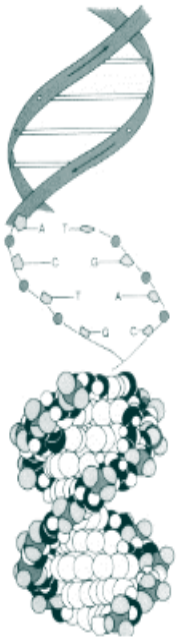
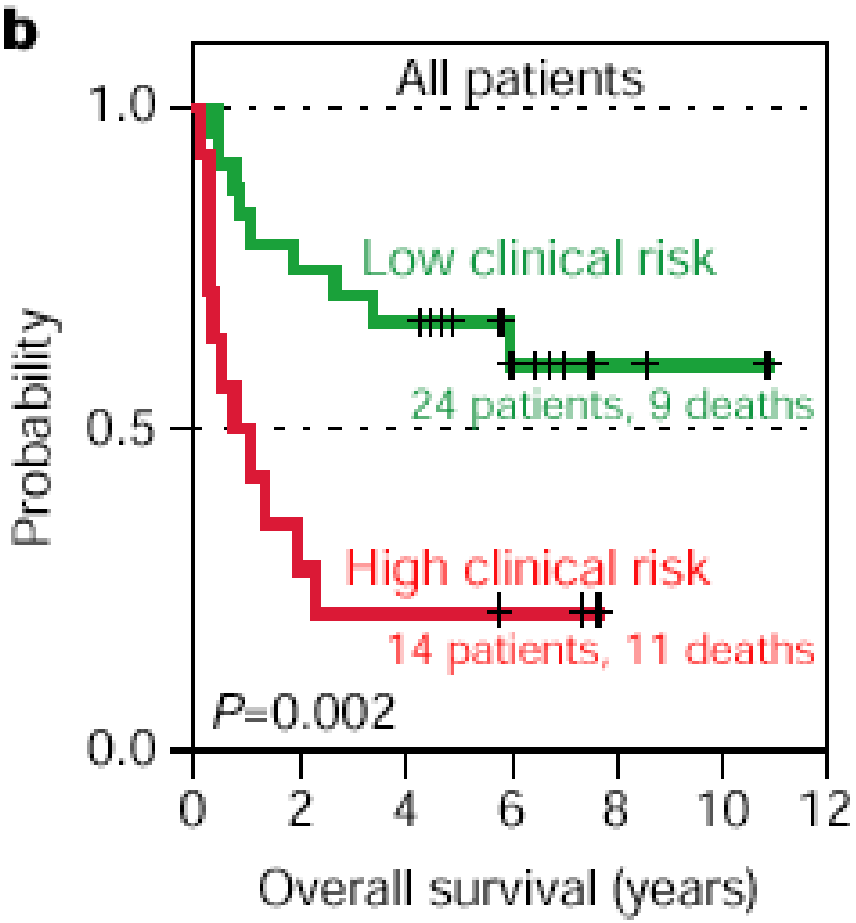


I2b2 - 10 Strategic Lessons in Instrumenting the Health Care Enterprise for Discovery Research in the Genomic Era

Shawn Murphy MD, Ph.D.

The paper that Launched at 100,000 chips: Alizadeh 2000



Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

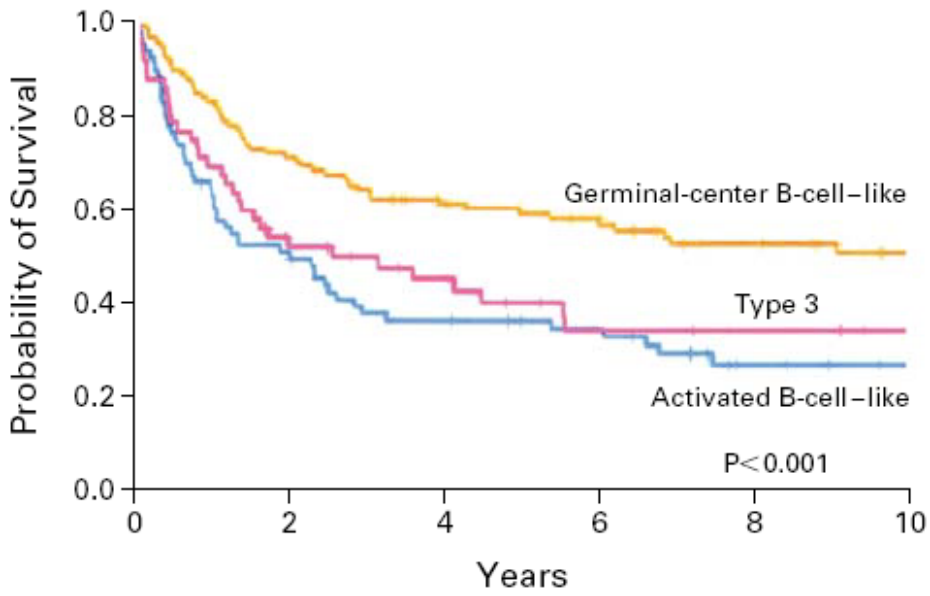
Ash A. Alizadeh[1,2], Michael B. Eisen[2,3,4], R. Eric Davis[5], Chi Ma[5], Izidore S. Lossos[6], Andreas Rosenwald[5], Jennifer C. Boldrick[1], Hajeer Sabet[5], Truc Tran[5], Xin Yu[5], John I. Powell[7], Liming Yang[7], Gerald E. Marti[8], Troy Moore[9], James Hudson Jr[9], Lisheng Lu[10], David B. Lewis[10], Robert Tibshirani[11], Gavin Sherlock[4], Wing C. Chan[12], Timothy C. Greiner[12], Dennis D. Weisenburger[12], James O. Armitage[13], Roger Warnke[14], Ronald Levy[6], Wyndham Wilson[15], Michael R. Grever[16], John C. Byrd[17], David Botstein[4], Patrick O. Brown[1,18] & Louis M. Staudt[5]

Departments of [1]Biochemistry, [3]Genetics, [14]Pathology, [6]Medicine, [10]Pediatrics and [11]Health Research & Policy and Statistics, and [18]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA
[5] Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[7] Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland 20892, USA
[8] CBER, FDA, Bethesda, Maryland 20892, USA
[9] Research Genetics, Huntsville, Alabama 35801, USA
Departments of [12]Pathology and Microbiology, and [13]Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA
[15] Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[16] Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA
[17] Walter Reed Army Medical Center, Washington, DC 20307, USA
[2] These authors contributed equally to this work

Alizadeh et al., Nature 2000

Example: Rosenwald

C



No. at Risk						
Germinal-center B-cell–like	115	81	60	46	32	19
Type 3	52	24	18	10	8	5
Activated B-cell–like	73	35	23			

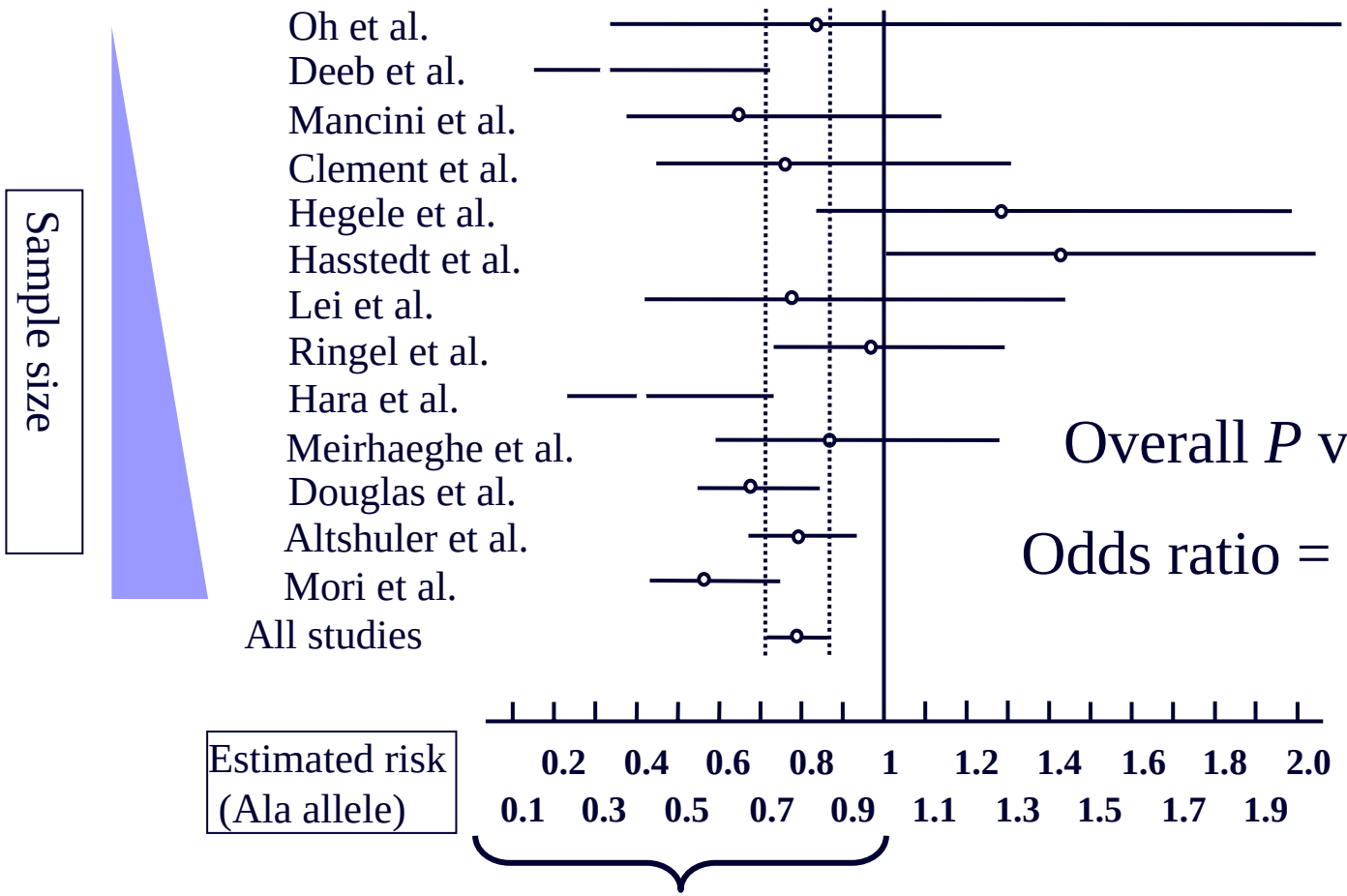
The New England Journal of Medicine

VOLUME 346 JUNE 20, 2002 NUMBER 25

THE USE OF MOLECULAR PROFILING TO PREDICT SURVIVAL AFTER CHEMOTHERAPY FOR DIFFUSE LARGE-B-CELL LYMPHOMA

ANDREAS ROSENWALD, M.D., GEORGE WRIGHT, PH.D., WING C. CHAN, M.D., JOSEPH M. CONNORS, M.D., ELIAS CAMPO, M.D., RICHARD I. FISHER, M.D., RANDY D. GASCOYNE, M.D., H. KONRAD MULLER-HERMELINK, M.D., ERLEND B. SMELAND, M.D., PH.D., AND LOUIS M. STAUDT, M.D., PH.D., FOR THE LYMPHOMA/LEUKEMIA MOLECULAR PROFILING PROJECT

Example: PPARγ Pro12Ala and Diabetes

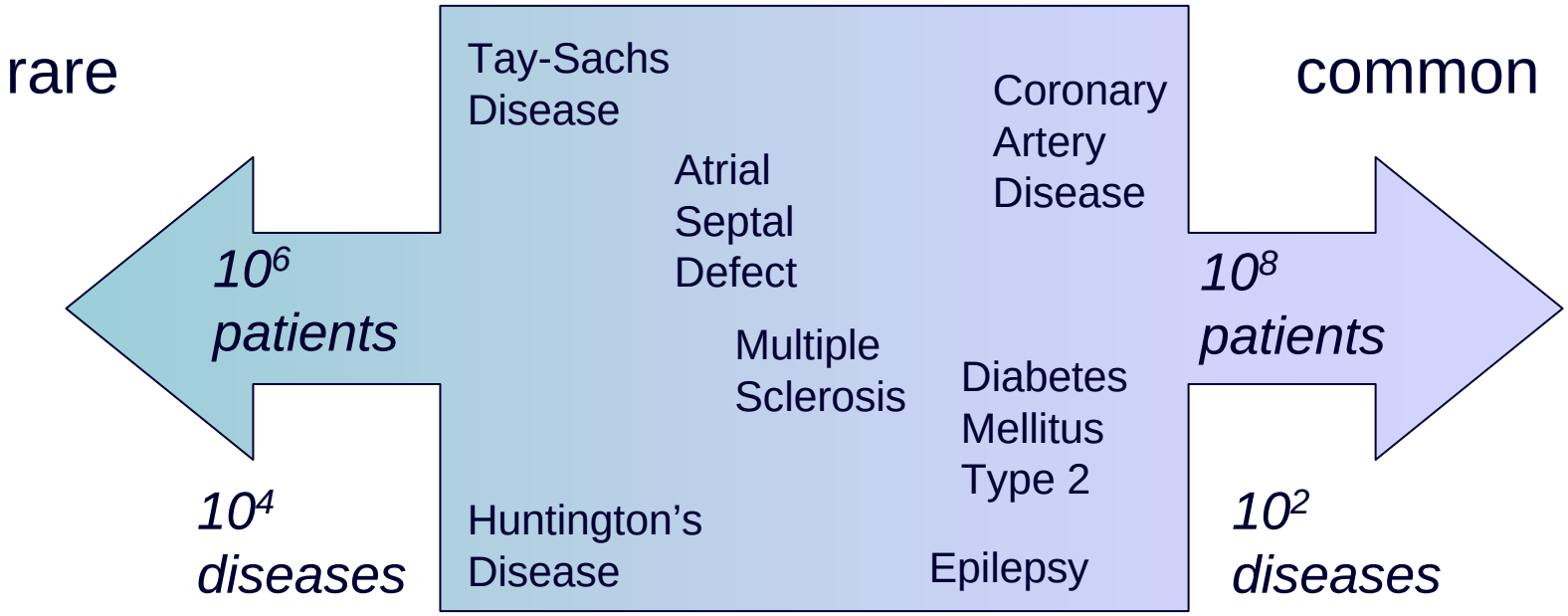


Overall *P* value = 2×10^{-7}

Odds ratio = 0.79 (0.72-0.86)

Courtesy J. Hirschhorn

Common-Rare: Weak-Strong Spectrum



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

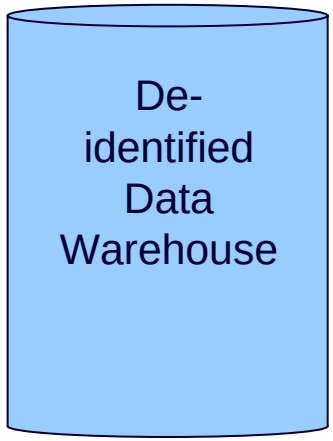
- Set of patients is selected from medical record data in a high throughput fashion
  - Partners Research Computing – Eugene Brunwald, John Glaser, Diane Keogh
- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical 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

Current Operations of the Research Patient Data Registry

Queries for aggregate patient numbers

- Warehouse of in & outpatient clinical data
- 4.6 million Partners Healthcare patients
- 1.2 billion diagnoses, medications, procedures, laboratories, physical findings, & genomics coupled to demographics & visits
- Authorized use by faculty status
- Clinicians can construct complex queries
- Queries cannot identify individuals, internally can produce identifiers for detailed data sets

Query construction in web tool



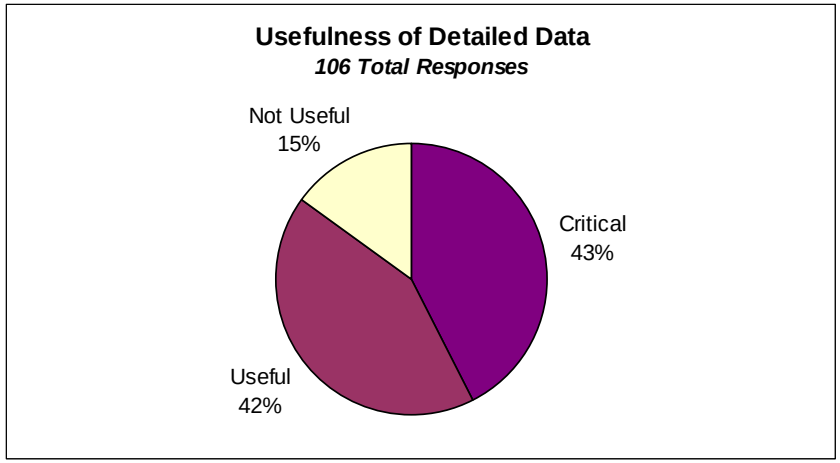
Z731984X
Z74902XX
...
..

Encrypted identifiers

1,970 registered users, 332 new in 2008

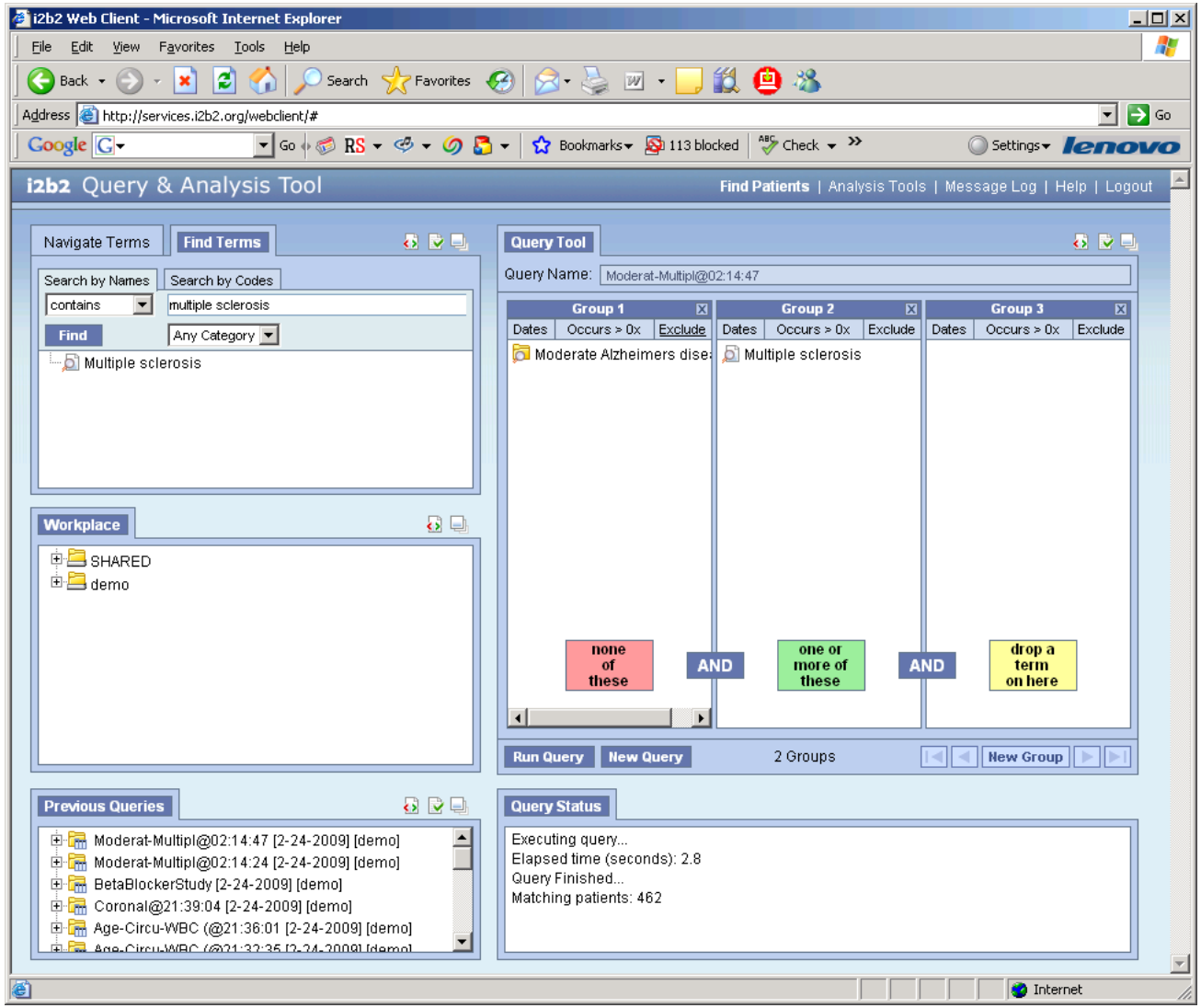
915 detailed data sets = research study returned to these teams, containing data for of 8.8 million patient records.

$94-136 million total research support critically dependent on RPDR from patient data received throughout life of funding.

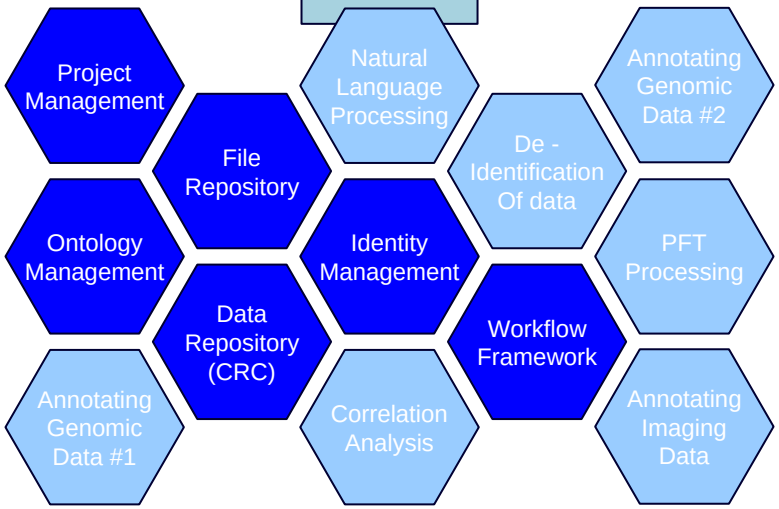


Enterprise web client

http://services.i2b2.org/webclient/



Enterprise-wide repurposing and distribution of medical record data for research

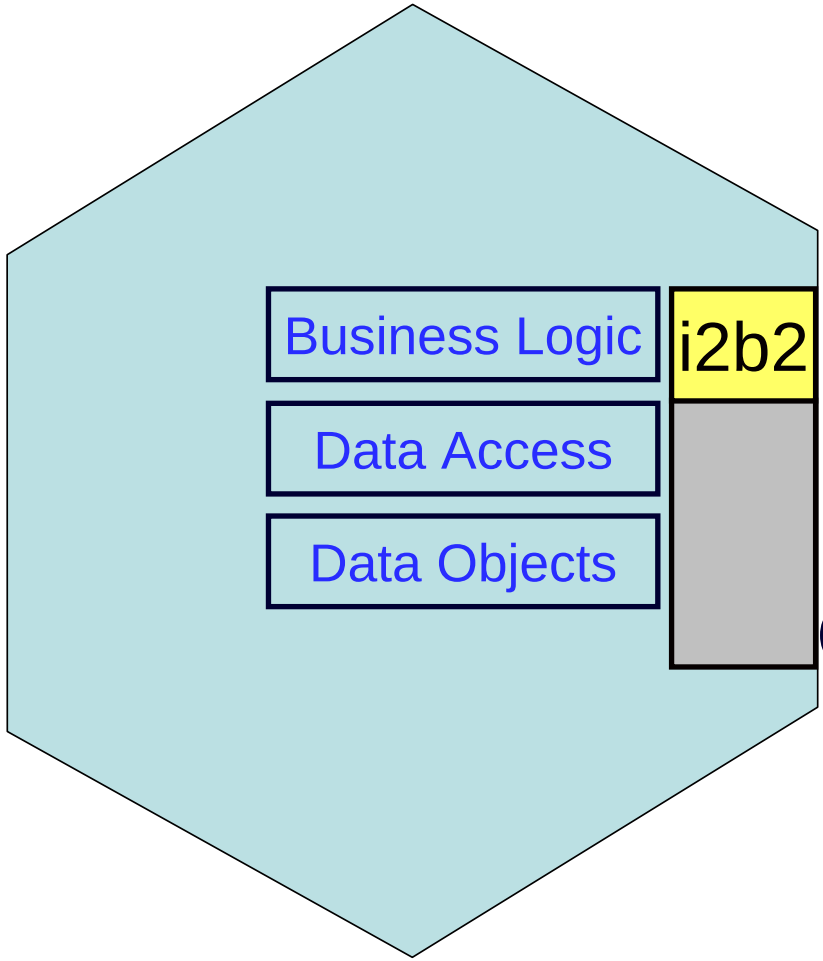


- Enable high performance collection of medical record data for querying and distribution
  - Enterprise web client
  - Create patient cohorts for further investigation
- Enable discovery within data on enterprise wide scale
  - Relevance networks
  - Pharmacovigilance

Technical Overview

- Formed as a collection of interoperable services provided by i2b2 Cells
- Loosely coupled
- Makes no assumptions about proximity
- Connected by Web services
- Activity can be directed manually or automatically

i2b2 Cell: Canonical Hive Unit

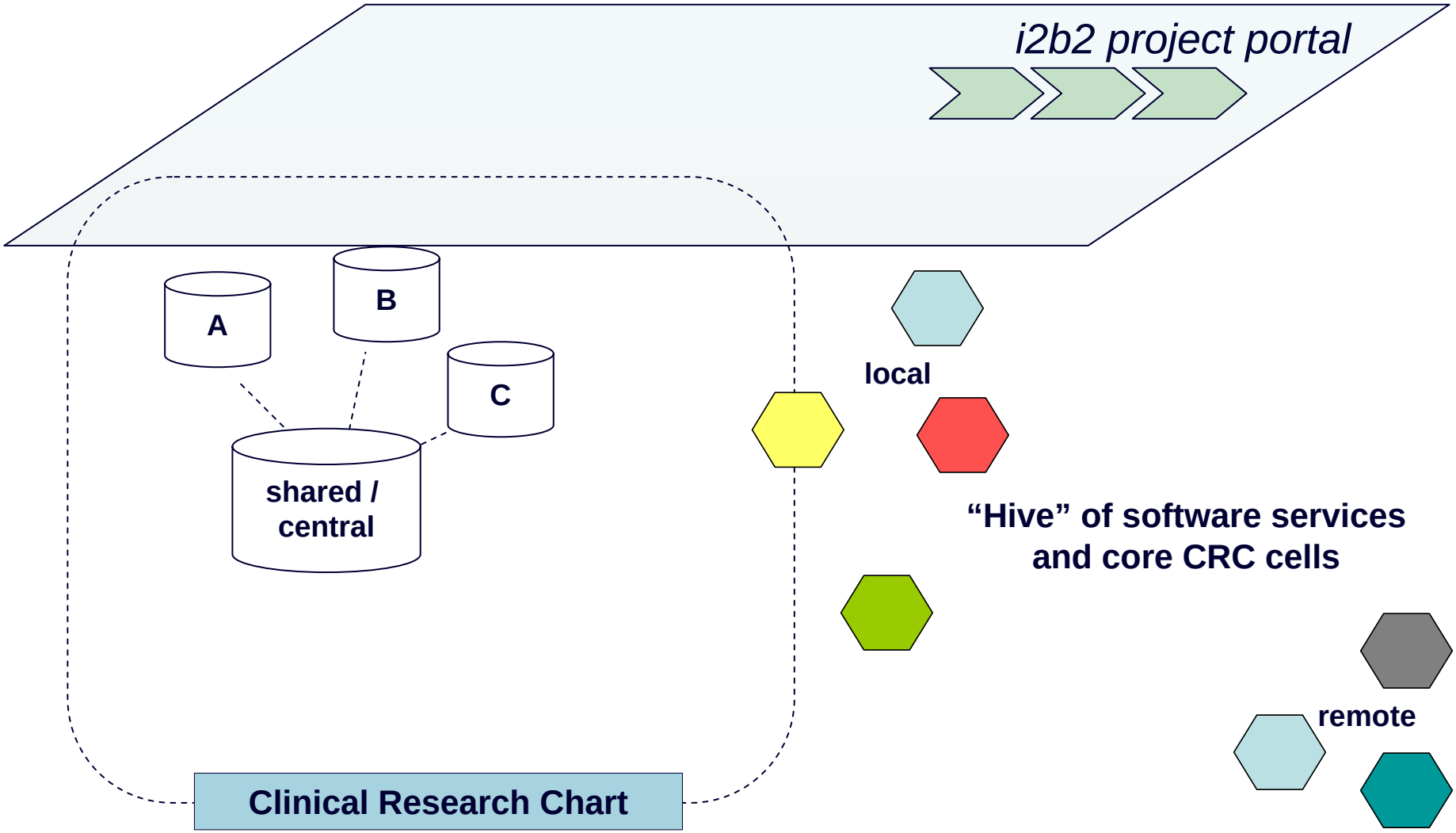


HTTP XML

(minimum: RESTful, others like SOAP optional)

Programmatic Access

i2b2 Environment



i2b2

Informatics for Integrating Biology & the Bedside

A National Center for Biomedical Computing

HealthMap

A CTSAs* Adopting i2b2

B CTSAs* Evaluating i2b2 platform

C Academic Medical Centers Adopting i2b2 Platform

D Foreign Medical Centers Adopting i2b2 Platform

Community

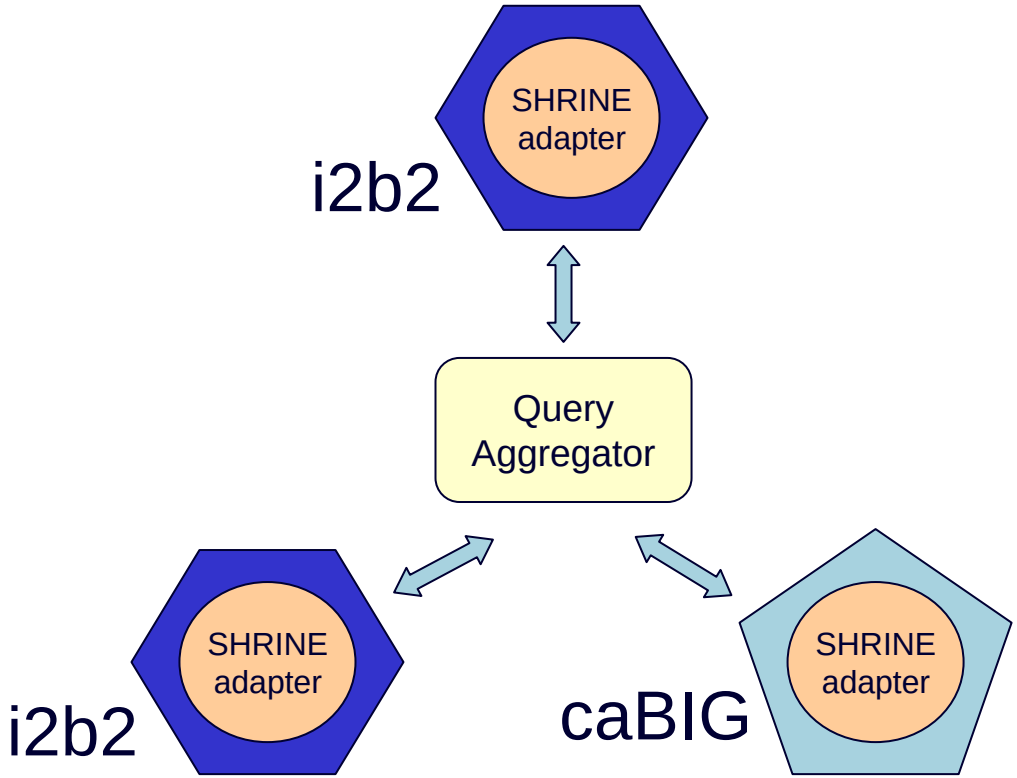
United States

- Beth Israel Deaconness Hospital, Boston, MA
- Boston University School of Medicine, Boston, MA
- Brigham and Women's Hospital, Boston, MA
- Children's Hospital, Boston, MA
- Denver Children's Hospital, Denver, CO
- Cincinnati Children's Hospital, Cincinnati, OH
- Cleveland Clinic, Cleveland, OH
- Weil Medical College of Cornell, NYC, NY
- Group Health Cooperative
- Harvard Medical School, Boston, MA
- Massachusetts General Hospital, Boston, MA
- Maine Medical Center, Portland, ME
- Marshfield Clinic, Wisconsin
- Morehouse School of Medicine, Atlanta, GA
- Oregon Health & Science University, Portland, OR
- Ohio State University Medical Center, Columbus, OH
- Philadelphia Children's Hospital, Philadelphia, PA
- Renaissance Computing Institute, Chapel Hill, NC
- Tufts New England Medical Center, Boston, MA
- University of California Davis, Davis, CA
- University of California San Francisco, SF, CA
- 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 Houston, Houston, TX
- University of Texas Health Sciences Center San Antonio, SA, TX
- University of Texas Health Sciences Center Southwestern,
- Utah Health Science Center, Salt Lake City, UT
- University of Washington, Seattle, WA

International

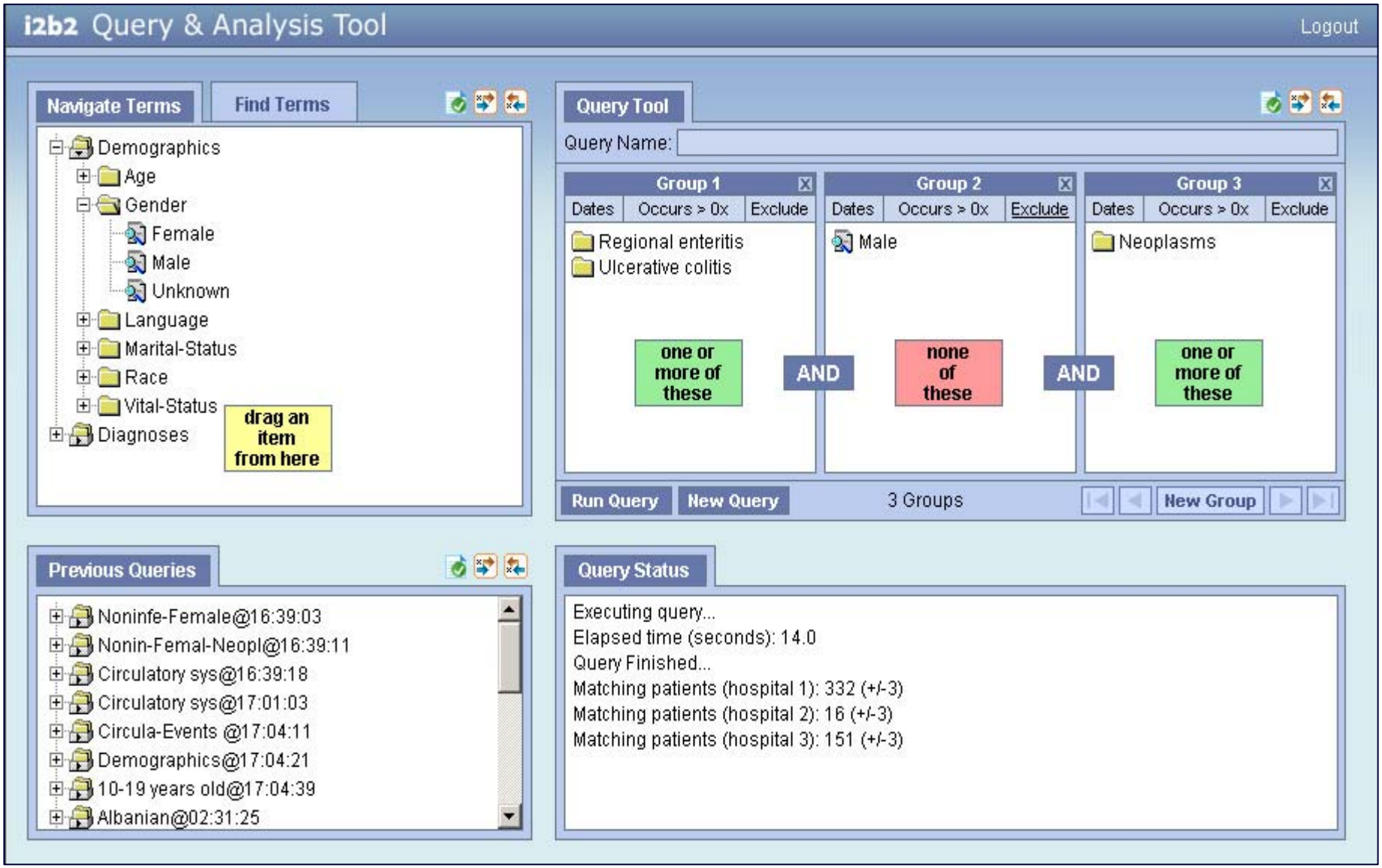
- Georges Pompidous Hospital, Paris, France
- University of Goettingen, Goettingen, Germany
- University of Pavia, Pavia, Italy
- University of Seoul, Seoul, Korea

SHRINE (Shared Research Informatics Network) = Distributed Queries



Central “aggregator” broadcasts query to local hospital “adaptors”, which return aggregate counts only

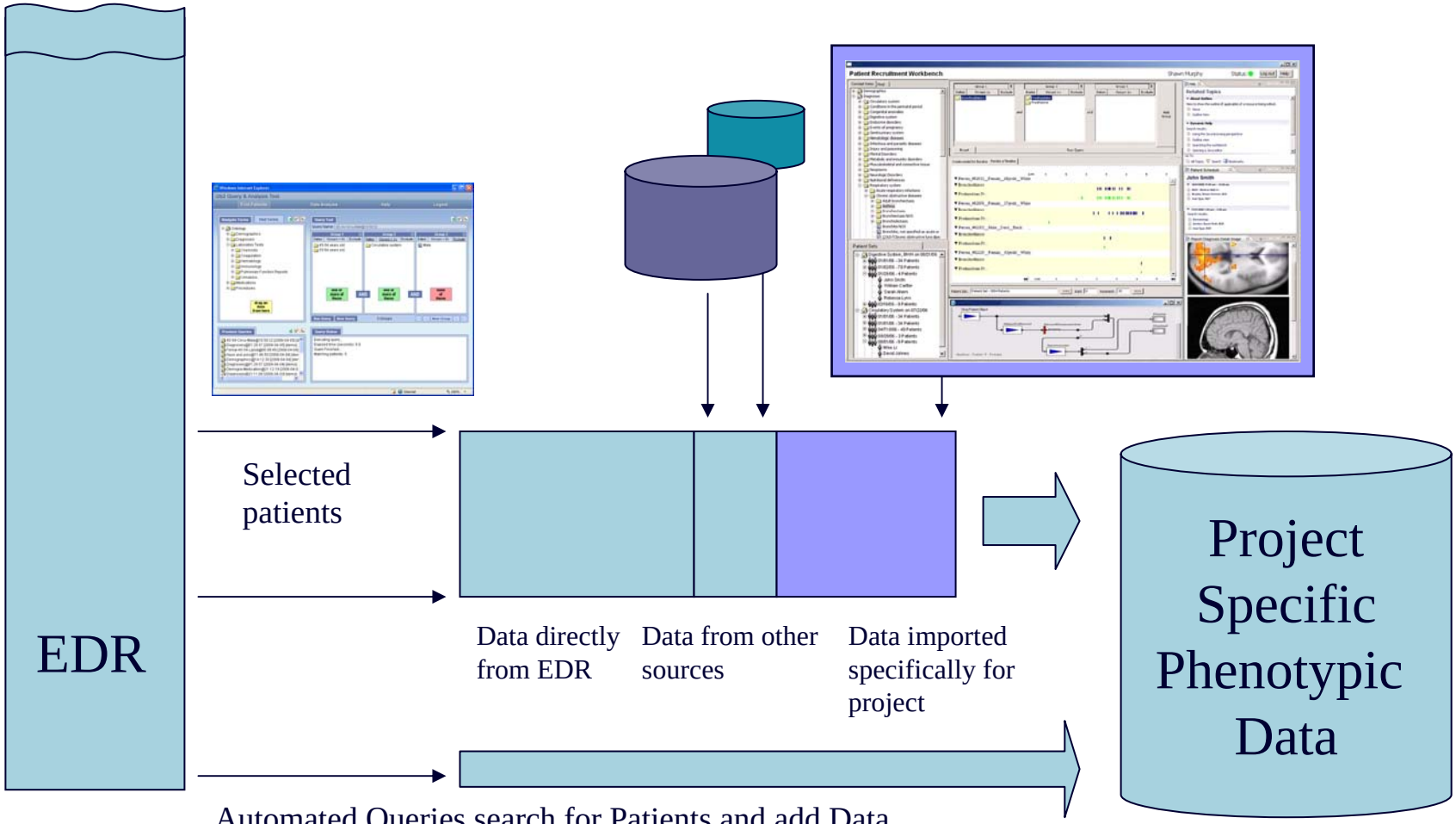
SHRINE


i2b2 Query & Analysis Tool
Logout
Navigate Terms
Find Terms
Demographics
Age
Gender
Female
Male
Unknown
Language
Marital-Status
Race
Vital-Status
Diagnoses
drag an item from here
Query Tool
Query Name:
Group 1
Dates
Occurs > 0x
Exclude
Regional enteritis
Ulcerative colitis
one or more of these
AND
Group 2
Dates
Occurs > 0x
Exclude
Male
none of these
AND
Group 3
Dates
Occurs > 0x
Exclude
Neoplasms
one or more of these
Run Query
New Query
3 Groups
New Group
Previous Queries
Noninfe-Female@16:39:03
Nonin-Femal-Neopl@16:39:11
Circulatory sys@16:39:18
Circulatory sys@17:01:03
Circula-Events @17:04:11
Demographics@17:04:21
10-19 years old@17:04:39
Albanian@02:31:25
Query Status
Executing query...
Elapsed time (seconds): 14.0
Query Finished...
Matching patients (hospital 1): 332 (+/-3)
Matching patients (hospital 2): 16 (+/-3)
Matching patients (hospital 3): 151 (+/-3)

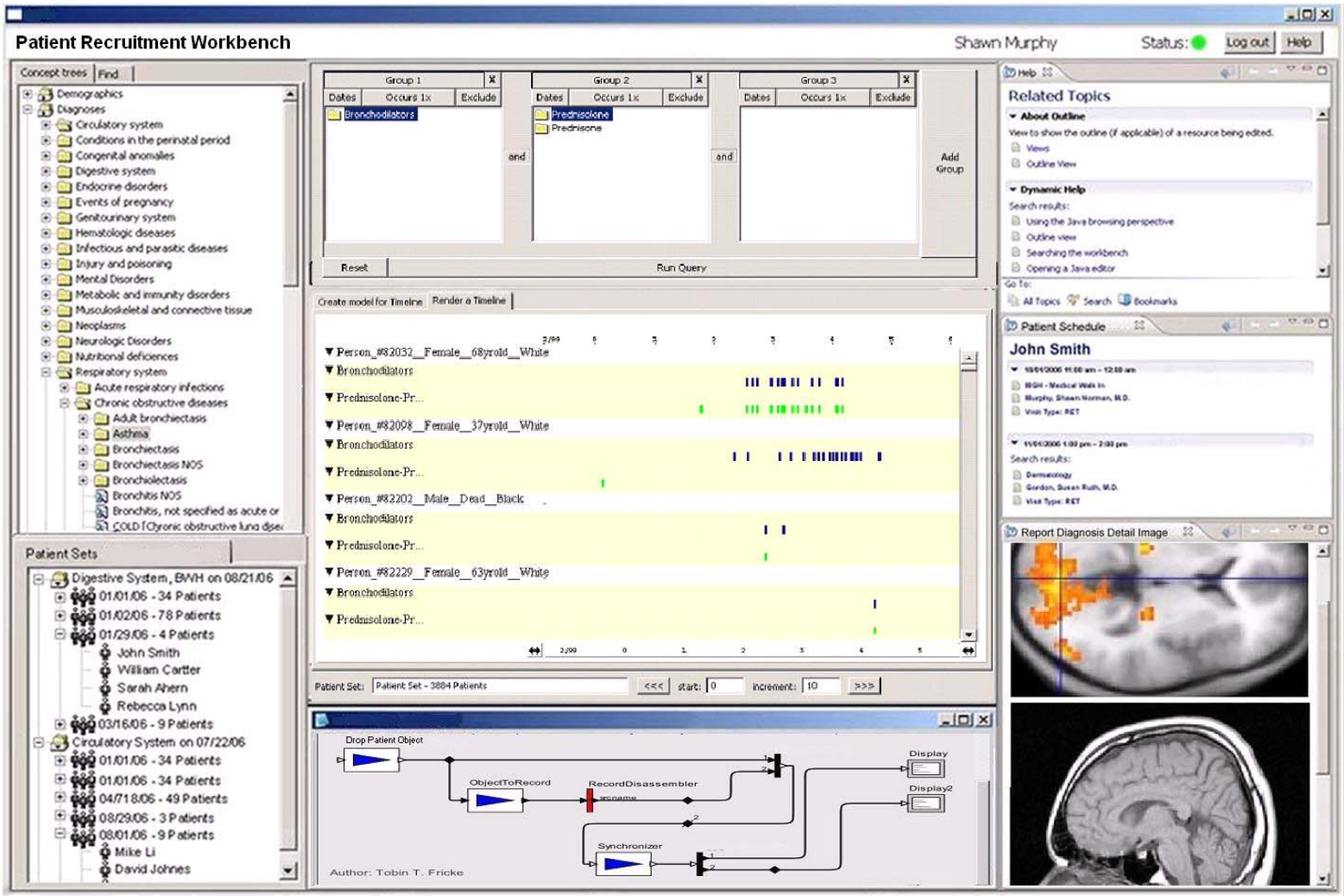
High Throughput Methods for supporting Research at Partners Healthcare

- 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
  - NIH/NCBC - Isaac Kohane
- 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

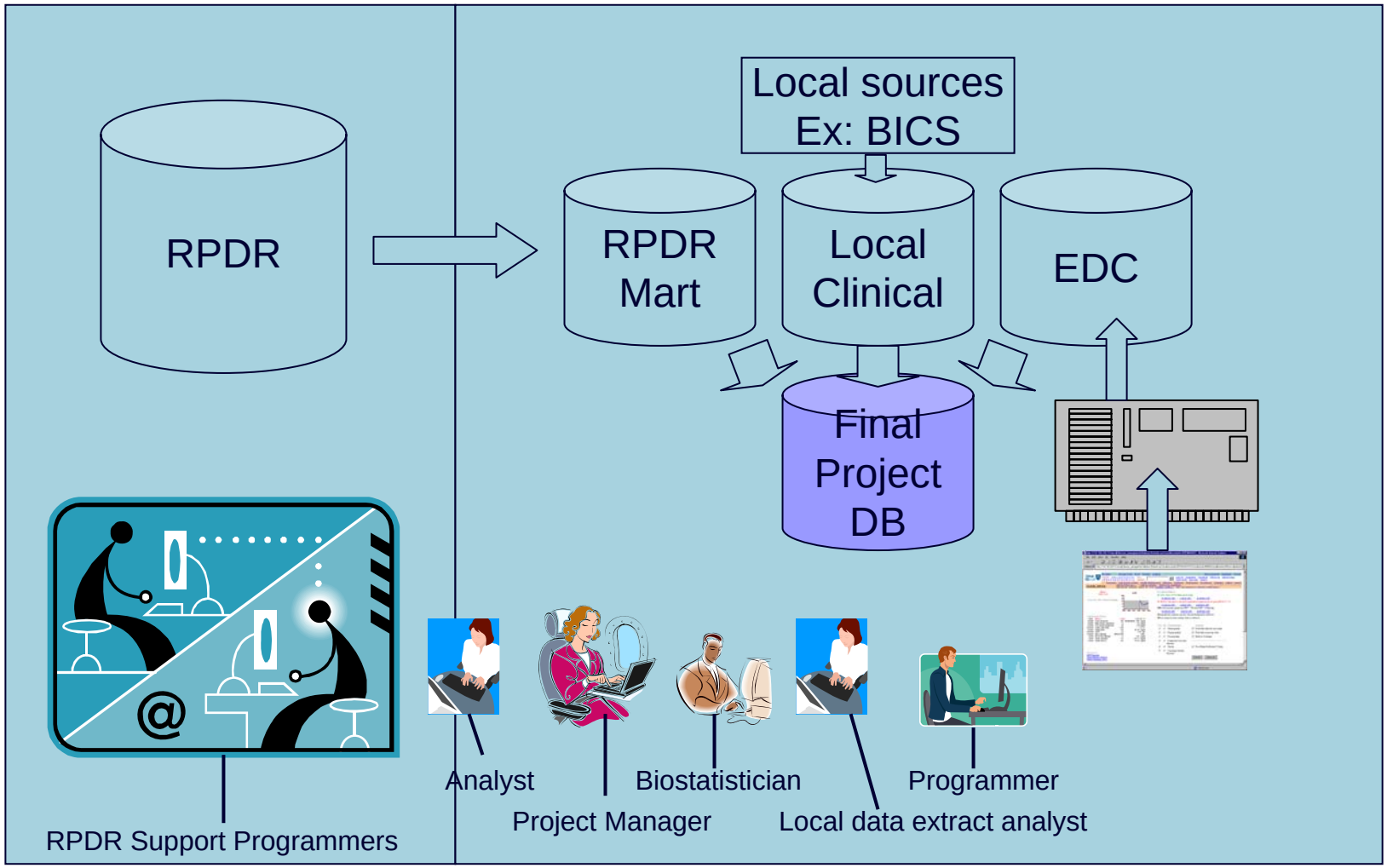
Set of patients is selected through Enterprise Repository and data is gathered into a data mart



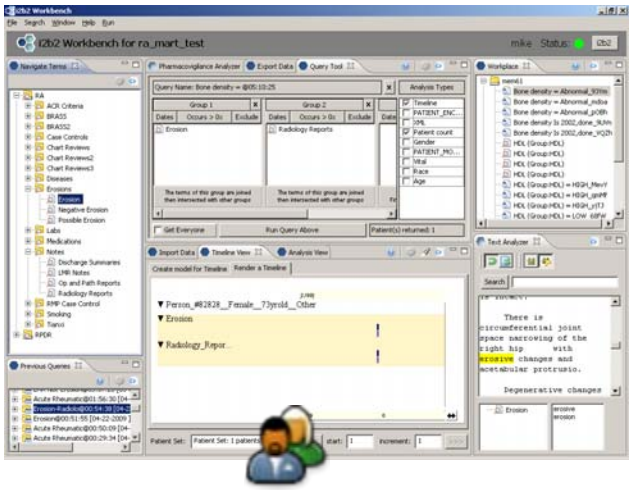
Data is available through the i2b2 Workbench



Team support for Projects



NLP Workflow



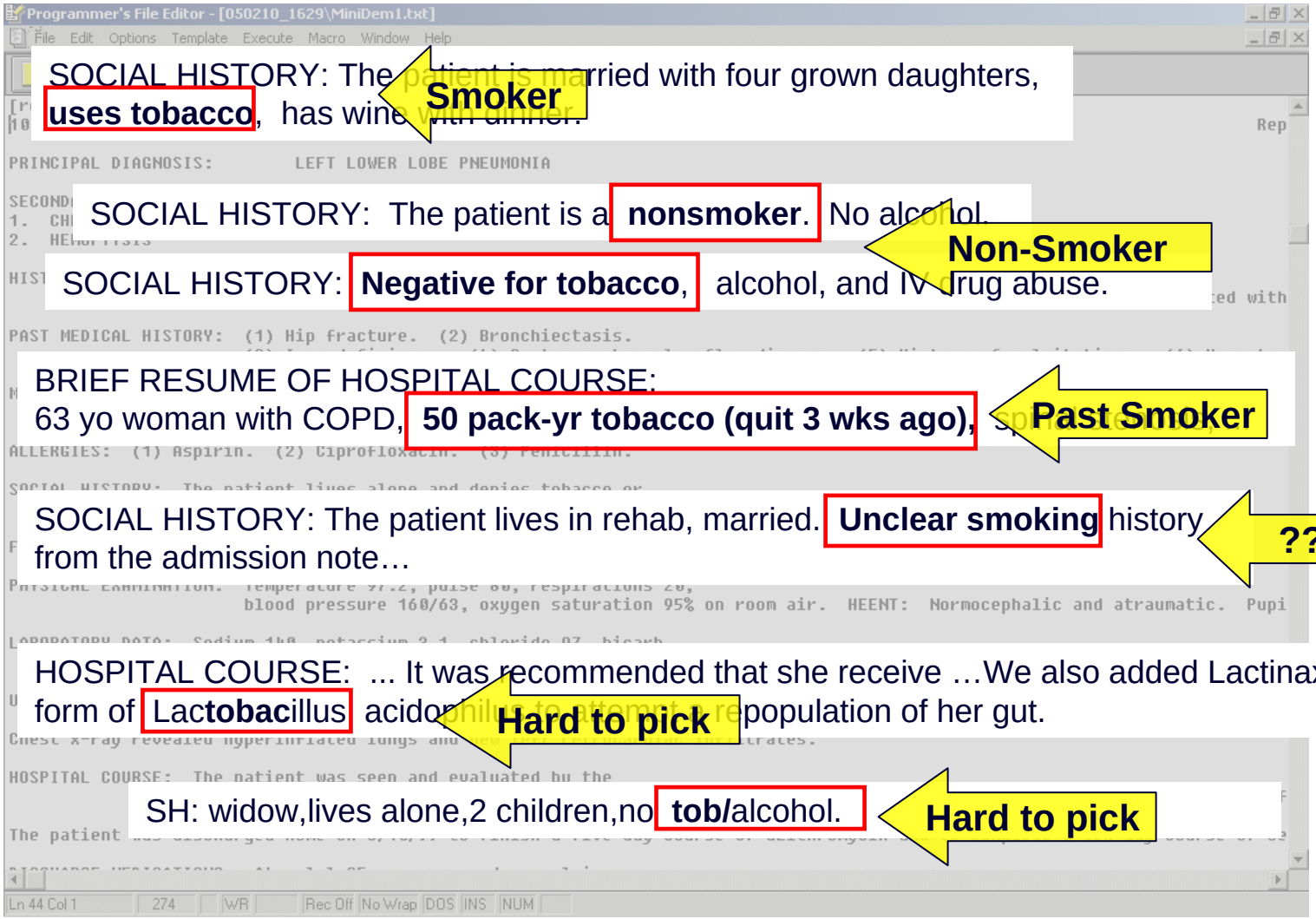
I2b2 Project Investigators

Communication

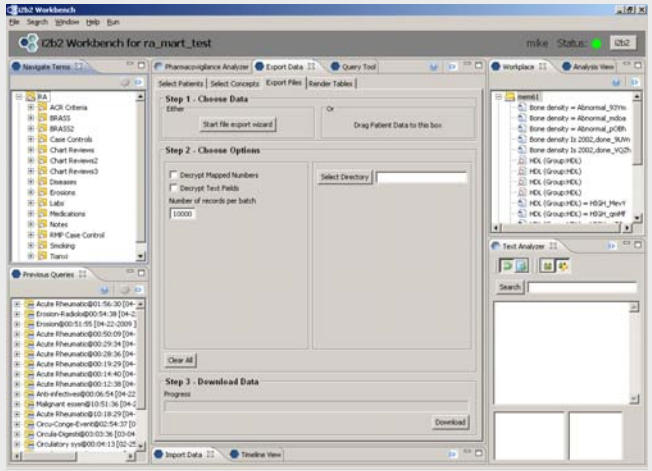
Results Delivery

NLP Specialists

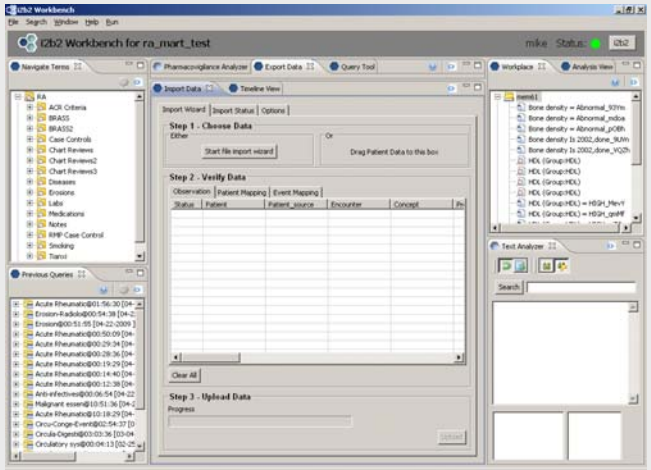
NLP (and comedy) is not pretty



NLP Specialists Workstation



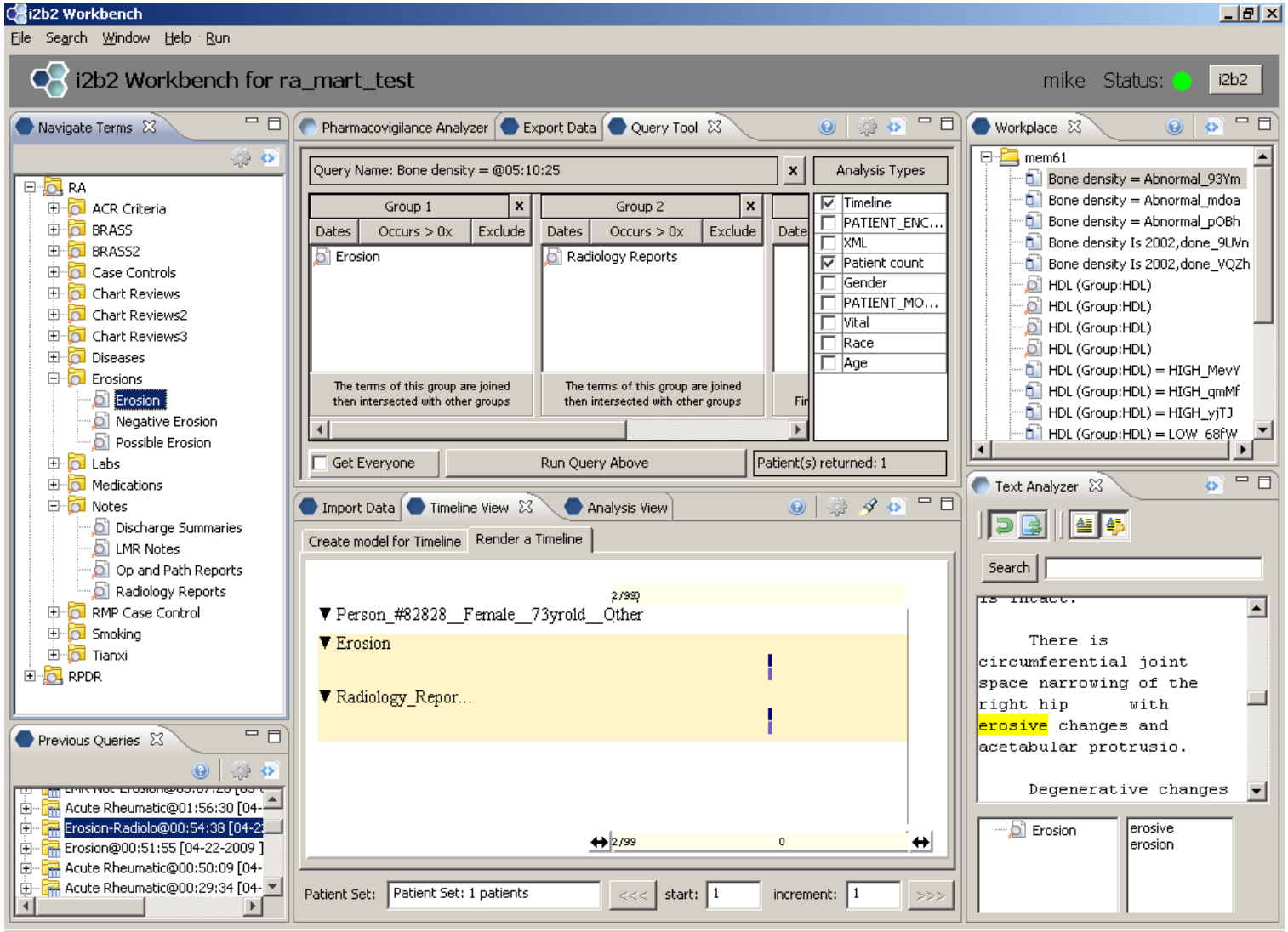
Export Notes



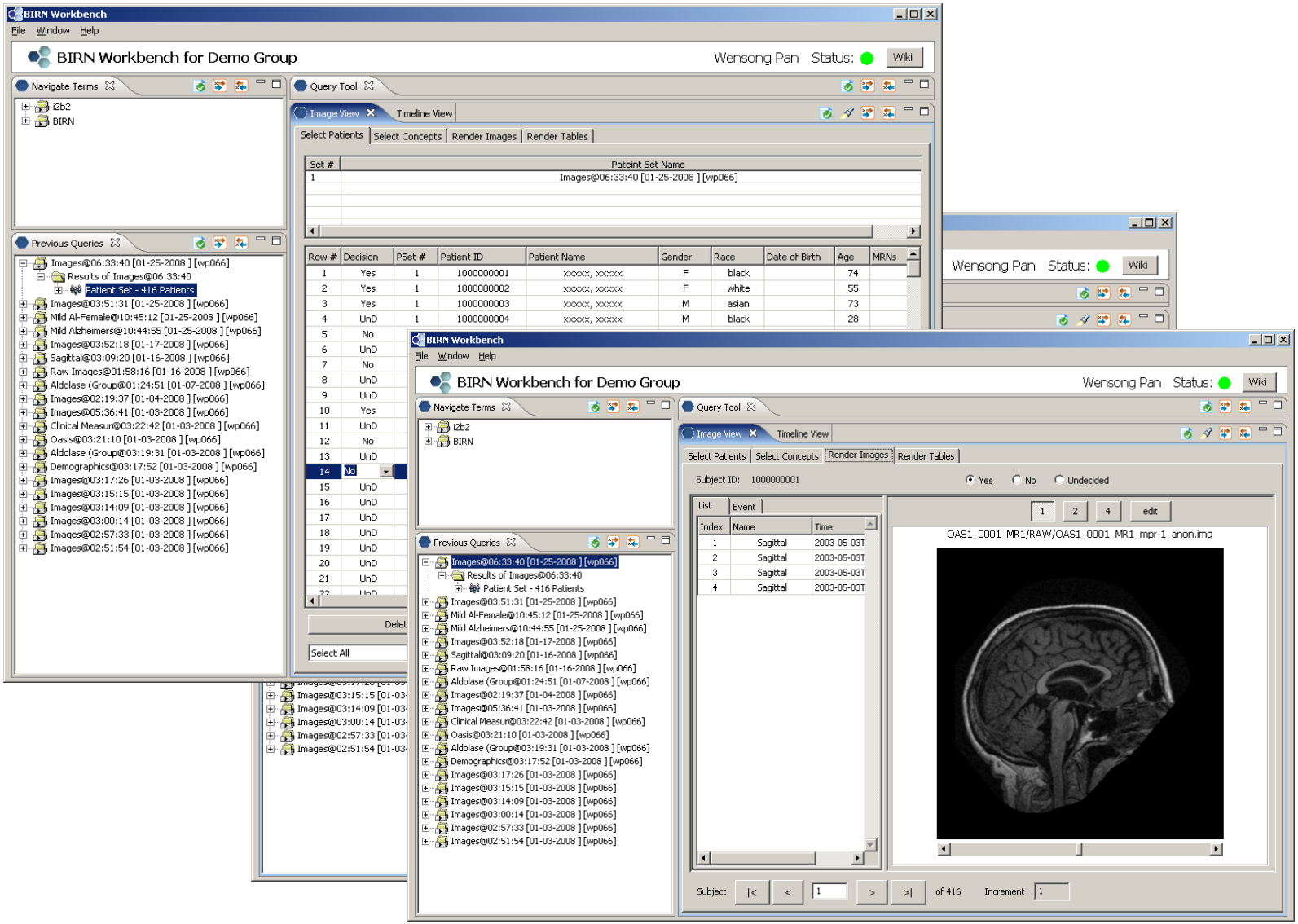
Import Derived Codes

NLP Specialists

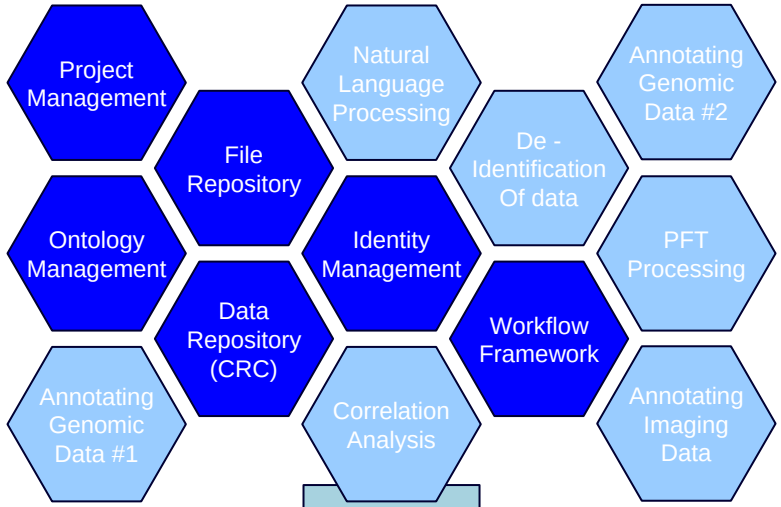
Investigator Review



Select patients for clinical trials



- Repurpose medical record information for research studies
  - I2b2 Workbench
  - Natural language processing
- Enable genomic studies
  - Tissue/blood selection
  - Genetic data integration

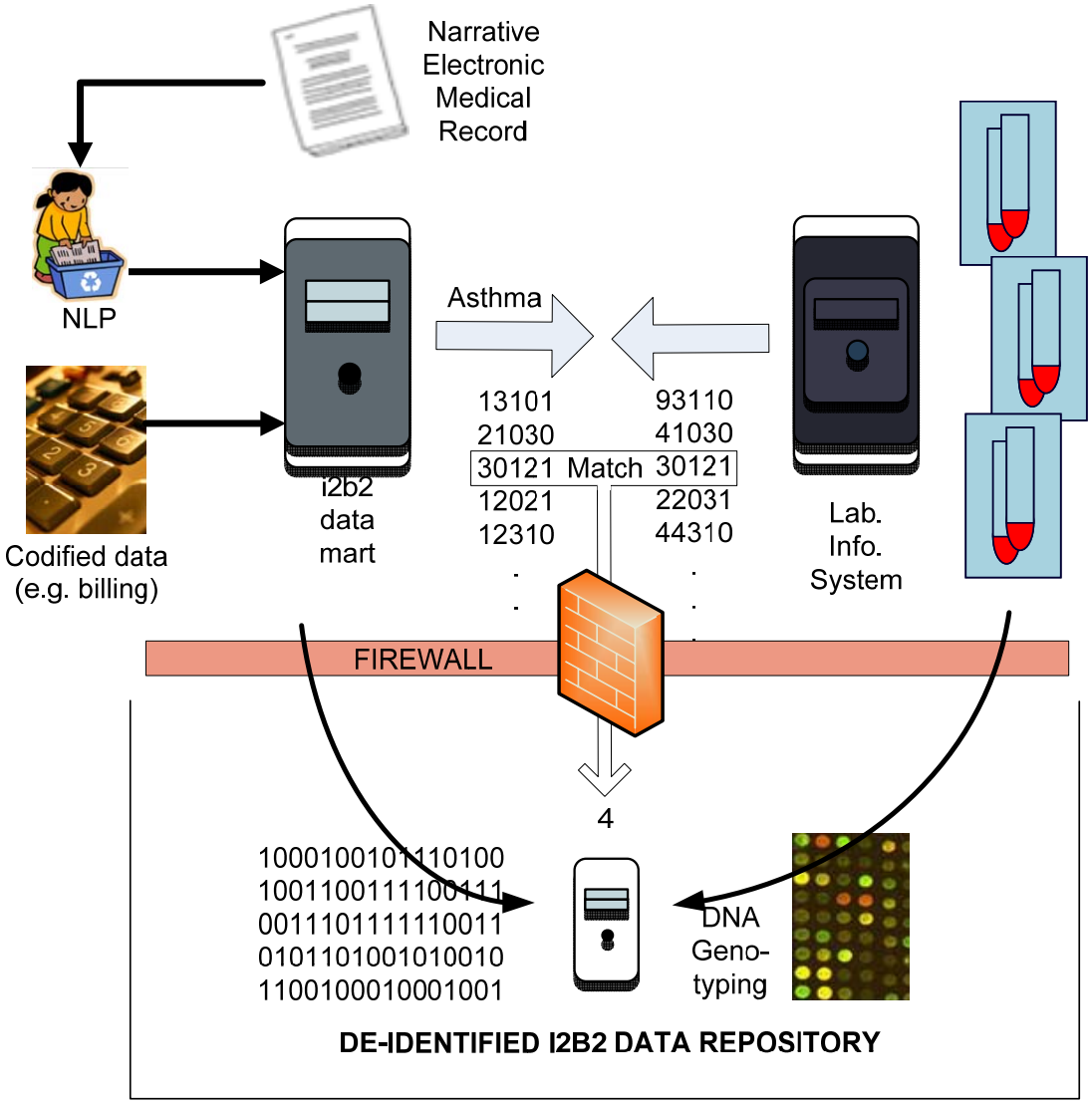


Use of medical record data in clinical studies focused upon teamwork and workflow

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
  - BWH Pathology –Lynn Bry
- Automated discovery can be created from these projects to support further hypothesis-driven research

Genotype samples and compare to controls



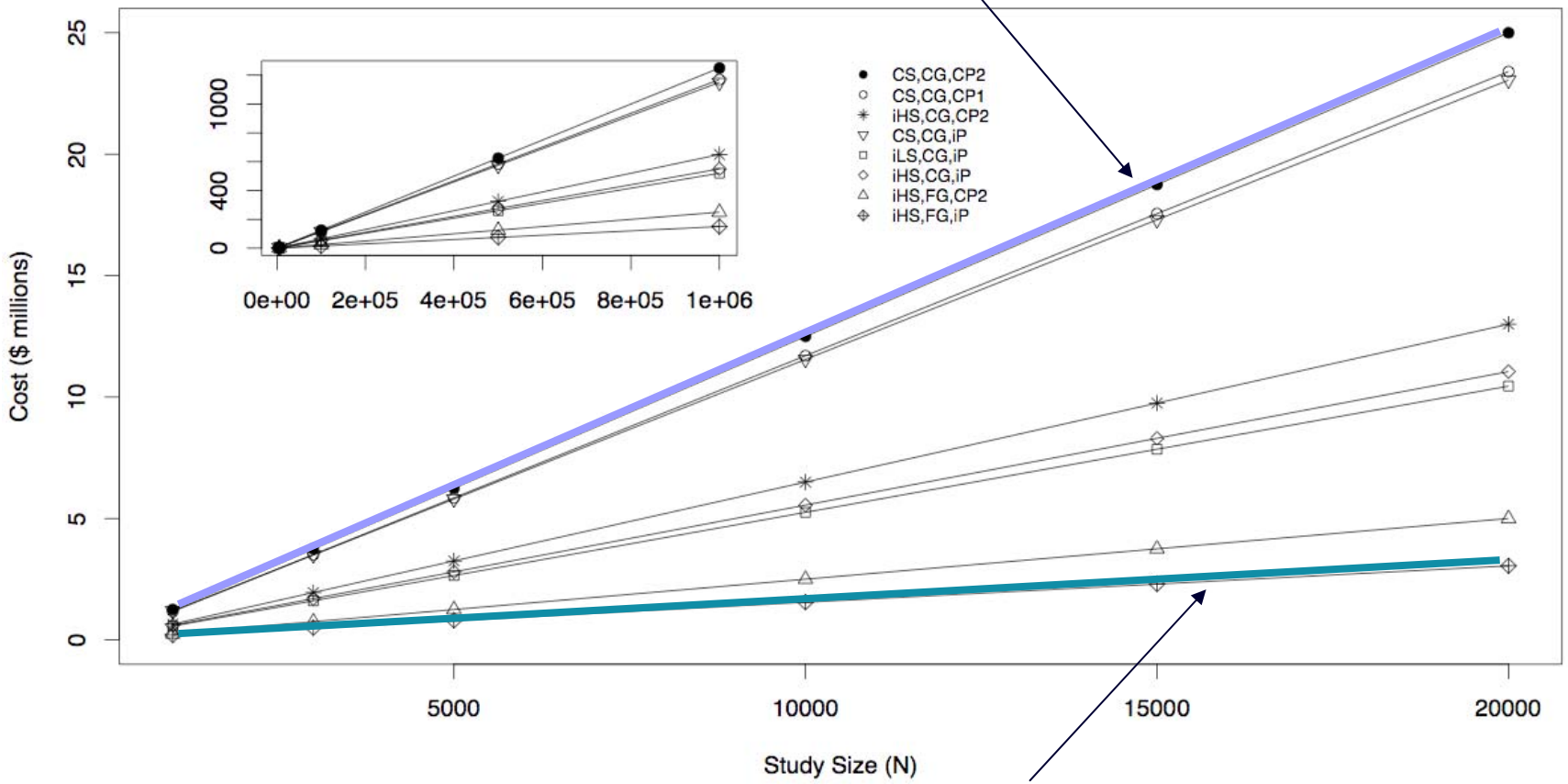
Cost and time benefit of Instrumenting with Sample Collection for Modest-size Study with 10,000 subjects (cases + controls)

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

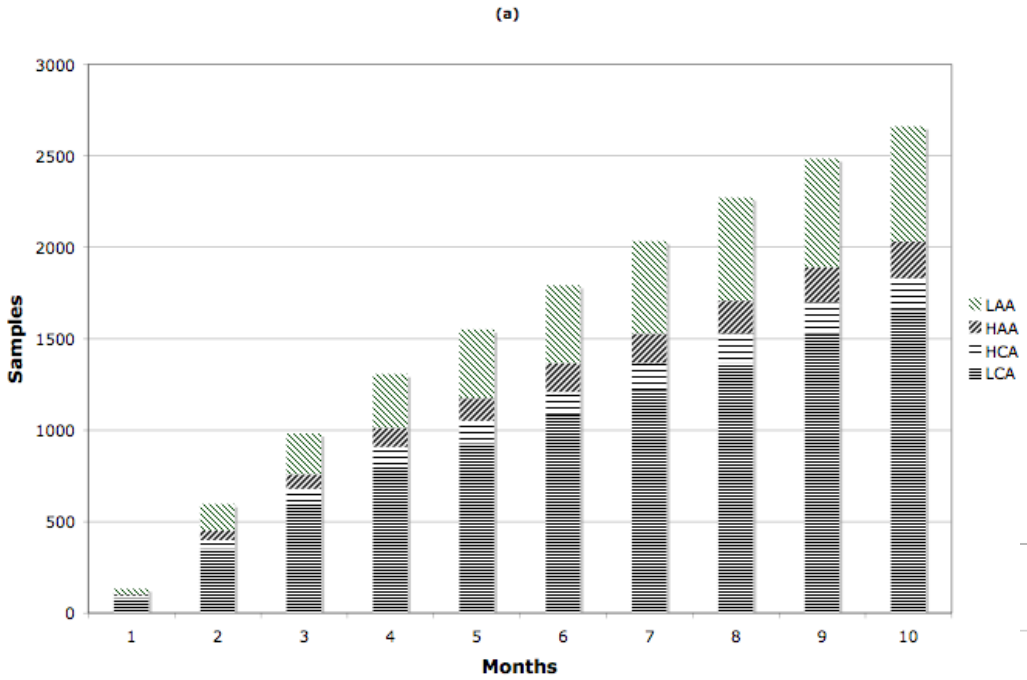
Escalating cost and time benefit of Instrumenting with Sample Collection

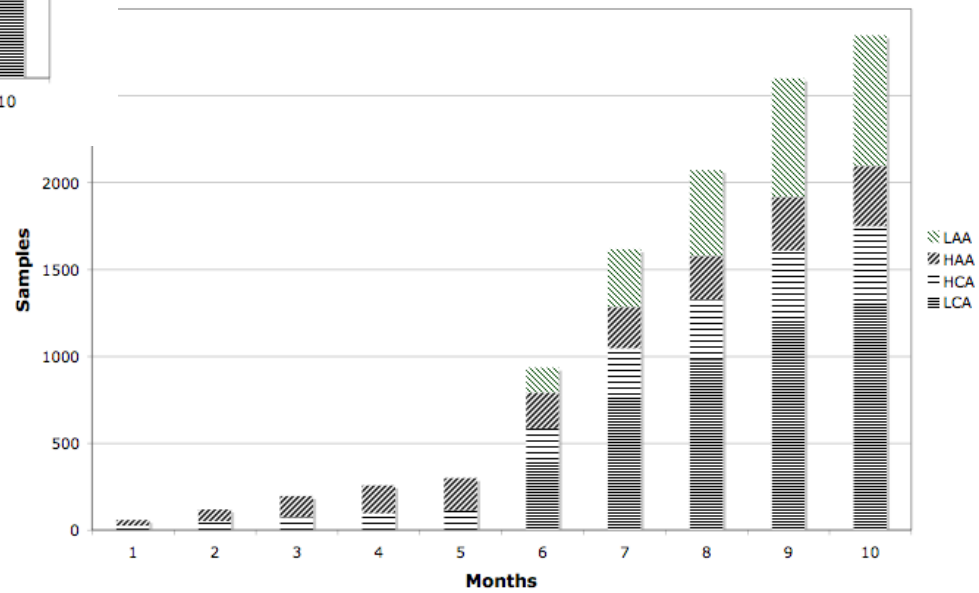
Previous model for collecting specimens



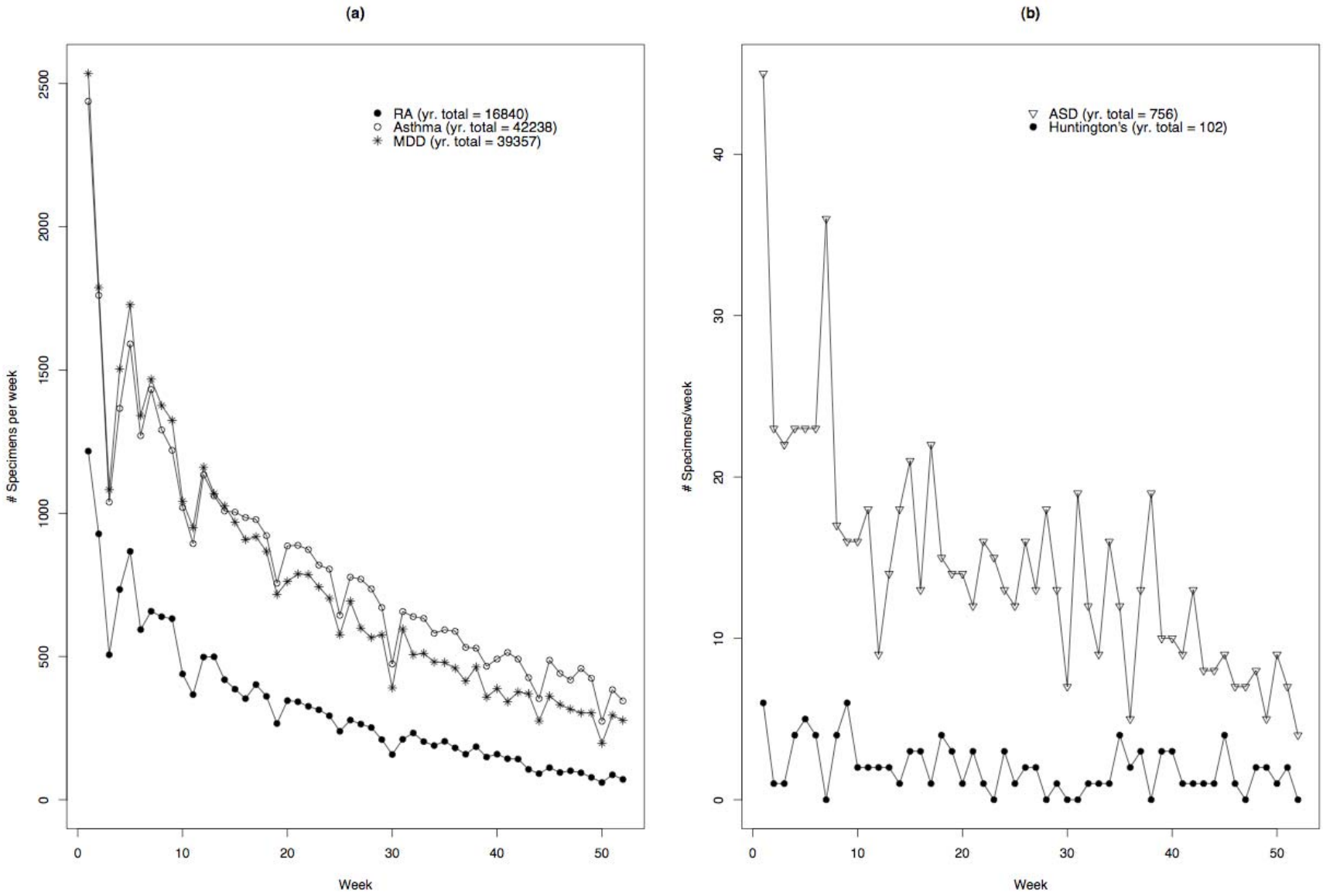
New model for collecting specimens

Meeting Expectations





Accrual Rates



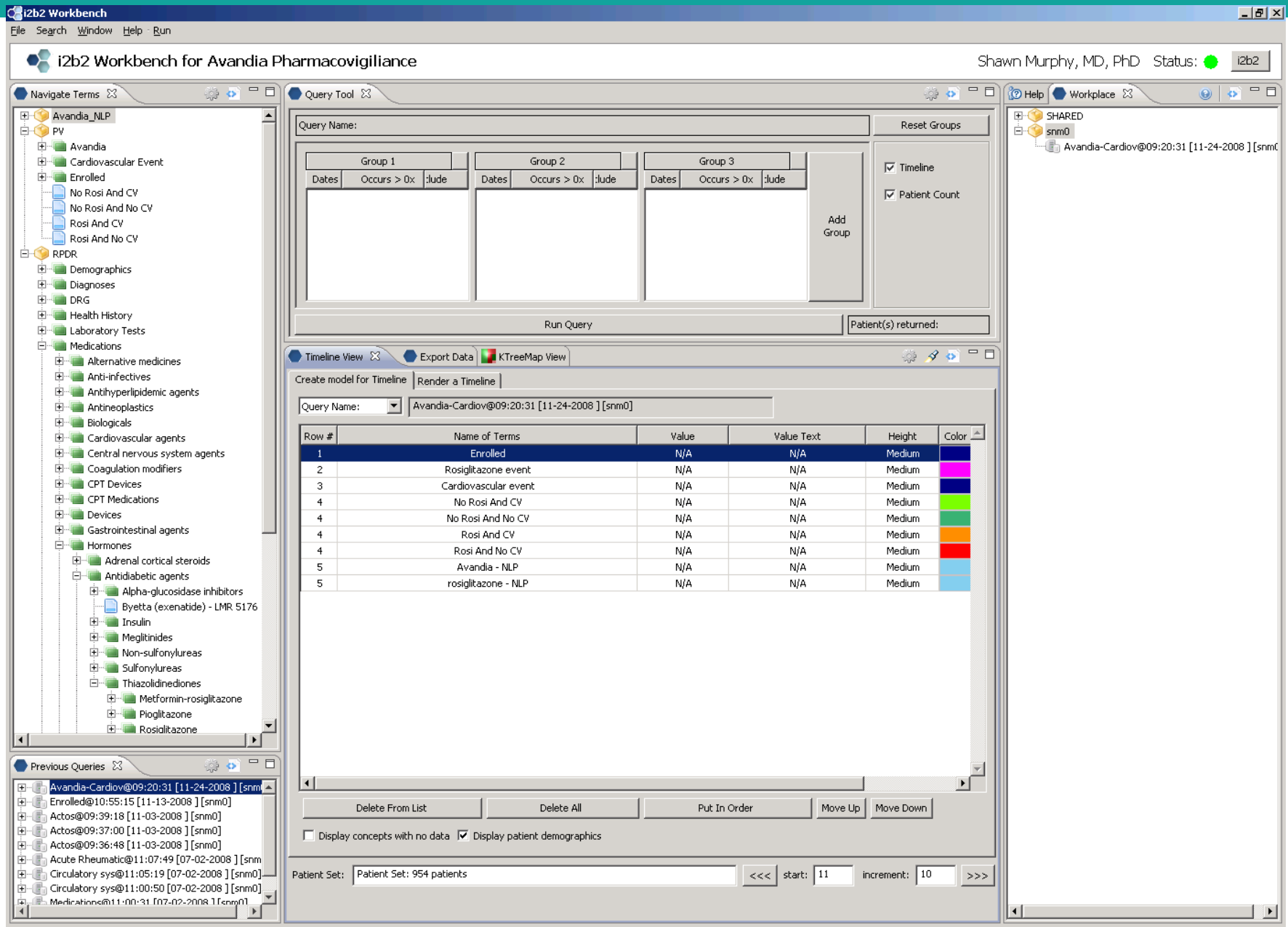
High Throughput Methods for supporting Research at Partners Healthcare

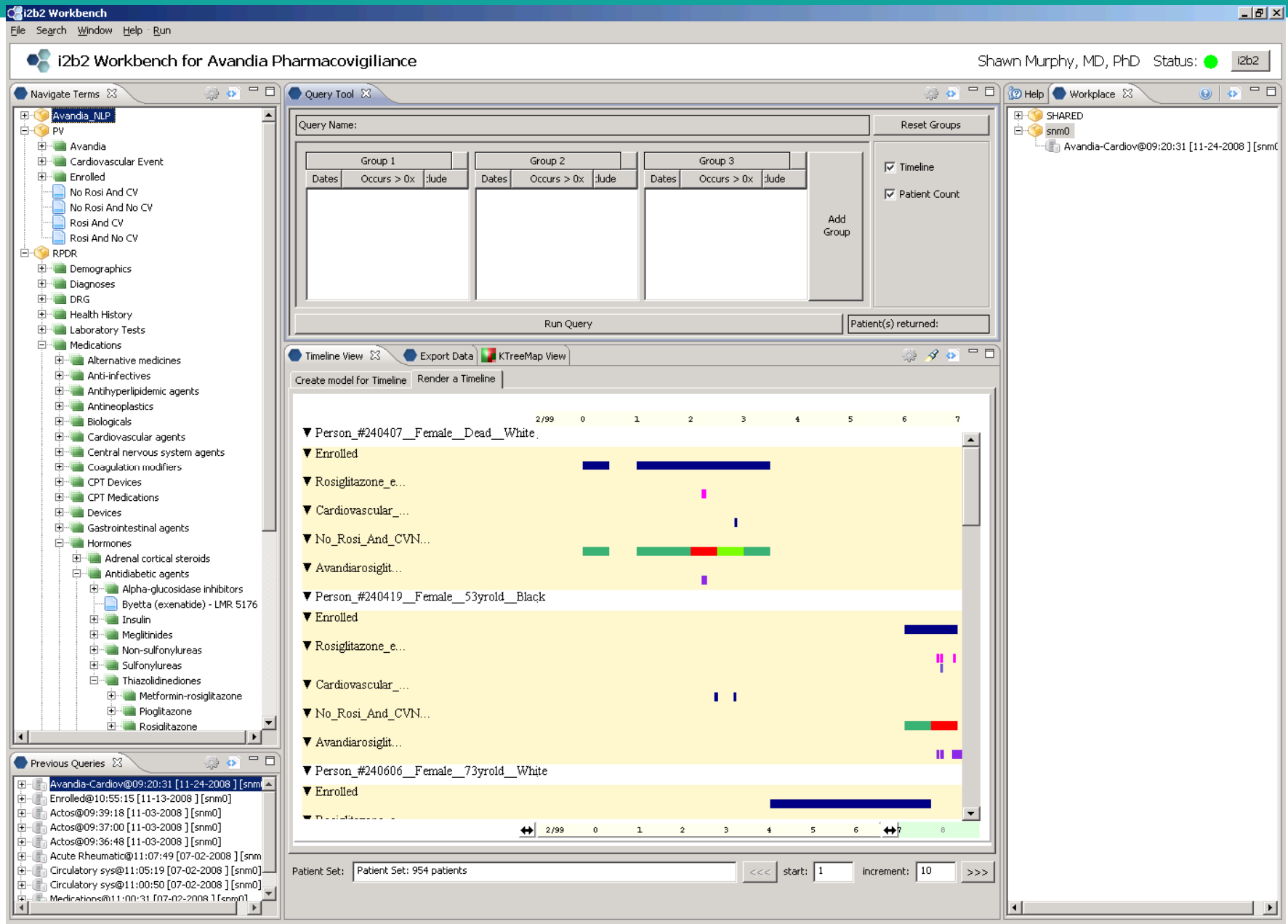
- Set of patients is selected from medical record data in a high throughput fashion
- Investigators work with the data of these patients using new i2b2 tools and a specialized team, both developed to work specifically with medical record data
- Using the Crimson system, tissues of these patients can be made available for genomic and biochemical analysis
- Automated discovery can be created from these projects to support further hypothesis-driven research
  - Pharmacovigilance – John Brownstein and Judy Colecchi

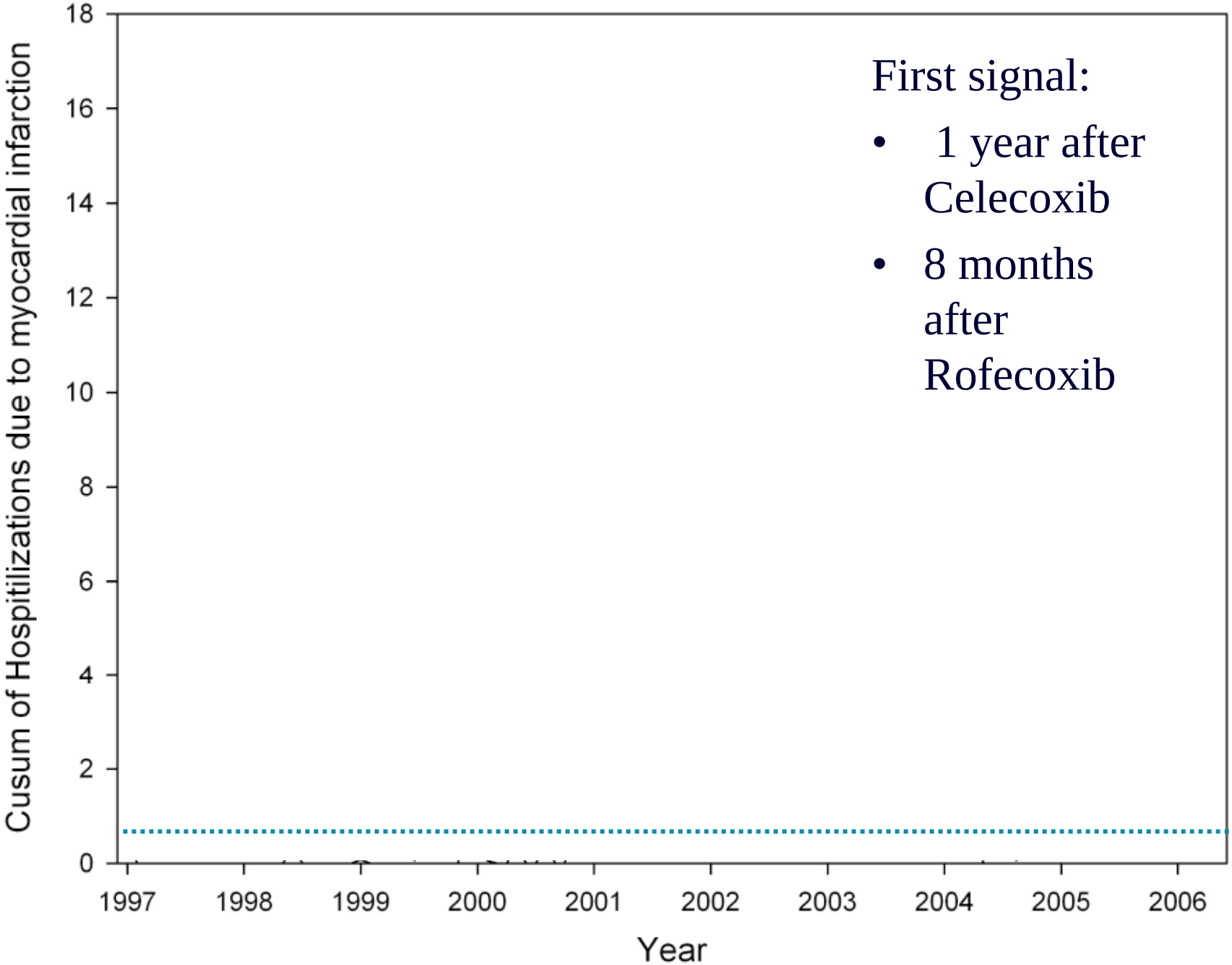
Centrally supported Automated Discovery Projects
Clinical trials performed in-silico

- Performing an observational, phase IV study is an expensive and complex process that can be potentially modeled in a retrospective database using groups of patients available in the large amounts of highly organized medical data.

- Fundamental problems complicate this approach:
  - Patients drift in and out of the system. Sophisticated statistical models using adequate control populations are necessary to compensate.
  - Confounding variables are not found in the database. Sophisticated natural language processing is needed to extract the confounders from textual reports to allow these confounders to be controlled.
  - Missing data disrupts typical statistical approaches


i2b2 Workbench
File Search Window Help Run
i2b2 Workbench for Avandia Pharmacovigiliance
Shawn Murphy, MD, PhD Status:
i2b2
Navigate Terms
Avandia_NLP
PV
Avandia
Cardiovascular Event
Enrolled
No Rosi And CV
No Rosi And No CV
Rosi And CV
Rosi And No CV
RPDR
Demographics
Diagnoses
DRG
Health History
Laboratory Tests
Medications
Alternative medicines
Anti-infectives
Antihyperlipidemic agents
Antineoplastics
Biologicals
Cardiovascular agents
Central nervous system agents
Coagulation modifiers
CPT Devices
CPT Medications
Devices
Gastrointestinal agents
Hormones
Adrenal cortical steroids
Antidiabetic agents
Alpha-glucosidase inhibitors
Byetta (exenatide) - LMR 5176
Insulin
Meglitinides
Non-sulfonylureas
Sulfonylureas
Thiazolidinediones
Metformin-rosiglitazone
Pioglitazone
Rosiglitazone
Previous Queries
Avandia-Cardiov@09:20:31 [11-24-2008] [snm0]
Enrolled@10:55:15 [11-13-2008] [snm0]
Actos@09:39:18 [11-03-2008] [snm0]
Actos@09:37:00 [11-03-2008] [snm0]
Actos@09:36:48 [11-03-2008] [snm0]
Acute Rheumatic@11:07:49 [07-02-2008] [snm0]
Circulatory sys@11:05:19 [07-02-2008] [snm0]
Circulatory sys@11:00:50 [07-02-2008] [snm0]
Query Tool
Query Name:
Reset Groups
Group 1
Group 2
Group 3
Dates
Occurs > 0x
:lude
Add Group
Timeline
Patient Count
Run Query
Patient(s) returned:
Timeline View
Export Data
KTreeMap View
Create model for Timeline
Render a Timeline
Query Name:
Avandia-Cardiov@09:20:31 [11-24-2008] [snm0]
Row #
Name of Terms
Value
Value Text
Height
Color
1 Enrolled N/A N/A Medium
2 Rosiglitazone event N/A N/A Medium
3 Cardiovascular event N/A N/A Medium
4 No Rosi And CV N/A N/A Medium
4 No Rosi And No CV N/A N/A Medium
4 Rosi And CV N/A N/A Medium
4 Rosi And No CV N/A N/A Medium
5 Avandia - NLP N/A N/A Medium
5 rosiglitazone - NLP N/A N/A Medium
Delete From List
Delete All
Put In Order
Move Up
Move Down
Display concepts with no data
Display patient demographics
Patient Set:
Patient Set: 954 patients
<<<
start: 11
increment: 10
>>>
Help
Workplace
SHARED
snm0
Avandia-Cardiov@09:20:31 [11-24-2008] [snm0]

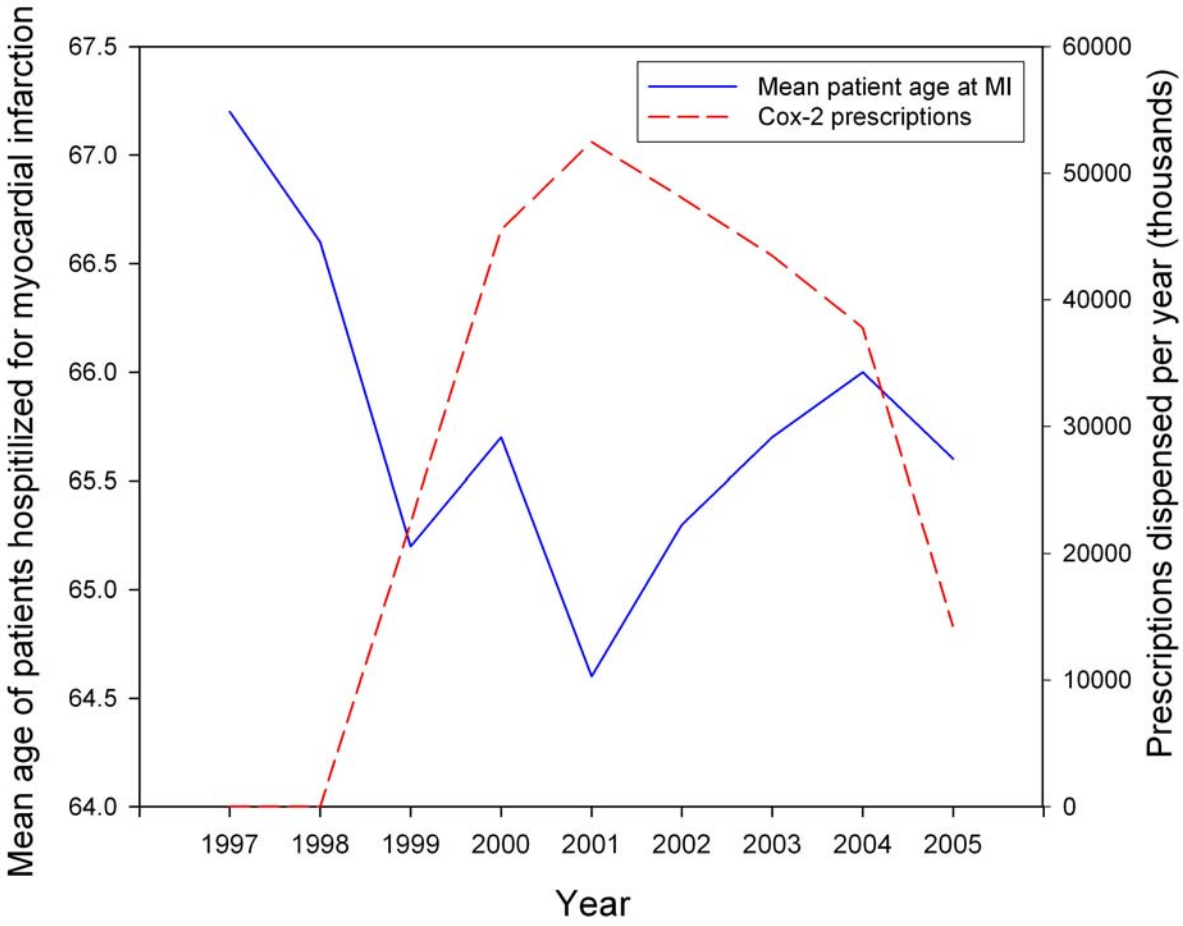

i2b2 Workbench
File Search Window Help Run
i2b2 Workbench for Avandia Pharmacovigiliance
Shawn Murphy, MD, PhD Status:
i2b2
Navigate Terms
Avandia_NLP
PV
Avandia
Cardiovascular Event
Enrolled
No Rosi And CV
No Rosi And No CV
Rosi And CV
Rosi And No CV
RPDR
Demographics
Diagnoses
DRG
Health History
Laboratory Tests
Medications
Alternative medicines
Anti-infectives
Antihyperlipidemic agents
Antineoplastics
Biologicals
Cardiovascular agents
Central nervous system agents
CPT Devices
CPT Medications
Devices
Gastrointestinal agents
Hormones
Adrenal cortical steroids
Antidiabetic agents
Alpha-glucosidase inhibitors
Byetta (exenatide) - LMR 5176
Insulin
Meglitinides
Non-sulfonylureas
Sulfonylureas
Thiazolidinediones
Metformin-rosiglitazone
Pioglitazone
Rosiglitazone
Previous Queries
Avandia-Cardiov@09:20:31 [11-24-2008] [snm
Enrolled@10:55:15 [11-13-2008] [snm0]
Actos@09:39:18 [11-03-2008] [snm0]
Actos@09:37:00 [11-03-2008] [snm0]
Actos@09:36:48 [11-03-2008] [snm0]
Acute Rheumatic@11:07:49 [07-02-2008] [snm
Circulatory sys@11:05:19 [07-02-2008] [snm0]
Circulatory sys@11:00:50 [07-02-2008] [snm0]
Query Tool
Query Name:
Reset Groups
Group 1
Group 2
Group 3
Dates
Occurs > 0x
:lude
Add Group
Timeline
Patient Count
Run Query
Patient(s) returned:
Timeline View
Export Data
KTreeMap View
Create model for Timeline
Render a Timeline
2/99 0 1 2 3 4 5 6 7
Person_#240407__Female__Dead__White
Enrolled
Rosiglitazone_e...
Cardiovascular_...
No_Rosi_And_CVN...
Avandiarosiglit...
Person_#240419__Female__53yrold__Black
Enrolled
Rosiglitazone_e...
Cardiovascular_...
No_Rosi_And_CVN...
Avandiarosiglit...
Person_#240606__Female__73yrold__White
Enrolled
Patient Set:
Patient Set: 954 patients
<<<
start: 1
increment: 10
>>>
Help
Workplace
SHARED
snm0
Avandia-Cardiov@09:20:31 [11-24-2008] [snm0



First signal:

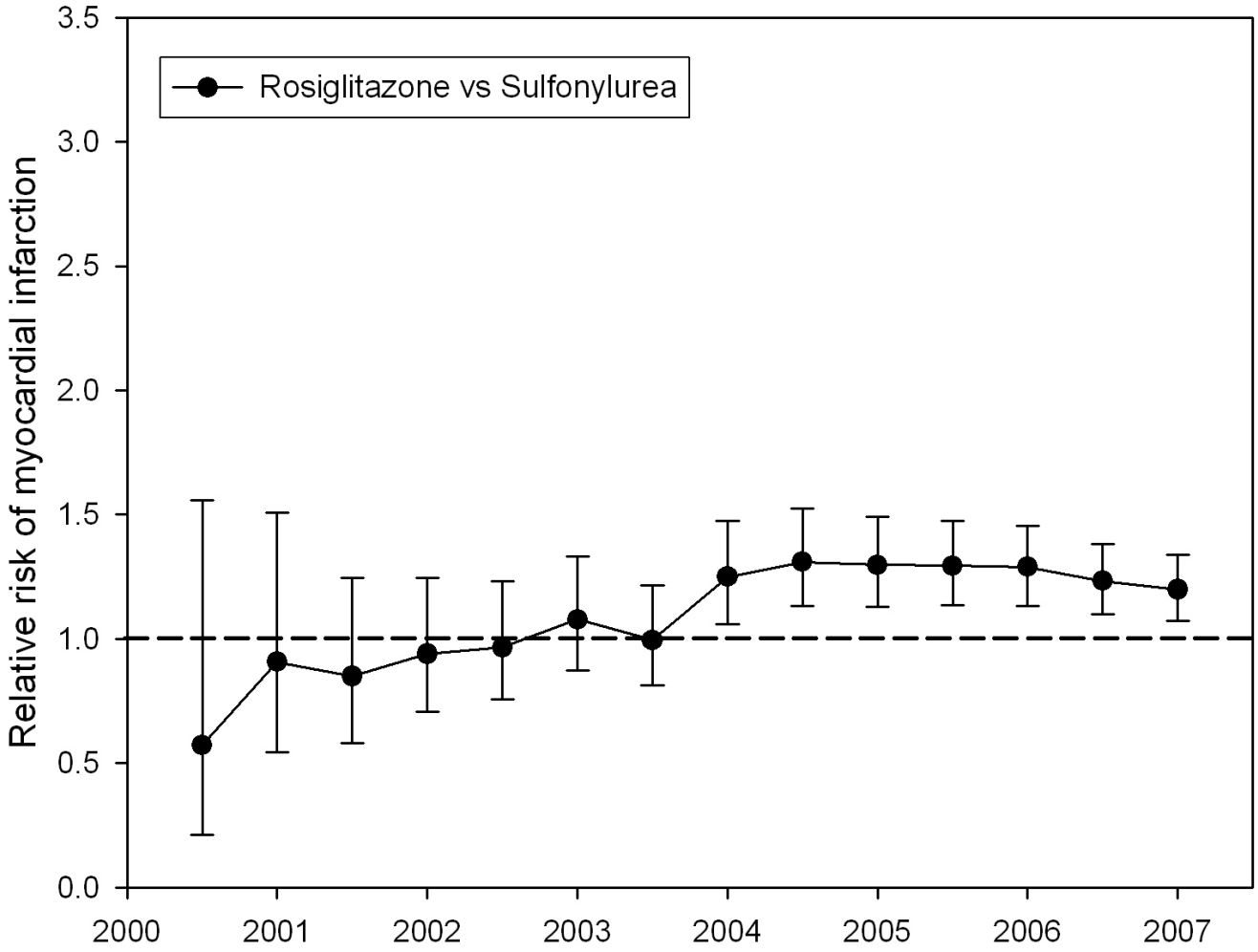
- 1 year after Celecoxib
- 8 months after Rofecoxib

Effect on patient age

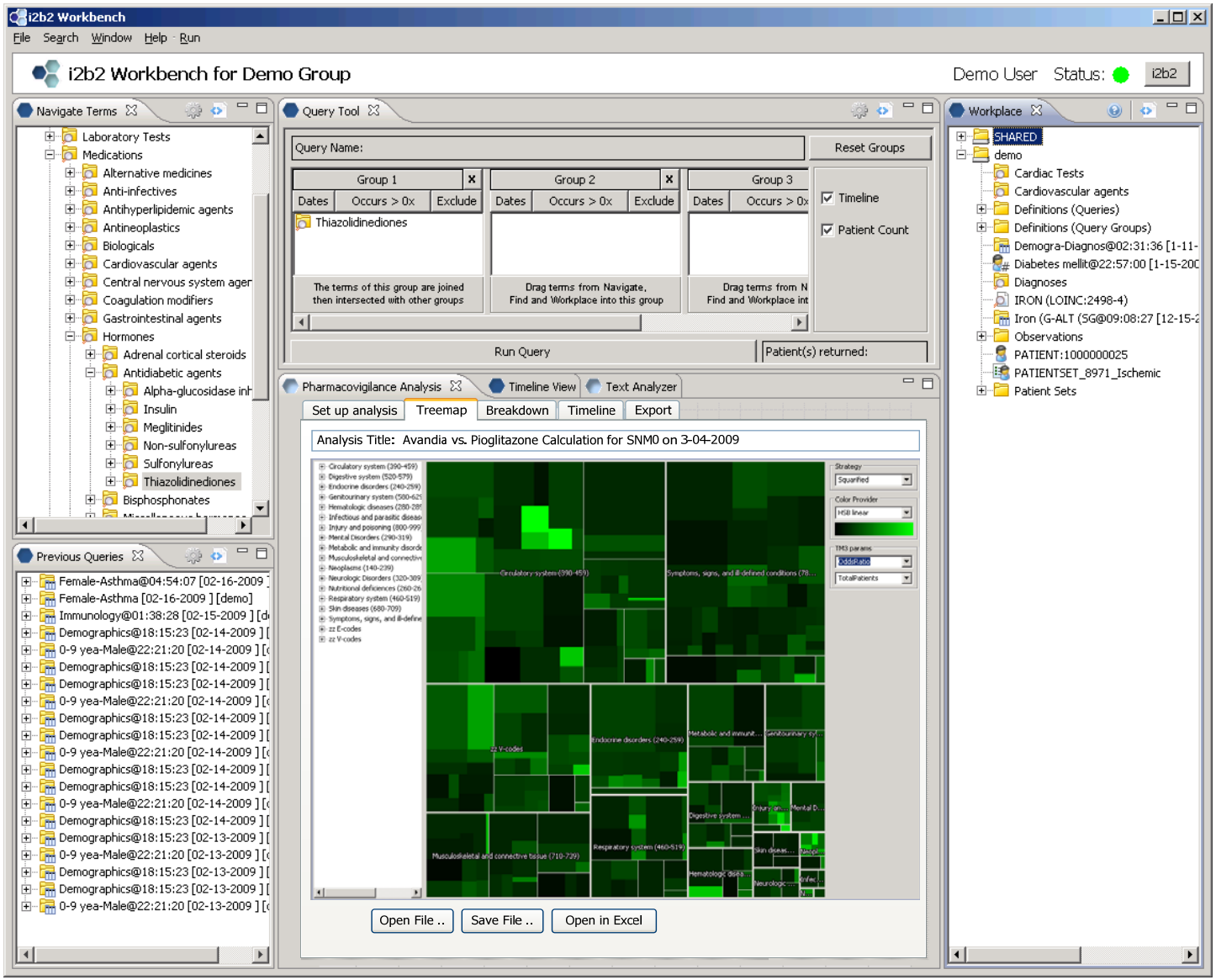


- Negative association between mean age at MI and prescription volume
- Spearman correlation -0.67, $P<0.05$

Prospective Surveillance of Cardiovascular Events while taking Avandia



Odds Ratios for Diseases expressed in comparing Rosiglitazone vs. Pioglitazone



10 Lessons from i2b2

- 1) Power is Numbers of patients recruited
- 2) Enable Enterprise use of patients for research
- 3) Extensible Architecture for developers
- 4) Enable Scientist Workflow
- 5) Enable Team Work with informaticians and researchers
- 6) Enable Natural Language Processing
- 7) Importance of Visualizations
- 8) High throughput Tissue acquisition for Genomic Research
- 9) Enable Health Surveillance
- 10) Enable Data Sharing