Health Informatics

Lecture 9

Samantha Kleinberg
samantha.kleinberg@stevens.edu

Next week: journal club

- Goal: evaluate and discuss work other than your own that may be outside your field, learn about new work
- For all papers: read, and prepare to comment on
- For your paper:
  - Read the article (+ other references if needed for context)
  - Prepare a brief summary
  - Prepare questions to lead discussion

Format

- Brief overview of topic
- Summary of paper
  - What's the main argument/hypothesis?
  - How is this supported?
    - Overview experiments, figures
- Discussion
  - Prepare questions

Note: All group members must participate in presentation and Q&A

Total of 25 min/paper

Project notes

- Papers due 5/2. Presentations same day
- Presentation: ~15-20 minute talk
  - Should cover introductory material/motivation (why did you do this?), methodology (what did you do?), results (+ limitations), conclusions + future directions (what would you do if you had more time?)

Phenotyping

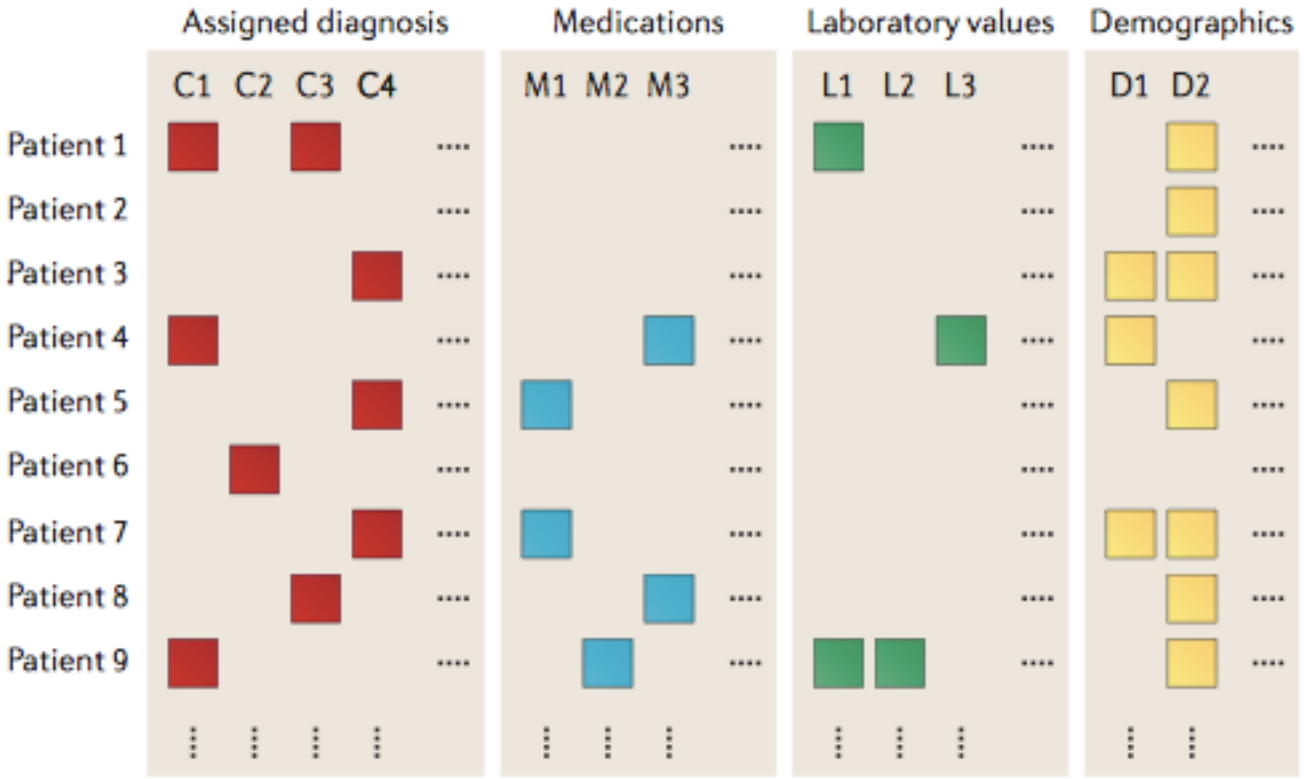
Data-> cohort with particular traits

Uses

- Retrospective research
- Clinical trials
- Epidemiology/population health
- Discovering subgroups

Using EHRs for research

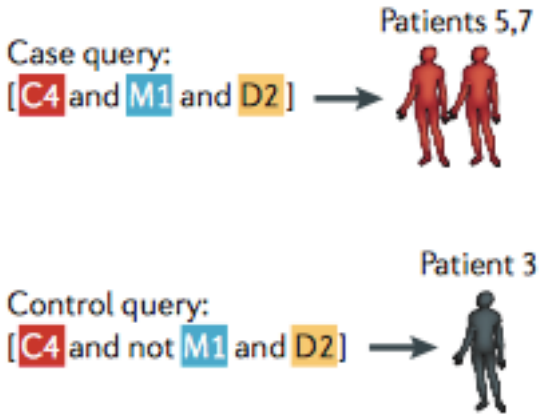
- First challenge is identifying cohort
  - Who has disease? Who doesn't?
  - Accuracy affects everything that follows
- Usually based on manually defined rules
  - Iterative process
  - Extremely time consuming
- Clinical trials
  - Usually manual identification of eligibility



TRANSLATIONAL GENETICS

Mining electronic health records: towards better research applications and clinical care

Peter B. Jensen[1], Lars J. Jensen[1] and Søren Brunak[1,2]



Usual process:

-Define criteria

-Evaluate results

-Iterate

Need some labeled data or ability to determine whether results are correct

on: 2008-04-21 meds for CHF: 2		
from: 2008-10-17 thru: 2008-10-19 meds for CHF: 3 terms: pulmonary edema, jvd, chf, cardiomegaly,		
2008-10-17	**X-Ray of Chest, Portable**	**cardiomegaly**
2008-10-17	Miscellaneous Nursing Note.	
2008-10-17	12-Lead Electrocardiogram	
2008-10-17	Nursing Adult Admission History	
2008-10-18	**Admission Note**	**jvd** **chf** **cardiomegaly**
2008-10-18	Miscellaneous Nursing Note.	
2008-10-18	**X-Ray of Chest, 2 Views**	**pulmonary edema** **cardiomegaly**
2008-10-18	Signout	
2008-10-18	Signout	
2008-10-19	SW Discharge Note	
2008-10-19	SW Progress Note-Adult	
2008-10-19	**Discharge Summary Note**	**jvd** **cardiomegaly**
2008-10-19	Signout	
2008-10-19	Nursing Discharge Note	
on: 2008-10-27		
2008-10-27	X-Ray of Chest, 2 Views	
on: 2008-11-17 meds for CHF: 2		

S. Kleinberg and N. Elhadad. Lessons learned in replicating data-driven experiments in multiple medical systems and patient populations. In AMIA Annual Symposium, 2013

Challenges

- Representation
  - How to describe/define a phenotype
- Complexity
  - Scale of data, complexity of definition
- Data quality/completeness/bias
  - Are the criteria we need there? Can we trust the data?
- Narrative text
  - Extracting criteria
- Portability
  - Does phenotype in one hospital work at another?

EXAMPLE: CHF case definition

Geisinger:

- 2 medication orders for CHF or
- 2 outpatient visits with CHF diagnosis or
- 1 medication order and 1 outpatient CHF diagnosis or
- CHF on problem list

Columbia:

- 2 ICD-9 codes for CHF or
- 1 ICD-9 code for CHF and 1 mention of CHF in note on the same date or
- 1 ICD-9 code for CHF and one medication typically indicated for CHF

Operational criteria	Number of Framingham criteria							Total
	none	1 minor, 0 major	2+ minor, 0 major	0 minor, 1 major	1 minor, 1 major	1 major, 2+ minor	2+ major	
1 outpatient visit or 1 medication order	331	111	15	262	160	56	368	1303
2+ medication orders only	299	81	10	195	99	22	203	909
2+ outpatient visits only	171	60	15	134	109	47	316	852
1 outpatient visit and 1 medication order only	44	18	7	45	28	7	73	222
Problem list only	124	47	8	81	56	26	139	481
Meet 2 of 3 criteria	281	114	32	303	224	89	826	1869
Meet all 3 criteria	72	44	8	103	88	42	504	861
Total	1322	475	95	1123	764	289	2429	**6497**

Meets operational criteria and Framingham criteria, N=2294 (35%)

Meets operational criteria but not Framingham criteria, N=2900 (45%)

Meets Framingham criteria but not operational criteria, N=424 (7%)

Does not meet operational criteria or Framingham criteria, N=879 (14%)

Wu et al. (2010). Prediction Modeling Using EHR Data: Challenges, Strategies, and a Comparison of Machine Learning Approaches. Med Care 48(6):S106-13.

Phenotyping algorithms



A review of approaches to identifying patient

Table 2 Summary of number of studies using different data sources across the top 10 phenotypes

	Demographics	Medications	Lab reports	Vitals	Clinical	Diagnosis	Treatment	Notes	Genomic	Other
Cancer	12	0	6	1	9	10	4	9	4	5
Diabetes	11	10	14	4	7	14	7	8	3	7
Heart failure	1	1	2	0	0	2	0	1	0	0
Rheumatoid arthritis	1	3	2	0	0	5	0	4	0	0
Cataract	3	1	1	1	1	3	2	2	0	3
Drug side effect	1	0	1	0	0	1	0	1	0	3
Pneumonia	0	0	0	0	0	0	0	4	0	0
Asthma	2	1	1	0	1	3	0	2	0	0
Peripheral arterial disease	2	1	1	1	1	2	2	1	0	2
Hypertension	1	0	0	0	0	0	0	2	0	1

University, Columbus, Ohio, USA
[3]Department of Biomedical Informatics, Columbia University, New York, New York, USA
[4]Center for Healthcare Informatics and Policy, Weill Cornell Medical College, New York, New York, USA

Journal of Biomedical Informatics, (3) *Proceedings of the Annual American Medical Informatics Association Symposium*, and (4) *Proceedings of Clinical Research Informatics Conference* within the past 3 years was assessed for inclusion in the review. Only articles using automated techniques were included.

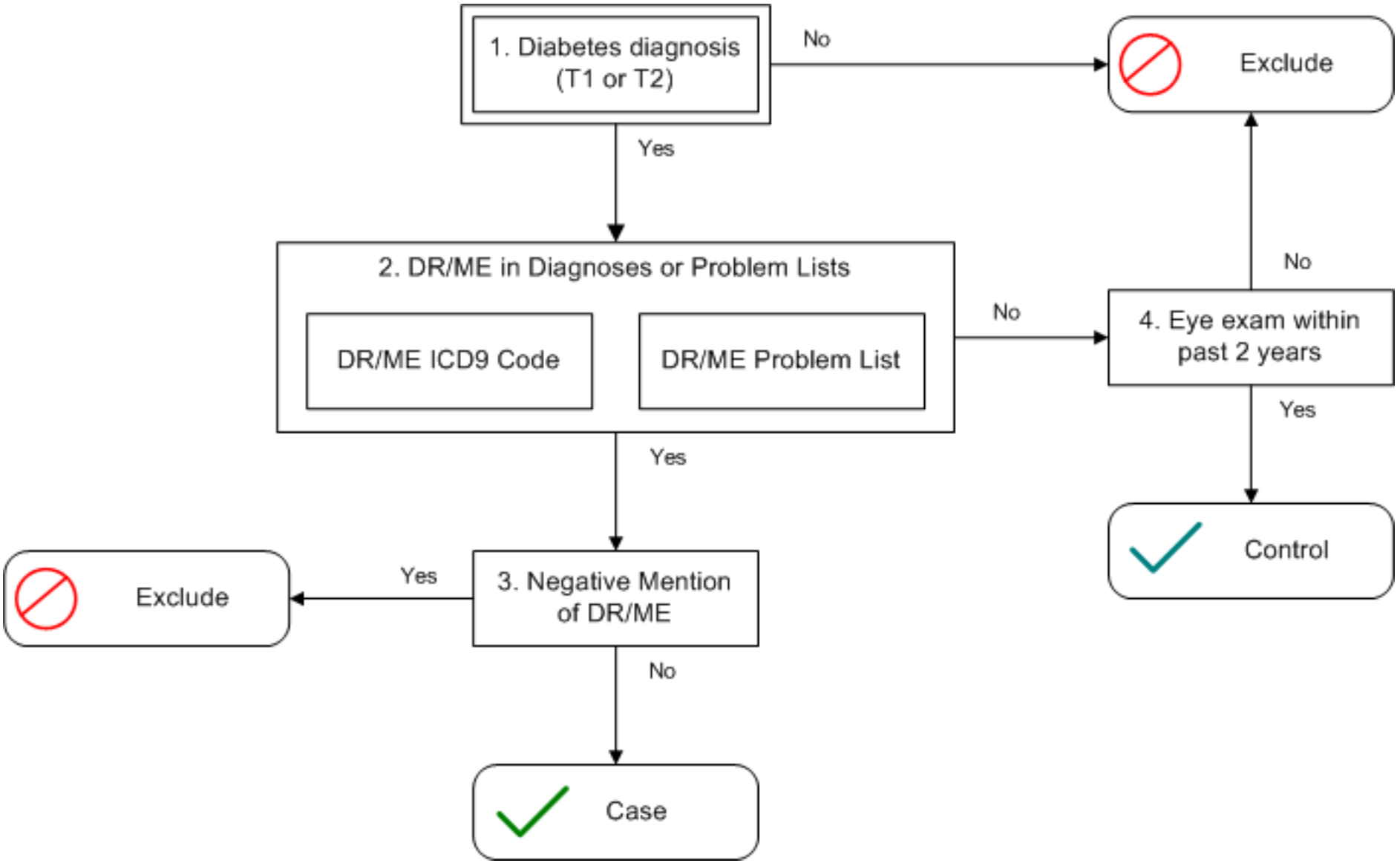
Results Ninety-seven articles met our inclusion criteria. Forty-six used natural language processing (NLP)-based techniques, 24 described rule-based systems, 41 used

ation strategies. The ability to extract meaningful pieces of information from the EHR and consolidate them into a coherent structure would provide great value for automatically identifying patient cohorts that satisfy complex criteria.

A large number of studies have been published describing automated phenotyping techniques employed by medical organizations across the USA. Owing to the sensitive nature of patient data,

Main approaches

- Logical rules
  - Where do rules come from? Guidelines, expert knowledge
- Text-based
  - Keywords, deeper semantic processing
- Statistical/ML(classification)


1. Diabetes diagnosis (T1 or T2)
No
Exclude
Yes
2. DR/ME in Diagnoses or Problem Lists
DR/ME ICD9 Code
DR/ME Problem List
No
4. Eye exam within past 2 years
No
Yes
Yes
Control
Exclude
Yes
3. Negative Mention of DR/ME
No
Case

eMERGE example - settings

Overview of participating institutions' EMR and biorepositories and recruitment models

Institution	Biorepository overview	Recruitment model	Repository size	EMR summary
Marshfield Clinic Research Foundation (Marshfield, Wisconsin, USA)	Personalized medicine research project: Geographically defined cohort within an integrated regional healthcare system	Population based	20 000 98% Caucasian	Comprehensive internally developed EMR since 1985 75% participants have 20+ years medical history
Northwestern University (Chicago, Illinois, USA)	Nugene project: Northwestern affiliated hospitals and outpatient clinics	Population based	10 000 12% AA 8% Hispanic	Comprehensive vendor-based inpatient (Cerner, Kansas City, Missouri, USA) and outpatient (Epic Systems, Verona, Wisconsin, USA) EMR since 2000 20+ years ICD-9 data
Vanderbilt University (Nashville, Tennessee, USA)	BioVU: Vanderbilt Clinic, diverse outpatient clinic	Population based	100 000 11% AA	35+ Years medical history data Comprehensive internally developed EMR since 2000
Group Health Cooperative (Seattle, Washington, USA)	GHC Biobank Alzheimer's disease patient registry and adult changes in thought study	Disease-specific cohort	4000 (>96% Caucasian)	Comprehensive vendor-based (Epic Systems) EMR since 2004 20+ years pharmacy data 15+ years ICD-9 data
Mayo Clinic (Rochester, Minnesota, USA)	Vascular diseases biorepository	Disease-specific cohort	3500 (>96% Caucasian)	Comprehensive internally developed EMR since 1995 40-year history of data extraction

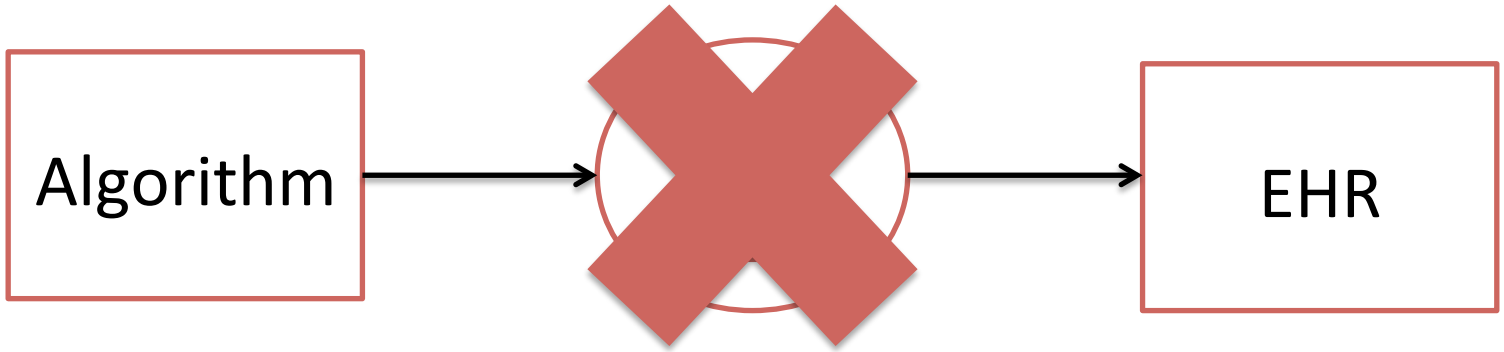
Table 4

Association results for rs7903146 in TCF7L2 with type 2 diabetes

	N	Allele frequencies (T)		OR	L95	U95	p Value
		Cases	Controls				
GHC EA	813	0.327	0.248	1.59	1.25	2.02	0.0002
Marshfield EA	930	0.323	0.243	1.49	1.15	1.93	0.0025
Mayo EA	1159	0.303	0.281	1.17	0.96	1.43	0.1177
NU EA	1229	0.353	0.274	1.51	1.23	1.86	9.27×10^{-5}
VU EA	396	0.334	0.263	1.42	0.99	2.04	0.0601
Cross cohort EA (Meta)		0.328	0.265	1.41			2.98×10^{-10}
NU AA	294	0.353	0.241	0.94	0.24	2.39	0.08672
VU AA	1021	0.353	0.260	1.35	0.11	2.07	2.25×10^{-6}
Cross cohort AA (Meta)		0.353	0.258	1.64			5.30×10^{-7}
Cross-cohort all (Meta)		0.334	0.263	1.46			2.05×10^{-15}

AA, African ancestry; EA, European ancestry; GHC, Group Health Cooperative; L95, 95% CI lower bound; NU, Northwestern University; U95, 95% CI upper bound; VU, Vanderbilt University

Automated phenotyping



4. Eye exam within the past 2 years

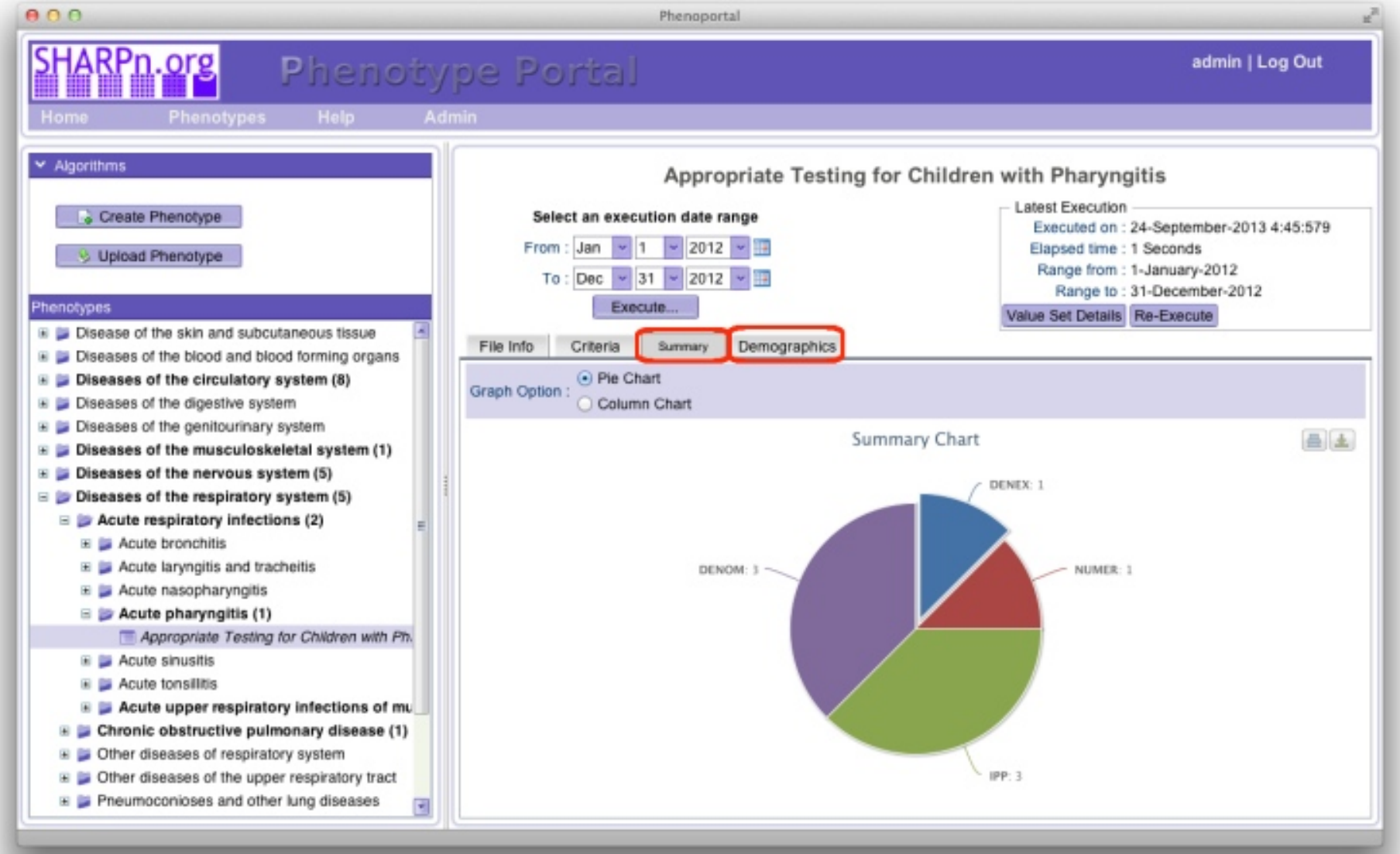
The patient is shown to have had at least one eye exam within the past two years of the reference date. If a patient is deceased, it will be within two years prior to date of death. For this algorithm an eye exam is defined as either:

- Any encounter with a provider in an Optometry or Ophthalmology department
    - If appointment data are available, identify appointments in an Optometry or Ophthalmology department, or an encounter with a provider with a specialty of Optometry or Ophthalmology.

Or

- Any of the following CPT codes for Evaluation & Management (E&M) that were submitted by a provider in an Optometry or Ophthalmology department.
    - New Patient
        - 92002 Intermediate
        - 92004 Comprehensive
        - 99201 Initial Office Visit,Level 1
        - 99202 Initial Office Visit,Level 2
        - 99203 Initial Office Visit,Level 3
        - 99204 Initial Office Visit,Level 4
        - 99205 Initial Office Visit,Level 5
    - Established Patient
        - 92012 Intermediate
        - 92014 Comprehensive
        - 99211 Office O/P Est 5 Min
        - 99212 Office Visit,Level 2
        - 99213 Office Visit,Level 3
        - 99214 Office Visit,Level 4
        - 99215 Office Visit,Level 5

SHARPn


Phenoportal
SHARPn.org
Phenotype Portal
admin | Log Out
Home
Phenotypes
Help
Admin
Algorithms
Create Phenotype
Upload Phenotype
Phenotypes
Disease of the skin and subcutaneous tissue
Diseases of the blood and blood forming organs
Diseases of the circulatory system (8)
Diseases of the digestive system
Diseases of the genitourinary system
Diseases of the musculoskeletal system (1)
Diseases of the nervous system (5)
Diseases of the respiratory system (5)
Acute respiratory infections (2)
Acute bronchitis
Acute laryngitis and tracheitis
Acute nasopharyngitis
Acute pharyngitis (1)
Appropriate Testing for Children with Ph
Acute sinusitis
Acute tonsillitis
Acute upper respiratory infections of mu
Chronic obstructive pulmonary disease (1)
Other diseases of respiratory system
Other diseases of the upper respiratory tract
Pneumoconioses and other lung diseases
Appropriate Testing for Children with Pharyngitis
Select an execution date range
From : Jan 1 2012
To : Dec 31 2012
Execute...
Latest Execution
Executed on : 24-September-2013 4:45:579
Elapsed time : 1 Seconds
Range from : 1-January-2012
Range to : 31-December-2012
Value Set Details
Re-Execute
File Info
Criteria
Summary
Demographics
Graph Option :
Pie Chart
Column Chart
Summary Chart
DENEX: 1
NUMER: 1
IPP: 3
DENOM: 3

SHARPn - eMERGE

Results of analysing eMERGE phenotyping algorithms

Algorithm	Logic				Temporal					Document Data							
	SB[a]	NB[b]	CR[c]	NG[d]	TP[e]	CT[f]	FH[g]	MT[h]	SM[i]	DM[j]	VS[k]	LR[l]	EC[m]	NL[n]	CL[o]	MK[p]	EQ[q]
1 Alzheimers	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−
2 Dementia	−	−	−	−	−	−	−	−	−	−	−	−	+	−	+	−	−
3 Diabetic retinopathy	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	−	−
4 Height	+	−	−	−	−	−	−	+	+	+	−	+	+	−	+	−	−
5 Hypothyroidism	+	+	+	−	+	+	−	−	+	−	+	+	+	+	+	−	+
6 Serum lipid level	+	−	−	−	+	+	−	+	+	−	−	+	+	+	+	−	−
7 Low HDL[r] cholesterol level	+	+	+	+	+	+	−	+	+	−	+	+	+	+	−	−	−
8 Peripheral arterial disease	+	+	+	+	−	−	−	−	+	−	−	+	+	+	+	−	+
9 QRS duration	+	+	−	+	+	+	+	−	+	−	−	+	+	+	+	−	−
10 Red blood cell indices	+	+	−	+	+	+	−	−	+	−	−	+	+	+	+	−	−
11 Resistant hypertension	+	+	+	+	+	+	−	−	−	−	−	+	+	+	+	−	+
12 Type 2 diabetes	+	+	−	+	+	+	+	+	+	+	+	+	+	−	+	−	−
13 White blood cell indices	+	−	−	−	+	+	−	+	+	+	+	+	+	−	+	−	−
14 Cataract	+	+	+	+	+	+	−	+	+	+	−	−	+	+	+	−	−

[a]**SB** (Simple Boolean) — are s
[b]**NB** (Nested boolean) — are
[c]**CR** (Cardinality) — is there e
[d]**NG** (Negation) — is there evi
[e]**TP** (Temporal proximity) —
from now
[f]**CT** (Complex temporal) — m
[g]**FH** (Family History) — does
[h]**MT** (Metadata) — includes d
[i]**SM** (Summary) — does the d
[j]**DM** (Demographic data) — i
[k]**VS** (Vital signs) — defined fo
[l]**LR** (Laboratory test results) –
[m]**EC** (External code) — are co
[n]**NL** (NLP or indicative keywo
or phrases used for pattern ma
[o]**CL** (Collections) — are basic
[p]**MK** Meta-document knowle
[q]**EQ** (Complex calculation) —
[r]High-density lipoprotein

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243189/

High throughput phenotyping

- Previously: laboriously define criteria for one phenotype at a time

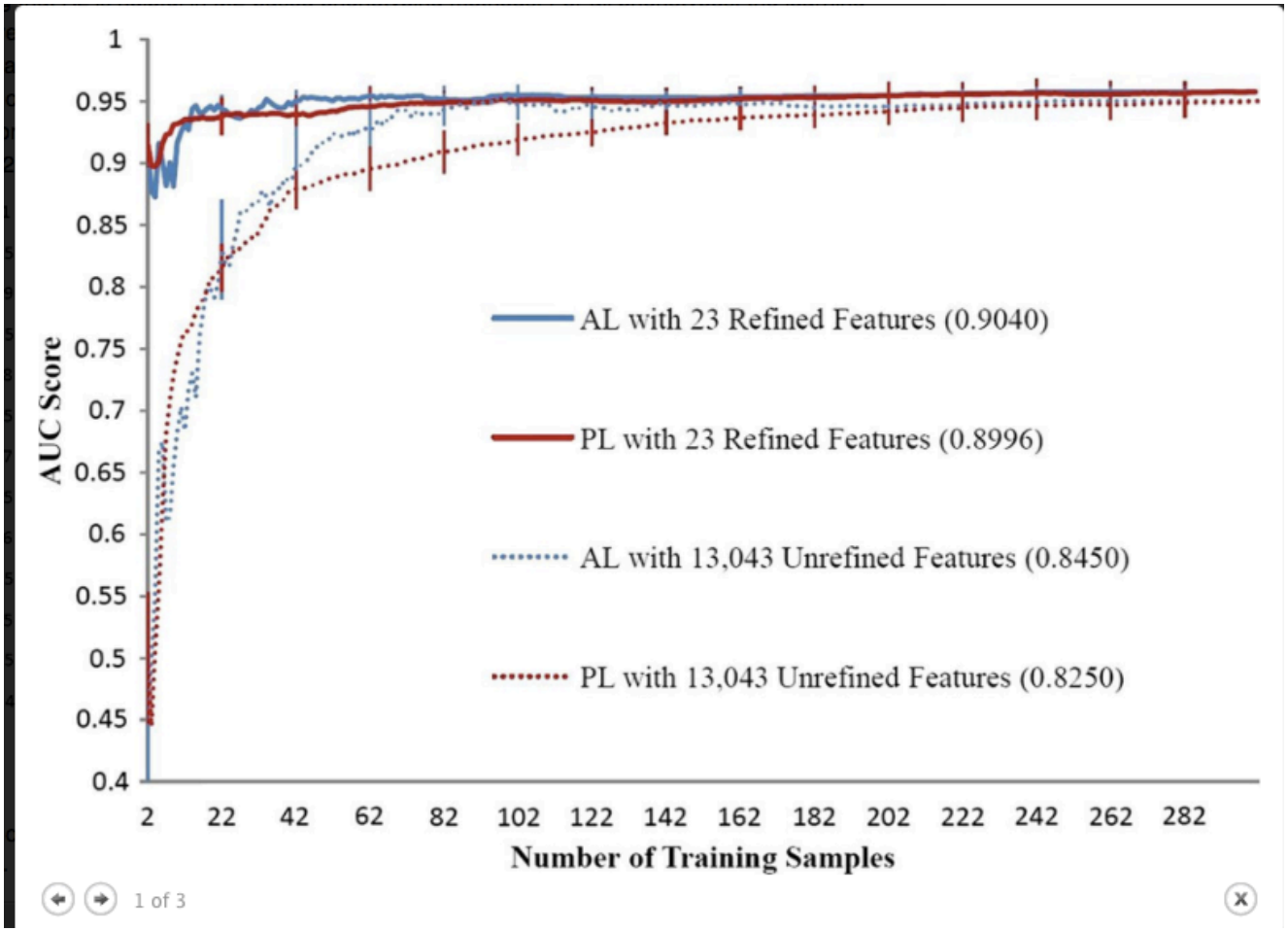
- Limitations
  - Large-scale GWAS
  - Discovering new phenotypes
  - Stratification
  - Only as good as prior knowledge

Table 2

Unrefined and refined feature sets for RA, CRC, and VTE

Task	Feature set	Description	No of features	Details
RA	Unrefined	All ICD-9 codes; all UMLS CUI, which are neither negated nor associated with other people; all medication names	13 043	Carroll *et al*9
	Refined	ICD-9 codes 714.*, 696.0 and 710.0; selected CUI of RA and related diseases; medications commonly used in treating RA	23	
CRC	Unrefined	All ICD-9 and CPT billing codes and all UMLS CUI, which are neither negated nor associated with other people	18 059	Xu *et al*12
	Refined	CRC related ICD-9 and CPT codes and selected UMLS CUI related to CRC	12	
VTE	Unrefined	All ICD-9 codes; all UMLS CUI, which are neither negated nor associated with other people; and all medication names	13 862	Online supplementary tables S1 and S2
	Refined	ICD-9 codes 453.*, 415.*, and V12.51; selected VTE-related CUI; medications including 'lovenox', 'heparin', 'warfarin', and 'coumadin'	56	

Results - RA


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
AUC Score
2
22
42
62
82
102
122
142
162
182
202
222
242
262
282
Number of Training Samples
AL with 23 Refined Features (0.9040)
PL with 23 Refined Features (0.8996)
AL with 13,043 Unrefined Features (0.8450)
PL with 13,043 Unrefined Features (0.8250)
1 of 3

Results - crc

Table 4
Performance comparisons of AL versus PL in annotation cost when AUC is fixed in RA, CRC, and VTE phenotyping tasks using both unrefined and refined feature sets

	AUC							
Phenotype+feature set	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
RA								
Unrefined set								
AL	7	9	13	16	21	27	44	96
PL	5	6	7	10	18	32	70	298
Refined set								
AL	–	–	–	–	–	–	8	42
PL	–	–	–	–	–	–	5	95
CRC								
Unrefined set								
AL	6	10	14	21	25	33	53	124
PL	4	9	11	16	21	32	68	161
Refined et								
AL	–	–	–	–	–	–	–	5
PL	–	–	–	–	–	–	–	13
VTE								
Unrefined set								

2 42 82 122 162 202 242 282 322 362 402 442 482 522 561.6

Challenges?

- Missing data
- Bias
- Error
- Standardization

Open problems

- Phenotype detection
- Time
  - Disease state changing over time, patient population changing over time, medical process changing over time
- Interpretation
  - Making automated phenotypes understandable
- Evaluation
  - What's the ground truth?
- Fragmented data
  - Nonclinical data
  - HIE

Pharmacovigilance

- Drug trials involve small populations
- What about
  - Rare side effects?
  - Interactions with other medications?
  - Interactions with foods?
  - Repurposing drugs?

- Postmarket analysis

How are ADEs reported?

- Clinical trials
  - Every event reported
- FDA AERS
  - Selected events submitted by patient, doctor

- But also indirectly
  - Searching for information
  - Evidence of side effect in EHR

Biomedical informatics goals

- **Drug side effects**
- **Drug-drug interactions**
- Dosing
- Ideal treatment
- Subgroup
- Finding new indications

Detecting Drug Interactions From Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels

NP Tatonetti[1,2,3], **JC Denny**[4,5], **SN Murphy**[6,7], **GH Fernald**[1,2,3], **G Krishnan**[8], **V Castro**[6], **P Yue**[9], **PS Tsau**[9], **I Kohane**[7,10,11], **DM Roden**[5], and **RB Altman**[2,3]

[1]Biomedical Informatics Training Program, Stanford University, Stanford, California, USA

[2]Department of Bioengineering, Stanford University, Stanford, California, USA

[3]Department of Genetics, Stanford University, Stanford, California, USA

[4]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA

[5]Department of Medicine, Vanderbilt University, Nashville, Tennessee, USA

[6]Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts, USA

[7]Harvard Medical School, Boston, Massachusetts, USA

[8]Stanford Center for Clinical Informatics, Stanford University, Stanford, California, USA

[9]Department of Medicine/Cardiovascular Medicine, Stanford University, Stanford, California, USA

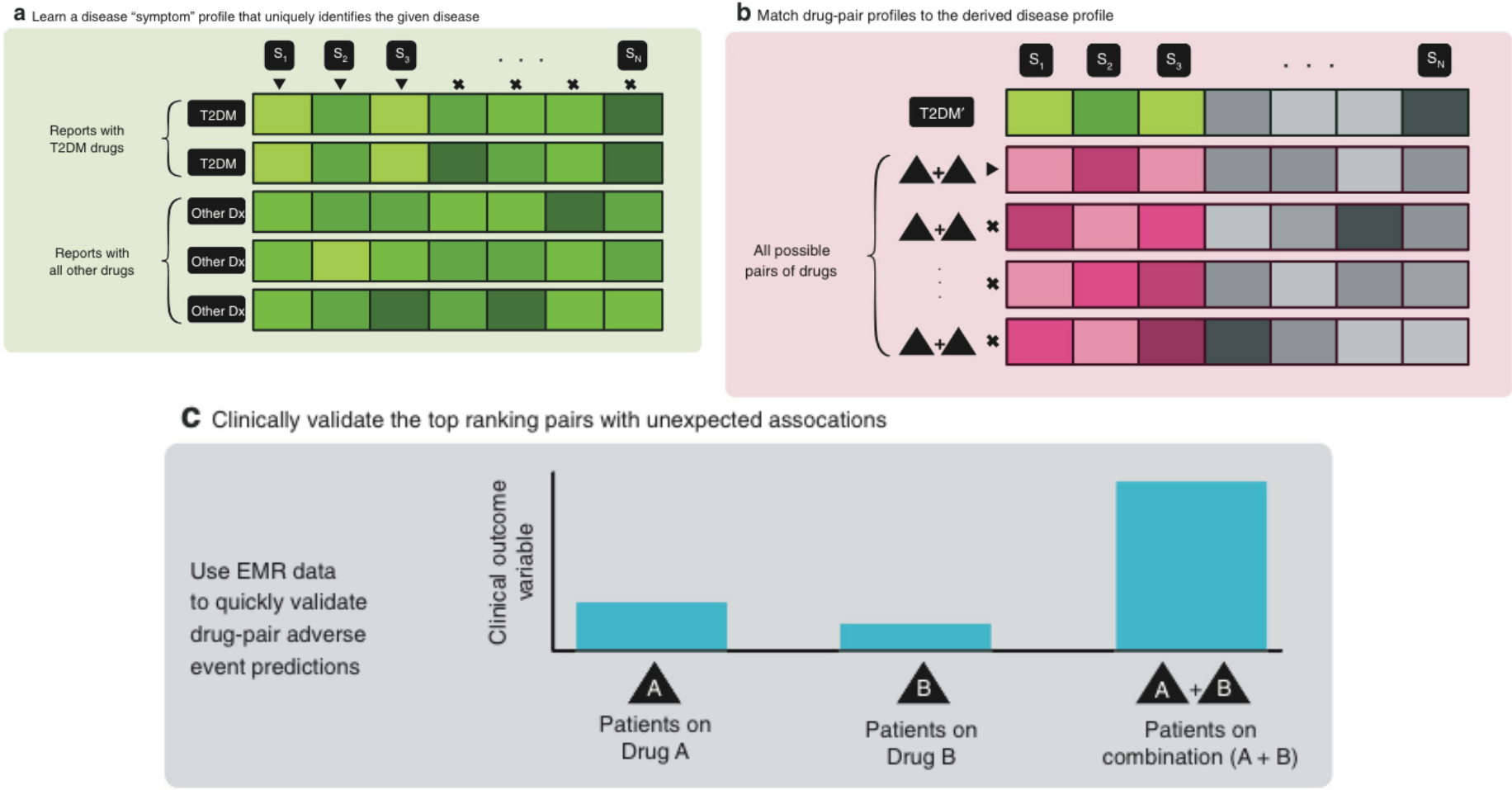
[10]Children's Hospital, Boston, Massachusetts, USA

[11]Brigham and Women's Hospital, Boston, Massachusetts, USA

- Many patients have comorbidities
  - Two common illnesses will lead to many drug-pairs in common
- One effect: raised glucose
  - T2DM increasing, major public health problem
- DDI may not be reported as such, need to look at co-occurrences

Methods

- 12,627 averse event reports
  - 37 drugs
- 4 pairs
  - 3 were infrequent combo
- 1 combination left: Paroxetine, Pravastatin
  - Took glucose measurements before/during treatment for patients at Stanford
    - 374 on paroxetine, 449 on pravastatin, 8 on both
  - Replicated at Vanderbilt, and Partners



Tatonetti, N.P., Denny, J.C., Murphy, S.N., Fernald, G.H., Krishnan, G., Castro, V., Yue, P., Tsau, P.S., Kohane, I., Roden, D.M. & Altman, R., 2011, Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels, *Clinical Pharmacology & Therapeutics*, 90(1), pp. 133-42

Today's paper

Downloaded from jamia.bmj.com on March 23, 2014 - Published by group.bmj.com

Brief communication

Web-scale pharmacovigilance: listening to signals from the crowd

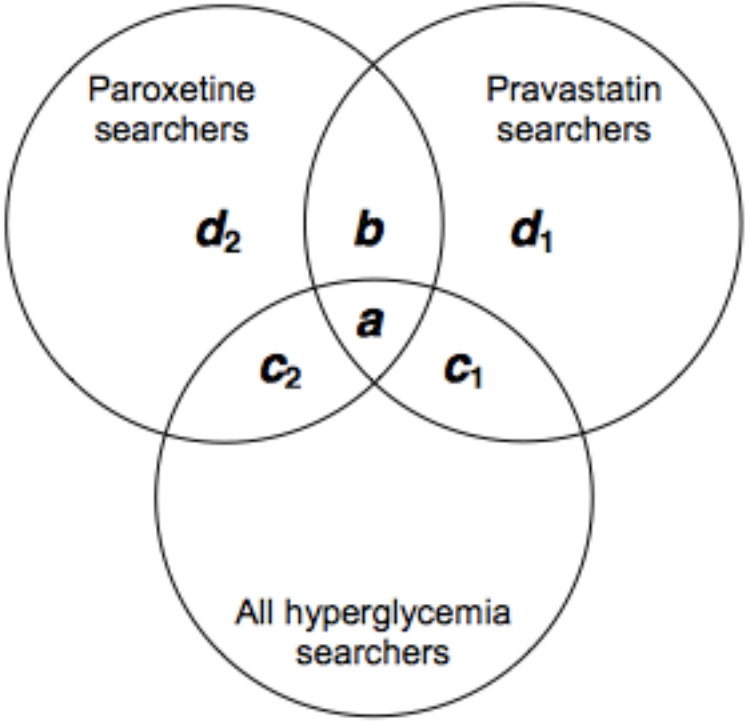
Ryen W White,[1] Nicholas P Tatonetti,[2] Nigam H Shah,[3] Russ B Altman,[4] Eric Horvitz[1]

ABSTRACT

Adverse drug events cause substantial morbidity and mortality and are often discovered after a drug comes to market. We hypothesized that Internet users may provide early clues about adverse drug events via their online information-seeking. We conducted a large-scale study of Web search log data gathered during 2010. We pay particular attention to the specific drug pairing of paroxetine and pravastatin, whose interaction was reported to cause hyperglycemia *after* the time period of the online logs used in the analysis. We also examine sets of drug pairs known to be associated with hyperglycemia and those not associated with hyperglycemia. We find that anonymized signals on drug

case an interaction between paroxetine (an antidepressant) and pravastatin (a cholesterol-lowering drug), which was recently reported to create hyperglycemia.[13 14] This association was extracted from the US Food and Drug Administration adverse event reporting system (AERS) using a data-mining algorithm that aggregates reports to identify drug–drug interactions.[13] The finding was confirmed in a retrospective analysis of the electronic health records of three regionally distinct medical institutions and confirmed in a mouse model.[14] We hypothesized that patients taking these two drugs might experience symptoms of hyperglycemia and may have conducted internet searches on these symptoms and concerns

- Can we use searches to find side effects?
  - People might search symptom
  - But might also search something they've heard on TV/read about
- Medications from previous study, before reported


Paroxetine searchers
Pravastatin searchers
d_2
b
d_1
a
c_2
c_1
All hyperglycemia searchers

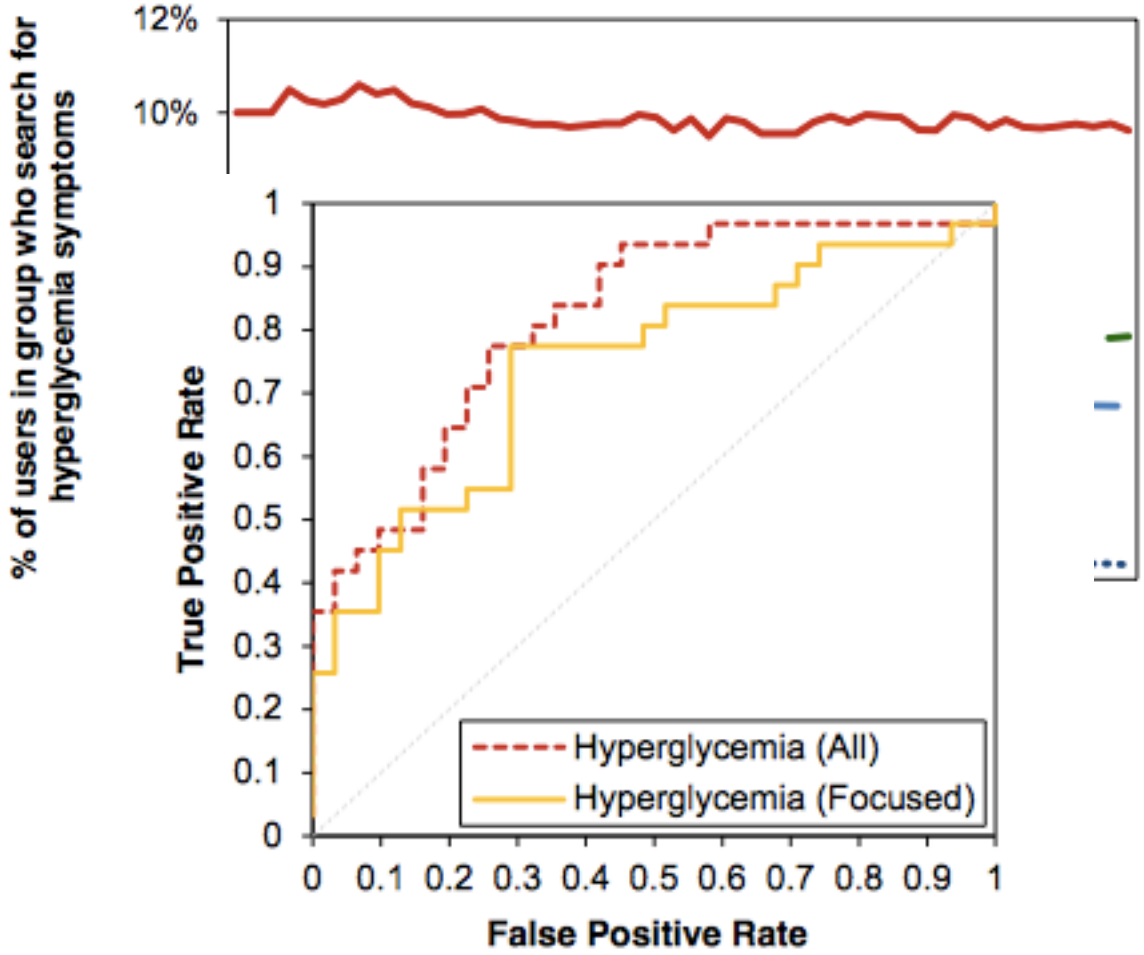


Fig. 2. Percentage of users in each of the three user groups searching for hyperglycemia-related terms. Percentage is computed per week over 12 months of search log data. *Background* refers to the fraction of all searchers who search for hyperglycemia-related symptoms or terminology independent of the presence of the drugs in the users' search histories.

Limitations?

- Not sure people necessarily on drug – might be searching for alternatives
- Might not be on meds at same time
- Multiple users of machine

Google flu

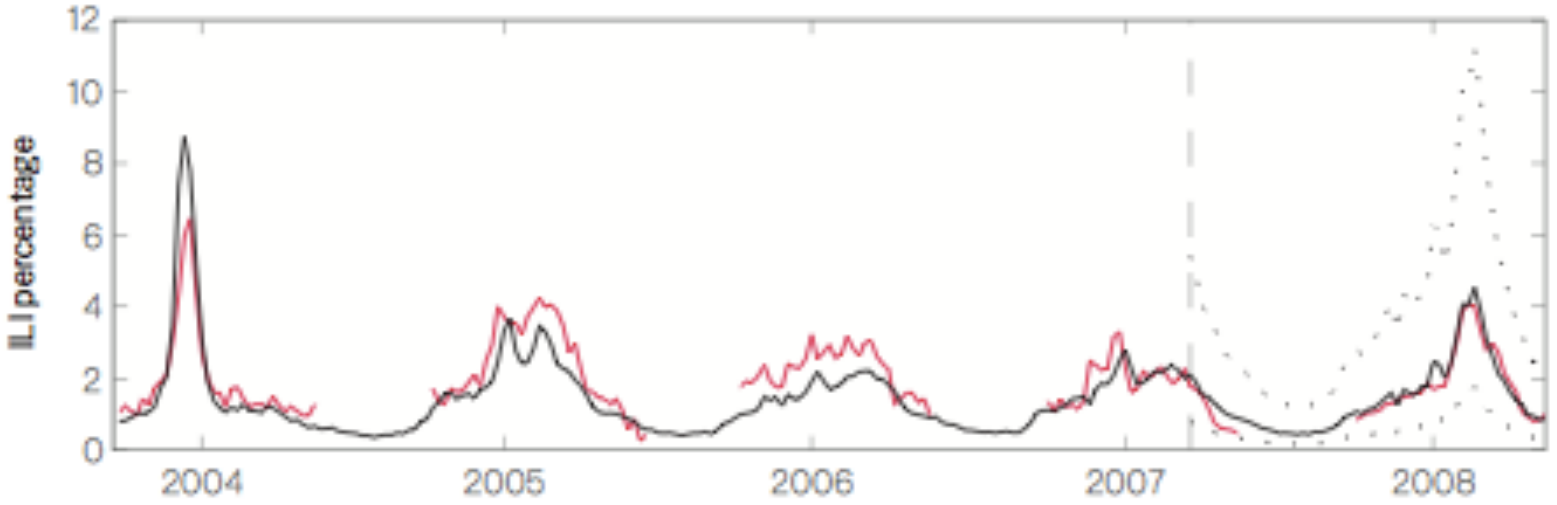
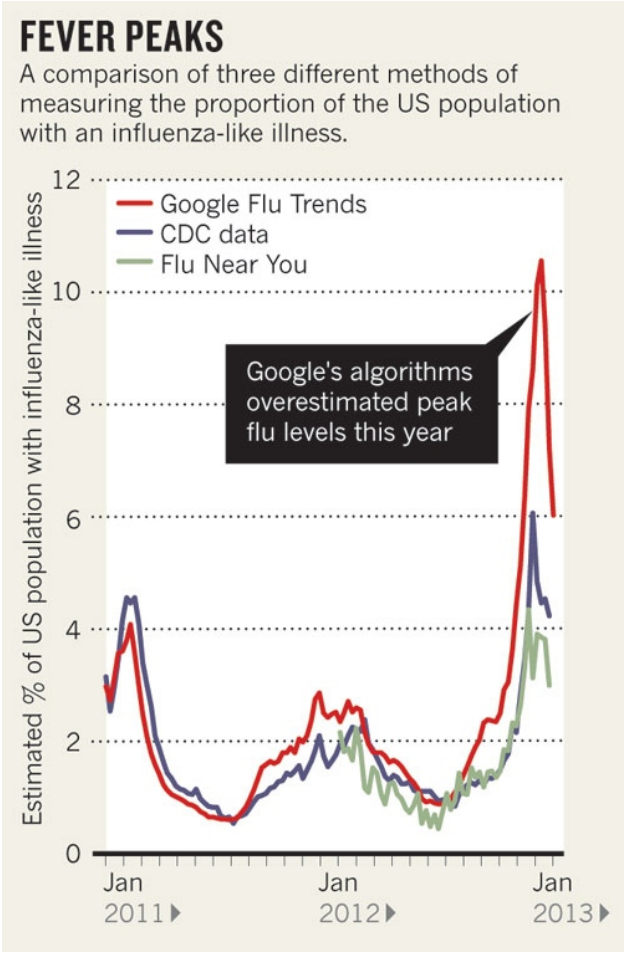


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

http://static.googleusercontent.com/media/research.google.com/en/us/archive/papers/detecting-influenza-epidemics.pdf



http://www.nature.com/news/when-google-got-flu-wrong-1.12413

You Are What You Tweet: Analyzing Twitter for Public Health

Michael J. Paul and **Mark Dredze**
Human Language Technology Center of Excellence
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
{mpaul,mdredze}@cs.jhu.edu

Ailment	**Allergies**	**Depression**	**Aches/Pains**	**Cancer**	**Obesity**	**Flu**	**Dental**
Prior	Allergies	Anxiety	Back Pain	Breast Cancer	Diabetes	Flu	Oral Health
Frequency	6.4%	5.8%	10.8%	8.0%	2.3%	8.1%	4.6%
General Words	allergies	help	body	cancer	blood	flu	meds
	stop	dont	head	pray	doctor	“swine flu”	killers
	eyes	body	need	mom	high	“flu shot”	dentist
	allergic	depression	hurts	shes	meds	dont	teeth
Symptoms	sneezing	pain	pain	pain	pressure	fever	pain
	cold	anxiety	aches	sad	“high blood	cold	toothache
	coughing	stomach	stomach	“breast cancer”	pressure”	“sore throat”	sore
Treatments	medicine	surgery	massage	surgery	hospital	hospital	braces
	benadryl	treatment	“hot bath”	hospital	diet	vaccine	surgery
	claritin	plastic	ibuprofen	“heart surgery”	exercise	medicine	antibiotics

Separating Fact from Fear: Tracking Flu Infections on Twitter

Alex Lamb, Michael J. Paul, Mark Dredze
Human Language Technology Center of Excellence
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
{alamb3,mpaul19,mdredze}@jhu.edu

Class Name	Words in Class
Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
Possession	bird, the flu, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
Vaccination	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
Past Tense	was, did, had, got, were, or verb with the suffix "ed"
Present Tense	is, am, are, have, has, or verb with the suffix "ing"

	Self	Other	Total
Awareness	23.15%	24.07%	47.22%
Infection	37.21%	15.57%	52.78%
Total	60.36%	39.64%	

Table 2: The distribution over labels of the data set. Infection tweets are more likely to be about the author (self) than those expressing awareness.

Data	System	2009	2011
Google	Flu Trends	0.9929	0.8829
Twitter	ATAM	0.9698	0.5131
	Keywords	0.9771	0.6597
	All Flu	0.9833	0.7247
	Infection	**0.9897**	**0.7987**
	Infection+Self	0.9752	0.6662

Table 4: Correlations against CDC ILI data: Aug 2009-Aug 2010, Dec 2011 to Aug 2012.

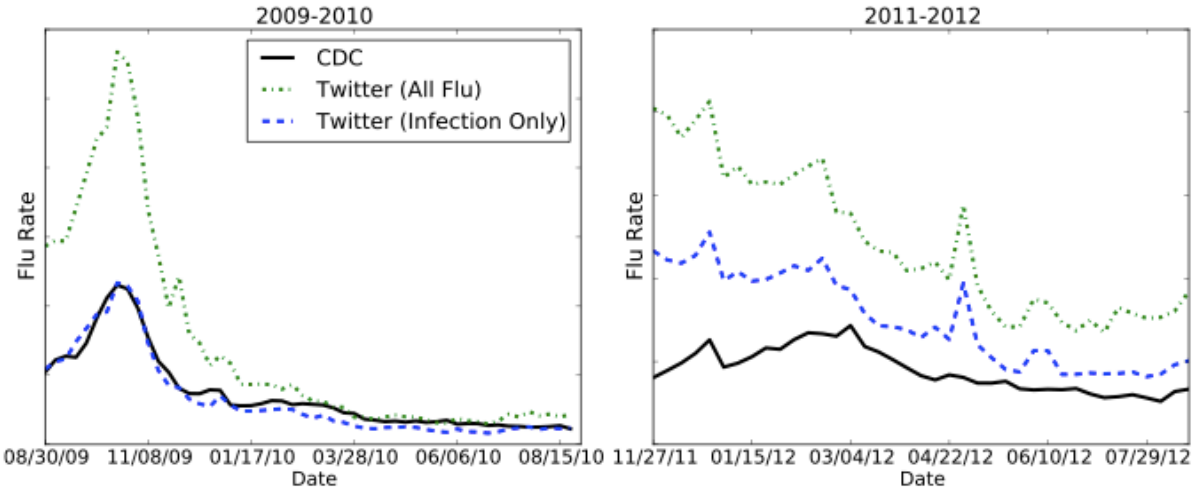


Figure 2: The Twitter flu rate for two years alongside the ILI rates provided by the CDC. The y-axes are not comparable between the two years due to differences in data collection, but we note that the 2011-12 season was much milder.

Main points

- Data is changing over time
- Publicizing results can change searcher behavior
- Correlation vs causation
  - Robustness
- Big data not necessarily good data

- How to discover new uses for drugs? (why important?)
- How to find out when drugs being used off-label?

RESEARCH ARTICLE

DRUG DISCOVERY

Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data

Marina Sirota,[1,2,3]* Joel T. Dudley,[1,2,3]* Jeewon Kim,[4] Annie P. Chiang,[1,2,3] Alex A. Morgan,[1,2,3] Alejandro Sweet-Cordero,[1,5] Julien Sage,[1,5,6] Atul J. Butte[1,3,5]†

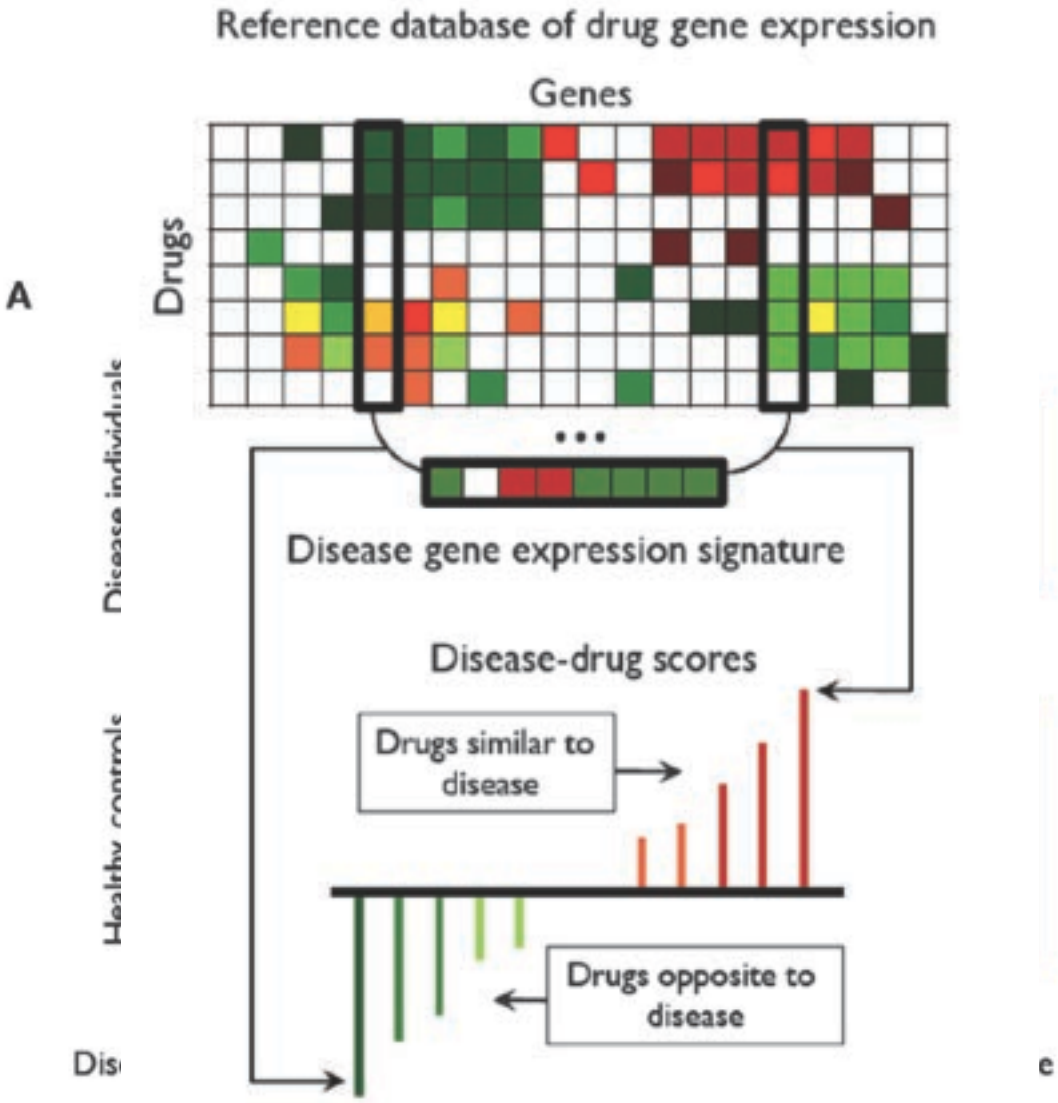
Published 17 August 2011; revised 28 September 2011

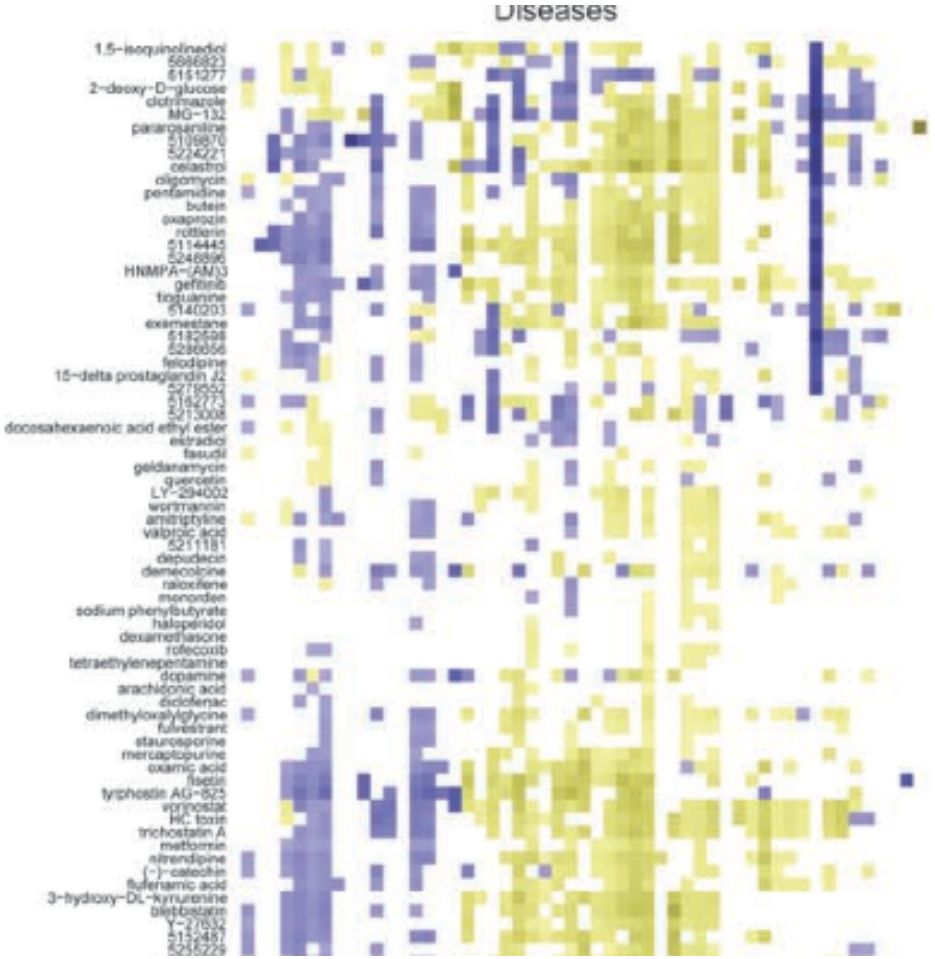
The application of established drug compounds to new therapeutic indications, known as drug repositioning, offers several advantages over traditional drug development, including reduced development costs and shorter paths to approval. Recent approaches to drug repositioning use high-throughput experimental approaches to assess a compound's potential therapeutic qualities. Here, we present a systematic computational approach to predict novel therapeutic indications on the basis of comprehensive testing of molecular signatures in drug-disease pairs. We integrated gene expression measurements from 100 diseases and gene expression measurements on 164 drug compounds, yielding predicted therapeutic potentials for these drugs. We recovered many known drug and disease relationships using computationally derived therapeutic potentials and also predict many new indications for these 164 drugs. We experimentally validated a prediction for the antiulcer drug cimetidine as a candidate therapeutic in the treatment of lung adenocarcinoma, and demonstrate its efficacy both in vitro and in vivo using mouse xenograft models. This computational method provides a systematic approach for repositioning established drugs to treat a wide range of human diseases.

- Lots of $$ and time to develop new drugs
- Can we repurpose old ones for new indications?

- 100 diseases, 164 drugs
  - 16000 pairings -> 2664 stat. sig

- Assumption: if drug leads to opposite expression from disease, could be useful treatment


A
Reference database of drug gene expression
Genes
Drugs
Disease gene expression signature
Disease-drug scores
Drugs similar to disease
Drugs opposite to disease


Diseases
1,5-isoquinolinediol
5986923
6151077
2-deoxy-D-glucose
clotrimazole
MG-132
pararosaniline
5109870
5224221
celastrol
oligomycin
pentamidine
butein
oxaprozin
rottlerin
5114445
5246696
HNMPA-(AM)3
gefitinib
toquamine
6140203
exemestane
5182598
5286056
felodipine
15-delta prostaglandin J2
5279552
5162773
5213008
docosahexaenoic acid ethyl ester
estradiol
fasudil
geldanamycin
quercetin
LY-294002
wortmannin
amitriptyline
valproic acid
5211181
depudecin
democolcine
raloxifene
monorden
sodium phenylbutyrate
haloperidol
dexamethasone
rofecoxib
tetraethylenepentamine
dopamine
arachidonic acid
diclofenac
dimethyloxalylglycine
fulvestrant
staurosporine
mercaptopurine
oxamic acid
fisetin
typhostin AG-825
vorinostat
HC toxin
trichostatin A
metformin
nitrendipine
(-)-catechin
flufenamic acid
3-hydroxy-DL-kynurenine
blebbistatin
Y-27632
5152487
5258229

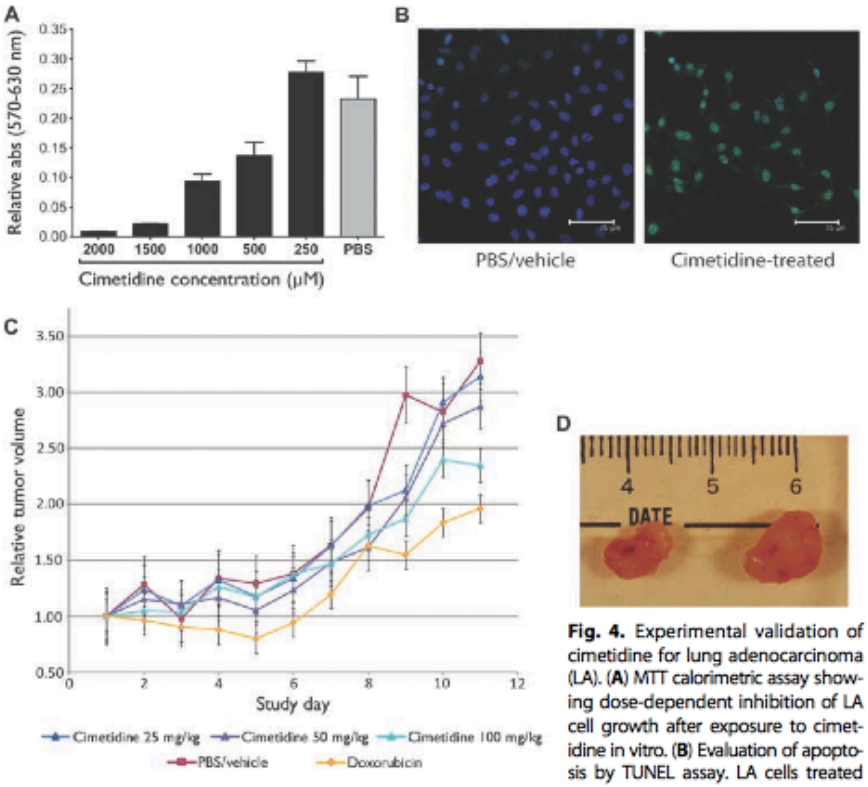


Fig. 4. Experimental validation of cimetidine for lung adenocarcinoma (LA). (**A**) MTT calorimetric assay showing dose-dependent inhibition of LA cell growth after exposure to cimetidine in vitro. (**B**) Evaluation of apoptosis by TUNEL assay. LA cells treated with 2000 μM cimetidine exhibit a significant increase in TUNEL-positive (green) nuclei compared to vehicle (PBS)–treated control. The scale bar is 75 μM. (**C**) Results from a tumor xenograft experiment testing the efficacy of H2 agonist cimetidine in inhibiting the growth of A549 LA cell line tumors in SCID mice. Three treatment groups (25, 50, or 100 mg/kg per injection) and one control group (PBS) were used. Another group was treated with doxorubicin as a positive control. (**D**) Representative images of tumors treated with high dose of cimetidine (left) and control (right).

Next week

Journal club!