High-throughput Population Phenotyping

Victor Castro & Vivian Gainer

Research Information Systems and Computing (RISC)

Partners Healthcare

What is a High-Quality Computed Phenotype?

Phenotype = The set of features that characterize a specific patient population.

A **Computed Phenotype** is one that uses data from EHRs, both structured and narrative, to come up with and calculate the set of features that define a condition.

Why are these phenotypes important?

Codes alone are not sufficient to tell us whether someone has a condition.

Computable phenotypes were developed to use secondary data to determine, with statistical significance, whether or not someone or a population has a given condition.

History

Types of computable phenotypes:

- Rules-based (eMERGE)
- Machine learning

We have worked on creating machine learning algorithms since the start of i2b2, beginning with our Driving Biology Projects (DBP). Our methods and tools have evolved and continue to evolve to streamline the process and create a phenotyping workflow that researchers can understand and use.

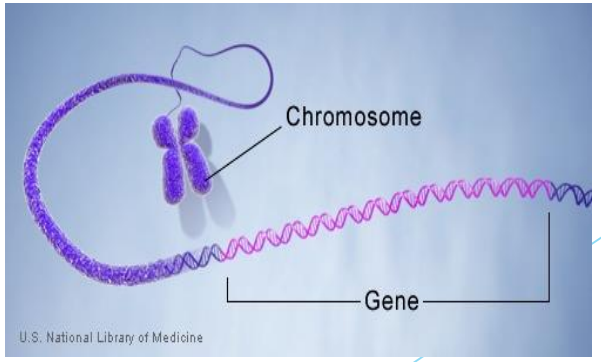
Partners Healthcare Biobank

- Repository of consented patient samples linked to the electronic medical record and supplemented with health information/family history from surveys.

- To date, **40,000+** patients have consented to participate and **30,000+** have provided samples. The target is 75,000 consented patients by 2018. 1,000+ patients are consented every month.

- Genomic data on **~10,000** patients is available, for free, to Partners investigators. Genomic data for another 15,000 patients (**total of 25,000**) will be released over the next 12-24 months.

- Supports $82M+ in grants across Partners institutions



Arthritis Care Res (Hoboken). 2010 Aug;62(8):1120-7. doi: 10.1002/acr.20184.

Electronic medical records for discovery research in rheumatoid arthritis.

Liao KP[1], Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM.

Am J Psychiatry. 2015 Apr;172(4):363-72. doi: 10.1176/appi.ajp.2014.14030423. Epub 2014 Dec 12.

Validation of electronic health record phenotyping of bipolar disorder cases and controls.

Castro VM[1], Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, Cai T, Hoffnagle AG, Dai Y, Block S, Weill SR, Nadal-Vicens M, Pollastri AR, Rosenquist JN, Goryachev S, Ongur D, Sklar P, Perlis RH, Smoller

BMJ. 2013 Jan 29;346:f288. doi: 10.1136/bmj.f288.

QT interval and antidepressant use: a cross sectional study of electronic health records.

Castro VM[1], Clements CC, Murphy SN, Gainer VS, Fava M, Weilburg JB, Erb JL, Churchill SE, Kohane IS, Iosifescu DV, Smoller JW, Perlis RH.

PLoS One. 2015 Aug 24;10(8):e0136651. doi: 10.1371/journal.pone.0136651. eCollection 2015.

Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts.

Liao KP[1], Ananthakrishnan AN[2], Kumar V[3], Xia Z[4], Cagan A[5], Gainer VS[5], Goryachev S[5], Chen P[6], Savova GK[7], Agniel D[8], Churchill S[9], Lee J[10], Murphy SN[11], Plenge RM[12], Szolovits P[13], Kohane I[7], Shaw SY[3], Karlson EW[1], Cai T[6].

Mol Psychiatry. 2014 Aug 26. doi: 10.1038/mp.2014.90. [Epub ahead of print]

Prenatal antidepressant exposure is associated with risk for attention-deficit hyperactivity disorder but not autism spectrum disorder in a large health system.

Clements CC[1], Castro VM[2], Blumenthal SR[1], Rosenfield HR[1], Murphy SN[3], Fava M[4], Erb JL[5], Churchill SE[6], Kaimal AJ[7], Doyle AE[1], Robinson EB[8], Smoller JW[9], Kohane IS[10], Perlis RH[1].

Inflamm Bowel Dis. 2013 Jun;19(7):1411-20. doi: 10.1097/MIB.0b013e31828133fd.

Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach.

Ananthakrishnan AN[1], Cai T, Savova G, Cheng SC, Chen P, Perez RG, Gainer VS, Murphy SN, Szolovits P, Xia Z, Shaw S, Churchill S, Karlson EW, Kohane I, Plenge RM, Liao KP.

Phenotyping the Biobank Population

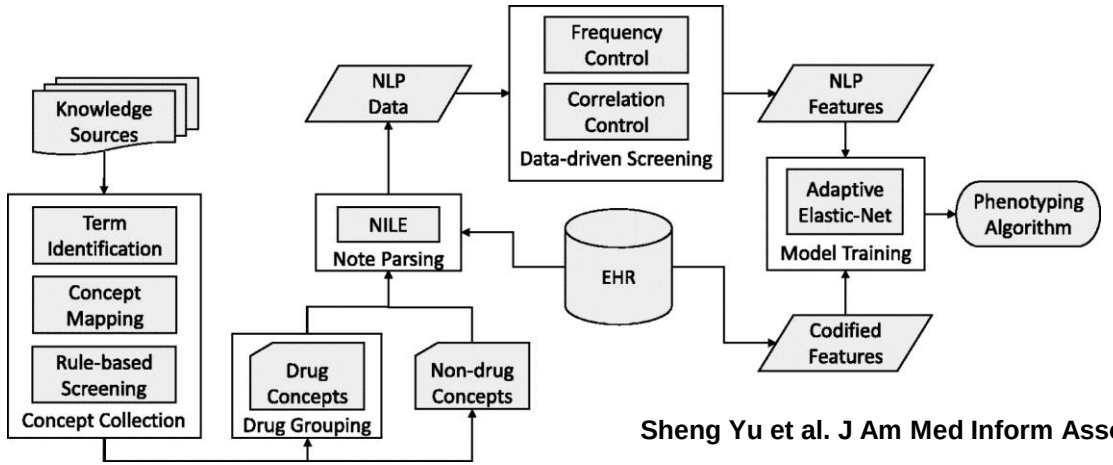
- **Goals**
  - Develop high-specificity algorithms for selected disease populations
    - Use case: genotype-phenotype association studies
  - Data-driven identification of relevant disease features
  - Algorithms should classify the entire population both Disease+ and Disease-
  - Algorithms will be computed on regular basis to include newly-consented individuals and new data from the EHR
  - Available to investigators inside the Biobank Portal i2b2 web client
    - Investigators can choose different PPVs depending on their algorithm

Phenotype Prevalence

Phenotype	Estimated Prevalence*	PPV of ≥1 ICD9/ICD10 Code
Asthma (AST)	12.0%	0.61
Bipolar Disorder (BD)	1.3%	0.39
Breast Cancer (BRCA)	4.0%	0.66
Chronic Obstructive Pulmonary Disease (COPD)	4.3%	0.33
Congestive Heart Failure (CHF)	4.4%	0.33
Coronary Artery Disease (CAD)	13.5%	0.43
Crohn's Disease (CD)	4.7%	0.57
Depression (DEPR)	16.0%	0.56
Epilepsy (EPIL)	3.9%	0.63
Gout (GOUT)	6.0%	0.84
Hypertension (HTN)	42.0%	0.77
Multiple Sclerosis (MS)	0.8%	0.52
Obesity (OBES)	48.9%	-
Rheumatoid Arthritis (RA)	3.8%	0.39
Schizophrenia (SCZ)	0.2%	0.16
Type-I Diabetes Mellitus (T1DM)	0.9%	0.16
Type-II Diabetes Mellitus (T2DM)	10.6%	-
Ulcerative Colitis (UC)	2.5%	0.48

* Prevalence is estimated based on clinician chart review of a random sample of Biobank participants.

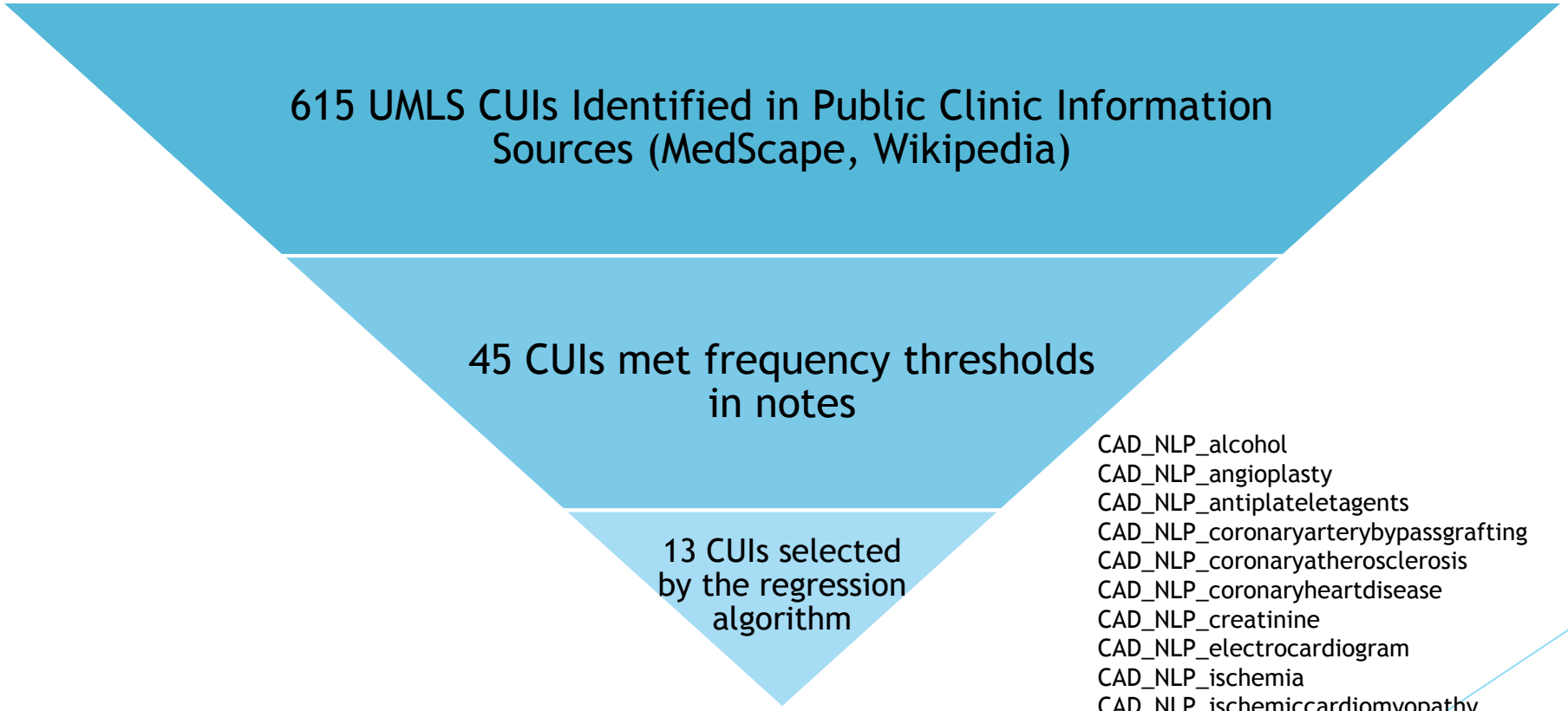
High-throughput Phenotype Training



Sheng Yu et al. J Am Med Inform Assoc 2015;22:993-1000

- Automated feature extraction
  - NLP terms identified from public knowledge sources (Medscape, Wikipedia) and mapped to UMLS CUIs
  - Terms are screened based on frequency and correlation in the data
  - Coded terms (COD) from the EHR also identified (i.e. ICD-10, CPT-4, RXNORM)

Example: Feature Selection in Coronary Artery Disease (CAD)



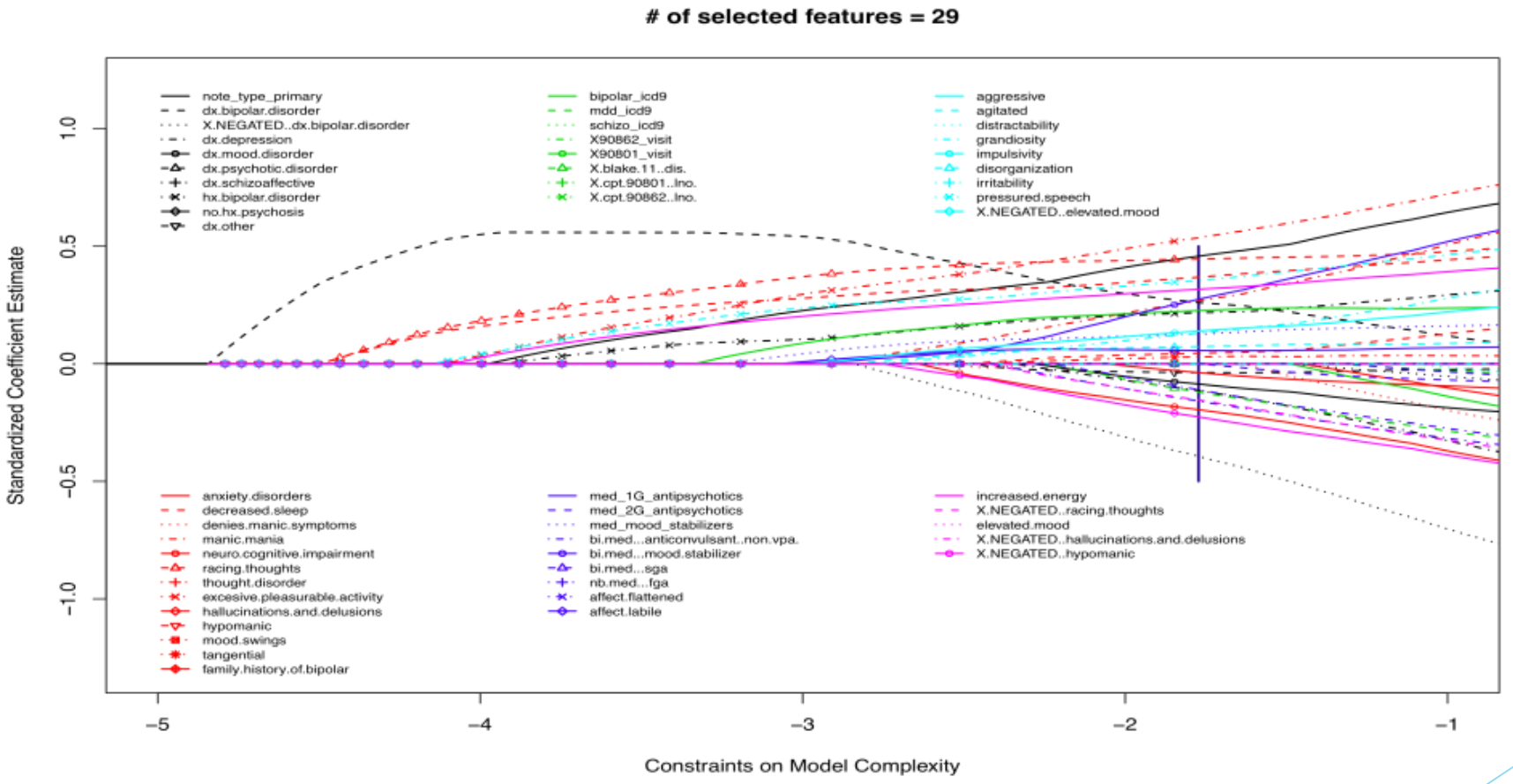
CAD_NLP_alcohol
CAD_NLP_angioplasty
CAD_NLP_antiplateletagents
CAD_NLP_coronaryarterybypassgrafting
CAD_NLP_coronaryatherosclerosis
CAD_NLP_coronaryheartdisease
CAD_NLP_creatinine
CAD_NLP_electrocardiogram
CAD_NLP_ischemia
CAD_NLP_ischemiccardiomyopathy
CAD_NLP_myocardialinfarction
CAD_NLP_nitroglycerin
CAD_NLP_plateletaggregationinhibitors

High-throughput Phenotype Training

- Minimize the chart-review bottleneck
  - Chart reviews conducted in the i2b2 workbench timeline view
  - Generate established chart review criteria
  - Concurrent chart reviews using prevalence-based sampling

Patient Number	Asthma	Breast cancer	Chronic airway obstruction (COPD)	Depression/MDD	Epilepsy	Hypertension	Ischemic stroke	Obesity	Schizophrenia	Type 1 diabetes
06150	N	Y	N	N	N	Y	N	Y	N	N
06249	N	N	N	N	N	Y	N	N	N	N
06391	N	N	N	N	N	Y	N	U	N	N
06395	N	N	N	Y	N	P	N	P	N	N
06500	Y	N	N	Y	N	Y	N	Y	N	N
06551	Y	N	P	N	N	Y	N	Y	N	N
06574	N	N	P	N	Y	Y	N	N	N	N
06580	N	N	N	N	Y	N	Y	N	N	N
06692	N	P	N	P	N	Y	N	Y	N	N
06769	N	N	N	N	N	Y	N	Y	N	N
06807	N	N	Y	N	N	Y	N	Y	N	N
06955	N	N	N	P	Y	Y	N	N	N	N
07018	N	N	P	N	N	Y	N	Y	N	N
07210	N	N	N	N	N	Y	Y	Y	N	N
07226	N	N	N	N	N	Y	N	Y	N	N
07471	N	N	N	N	N	Y	N	U	N	N

LASSO Regression



High-throughput Phenotype Training

- Standardize approach to training the phenotype models
  - Features are mapped and grouped in i2b2 schema and are defined based on C_FULLNAME
  - Standardized naming convention for NLP and Coded (COD) features
  - Simple and interpretable machine language techniques for feature shrinking and building a model.

Feature_ID	Beta (weight)	Feature Description
(Intercept)	0.548	Model Intercept (beta 0)
Epilepsy_COD_DX_Epilepsy	2.414	Count of coded diagnosis of epilepsy
Epilepsy_COD_MED_lamotrigine	0.129	Count of prescriptions for lamotrigine
Epilepsy_COD_MED_phenytoin	0.124	Count of prescriptions for phenytoin
Epilepsy_COD_PRC_HeadCT	-0.339	Count of Head CT scan procedures
patient_dxenct	-1.022	Total number of visits with a coded diagnosis

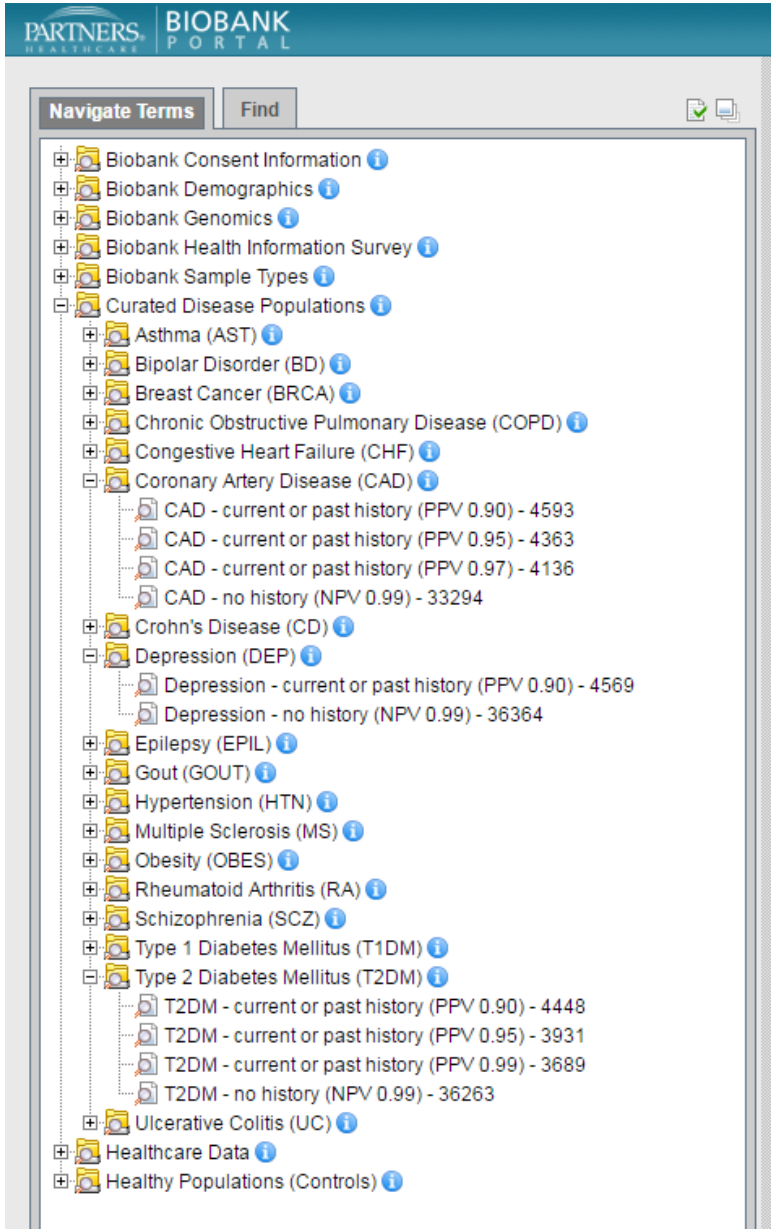
Epilepsy Algorithm Final Feature Betas

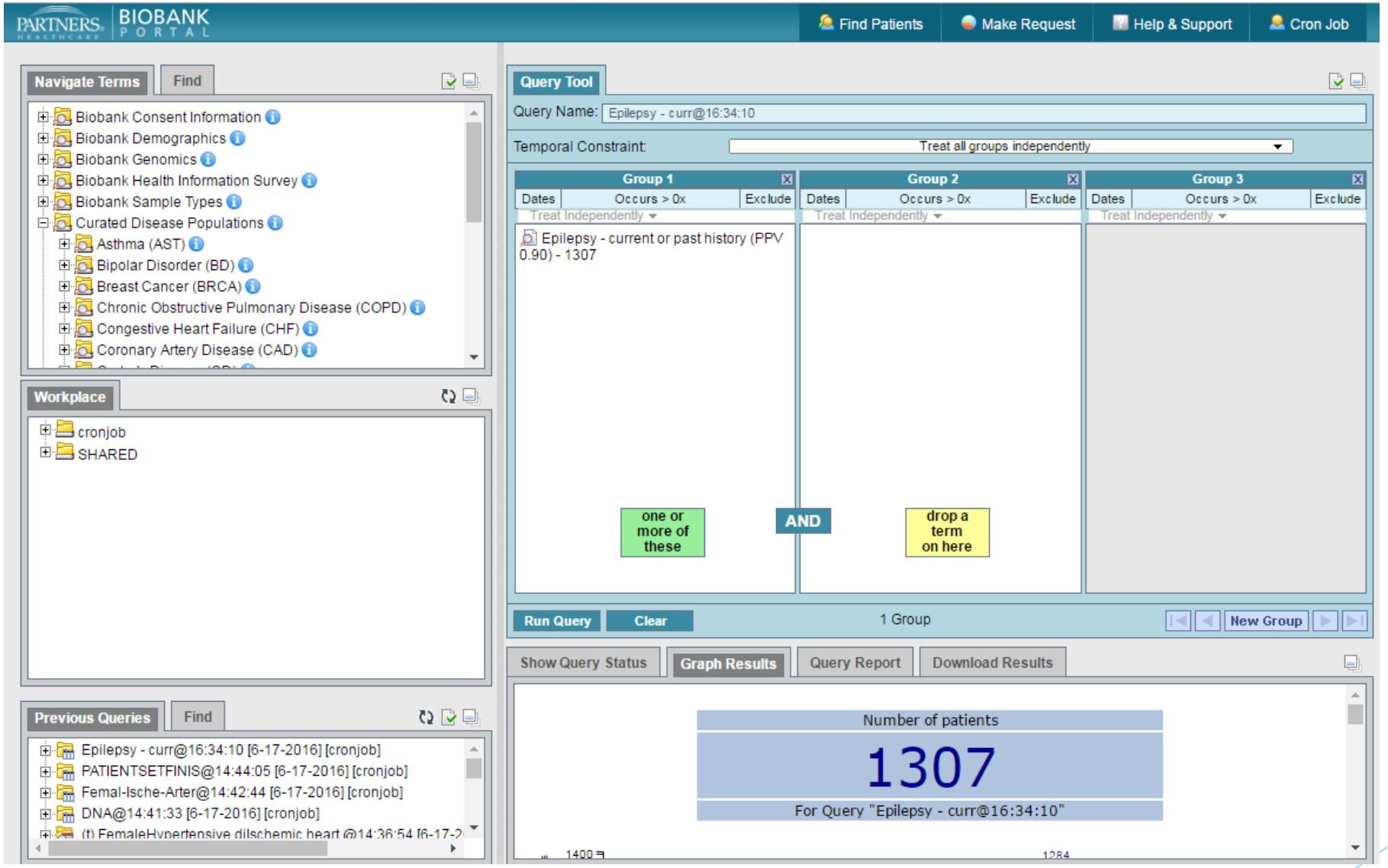
Algorithm Training Results

Phenotype	COD+NLP Algorithms		COD Algorithms	
	AUC	Sensitivity @PPV~0.9	AUC	Sensitivity @PPV~0.9
Asthma (AST)	0.889	0.70	0.893	0.76
Bipolar Disorder (BD)	0.920	0.28	0.822	0.23
Breast Cancer (BRCA)	0.982	0.97	0.951	0.94
Chronic Obstructive Pulmonary Disease (COPD)	0.851	0.43	0.768	0.23
Congestive Heart Failure (CHF)	0.921	0.53	0.897	0.42
Coronary Artery Disease (CAD)	0.989	0.97	0.953	0.82
Crohn's Disease (CD)	0.971	0.96	0.973	0.94
Depression (DEPR)	0.935	0.87	0.908	0.80
Epilepsy (EPIL)	0.951	0.91	0.957	0.93
Gout (GOUT)	0.848	0.95	0.870	0.93
Hypertension (HTN)	0.946	0.98	0.912	0.95
Multiple Sclerosis (MS)	0.947	0.81	0.925	0.79
Obesity (OBES)	0.954	0.85	0.948	0.87
Rheumatoid Arthritis (RA)	0.948	0.76	0.928	0.69
Schizophrenia (SCZ)	0.980	0.83	0.921	0.29
Type-I Diabetes Mellitus (T1DM)	0.990	0.84	0.972	0.78
Type-II Diabetes Mellitus (T2DM)	0.977	0.88	0.952	0.77
Ulcerative Colitis (UC)	0.962	0.87	0.967	0.88

Comparing Coded-only (COD) vs COD+NLP Algorithms

Phenotype	Sensitivity Difference COD vs COD+NLP	COD+NLP N @40k patients	COD N @40k patients	Weeks to Catch Up
Asthma (AST)	0.09	3,346	3,653	-11
Bipolar Disorder (BD)	-0.18	146	120	29
Breast Cancer (BRCA)	-0.04	1,555	1,498	5
Chronic Obstructive Pulmonary Disease (COPD)	-0.47	746	394	119
Congestive Heart Failure (CHF)	-0.20	933	744	34
Coronary Artery Disease (CAD)	-0.16	5,238	4,423	25
Crohn's Disease (CD)	-0.02	1,805	1,762	3
Depression (DEPR)	-0.08	5,587	5,126	12
Epilepsy (EPIL)	0.03	1,412	1,454	-4
Gout (GOUT)	-0.02	2,273	2,234	2
Hypertension (HTN)	-0.03	16,414	15,994	4
Multiple Sclerosis (MS)	-0.02	259	253	3
Obesity (OBES)	0.03	16,606	17,037	-3
Rheumatoid Arthritis (RA)	-0.09	1,155	1,052	13
Schizophrenia (SCZ)	-0.65	67	23	249
Type-I Diabetes Mellitus (T1DM)	-0.07	304	282	10
Type-II Diabetes Mellitus (T2DM)	-0.13	3,731	3,248	20
Ulcerative Colitis (UC)	0.02	870	884	-2


PARTNERS HEALTHCARE
BIOBANK PORTAL
Navigate Terms
Find
Biobank Consent Information
Biobank Demographics
Biobank Genomics
Biobank Health Information Survey
Biobank Sample Types
Curated Disease Populations
Asthma (AST)
Bipolar Disorder (BD)
Breast Cancer (BRCA)
Chronic Obstructive Pulmonary Disease (COPD)
Congestive Heart Failure (CHF)
Coronary Artery Disease (CAD)
CAD - current or past history (PPV 0.90) - 4593
CAD - current or past history (PPV 0.95) - 4363
CAD - current or past history (PPV 0.97) - 4136
CAD - no history (NPV 0.99) - 33294
Crohn's Disease (CD)
Depression (DEP)
Depression - current or past history (PPV 0.90) - 4569
Depression - no history (NPV 0.99) - 36364
Epilepsy (EPIL)
Gout (GOUT)
Hypertension (HTN)
Multiple Sclerosis (MS)
Obesity (OBES)
Rheumatoid Arthritis (RA)
Schizophrenia (SCZ)
Type 1 Diabetes Mellitus (T1DM)
Type 2 Diabetes Mellitus (T2DM)
T2DM - current or past history (PPV 0.90) - 4448
T2DM - current or past history (PPV 0.95) - 3931
T2DM - current or past history (PPV 0.99) - 3689
T2DM - no history (NPV 0.99) - 36263
Ulcerative Colitis (UC)
Healthcare Data
Healthy Populations (Controls)


PARTNERS HEALTHCARE
BIOBANK PORTAL
Find Patients
Make Request
Help & Support
Cron Job
Navigate Terms
Find
Biobank Consent Information
Biobank Demographics
Biobank Genomics
Biobank Health Information Survey
Biobank Sample Types
Curated Disease Populations
Asthma (AST)
Bipolar Disorder (BD)
Breast Cancer (BRCA)
Chronic Obstructive Pulmonary Disease (COPD)
Congestive Heart Failure (CHF)
Coronary Artery Disease (CAD)
Workplace
cronjob
SHARED
Previous Queries
Find
Epilepsy - curr@16:34:10 [6-17-2016] [cronjob]
PATIENTSETFINIS@14:44:05 [6-17-2016] [cronjob]
Femal-Ische-Arter@14:42:44 [6-17-2016] [cronjob]
DNA@14:41:33 [6-17-2016] [cronjob]
Query Tool
Query Name: Epilepsy - curr@16:34:10
Temporal Constraint:
Treat all groups independently
Group 1
Group 2
Group 3
Dates
Occurs > 0x
Exclude
Treat Independently
Epilepsy - current or past history (PPV 0.90) - 1307
one or more of these
AND
drop a term on here
Run Query
Clear
1 Group
New Group
Show Query Status
Graph Results
Query Report
Download Results
Number of patients
1307
For Query "Epilepsy - curr@16:34:10"
1400
1284

PARTNERS HEALTHCARE | BIOBANK PORTAL

Find Patients | Make Request | Help & Support | Cron Job

Navigate Terms | Find

- Biobank Consent Information
- Biobank Demographics
- Biobank Genomics
- Biobank Health Information Survey
- Biobank Sample Types
- Curated Disease Populations
  - Asthma (AST)
  - Bipolar Disorder (BD)
  - Breast Cancer (BRCA)
  - Chronic Obstructive Pulmonary Disease (COPD)
  - Congestive Heart Failure (CHF)
  - Coronary Artery Disease (CAD)

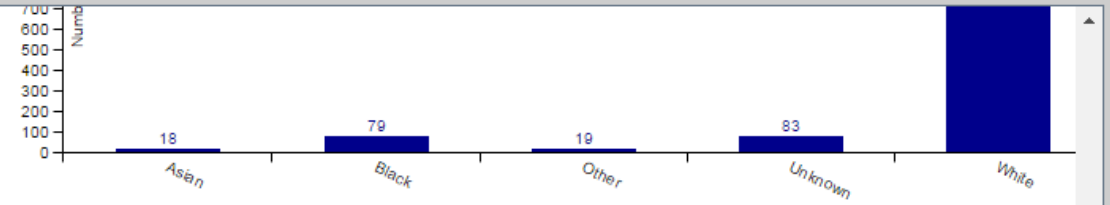
Workplace

- cronjob
- SHARED

Previous Queries | Find

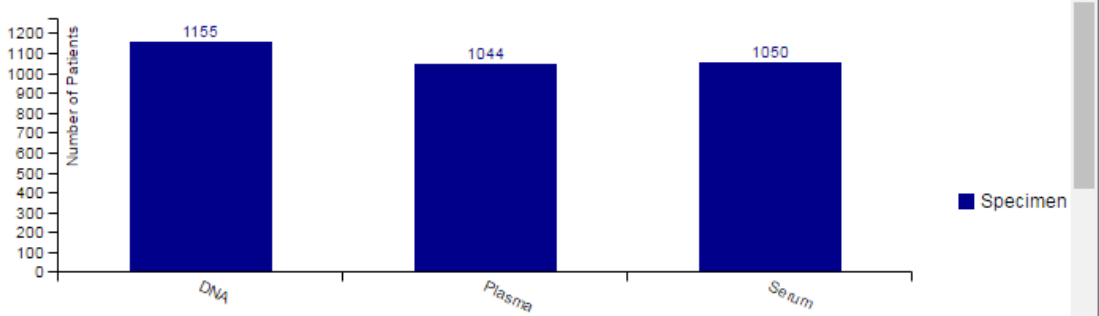
- Epilepsy - curr@16:34:10 [6-17-2016] [cronjob]
- PATIENTSETFINIS@14:44:05 [6-17-2016] [cronjob]
- Femal-Ische-Arter@14:42:44 [6-17-2016] [cronjob]
- DNA@14:41:33 [6-17-2016] [cronjob]
- (t) FemaleHypertensive dilschemic heart @14:36:54 [6-17-2

Show Query Status | Graph Results | Query Report | Download Results



Total Unique Patients by Specimen type

Specimen type	Counts
DNA	1155
Plasma	1044
Serum	1050



Total Unique Patients by Ethnicity

Ethnicity	Counts

Custom Specificity / Sensitivity

- Investigators can choose different algorithm predicted probability cutoffs of the phenotype corresponding to different levels of PPV and sensitivity
- For example, a study seeking to recruit patients for a study might choose a lower cut-off since they will be screening the patients.
- Predicted probability can also be used as a continuous measure in genotype-phenotype study to adjust for phenotype uncertainty.

	encounter_num	patient_num	CONCEPT_CD	provider_id	INSTANCE_NUM	VALTYPE_CD	NVAL_NUM
1	32930974	41144	Epilepsy_filterpositive_CODonly_31Mar16_yes_ppv090	@	1	N	0.92323
2	33763593	41409	Epilepsy_filterpositive_CODonly_31Mar16_yes_ppv090	@	1	N	0.94719
3	33169757	41453	Epilepsy_filterpositive_CODonly_31Mar16_yes_ppv090	@	1	N	0.81372
4	33741751	41840	Epilepsy_filterpositive_CODonly_31Mar16_yes_ppv090	@	1	N	0.97063
5	33312038	40970	Epilepsy_filterpositive_CODonly_31Mar16_yes_ppv090	@	1	N	0.95071

Summary

- Machine learning algorithms can be effectively and efficiently applied to a large population to accurately phenotype patients
- Algorithms provide flexibility to adjust sensitivity and specificity to varied use cases compared to pre-defined rules-based algorithms
- Methods and tools to optimize the building of gold-standard training sets can generate significant time-savings
- Future work:
  - Develop additional algorithms
  - Examine portability of algorithms in larger population (i.e. all patients in EHR)
  - Enhance i2b2 UI to allow users to “customize” their algorithm PPV / Sensitivity
  - Work towards a “Phenotyping Workbench” to optimize algorithm building process within the i2b2 framework.

Thanks!

Biobank Portal Team
Bhaswati Ghosh
Barbara Benoit
Andy Cagan
Tianxi Cai
Victor Castro
Stacey Duey
Alyssa Goodson
Sergey Goryachev
Reeta Metta
Pourab Roy
Nich Wattanasin
David Wang
Sheng Yu

Biobank Team
Jackie Aldama
Nicole Allen
Sami Amr
Ashley Blau
Natalie Boutin
Xander Cerretani
Kim Durniak
Kevin Embree
Ana Holzbach
Irene Leon
Lisa Mahanta
Neeta Rathi
Matilde Vickers
Ellen Tsai
Matt Lebo

Biobank Principal Investigators
Scott Weiss, Principal Investigator
Lynn Bry, MD, PhD (BWH)
Elizabeth Karlson, MD (BWH)
Sue Slaugenhaupt, PhD (MGH)
Jordan Smoller, MD, ScD (MGH)

Biobank Senior Leadership
Scott Weiss, M.D, Chief, Partners Personalized Medicine
Anne Klibanski, MD, PHS CAO
Paul Anderson, MD, PhD, BWH VP
Jeff Golden, MD, BWH Pathology
David Louis, MD, MGH Pathology
Harry Orf, PhD, MGH VP
Pearl O'Rourke, MD, PHS IRB
Shawn Murphy, MD, PhD