

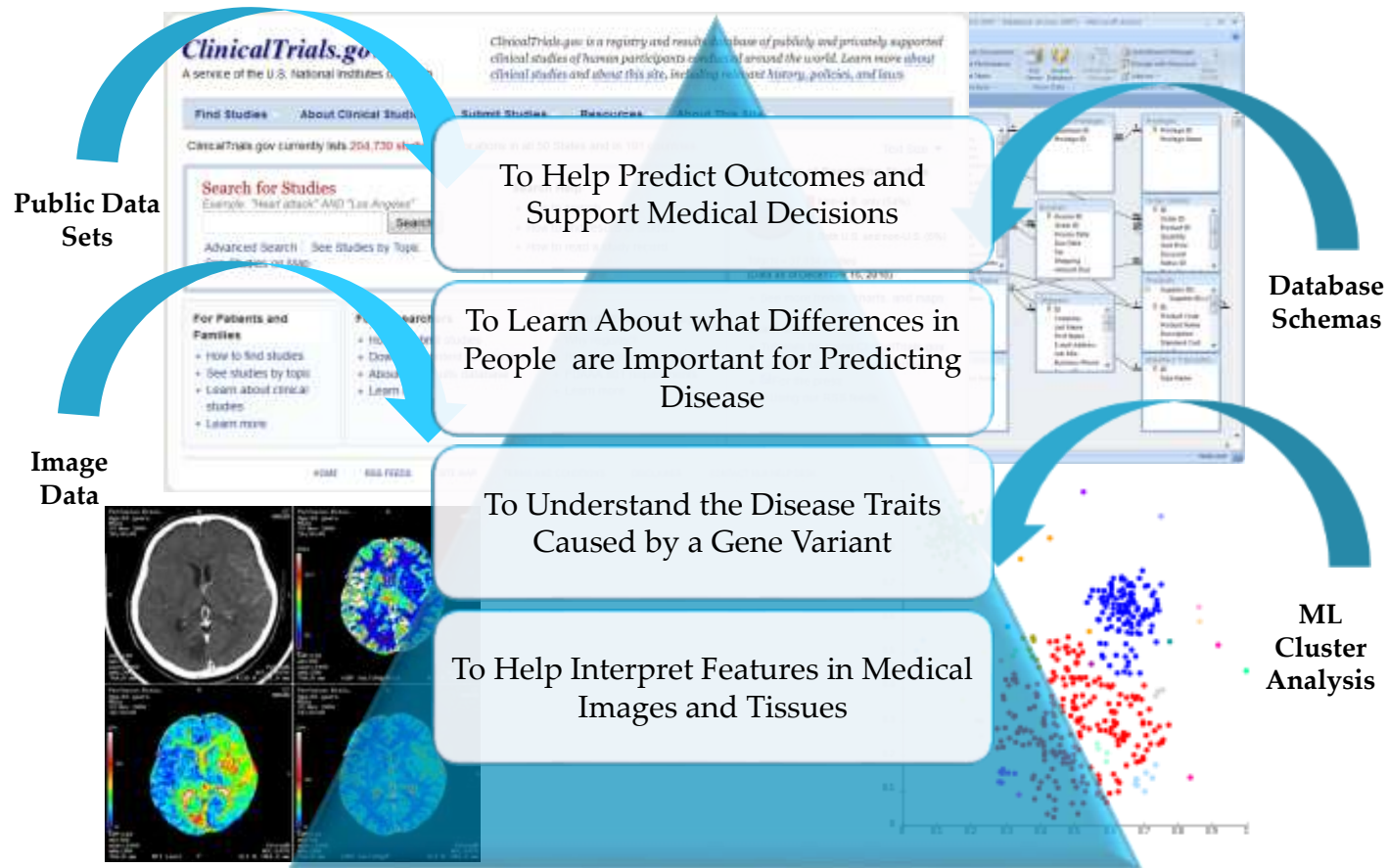


i2b2

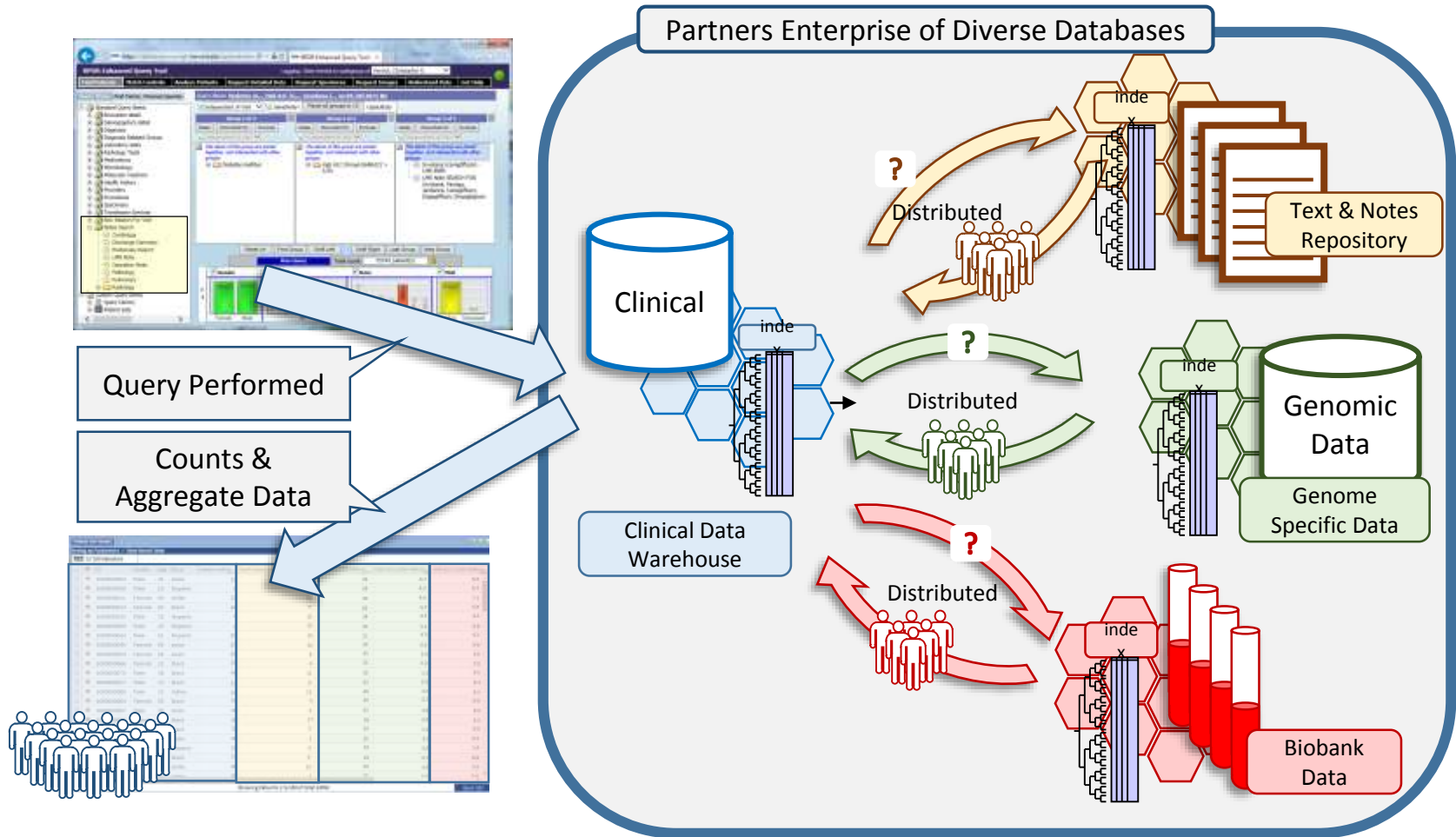
Joining with Big Data to Improve Healthcare Research Quality

Shawn Murphy MD, Ph.D.

Using Big Data to Improve Healthcare



The Researcher Querying the System interacts with a Simple Query Tool



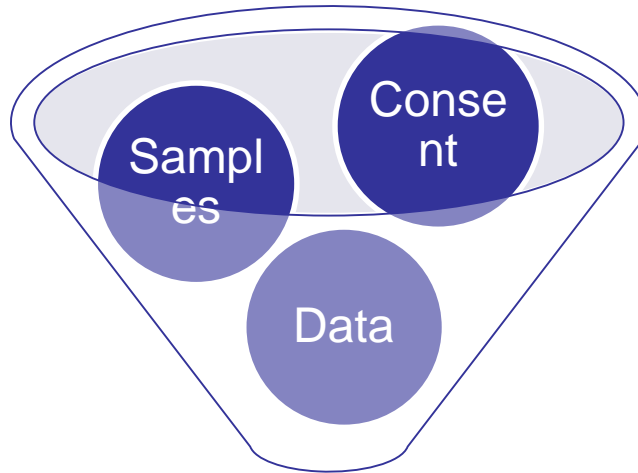
Can Find Patients and Gather Data Based on New Types of Searches ...

Nothing new for investigators to learn →

The screenshot displays the RPDR Enhanced Query Tool interface. The main window title is "RPDR Enhanced Query Tool" with a URL of "https://rpdresecure.mgh.harvard.edu/rpdrwebclien". The user is logged in as "Chris Herrick in workgroup of Herrick, Christopher D.". The interface features a navigation menu on the left with categories like "Standard Query Items" and "Notes Search". The main area shows a query configuration for "Diabetes m..., Hgb A1C (G..., Invokana (... on 09/28/2015 #6". It includes three groups of criteria, each with "Independent of Visit" and "Sensitivity" settings. The "Run Query" button shows a total count of 727±3 patient(s). Summary statistics are provided for Gender, Age, Race, and Vital status.

Category	Item	Count
Gender	Female	358±3
	Male	371±3
Age	Age 0-9 <3	6±3
	Age 9-19 <3	22±3
	40	95±3
	20±3	239±3
	80	107±3
	18±3	3±3
Race	A. Indian <3	23±3
	A	68±3
	B	20±3
	H	569±3
	W	27±3
	O U	22±3
Vital	Alive	719±3
	Deceased	6±3

The Partners Biobank



- The Partners Biobank provides samples (plasma, serum, and DNA) collected from consented patients.
- 40,000 patients have consented to date, 10,000 have been genotyped.
- Samples are available for distribution to Partners investigators* to help identify novel Personalized Medicine opportunities that reduce cost and provide better care

**with required approval from the Partners Institutional Review Board (IRB).*

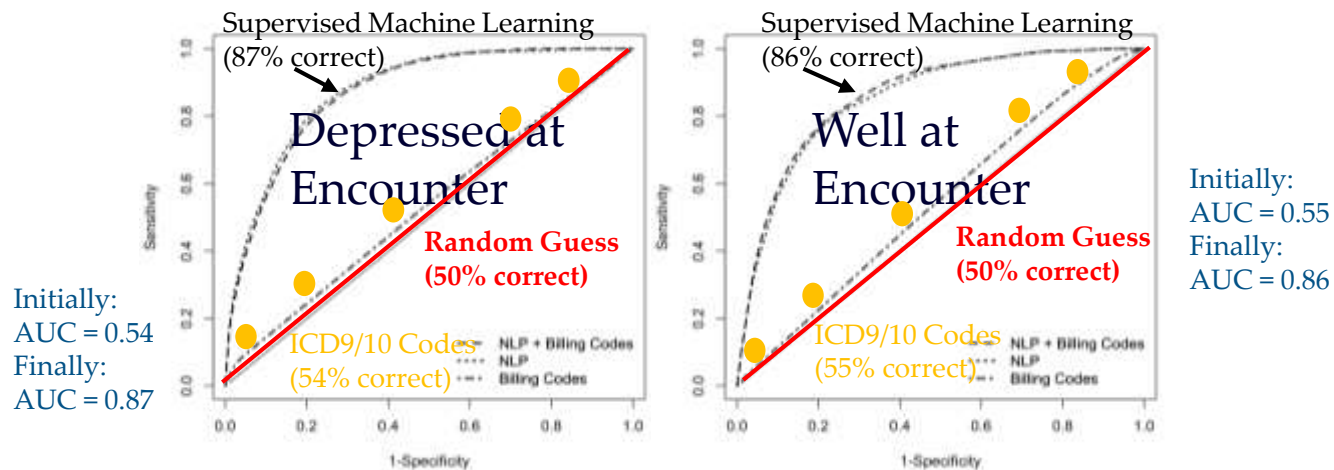
Improved Clinical Care for All Patients

Unpredictable Quality Using Raw ICD9/10 Codes

Phenotype	Count with ICD-9/ICD-10 Code	Count (90% positive predictive value)	Count with Genotype Data
Asthma	7618	3322	805
Bipolar Disorder	1754	219	84
Breast Cancer	2101	1711	378
Congestive Heart Failure	10160	4597	1859
Coronary Artery Disease	1435	803	236
Crohn's Disease	5177	700	350
Depression	11154	4273	1074
Epilepsy	2351	1211	381
Gout	2464	1828	566
Hypertension	20788	16995	4553
Multiple Sclerosis	602	320	58
Obesity	10245	12179	3191
Rheumatoid Arthritis	3475	878	261
Schizophrenia	509	83	14
Type 1 Diabetes	2196	232	61
Type 2 Diabetes	7123	4385	1268
Ulcerative Colitis	1359	624	157

May 4, 2016, n ~ 40,000

Phenotyping Algorithms to define cohorts of treatment-resistant and treatment-responsive depression



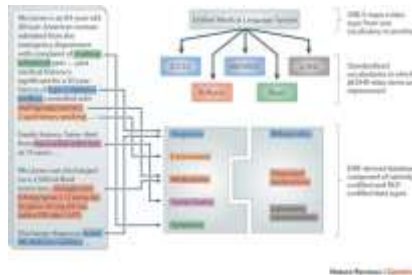
Clinical Status	Model	Specificity	Sensitivity	Precision	AUC
Depressed	Billing Codes	0.95	0.09 (0.03)	0.57 (0.14)	0.54 (0.02)
Depressed	NLP	0.95	0.42 (0.05)	0.78 (0.02)	0.88 (0.02)
Depressed	NLP + Billing Codes	0.95	0.39 (0.06)	0.78 (0.02)	0.87 (0.02)
Well	Billing Codes	0.95	0.06 (0.02)	0.26 (0.27)	0.55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

Creating Quality Data with Supervised Machine Learning

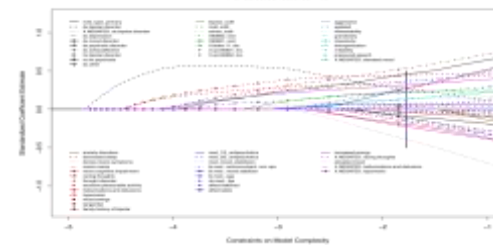
1. Create a gold standard training set.



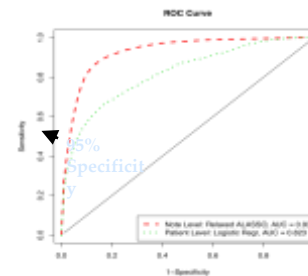
2. Create a comprehensive list of features (concepts/variables) that describe the phenotype of interest



3. Develop the classification algorithm. Using the data analysis file and the training set from step 1, assess the frequency of each variable. Remove variables with low prevalence. Apply adaptive LASSO penalized logistic regression to identify highly predictive variables for the algorithm



4. Apply the algorithm to all subjects in the superset and assign each subject a probability of having the phenotype



Predictably Quality - Biobank Portal Computed Phenotypes

Phenotype	Count with ICD-9/ICD-10 Code	Count (90% positive predictive value)	Count with Genotype Data
Asthma	7618	3322	805
Bipolar Disorder	1754	219	84
Breast Cancer	2101	1711	378
Congestive Heart Failure	10160	4597	1859
Coronary Artery Disease	1435	803	236
Crohn's Disease	5177	700	350
Depression	11154	4273	1074
Epilepsy	2351	1211	381
Gout	2464	1828	566
Hypertension	20788	16995	4553
Multiple Sclerosis	602	320	58
Obesity	10245	12179	3191
Rheumatoid Arthritis	3475	878	261
Schizophrenia	509	83	14
Type 1 Diabetes	2196	232	61
Type 2 Diabetes	7123	4385	1268
Ulcerative Colitis	1359	624	157

May 4, 2016, n ~ 40,000

Definitive Disease States Associated with Genomic Data

Phenotype	Count with ICD-9/ICD-10 Code	Count (90% positive predictive value)	Count with Genotype Data
Asthma	7618	3322	805
Bipolar Disorder	1754	219	84
Breast Cancer	2101	1711	378
Congestive Heart Failure	10160	4597	1859
Coronary Artery Disease	1435	803	236
Crohn's Disease	5177	700	350
Depression	11154	4273	1074
Epilepsy	2351	1211	381
Gout	2464	1828	566
Hypertension	20788	16995	4553
Multiple Sclerosis	602	320	58
Obesity	10245	12179	3191
Rheumatoid Arthritis	3475	878	261
Schizophrenia	509	83	14
Type 1 Diabetes	2196	232	61
Type 2 Diabetes	7123	4385	1268
Ulcerative Colitis	1359	624	157

May 4, 2016, n ~ 40,000

Query 1.68 billion rows of Genomic Data for Specific Variants

The screenshot displays the Partners Biobank Portal interface. The browser address bar shows the URL <https://biobankportaldev.partners.org/webclient/>. The page header includes the Partners Biobank Portal logo and navigation buttons for "Find Patients", "Make Request", "Help & Support", and the user name "Shawn Murphy, MD".

The main content area is divided into two panels. The left panel, titled "Navigate Terms", shows a tree view of data categories. The right panel, titled "Query Tool", contains the query configuration and results.

Query Tool Configuration:

- Query Name: SNP DT:GENOTYPE@09:39:53
- Temporal Constraint: Treat Independently
- Group 1: SNP DT:GENOTYPE = ("rs9574622 AND Heterozygous AND A_to_C")

Results:

Number of patients: **315**
For Query "SNP DT:GENOTYPE@09:39:53"

High Quality Data Available for Genomics Queries

The screenshot displays the Partners Biobank Portal interface. On the left, a navigation tree shows various data categories, including Biobank Genomics, with a sub-entry for 'RA - current or past history (PPV 0.90) - 717'. The central 'Query Tool' window shows a query named 'RA -- Adali-Illum@15:46:12' with a temporal constraint of 'Treat all groups independently'. It defines three groups: Group 1 (RA - current or past history), Group 2 (Adalimumab), and Group 3 (Illumina Multi-Ethnic GWAS/Exome SNP Array). Below the query tool, a 'Graph Results' tab shows a bar chart with the number '70' and the text 'For Query "RA -- Adali-Illum@15:46:12"'. On the right, a 'Biobank Portal Query Report' window provides a detailed definition of the query, listing the three groups and their associated phenotypes and data sources.

Query Name: RA -- Adali-Illum@15:46:12

Temporal Constraint: Treat all groups independently

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Treat independently			Treat independently			Treat independently		
RA - current or past history (PPV 0.90) - 717			Adalimumab - 689			Illumina Multi-Ethnic GWAS/Exome SNP Array - 4930		

Run Query **Clear** 3 Groups **New Group**

Show Query Status **Graph Results** **Query Report** **Download Data**

Number of patients

70

For Query "RA -- Adali-Illum@15:46:12"

Biobank Portal Query Report

The query is entitled "RA -- Adali-Illum@15:46:12"

Query Definition

Temporal constraint: Treat all groups independently

All Groups

- RA - current or past history (PPV 0.90)
Phenotypes: Rheumatoid Arthritis (RA) / RA - current or past history (PPV 0.90)
Independent of Visit
From earliest date available to latest date available
of times an item is recorded is: 2

AND

- Adalimumab
Medications / Immunologic agents / monoclonal antibodies / adalimumab
Independent of Visit
From earliest date available to latest date available
of times an item is recorded is: 0

AND

- Illumina Multi-Ethnic GWAS/Exome SNP Array
Illumina Multi-Ethnic GWAS/Exome SNP Array
Independent of Visit
From earliest date available to latest date available
of times an item is recorded is: 0

Partners Biobank Portal

biobankportaldev.partners.org/devclient/

PARTNERS HEALTHCARE BIOBANK PORTAL Find Patients Make Request Help & Support Wattanasin, Nich

Navigate Terms Find

- Biobank Consent Information
- Biobank Demographics
- Biobank Genomics
 - People with genomic data - 4930
 - Illumina Multi-Ethnic GWAS/Exome SNP Array - 4930
- Biobank Health Information Survey
- Biobank Sample Types
- Curated Disease Populations
 - Bipolar Disorder (BD)
 - Congestive Heart Failure (CHF)
 - Coronary Artery Disease (CAD)
 - Crohn's Disease (CD)
 - Multiple Sclerosis (MS)
 - Rheumatoid Arthritis (RA)
 - RA - current or past history (PPV 0.90) - 717
 - RA - no history (NPV 0.99) - 24086
 - Type 2 Diabetes Mellitus (T2DM)
 - Ulcerative Colitis (UC)
- Healthcare Data
- Healthy Populations (Controls)

Query Tool


Query Name: RA - --Adali-Illum@15:46:12

Temporal Constraint: Treat all groups independently

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Treat Independently			Treat Independently			Treat Independently		
RA - current or past history (PPV 0.90) - 717			Adalimumab - 689			Illumina Multi-Ethnic GWAS/Exome SNP Array - 4930		
one or more of these			AND			one or more of these		
			AND			one or more of these		

Run Query Clear 3 Groups New Group

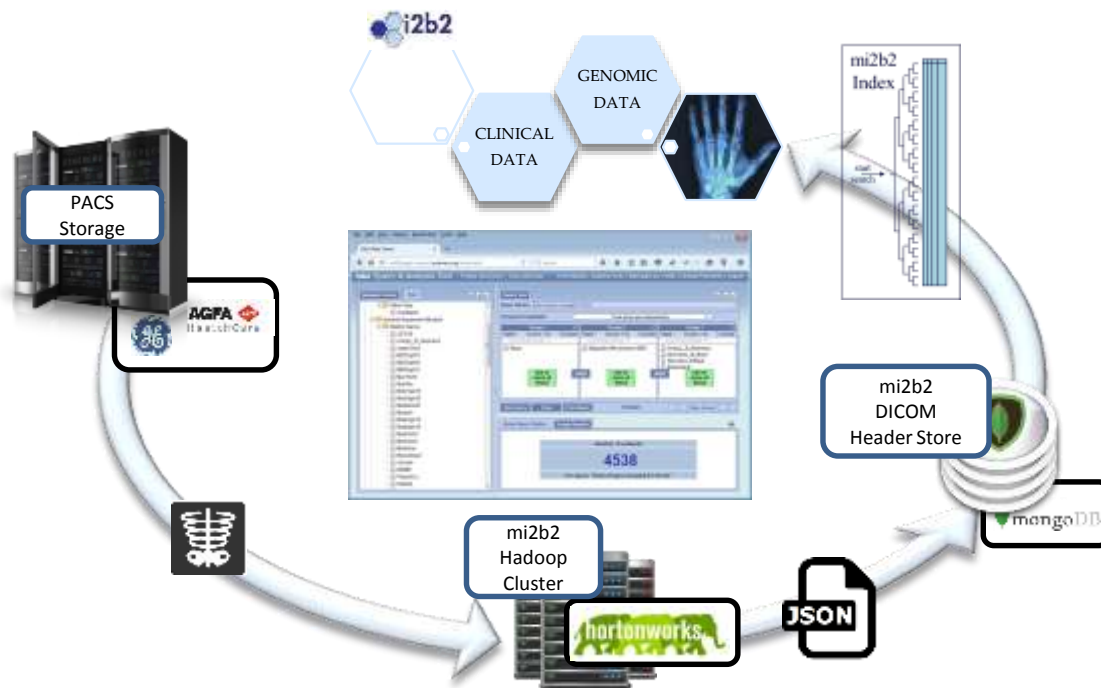
Show Query Status Graph Results Query Report **Download Data**

 You can download the de-identified data for this query as an Excel/CSV file. [Proceed to the Download](#)

Partners Biobank Portal – Download De-Identified Data

Imaging – DICOM Image Index allows imaging data to join other clinical and genomic data in queries

- Investigators will be able to define sets of patients who are relevant to their research by defining the specific type of image required for their analysis (e.g. high resolution).
- Through the Big Data Commons, Investigators will be able to link this patient cohort to other available data (genomic data, biobank samples, other research data, EHR data, etc)



Impact to Clinical care

**Linking to EMR with
SMART “Apps”**

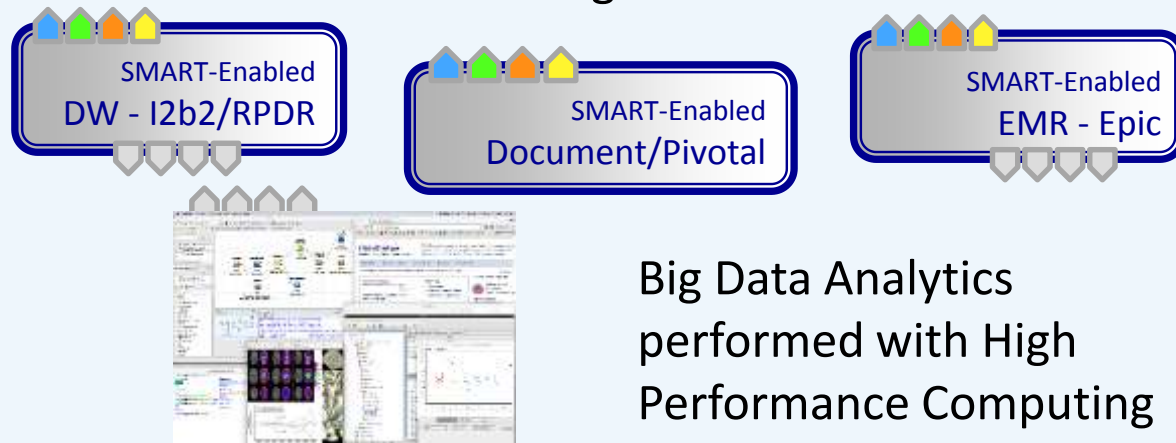
Published 2011

Enabling Innovation to reach into EMR

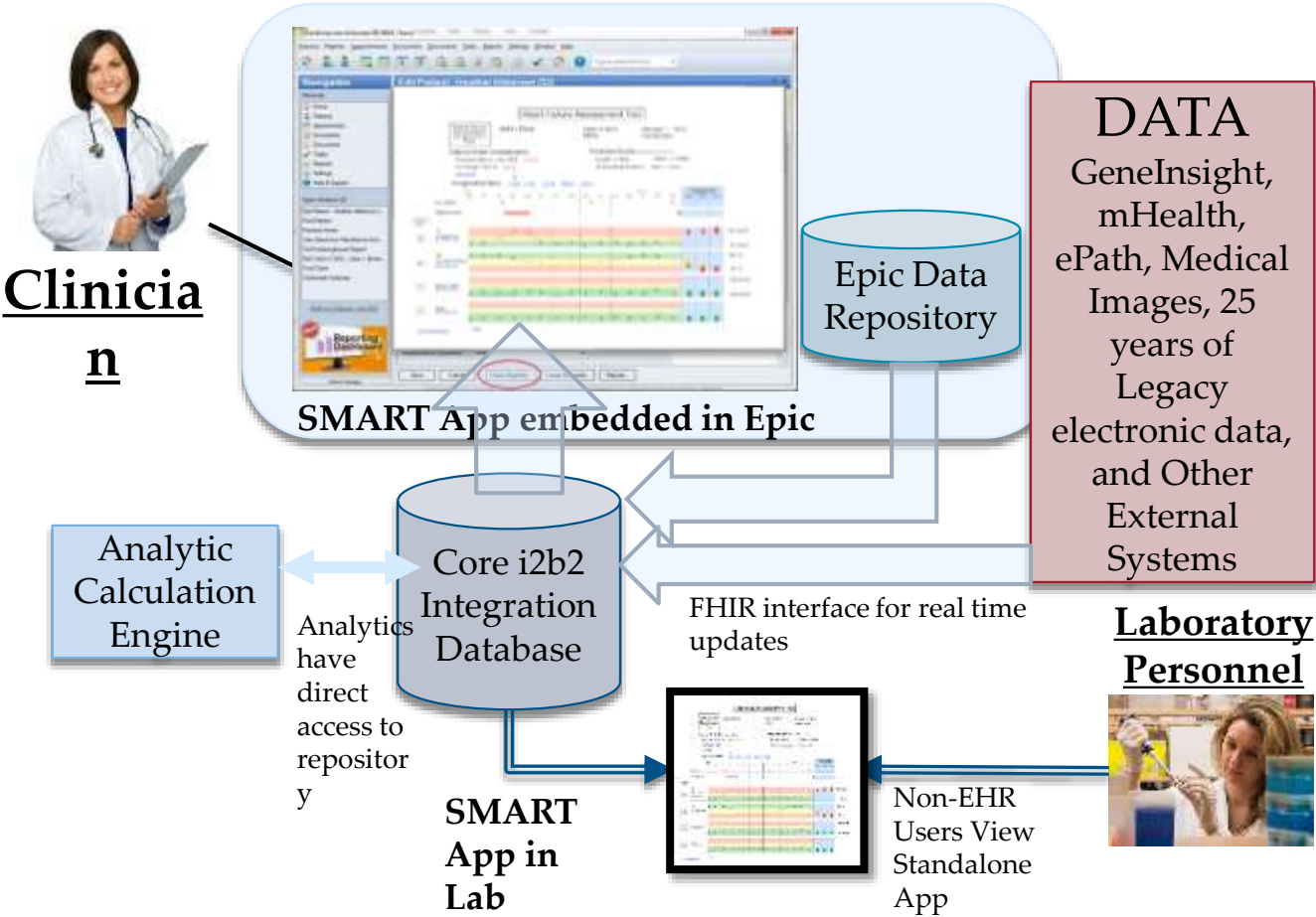
SMART
Apps run
in EMR



Partners Big Data Commons

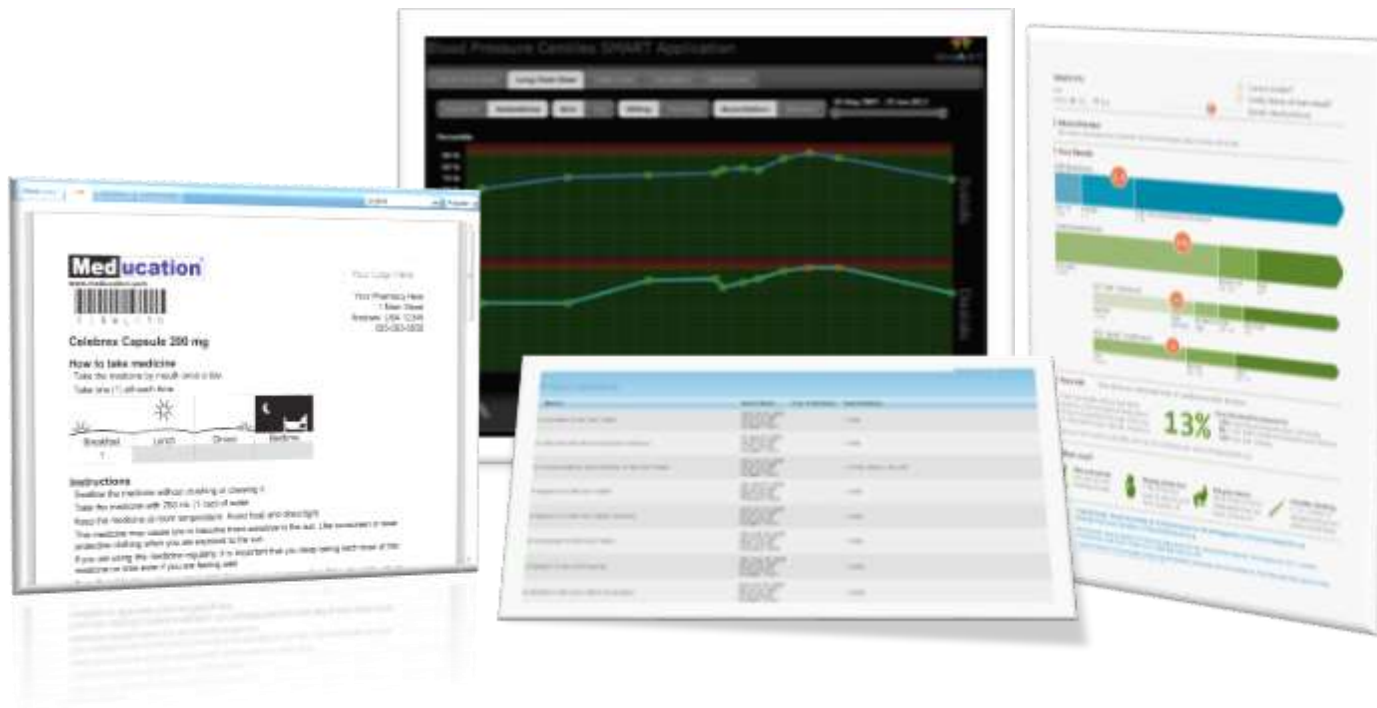


Bringing Big Data into Clinical Care with Open App Development



Out of the Box - SMART Apps link Big Data to the EMR

- Substitutable Medical Application and Reusable Technology – Started with grant from the Office of the National Coordinator
- Paradigm is similar to Mobile Apps with a proposed standard interface using FHIR (Fast Healthcare Interoperable Resource)



What Big Data can do for the Everyday Clinician - Finding Similar Patients

- Looking at similar patients can help predict:
 - Future outcomes and responses to therapy
 - Course of disease
 - Penetrance of genetic variants
 - Likelihood that a diagnostic pathway might be fruitful
- Finding similar patients is very computationally intensive, but a perfect opportunity for combining data from the Electronic Health Record, Specialized Health Databases, Analytics from Big Data Queries, and presentation in SMART Apps
- Presentation of results can be greatly enhanced with engaging visualizations for the provider making difficult, complex decisions

Growth Charts reinvented as a SMART app – is this child similar to other children?

<https://gallery.smarthealthit.org/boston-childrens-hospital/growth-chart>



SMART App Gallery - Growth Charts Application

https://app-data.unahealth.org/growth-chart/

Paul Luttrell sex: male dob: 01Aug2003 age: 12y 8m corrected age: 12y 8m

App Version: 0.0.1-BETA

GRAPHS TABLE PARENT

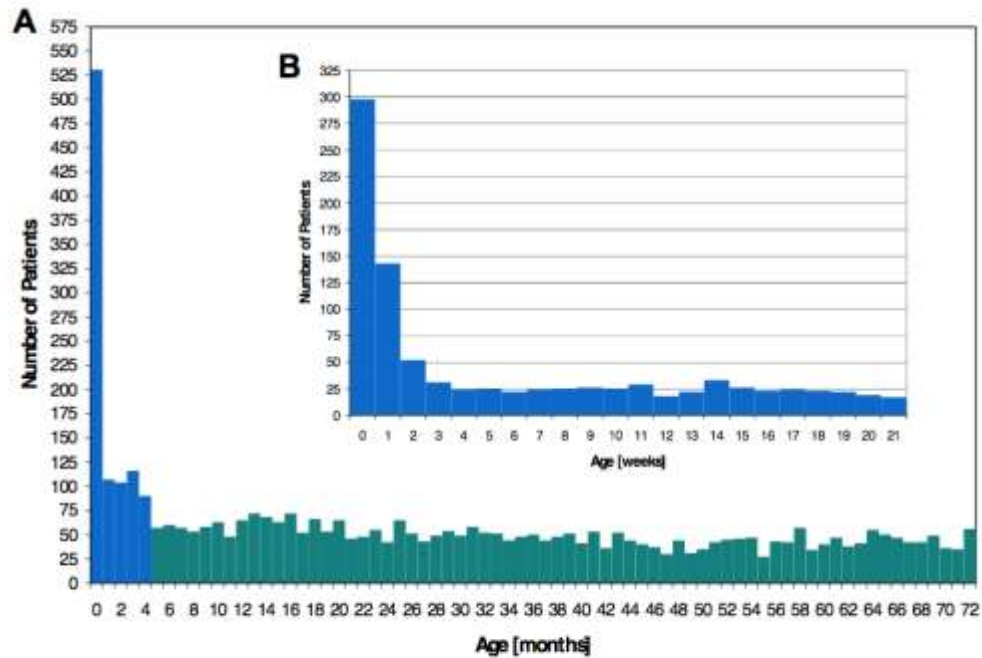
0 - 13 Weeks 0 - 6 Months 0 - 2 Years 0 - 20 Years Fit to Age

Last recording: 04Feb2013 By 6m

Entry Date	08Aug2006	08Aug2007	07Aug2008	05Feb2010	06Aug2010	14Mar2011	06Aug2011	13Mar2012	12Sep2012	04Feb2013	
Age	3y 6d	4y 6d	5y 5d	6y 6m	7y 5d	7y 7m	8y 5d	9y 7m	9y 1m	9y 6m	
Annotation	See at	—	—	—	—	—	—	—	—	—	
Length	cm	87.6	95	100.2	108.2	111.1	114	118.1	119.6	122.3	122.9cm
Percentile	%	2	4	5	2	2	2	4	2	5	2
Z Score	Z	-2.1	-1.8	-1.9	-2	-2	-2.1	-1.7	-2	-2	-2
Velocity	cm/y	5.6	5.3	6.3	5.2	5.1	5.2	3.9	4.8	4	To here
Weight	kg	12.1	13.7	16.1	17.9	18.7	20.3	21.7	23.6	24.2	24.4kg
Percentile	%	6	6	4	6	5	7	11	15	12	6
Z Score	Z	-1.6	-1.6	-1.7	-1.6	-1.6	-1.4	-1.2	-1	-1.2	-1.4
Velocity	kg/y	1.8	1.9	2.1	2.2	2.3	2.2	1.6	0.9	0.5	To here
Head C	cm	—	—	—	—	—	—	—	—	—	—
Percentile	%	—	—	—	—	—	—	—	—	—	—
Z Score	Z	—	—	—	—	—	—	—	—	—	—
Velocity	cm/y	—	—	—	—	—	—	—	—	—	To here
BMI	kg/m ²	14.4	14	14	14.4	14.1	14.6	14.4	14.6	14.9	14.6

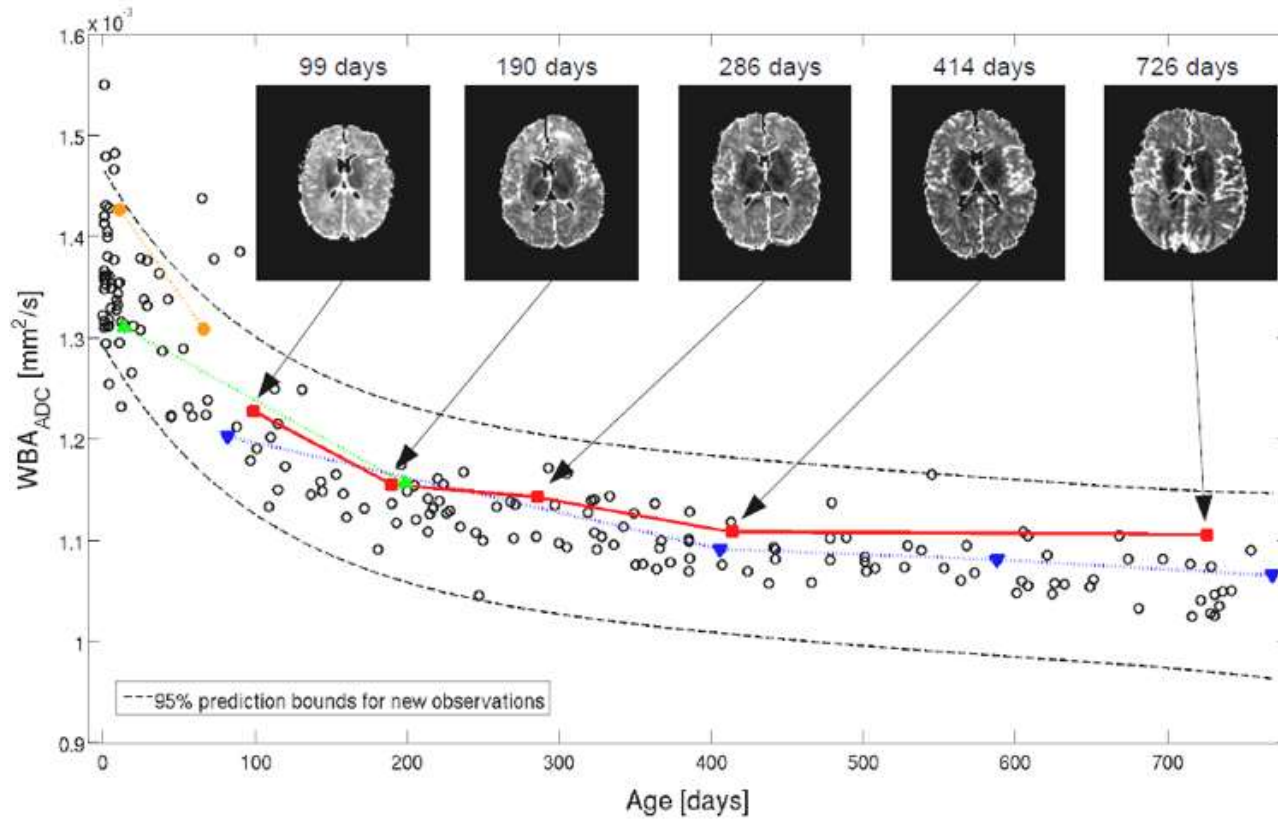


Find Normal MRI's at All Ages 0-6 y/o



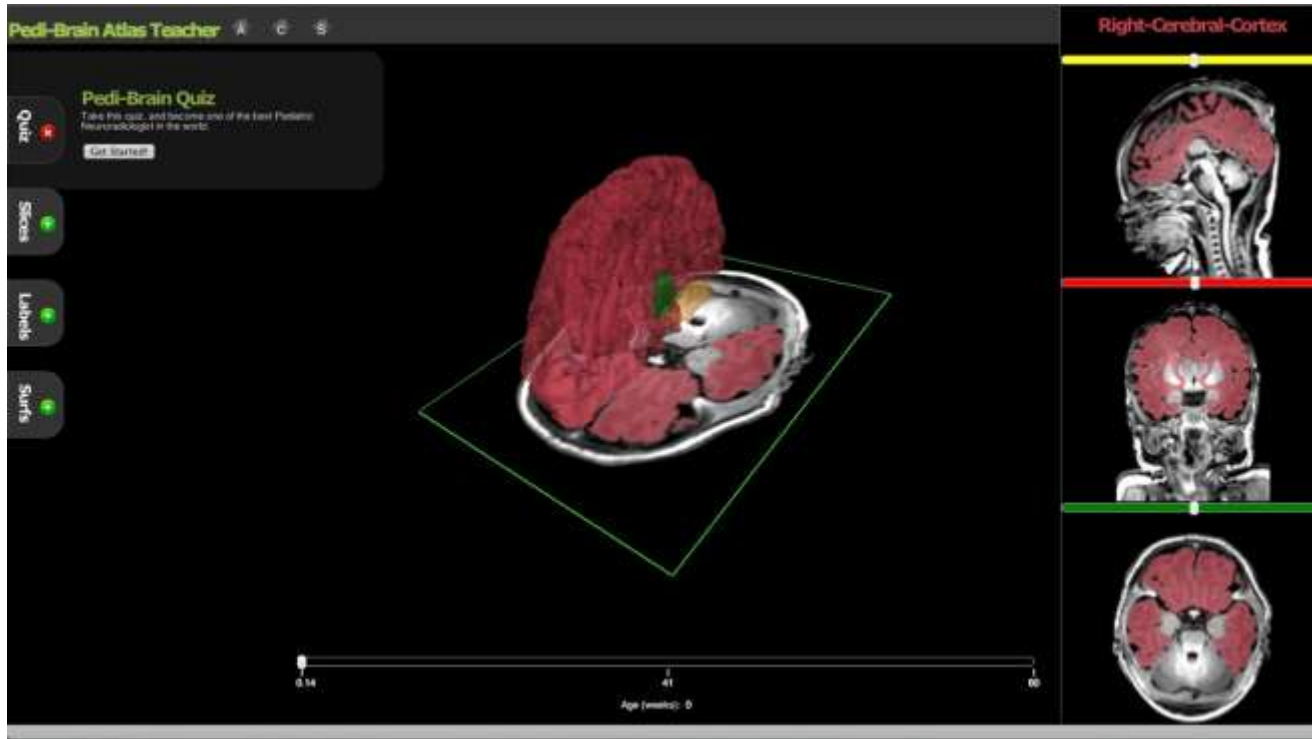
Number of patients who had a brain MRI scan at a particular age in months from 0 to 6 years (A) and in weeks from 0 to 4 months (B)

Determining a Normal Child's MRI



Normative Pediatric MRI

Generating quantitative atlases for regular intervals in pediatric development to be used for clinical brain MRI analysis



Comparing Medication Use Patterns to Determine Non-Compliance

<https://mpr-monitor.smarthealthit.org/fhir-app/risk.html>

The screenshot shows a web browser displaying the Medication Possession Ratio Monitor interface. The page title is "Medication Possession Ratio Monitor" and the patient name is "Carol G. Allen". The interface is divided into sections for different medication classes: ANTIHYPERLIPIDEMICS, ANTIHYPERTENSIVES, and ORAL_HYPOGLYCEMICS. Each section contains a table with columns for Medication, First fill, Last fill, and Adherence.

Medication Possession Ratio Monitor

Home About Med details All meds

Patient: Carol G. Allen

ANTIHYPERLIPIDEMICS

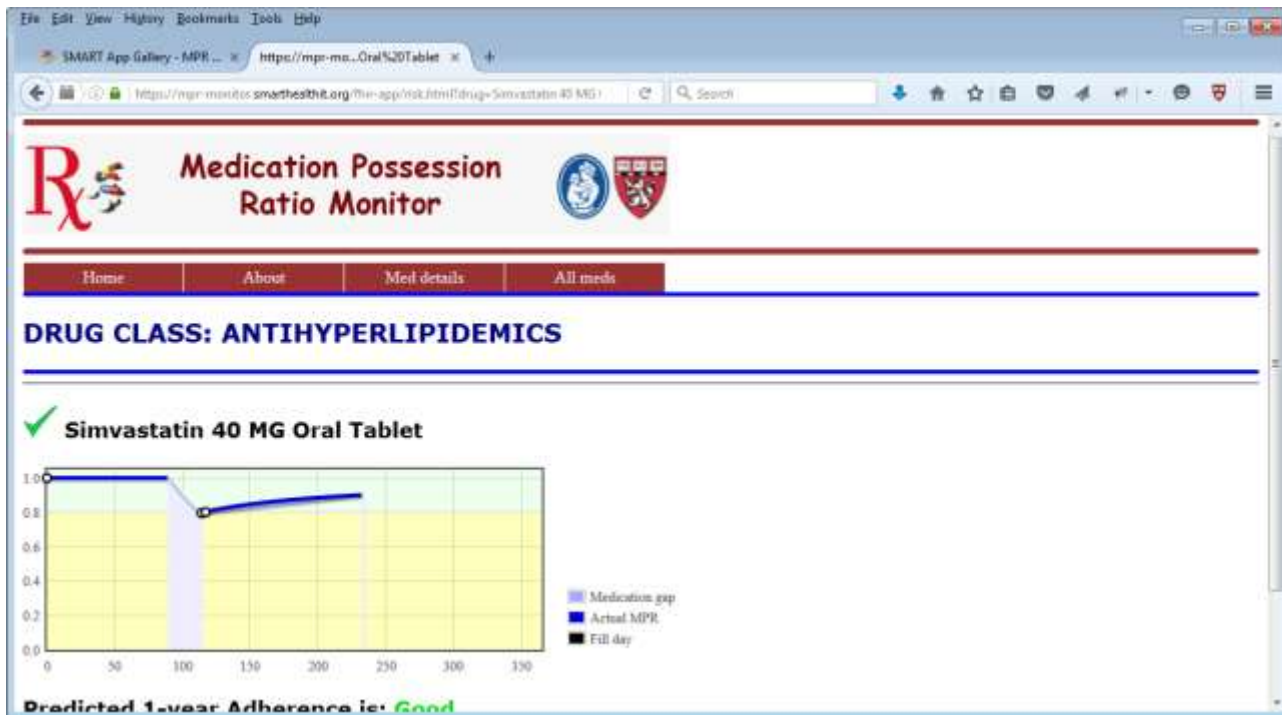
Medication	First fill	Last fill	Adherence
Simvastatin 40 MG Oral Tablet	March 23, 2009	July 18, 2009	✓

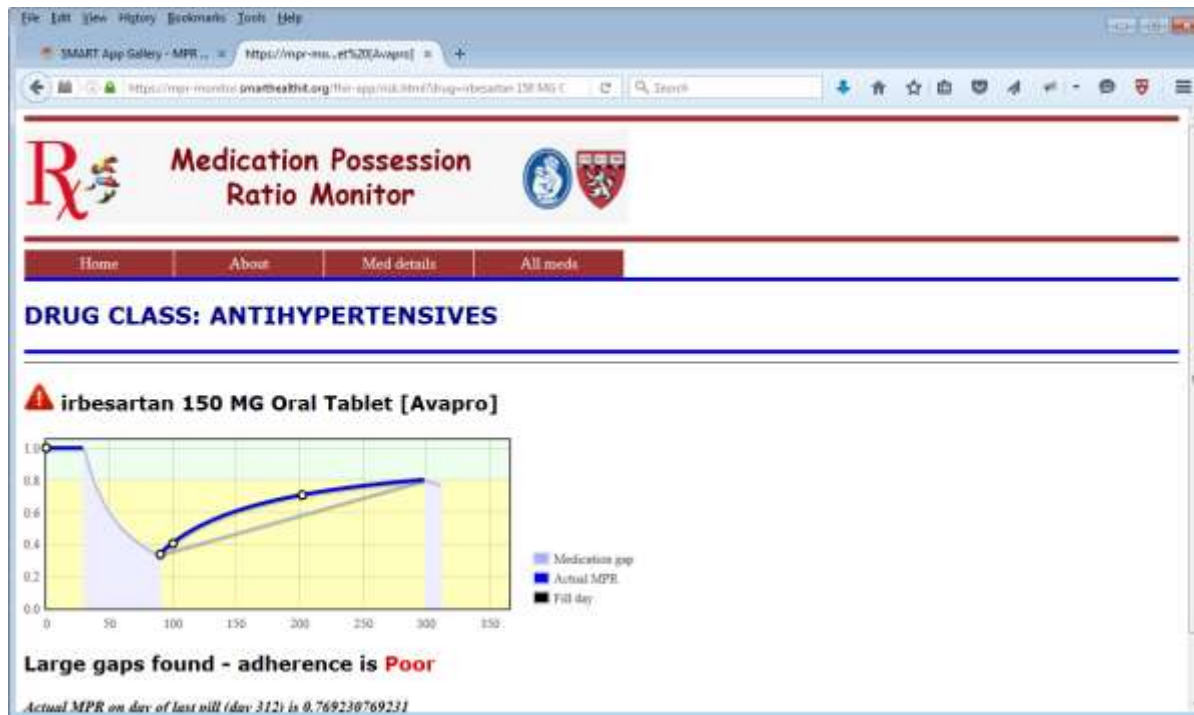
ANTIHYPERTENSIVES

Medication	First fill	Last fill	Adherence
irbesartan 150 MG Oral Tablet [Avalirel]	Feb. 10, 2009	Aug. 31, 2009	⚠

ORAL_HYPOGLYCEMICS

Medication	First fill	Last fill	Adherence
------------	------------	-----------	-----------





Tribute to...

- RPDR/I2b2 Core Team

- Christopher Herrick
- Michael Mendis
- Lori Phillips
- Janice Donahoe
- Nich Wattanasin
- Wayne Chan
- Vivian Gainer
- Alyssa Goodson
- Mariah Mitchell
- Martin Rees
- Charles Wang
- Laurie Bogosian
- Stacey Duey
- Andrew Cagan
- David Wang

- Biobank Team

- Natalie Boutin
- Victor Castro
- Scott Weiss
- Beth Karlson

- SMART Team

- Ken Mandl
- Josh Mandel
- Kavi Wagholikar

- Genomics Innovation Team

- Sandy Aronson
- Heidi Rehm
- Calum MacRea

... and many more.



i2b2

I2b2 and SMART Information and Software on the Web

i2b2 Homepage (<https://www.i2b2.org>)

i2b2 Software (<https://www.i2b2.org/software>)

i2b2 Community Site (<https://community.i2b2.org>)

SMART Platforms Homepage (<http://smarthealthit.org>)