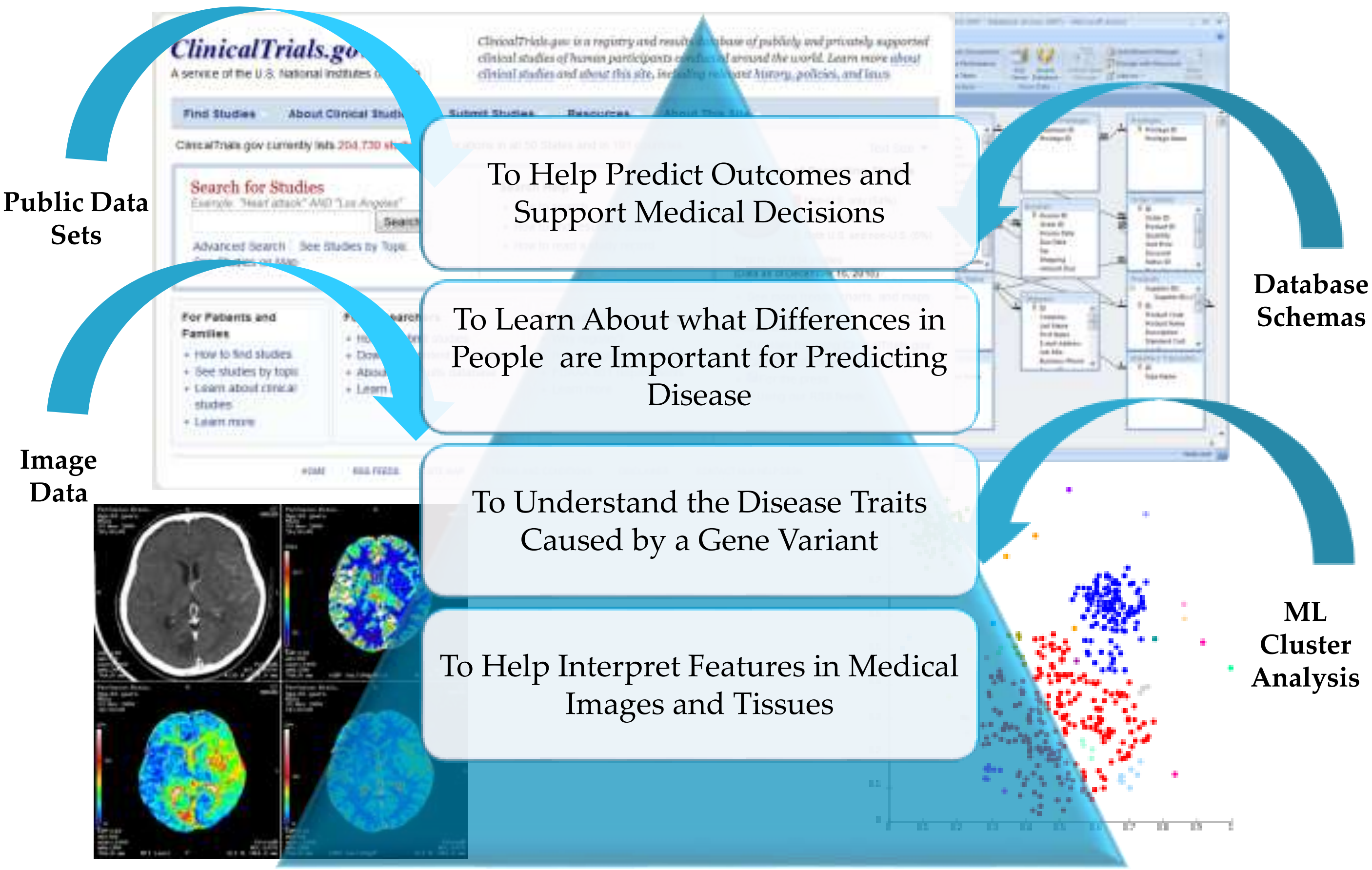
i2b2

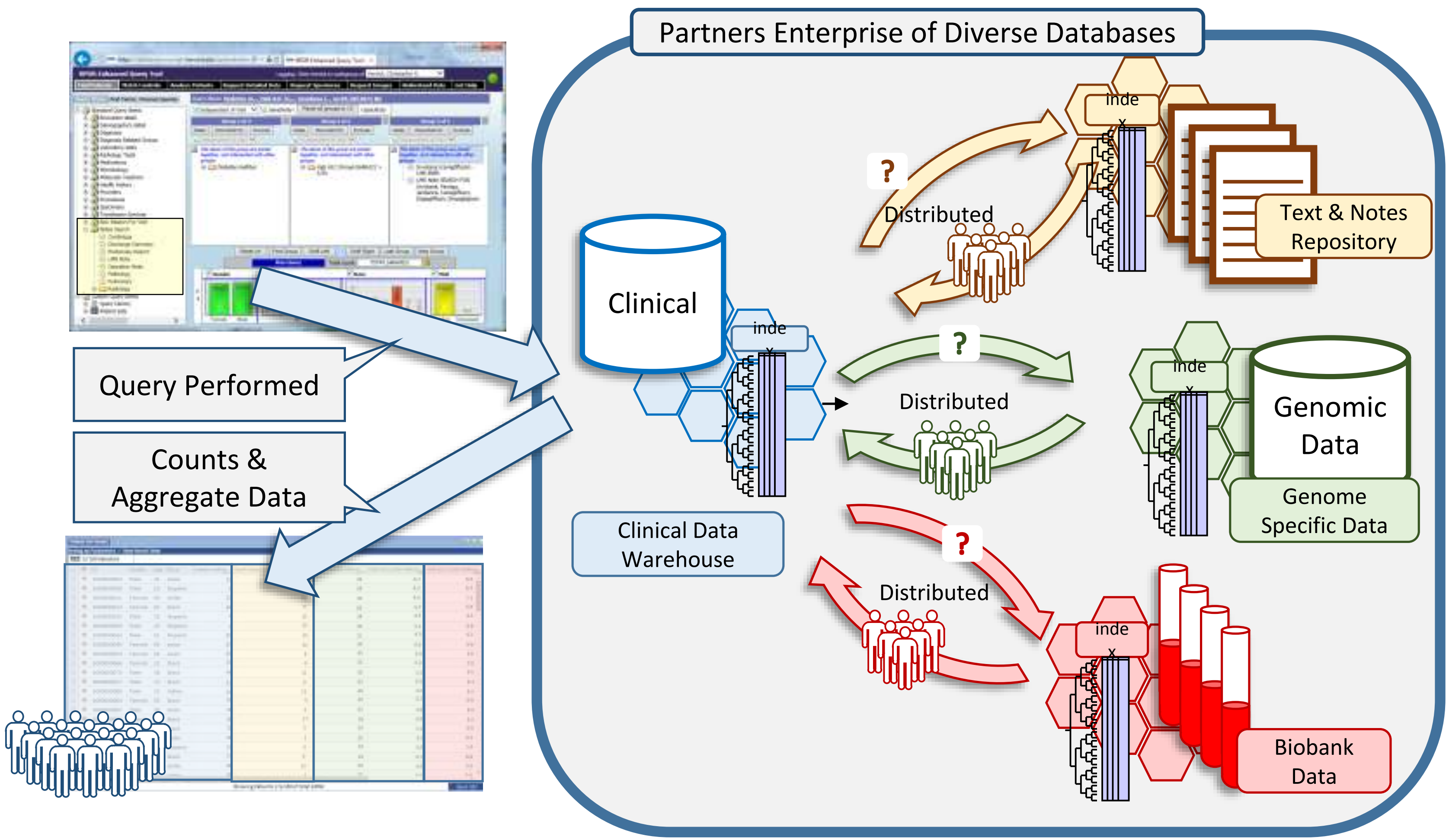
Joining with Big Data to Improve Healthcare Research Quality

Shawn Murphy MD, Ph.D.

Using Big Data to Improve Healthcare

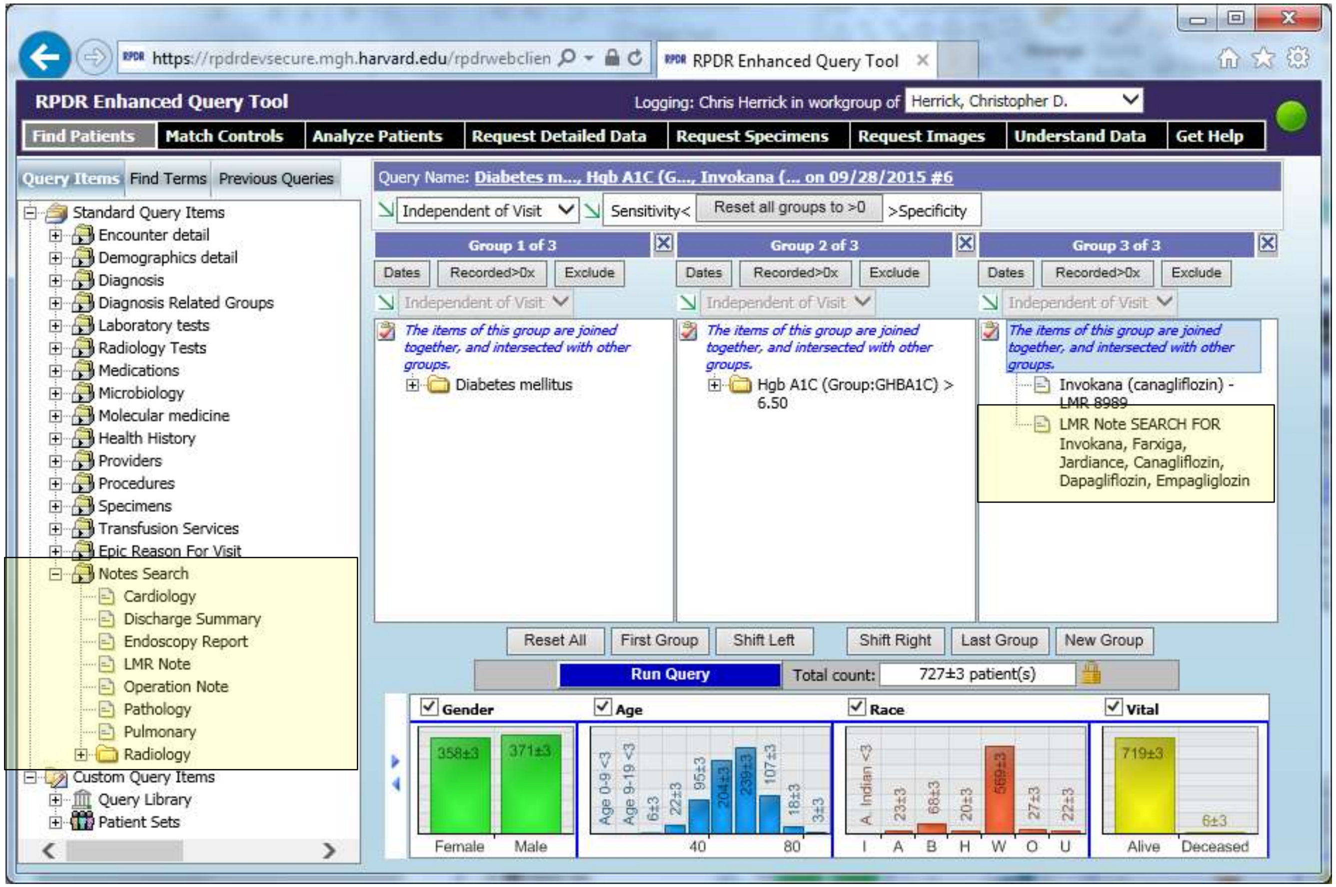


The Researcher Querying the System interacts with a Simple Query Tool



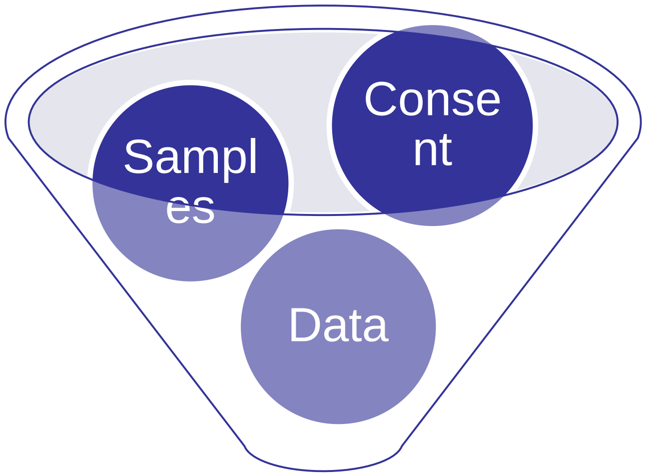
PARTNERS
HEALTHCARE

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



Nothing new for investigators to learn

The Partners Biobank



Research Discoveries

Improved Clinical Care for All Patients

- The Partners Biobank provides samples (plasma, serum, and DNA) collected from consented patients.
- 40,000 patients have consented to date, 10000 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).*

PARTNERS HEALTHCARE

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

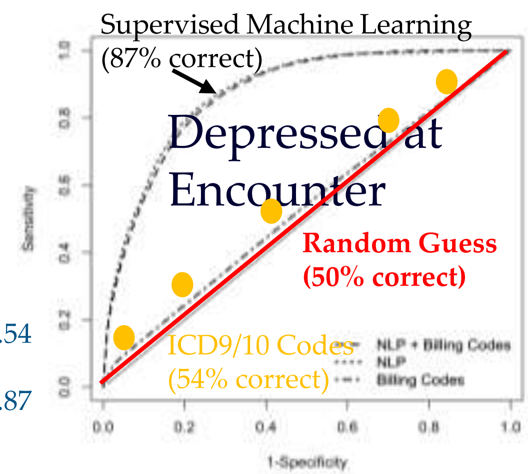
Psychological Medicine (2012), 42, 41–50. © Cambridge University Press 2011
doi:10.1017/S0033291711000997

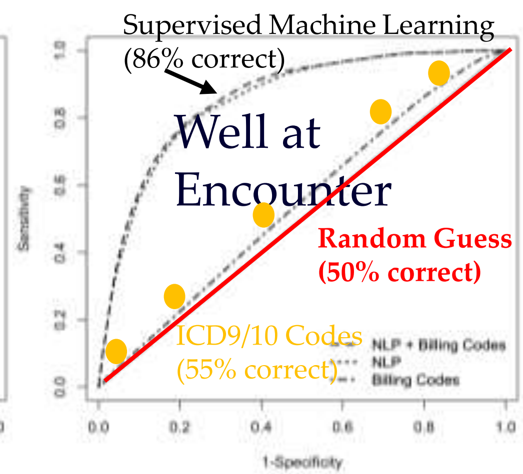
ORIGINAL ARTICLE

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

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

Initially:
AUC = 0.54
Finally:
AUC = 0.87





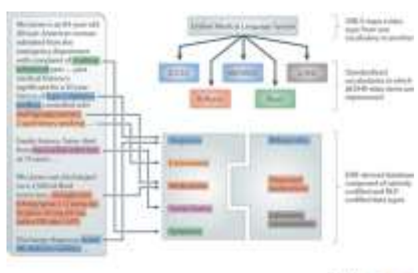
Initially:
AUC = 0.55
Finally:
AUC = 0.86

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)

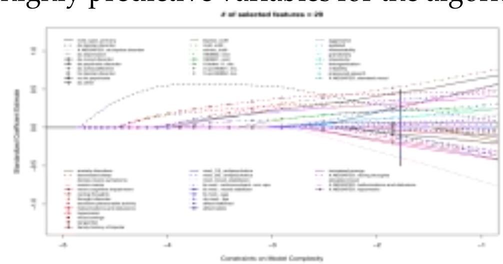
Creating Quality Data with Supervised Machine Learning

1. **Create a gold standard training set.**

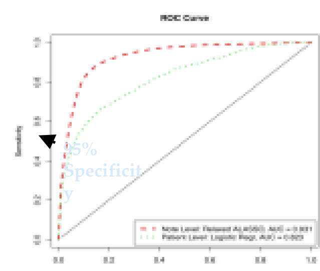
2. **Create a comprehensive list of features** (concepts/variables) that describe the phenotype of interest



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

PARTNERS HEALTHCARE

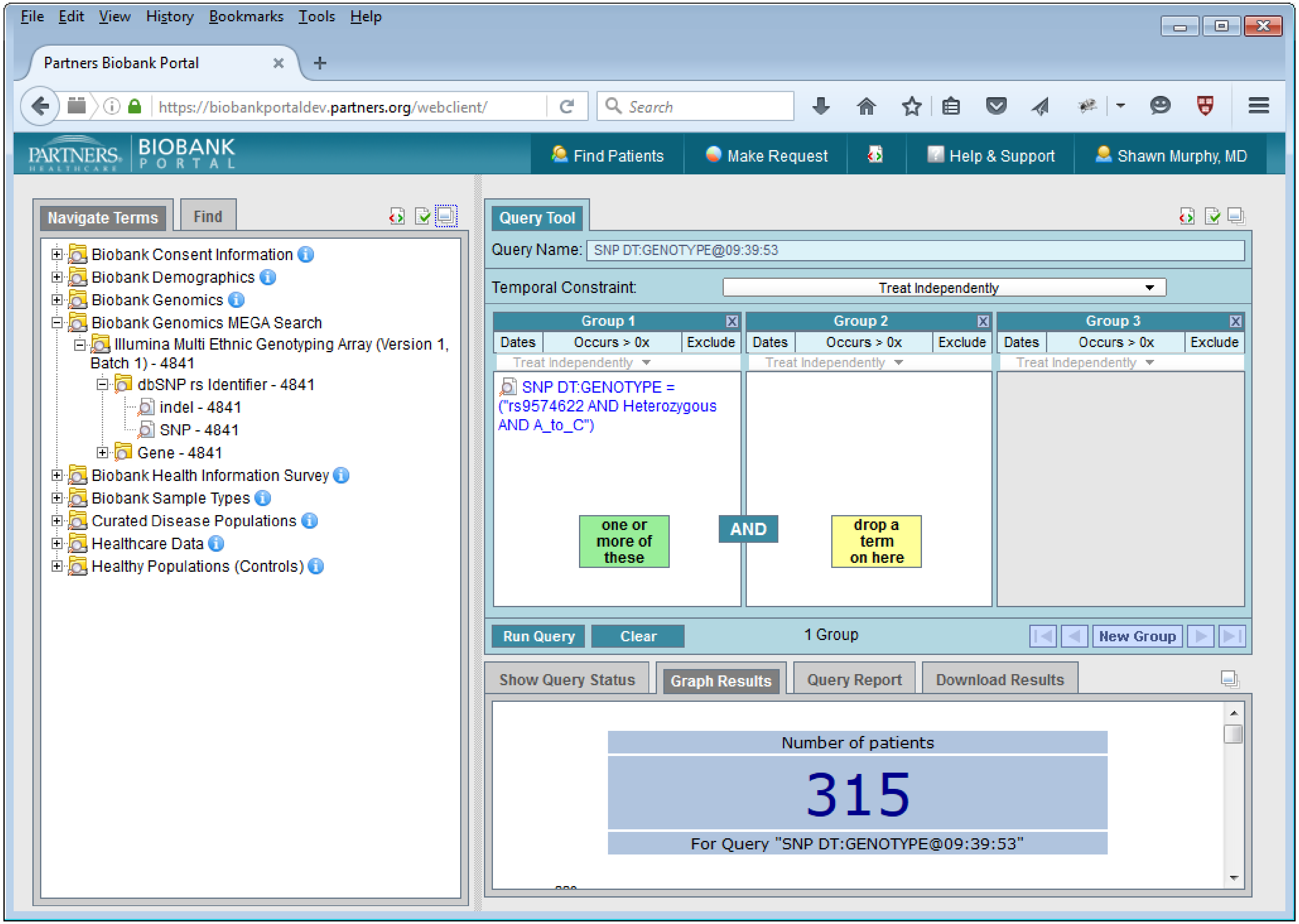
Definitive Disease States Associated with Genomic Data

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

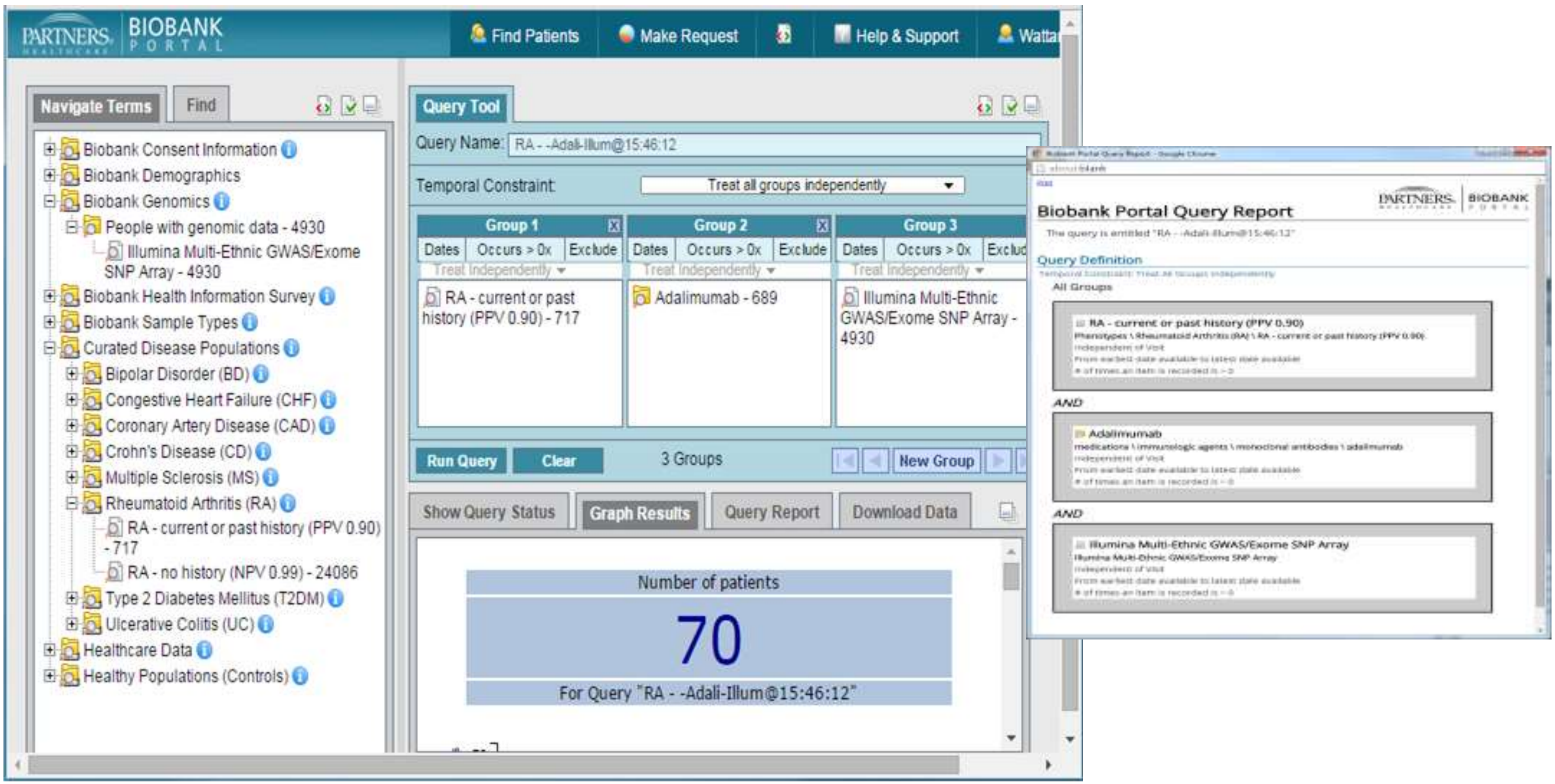
May 4, 2016, n ~ 40,000

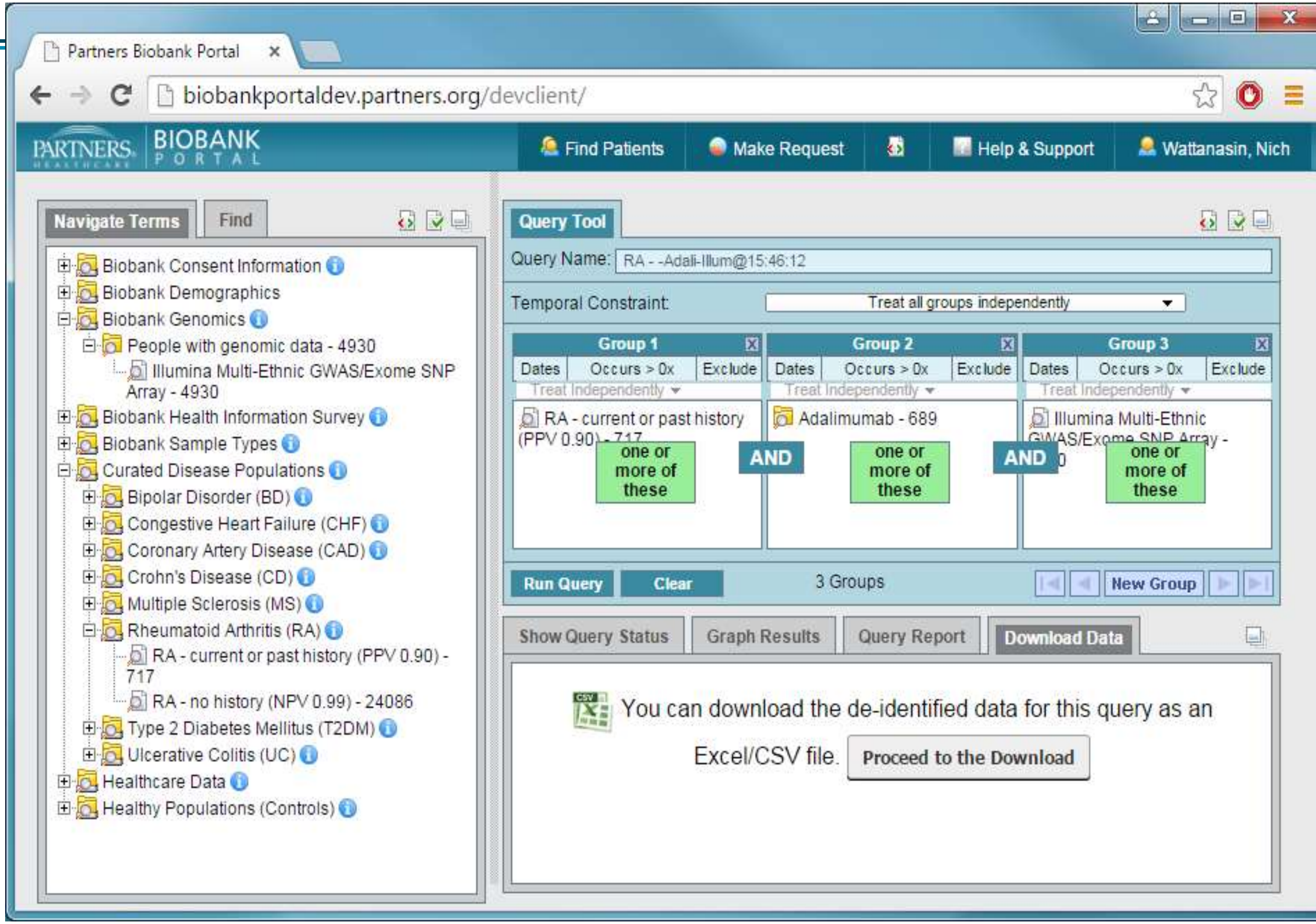
PARTNERS HEALTHCARE

Query 1.68 billion rows of Genomic Data for Specific Variants



High Quality Data Available for Genomics Queries

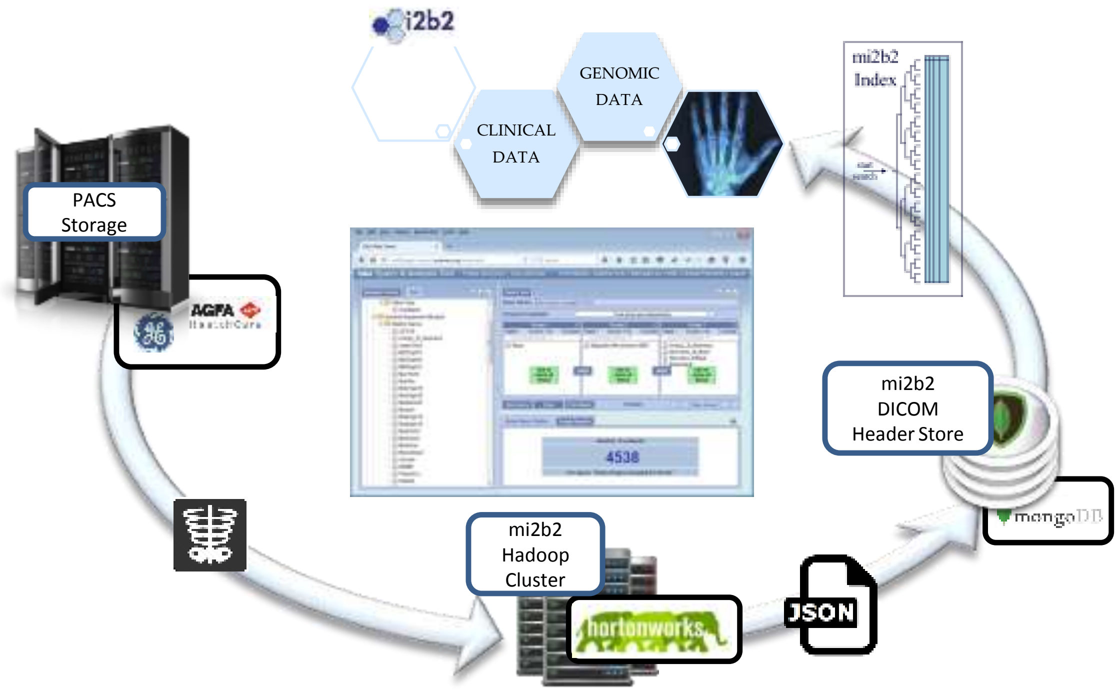




Partners Biobank Portal – Download De-Identified Data

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)

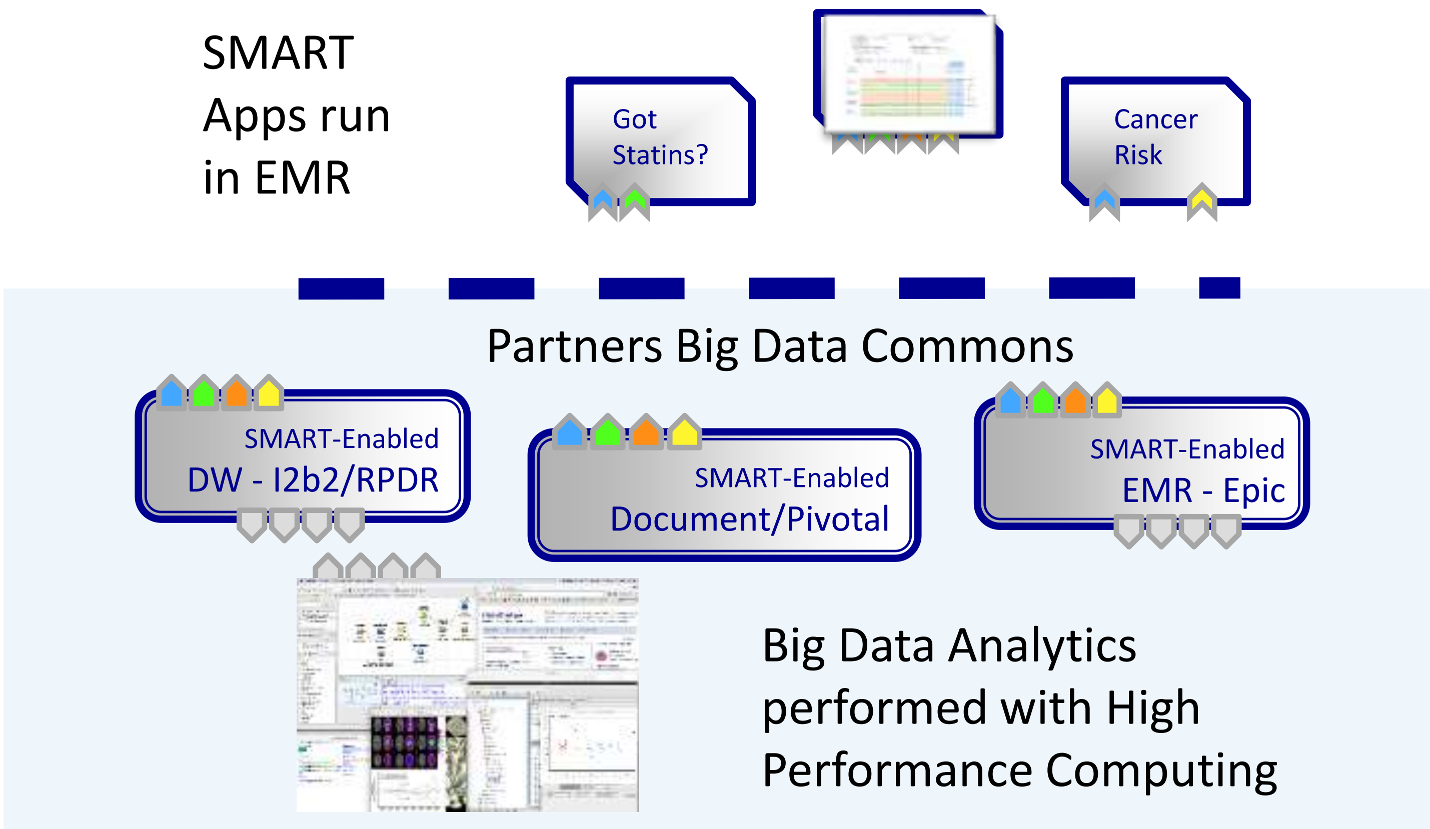


Impact to Clinical care

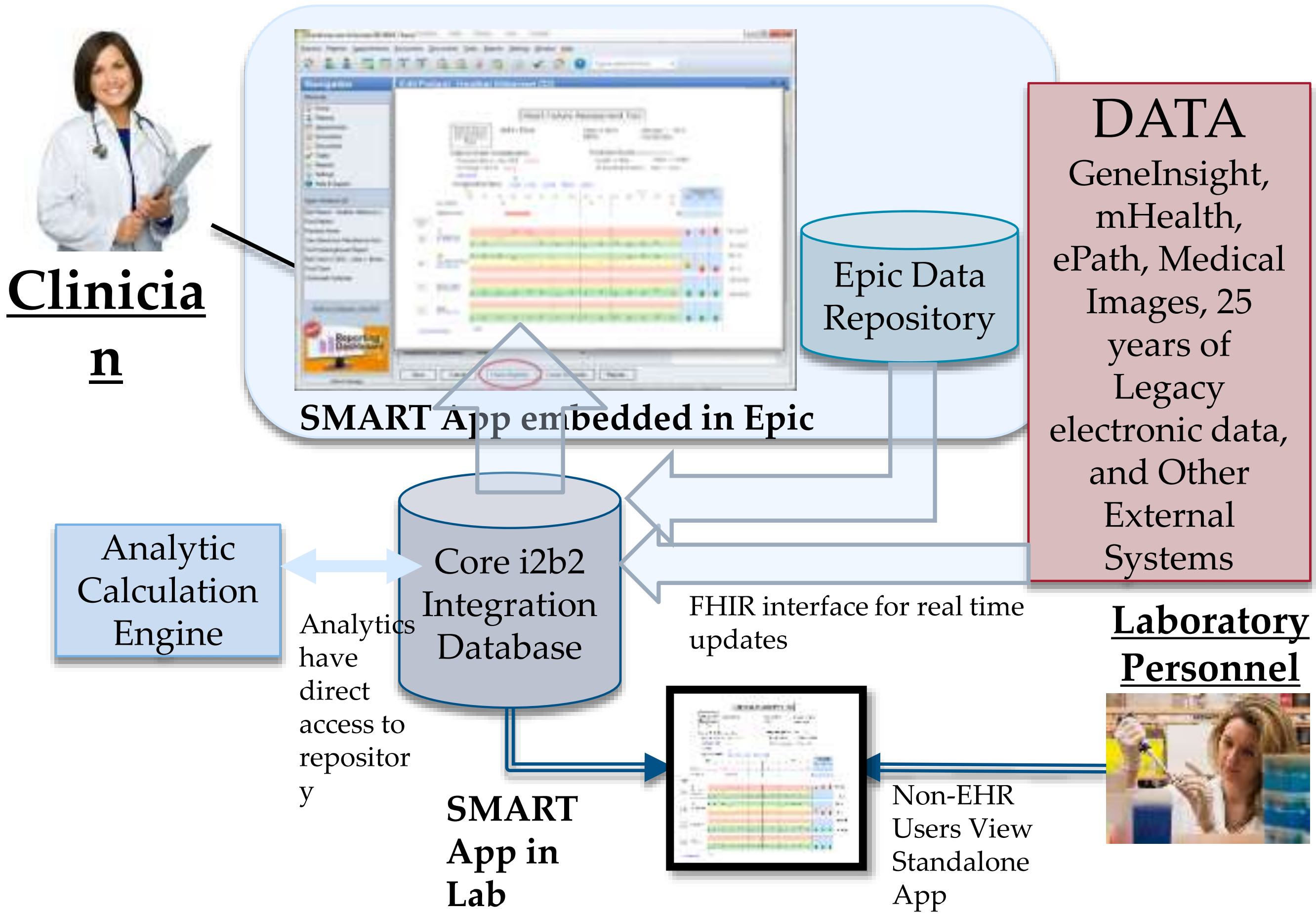
Linking to EMR with SMART "Apps"

Published 2011

Enabling Innovation to reach into EMR

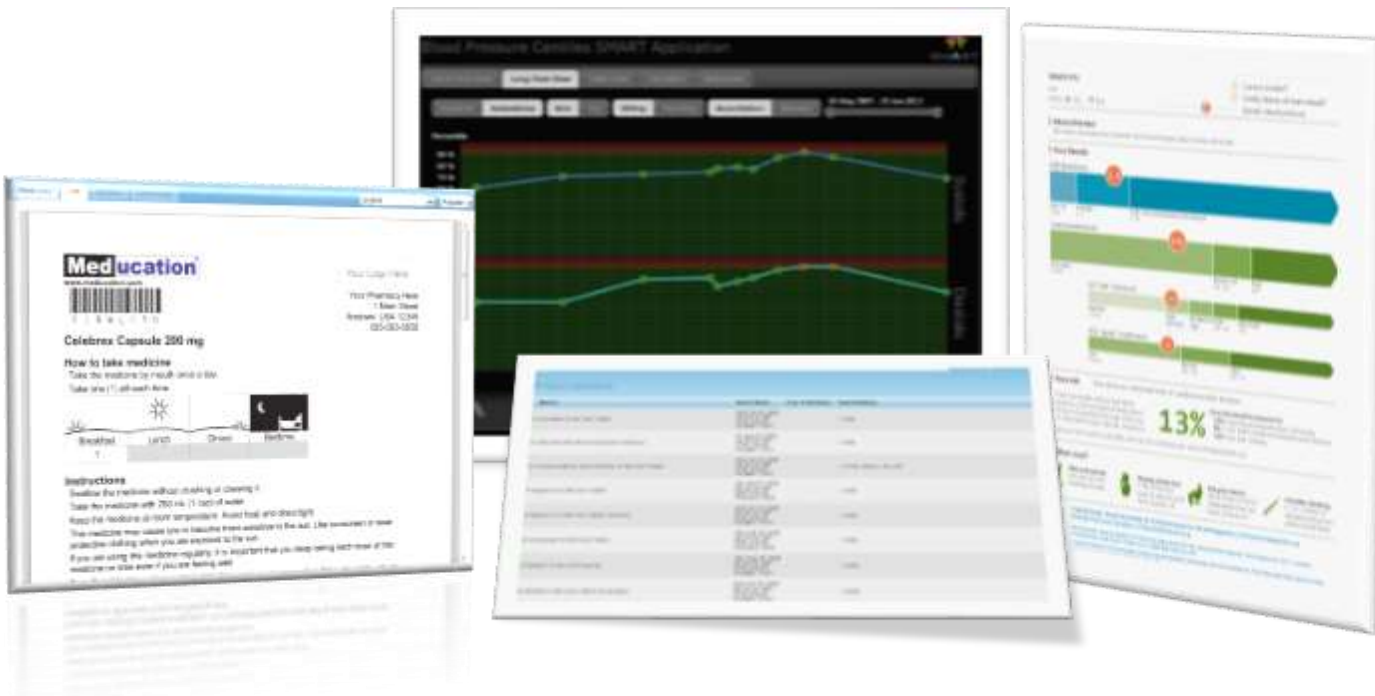


Bringing Big Data into Clinical Care with Open App Development



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

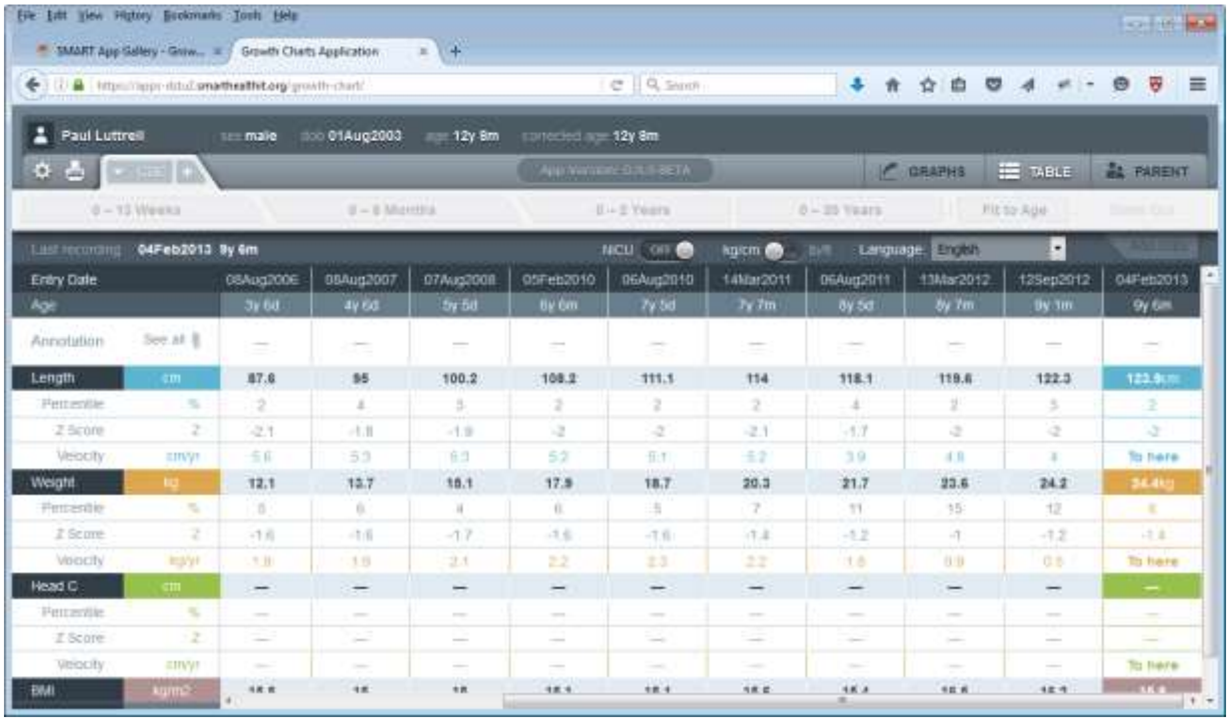
- Looking at similar patients can help predict:
  - Future outcomes and responses to therapy
  - Course of disease
  - Penetrance of genetic variants
  - Likelihood that a diagnostic pathway might be fruitful

- 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

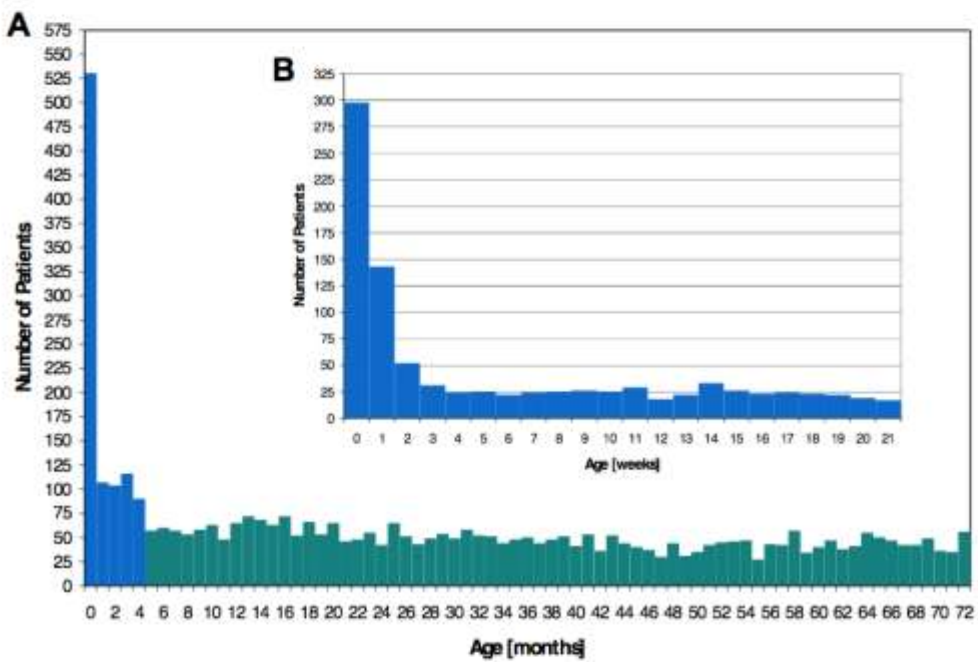
Growth Charts reinvented as a SMART app – is this child similar to other children?

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


Growth Charts Application
Paul Luttrell
male
01Aug2003
12y 8m
GRAPHS
TABLE
PARENT
04Feb2013 9y 6m
NICU
kg/cm
Language English
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 8y 7m 9y 1m 9y 6m
Annotation
Length cm 87.6 95 100.2 108.2 111.1 114 118.1 119.6 122.3
Percentile
Z Score
Velocity cm/yr
Weight kg 12.1 13.7 16.1 17.9 18.7 20.3 21.7 23.6 24.2
Percentile
Z Score
Velocity kg/yr
Head C cm
Percentile
Z Score
Velocity cm/yr
BMI kg/m2

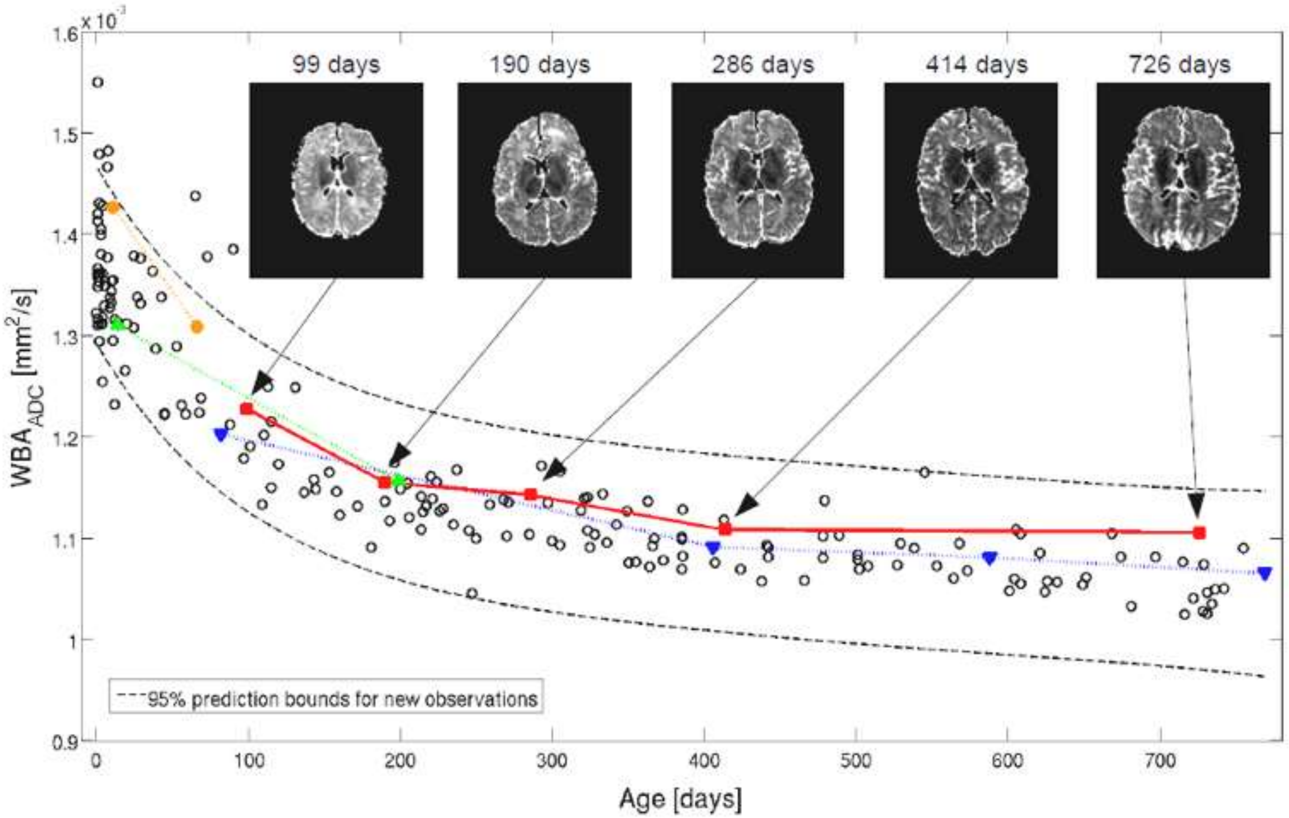

Growth Charts Application
Paul Luttrell
male
01Aug2003
12y 8m
12y 8m
GRAPHS
TABLE
PARENT
04Feb2013 9y 6m
NICU
Language
English
Height
Weight
Head C
BMI
123.9cm | 2%
24.4kg | 8%
N/A
15.8 | 37%
Mother
163 cm
Father
182 cm
Paul Luttrell
Underweight
Healthy
Overweight
Obese
24.4kg

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



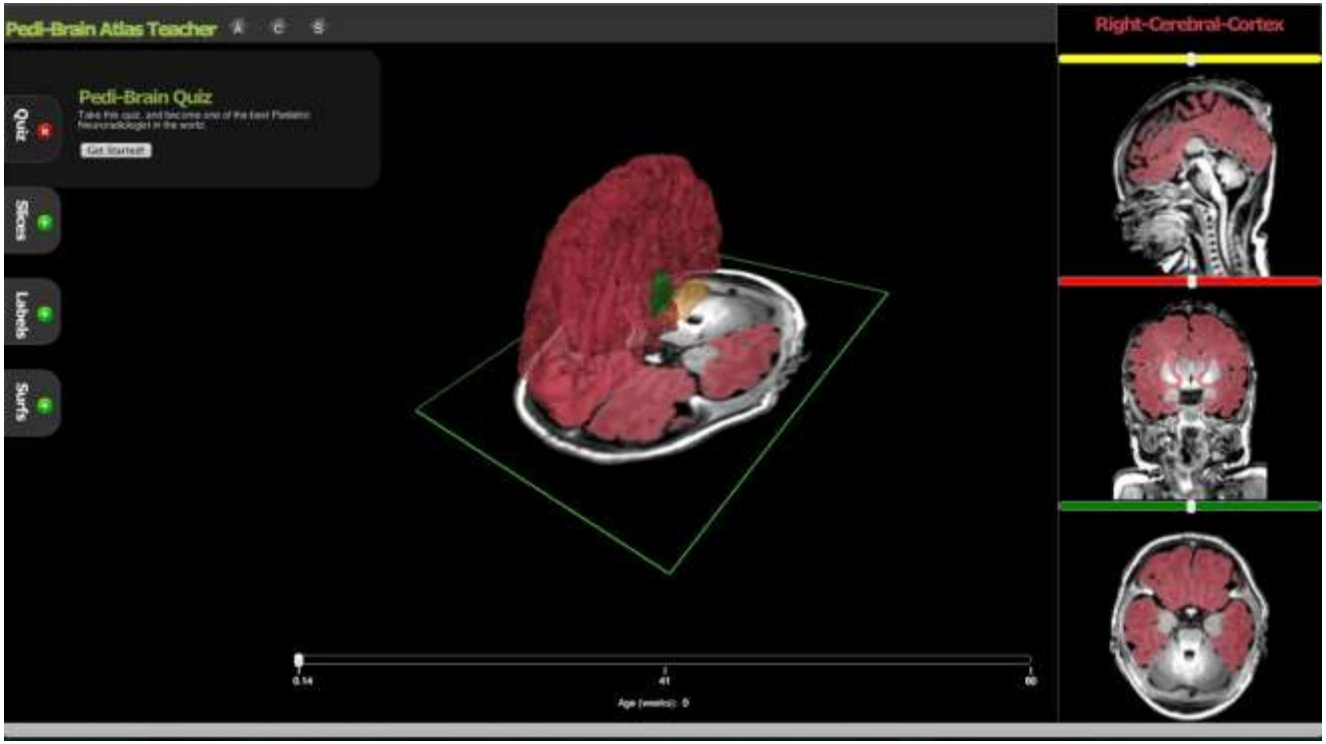
Number of patients who had a brain MRI scan at a particular age in months from 0 to 6 years (A) and in weeks from 0 to 4 months (B)

Determining a Normal Child's MRI



Normative Pediatric MRI

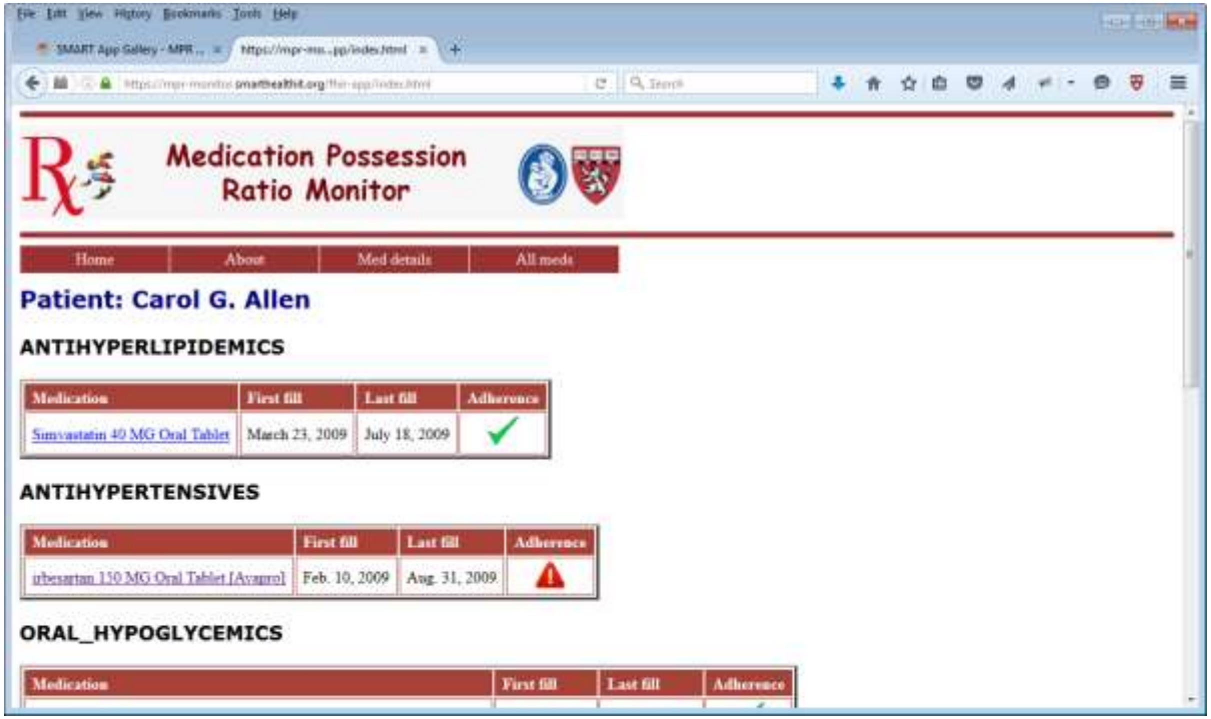
Generating quantitative atlases for regular intervals in pediatric development to be used for clinical brain MRI analysis

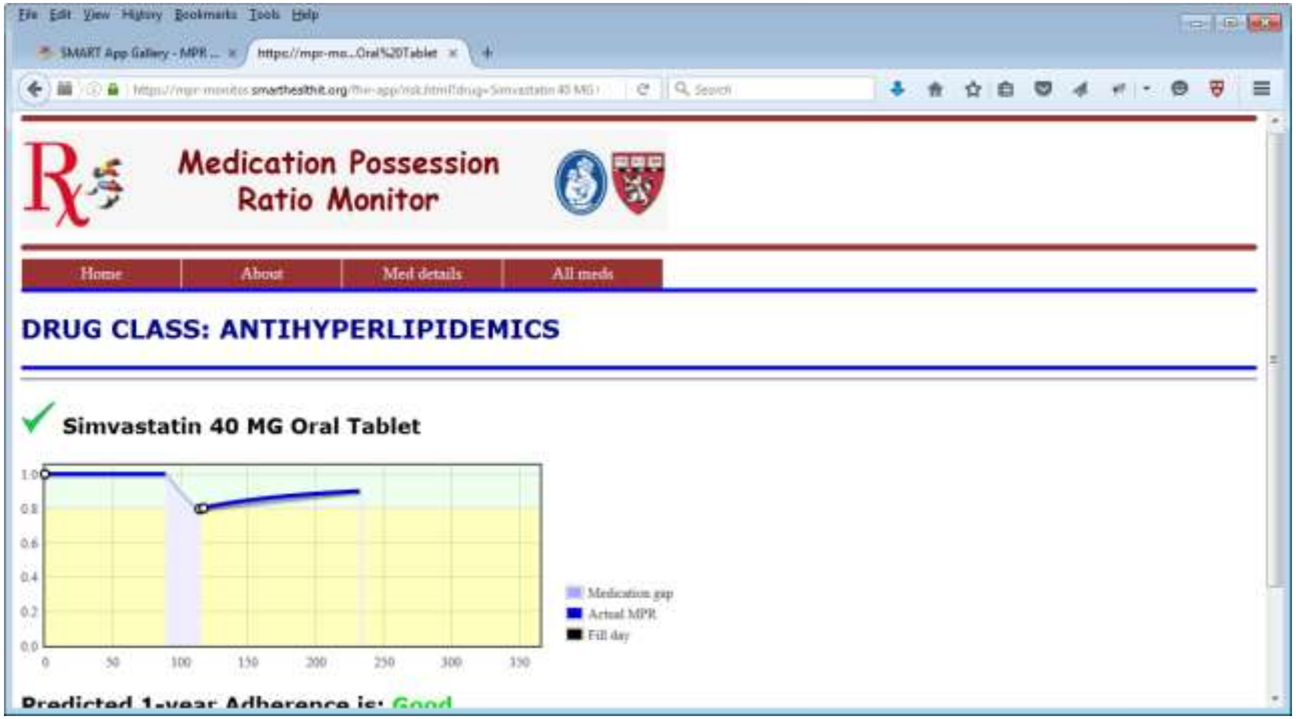


PARTNERS HEALTHCARE

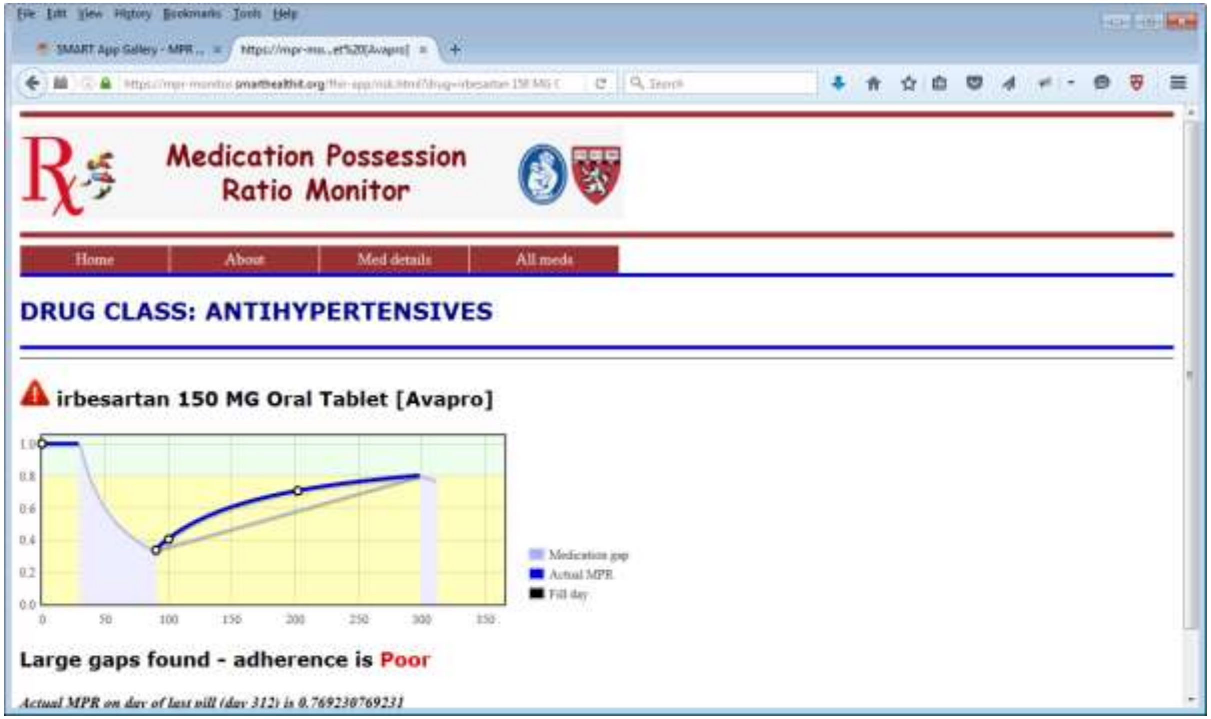
Comparing Medication Use Patterns to Determine Non-Compliance

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





File Edit View History Bookmarks Tools Help
SMART App Gallery - MPR ...
https://mpr-mo...Oral%20Tablet
Search
Medication Possession Ratio Monitor
Home
About
Med details
All meds
DRUG CLASS: ANTIHYPERLIPIDEMICS
Simvastatin 40 MG Oral Tablet
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
Medication gap
Actual MPR
Fill day
Predicted 1-year Adherence is: Good


File Edit View History Bookmarks Tools Help
SMART App Gallery - MPR ...
Medication Possession
Ratio Monitor
Home
About
Med details
All meds
DRUG CLASS: ANTIHYPERTENSIVES
irbesartan 150 MG Oral Tablet [Avapro]
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
Medication gap
Actual MPR
Fill day
Large gaps found - adherence is Poor
Actual MPR on day of last pill (day 312) is 0.769230769231

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 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://smarthealtit.org)