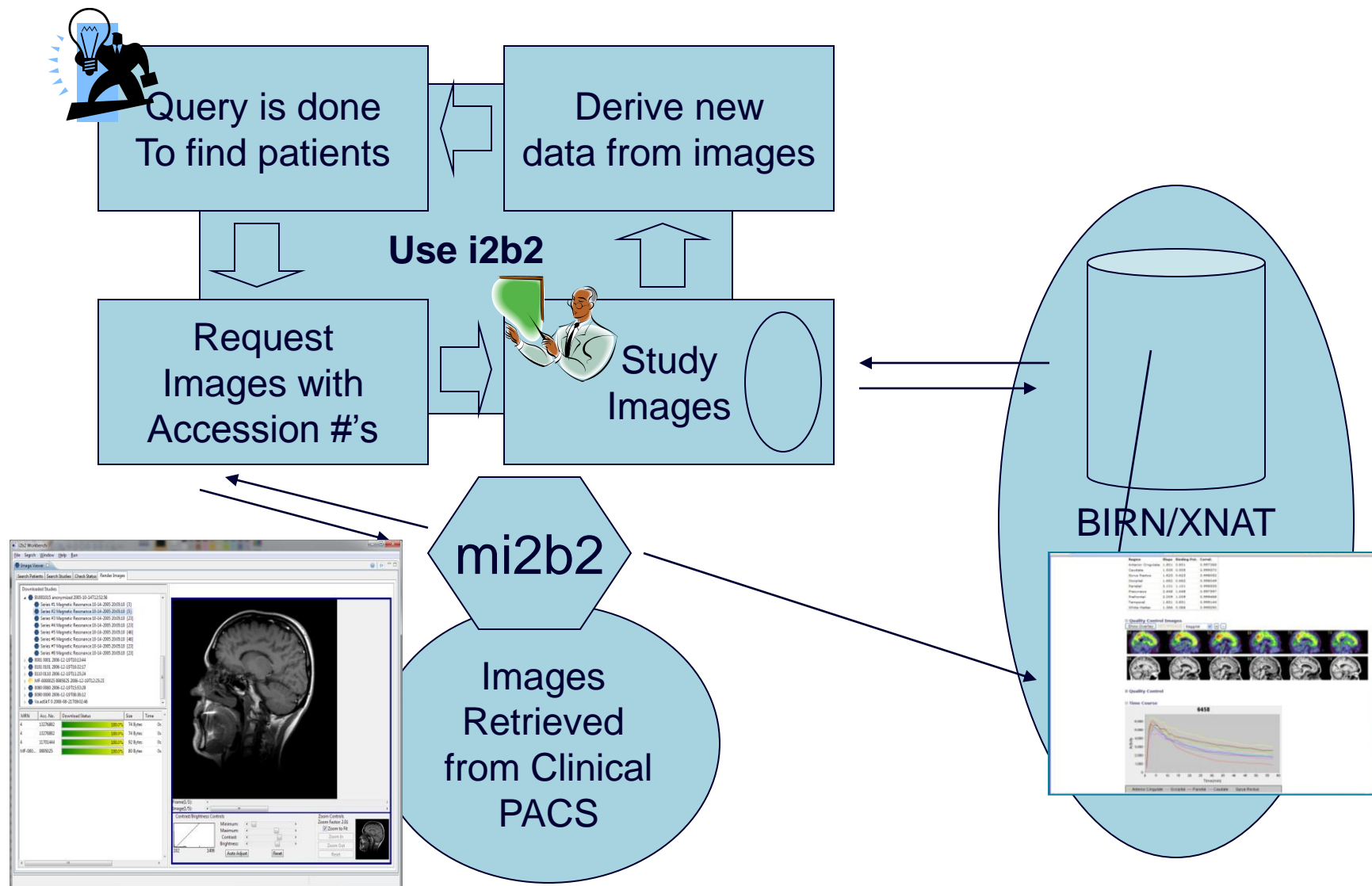
Set of patients is selected through Enterprise Repository and data is gathered into a data mart



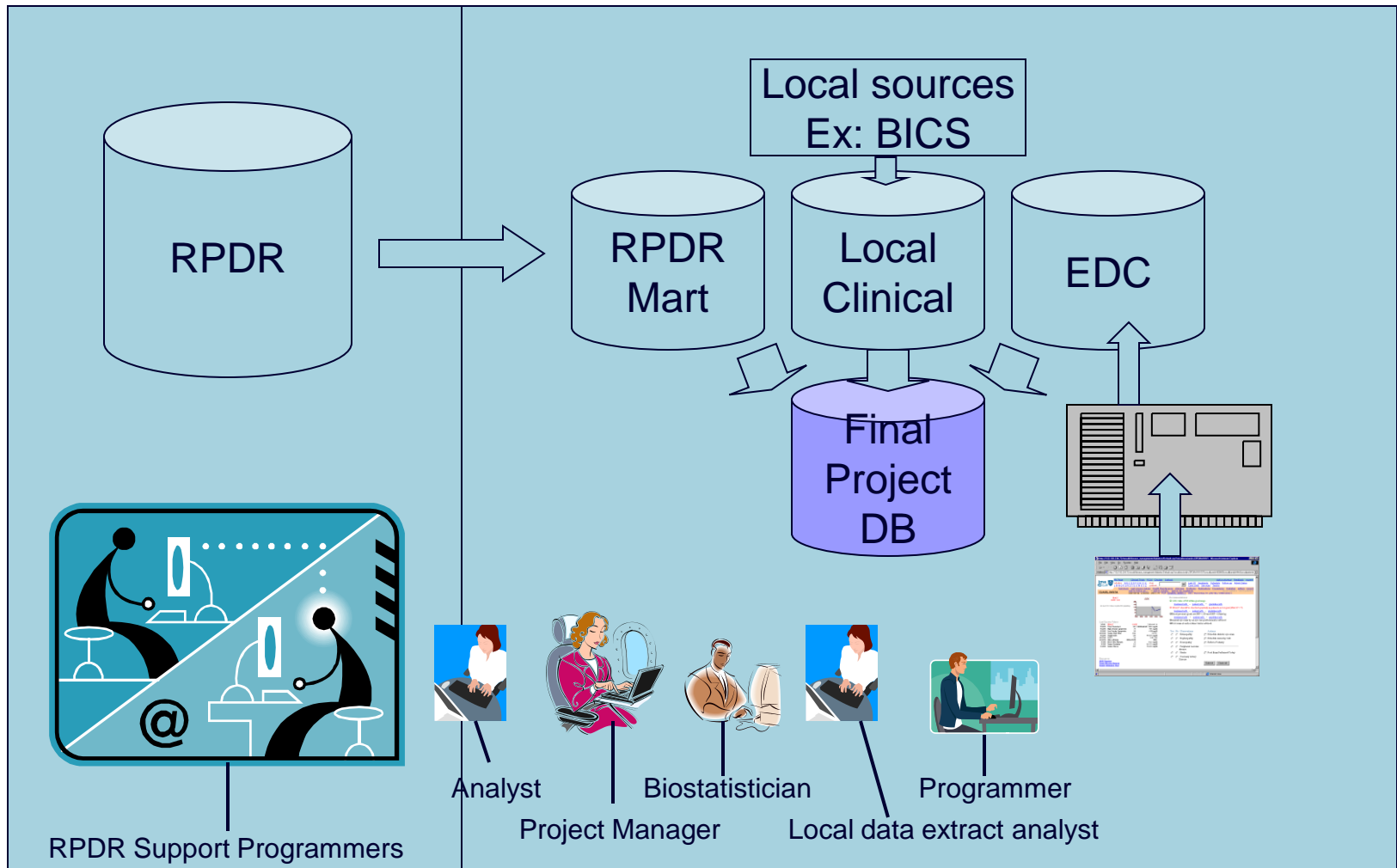
Data is available through the i2b2 Workbench



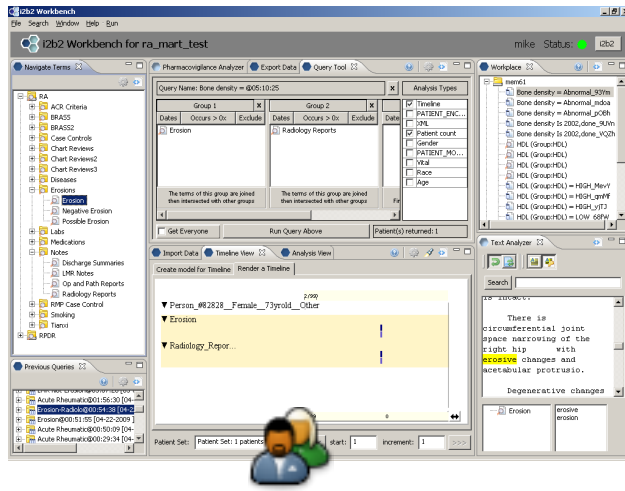
Research Investigator Workflow enabled by mi2b2



Team support for Projects



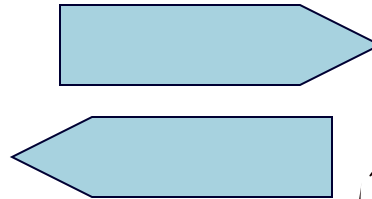
NLP Workflow



I2b2 Project Investigators

Results Delivery

Communication



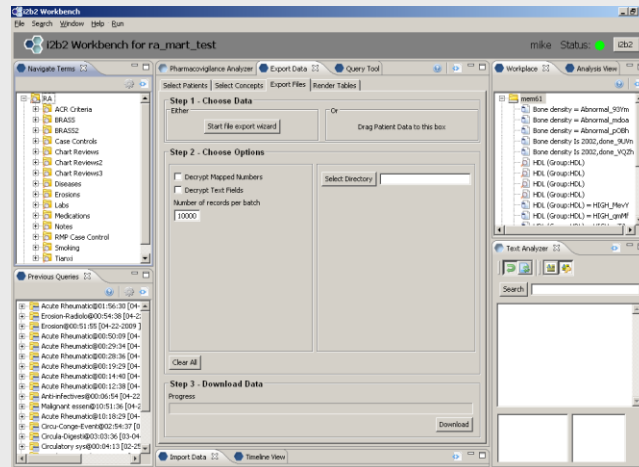
NLP Specialists

NLP (and comedy) is not pretty

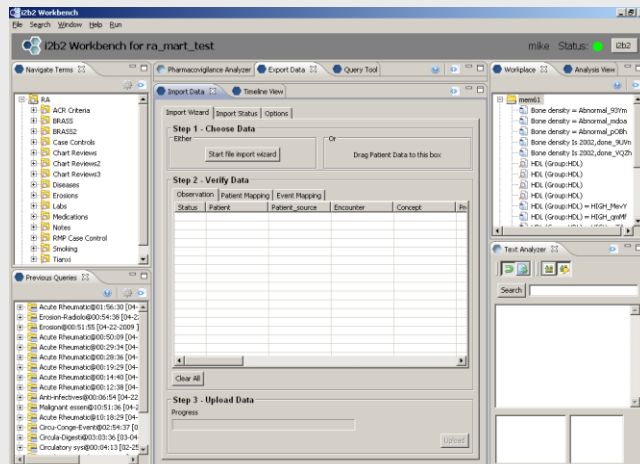
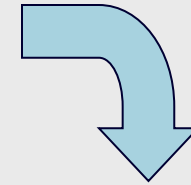
The image shows a medical chart with several text boxes and yellow arrows pointing to specific information:

- SOCIAL HISTORY:** The patient is married with four grown daughters, **uses tobacco**, has wine with dinner. (Arrow: **Smoker**)
- SOCIAL HISTORY:** The patient is a **nonsmoker**. No alcohol. (Arrow: **Non-Smoker**)
- SOCIAL HISTORY:** **Negative for tobacco**, alcohol, and IV drug abuse. (Arrow: **Non-Smoker**)
- BRIEF RESUME OF HOSPITAL COURSE:** 63 yo woman with COPD, **50 pack-yr tobacco (quit 3 wks ago)**, SpO2 92% on 2L. (Arrow: **Past Smoker**)
- SOCIAL HISTORY:** The patient lives in rehab, married. **Unclear smoking** history from the admission note... (Arrow: **???**)
- HOSPITAL COURSE:** ... It was recommended that she receive ... We also added Lactinax, oral form of **Lactobacillus** acidophilus to attempt to re-populate her gut. (Arrow: **Hard to pick**)
- SH:** widow, lives alone, 2 children, no **tob/alcohol**. (Arrow: **Hard to pick**)

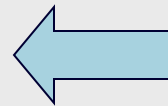
NLP Specialists Workstation



Export Notes



Import
Derived
Codes



NLP Specialists

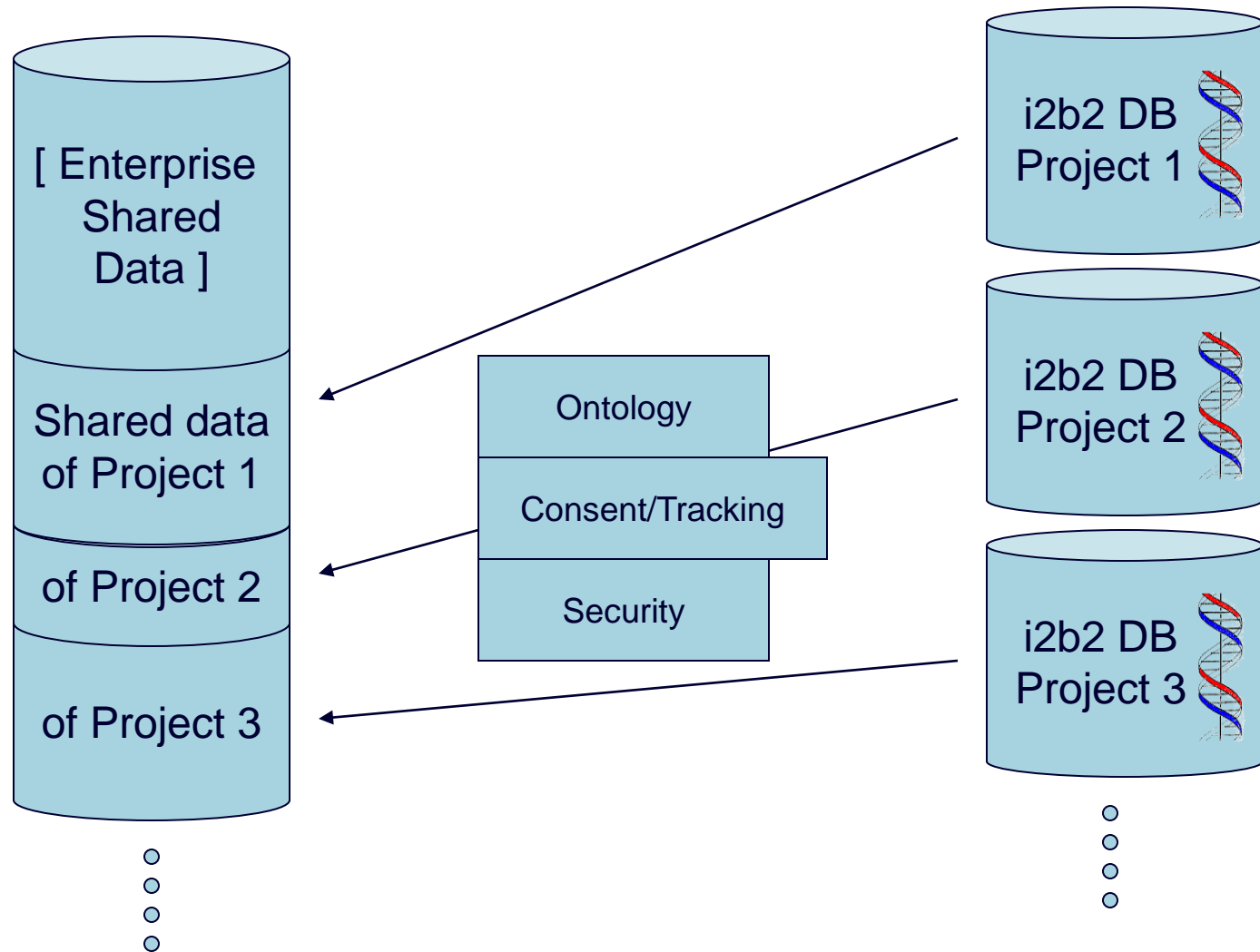
Investigator Review

The screenshot displays the i2b2 Workbench interface for a user named 'mike'. The main window is titled 'i2b2 Workbench for ra_mart_test'. The interface is divided into several panes:

- Navigate Terms:** A tree view on the left showing a hierarchy of medical terms. The 'Erosions' category is expanded, showing sub-terms like 'Erosion', 'Negative Erosion', and 'Possible Erosion'.
- Pharmacovigilance Analyzer:** The central pane shows a query named 'Bone density = @05:10:25'. It includes a table for 'Analysis Types' with columns for 'Dates', 'Occurs > 0x', and 'Exclude'. The 'Erosion' term is selected in the 'Dates' column. Below the table, there are buttons for 'Get Everyone', 'Run Query Above', and 'Patient(s) returned: 1'.
- Workplace:** A pane on the right showing a list of terms related to the query, including 'Bone density = Abnormal_93Ym', 'Bone density = Abnormal_mdoa', 'Bone density = Abnormal_p0Bh', 'Bone density Is 2002,done_9UvN', 'Bone density Is 2002,done_VQZh', 'HDL (Group:HDL)', 'HDL (Group:HDL)', 'HDL (Group:HDL)', 'HDL (Group:HDL) = HIGH_MevY', 'HDL (Group:HDL) = HIGH_qmMf', 'HDL (Group:HDL) = HIGH_yJTJ', and 'HDL (Group:HDL) = LOW_68FW'.
- Text Analyzer:** A pane at the bottom right showing a search for 'erosive' and 'erosion' in a text document. The search results show 'erosive' and 'erosion' in a list.
- Timeline View:** The bottom pane shows a timeline view of the query results. It includes a search bar, a list of terms, and a timeline plot. The plot shows a single data point for 'Erosion' at a specific time point. The timeline is labeled 'Person_#82828_Female_73yroid_Other'.

The bottom of the interface shows a 'Patient Set' section with a dropdown menu set to 'Patient Set: 1 patients', a 'start' field set to 1, and an 'increment' field set to 1.

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
- 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

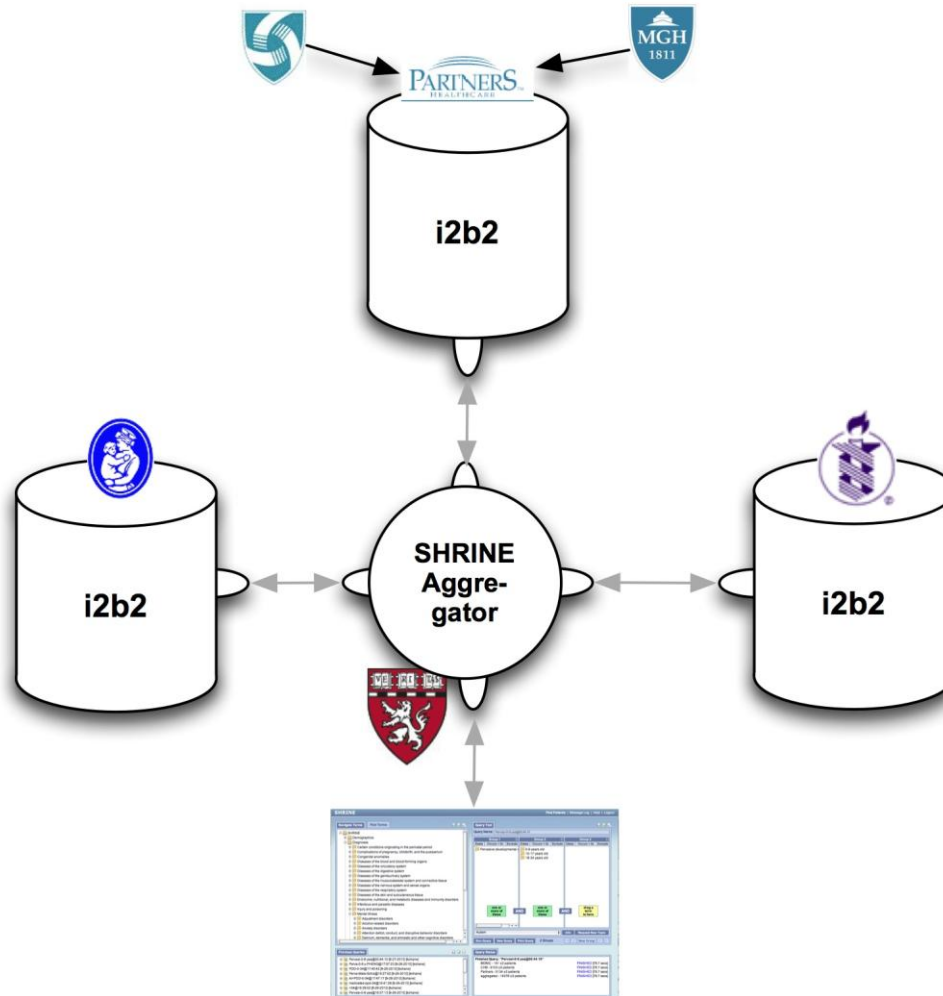
International

- Georges Pompidou 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)

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 amnesic and other cognitive disorders

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 Tool



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

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Pervasive developmental c			0-9 years old 10-17 years old 18-34 years old					

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

Query Status

Finished Query: "Pervasi-0-9 yea@00:44:10"

BIDMC - 141 ±3 patients

FINISHED [78.7 secs]

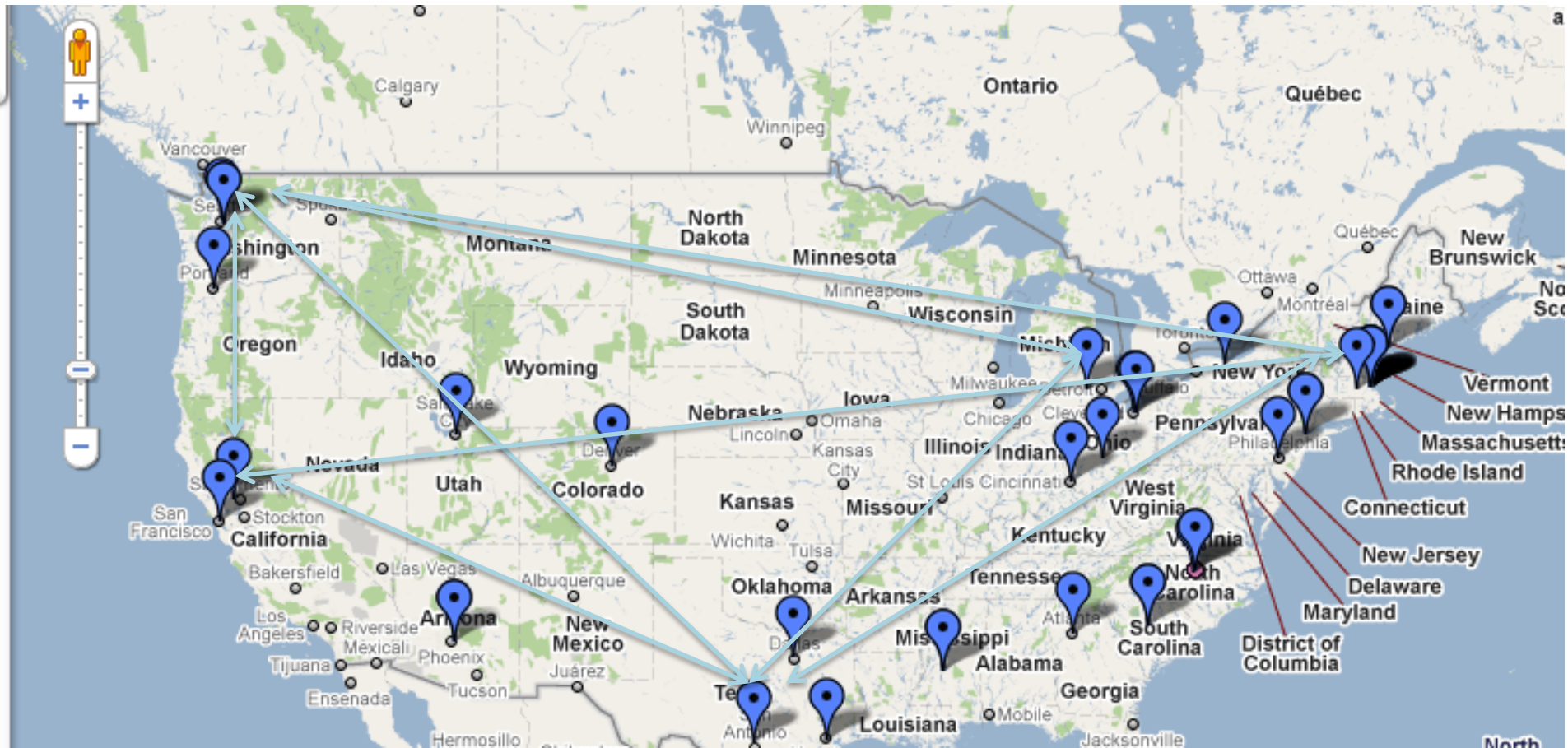
CHB - 9103 ±3 patients

FINISHED [78.7 secs]

Partners - 5134 ±3 patients

FINISHED [78.7 secs]

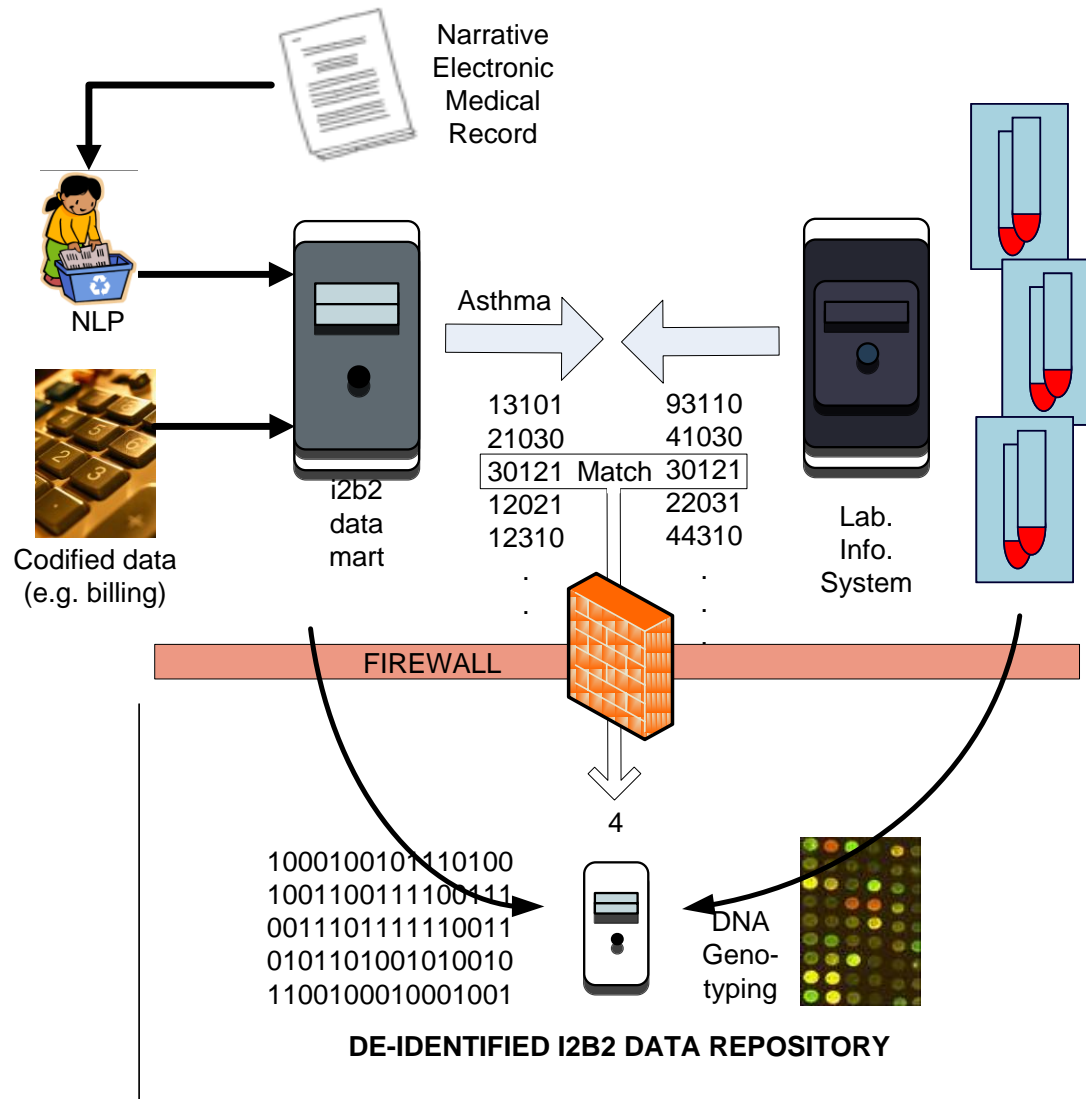
© 2013 Pearson Education, Inc. or its affiliate(s). All rights reserved. This material is intended solely for the personal use of the individual user and is not to be disseminated broadly.



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



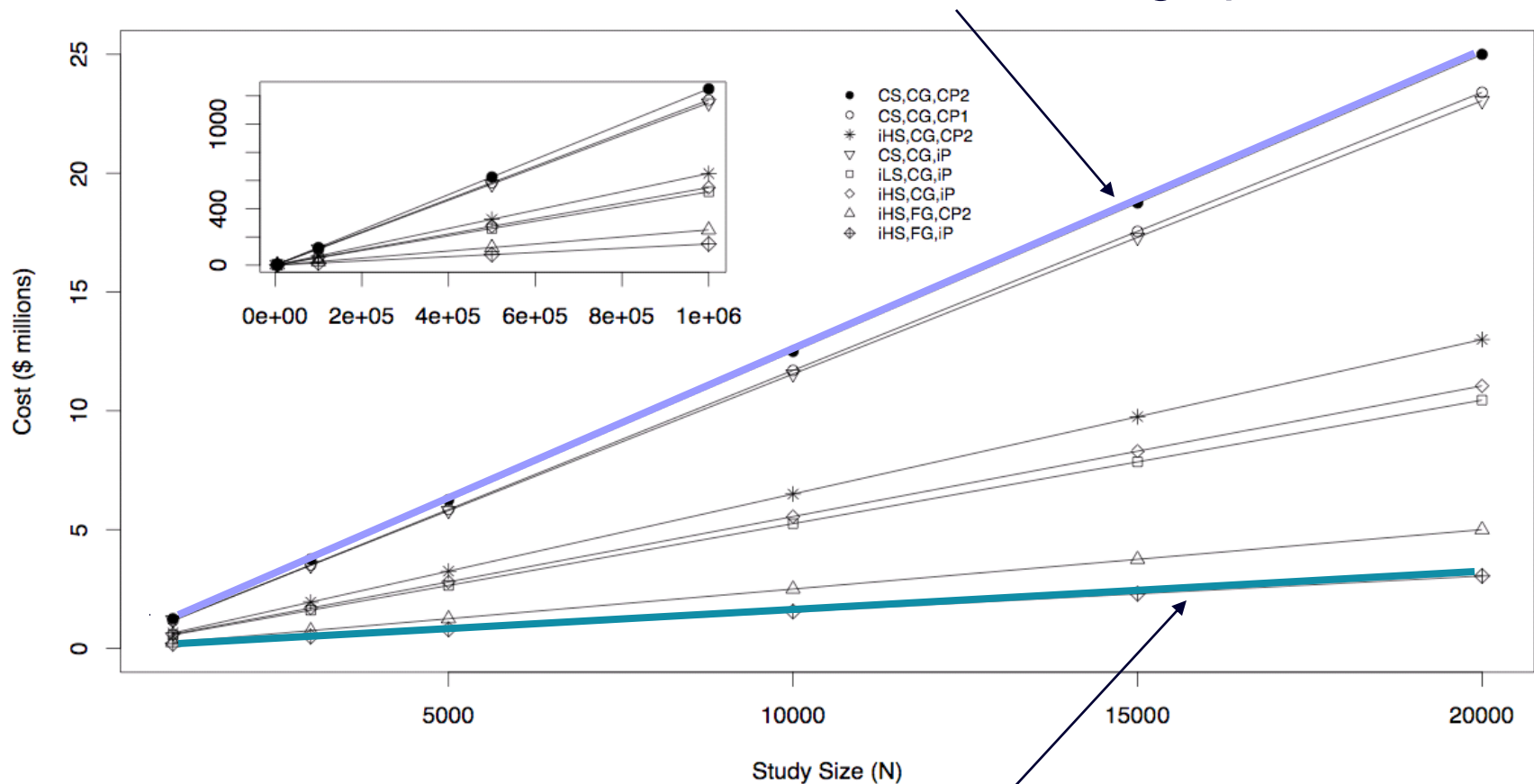
Cost and time benefit of Instrumenting with Sample Collection for Modest-size Study with 10,000 subjects (cases + controls)

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 /week²

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

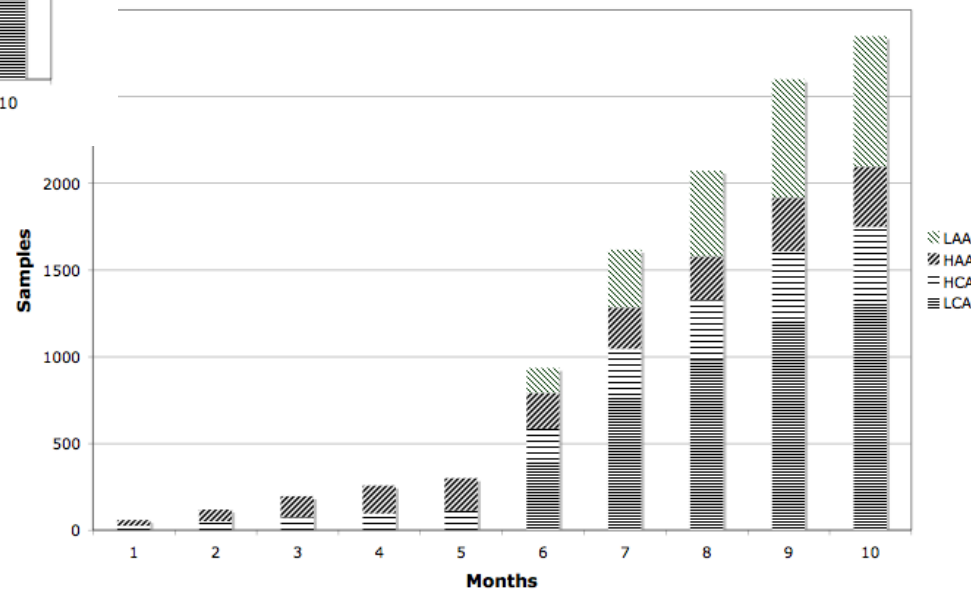
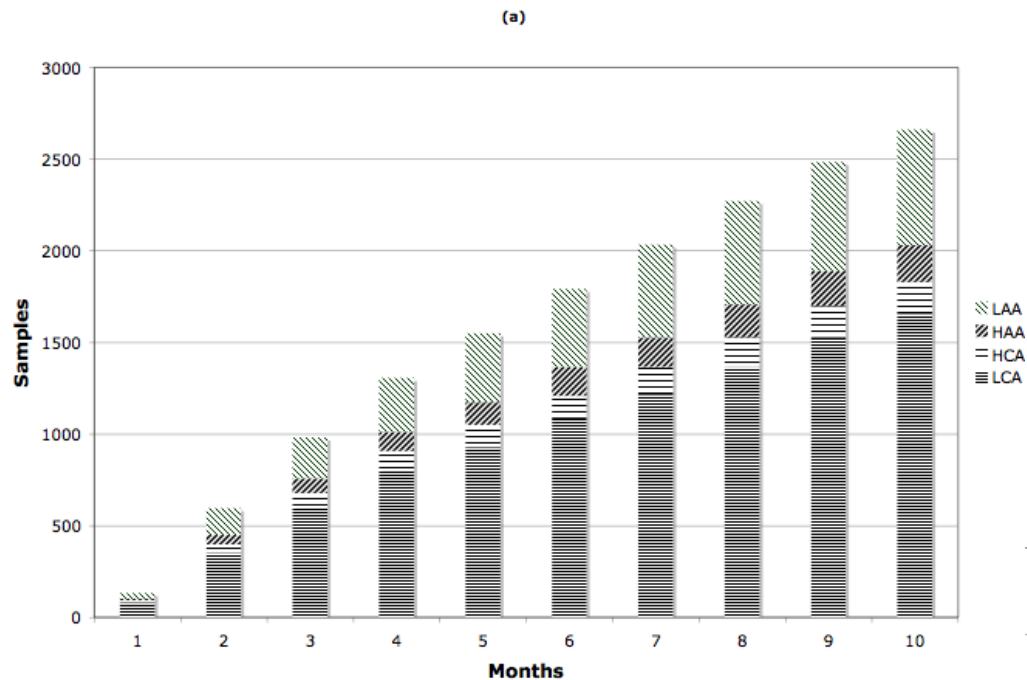
Escalating cost and time benefit of Instrumenting with Sample Collection

Previous model for collecting specimens

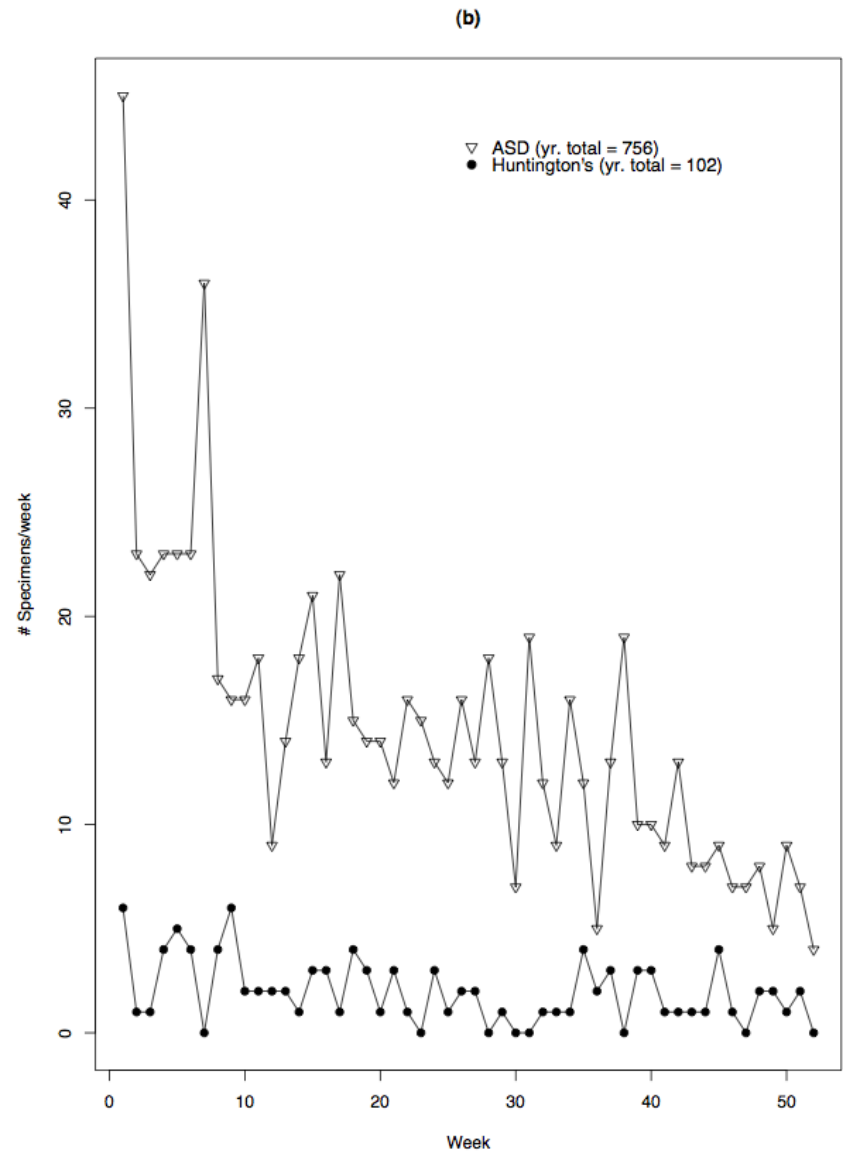
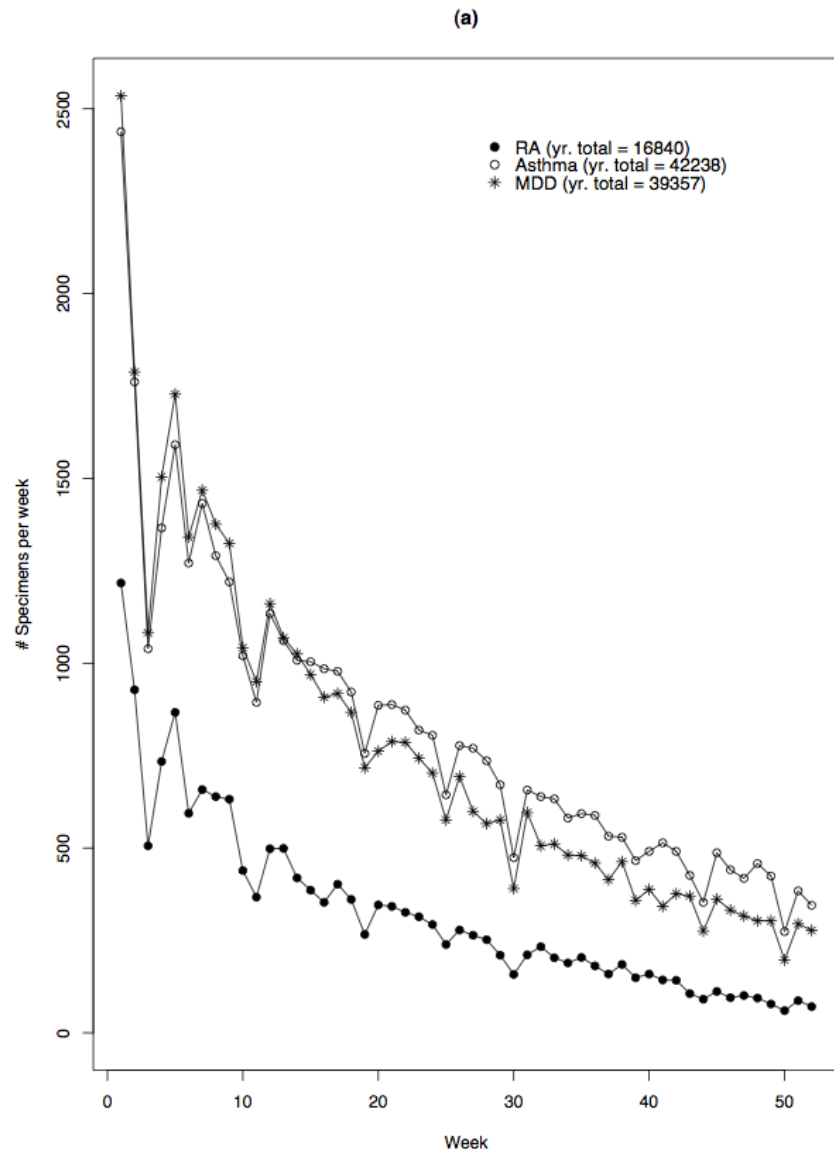


New model for collecting specimens

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

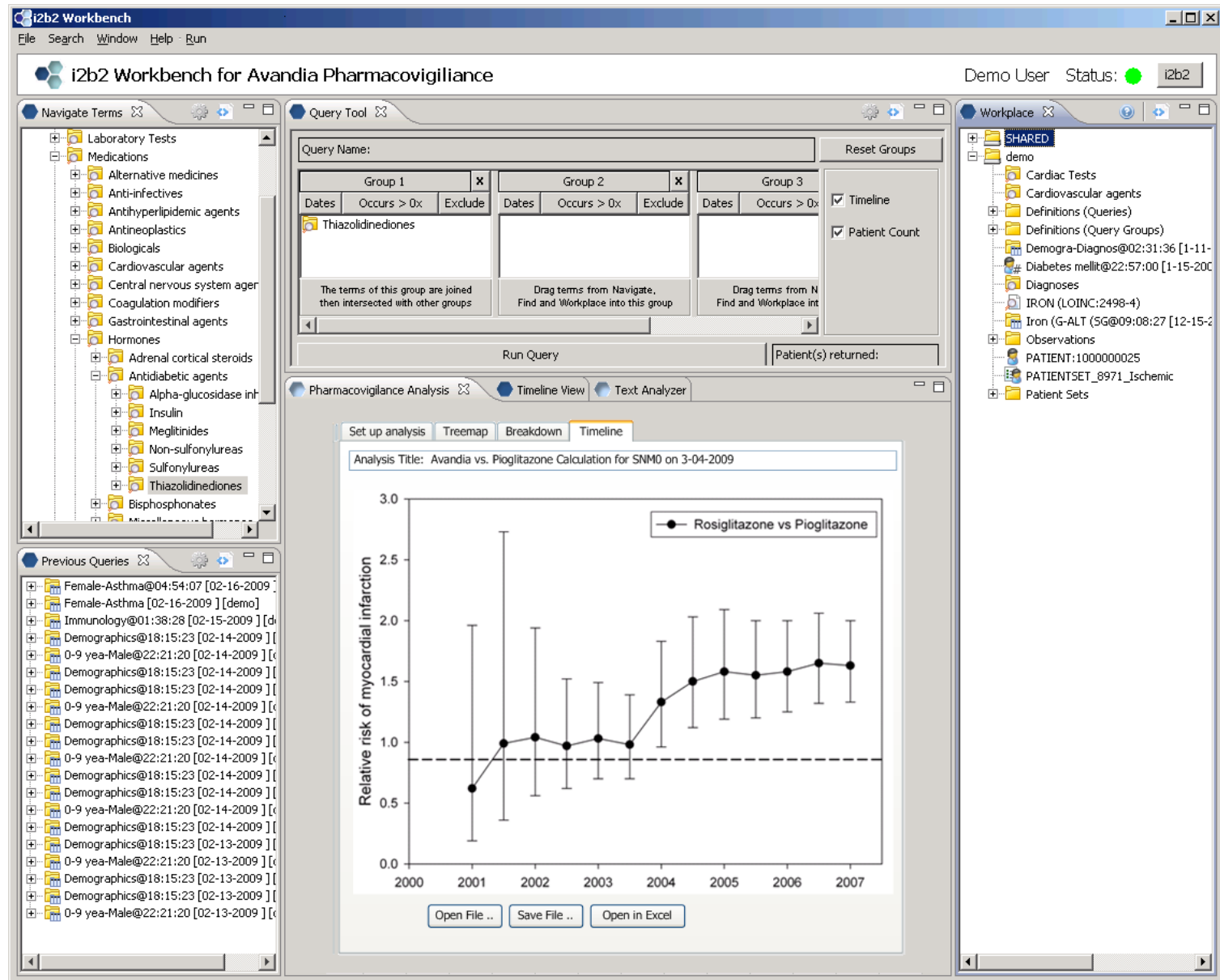
The screenshot displays the i2b2 Workbench interface for Avandia Pharmacovigilance. The top menu bar includes File, Search, Window, Help, and Run. The main window is titled "i2b2 Workbench for Avandia Pharmacovigilance" and shows a Demo User status.

The interface is divided into several panes:

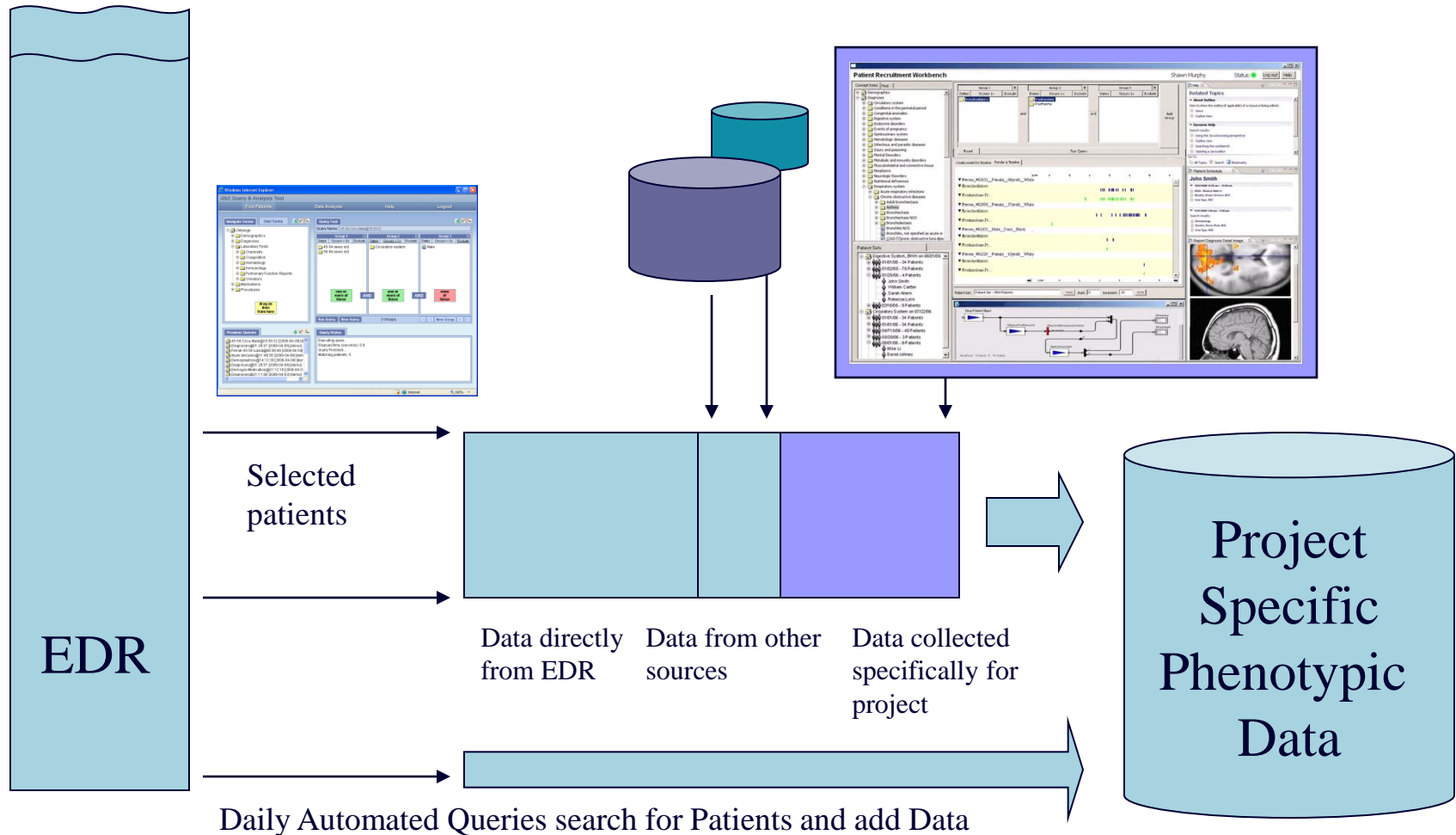
- Navigate Terms:** A tree view on the left showing categories like Laboratory Tests, Medications, and various drug classes (e.g., Alternative medicines, Anti-infectives, Antihyperlipidemic agents).
- Query Tool:** A central pane for building queries. It includes a "Query Name" field, "Reset Groups" button, and three groups (Group 1, Group 2, Group 3) for selecting terms. It also has checkboxes for "Timeline" and "Patient Count", and a "Run Query" button.
- Workplace:** A right-hand pane showing a list of shared queries and patient sets, including "Cardiac Tests", "Cardiovascular agents", "Definitions (Queries)", "Demogra-Diagnos@02:31:36 [1-11-2009]", "Diabetes mellit@22:57:00 [1-15-2009]", "Diagnoses", "IRON (LOINC:2498-4)", "Iron (G-ALT (SG@09:08:27 [12-15-2009]", "Observations", "PATIENT:1000000025", "PATIENTSET_8971_Ischemic", and "Patient Sets".
- Pharmacovigilance Analysis:** A bottom pane showing analysis results. It includes tabs for "Set up analysis", "Treemap", "Breakdown", and "Timeline". The "Treemap" view is active, displaying a hierarchical heatmap of analysis results. The analysis title is "Avandia vs. Pioglitazone Calculation for SNM0 on 3-04-2009". The heatmap shows various medical categories (e.g., Circulatory system, Digestive system, Endocrine disorders) and their associated counts. A color scale on the right indicates the magnitude of the signals.

At the bottom of the Pharmacovigilance Analysis pane, there are buttons for "Open File...", "Save File...", and "Open in Excel".

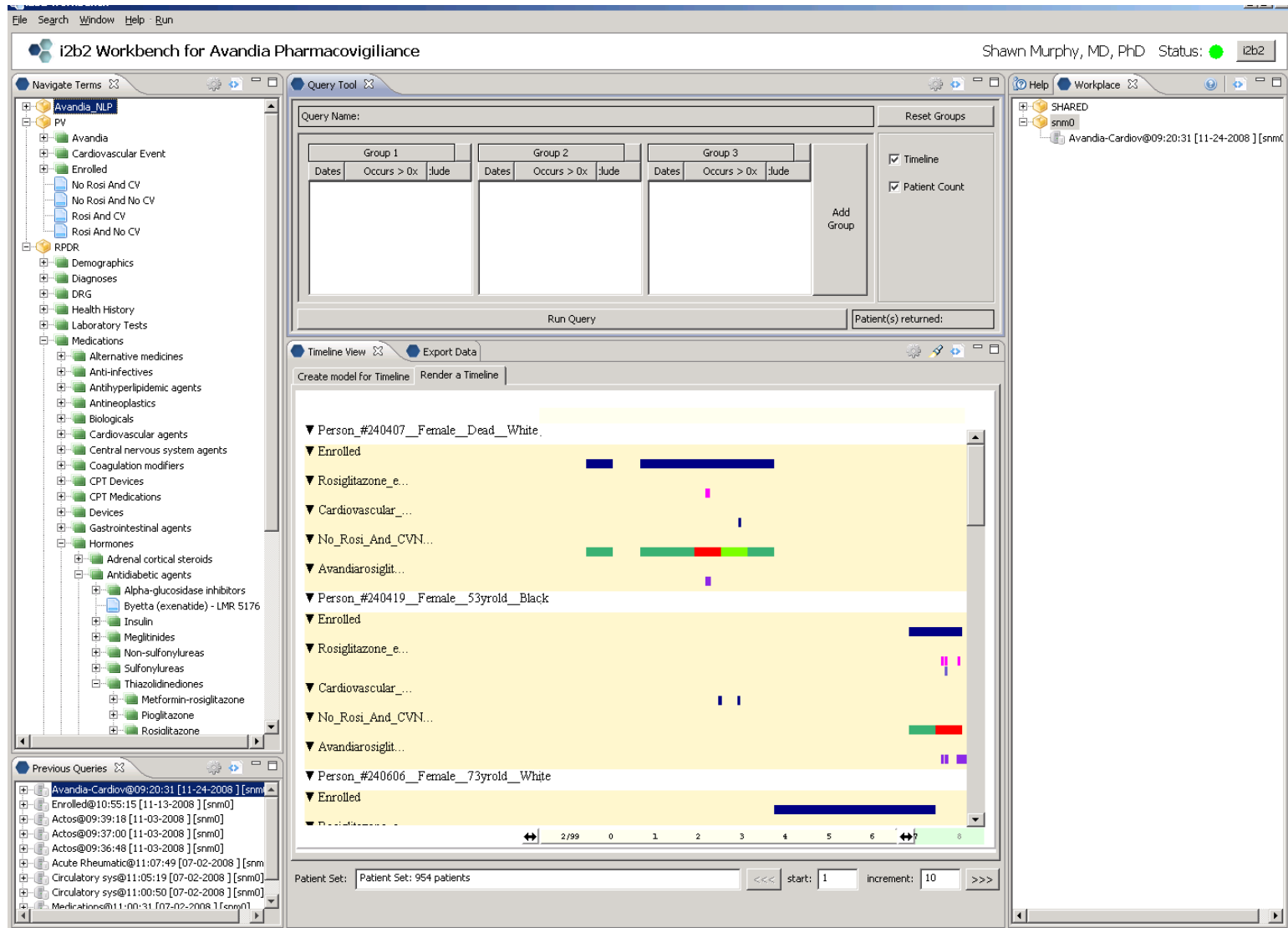
Dashboard used to observe high-level signals



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