VANDERBILT UNIVERSITY School of Medicine

Using Abstraction to Overcome Problems of Sparsity, Irregularity, and Asynchrony in Structured Medical Data

Slides from Jacob P. VanHouten, PhD MS

Overview

Sparsity, irregularity, and asynchrony pose challenges to using EHR data for research

Abstracting clinical data into models and using elements from those models is one way to overcome these problems

This presentation provides insight into the use of different models for this purpose

EHRs are important clinical tools

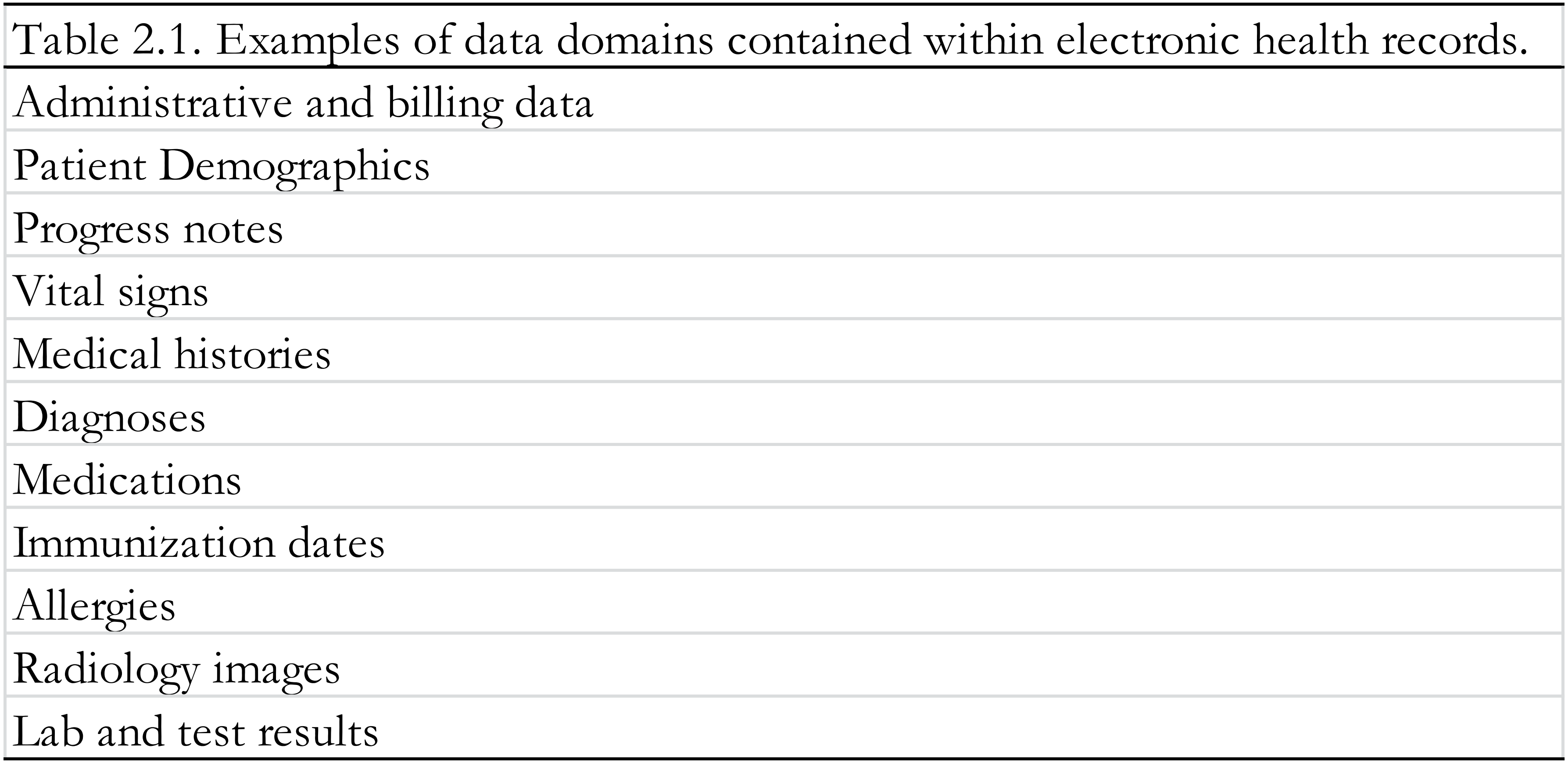


Table 2.1. Examples of data domains contained within electronic health records.

Administrative and billing data
Patient Demographics
Progress notes
Vital signs
Medical histories
Diagnoses
Medications
Immunization dates
Allergies
Radiology images
Lab and test results

EHRs are rich sources of data for secondary research

www.dreamstime.com

Cheap

Longer timeframes

More representative of clinical population and practice

Learning Health System

Approaches to learning from EHRS

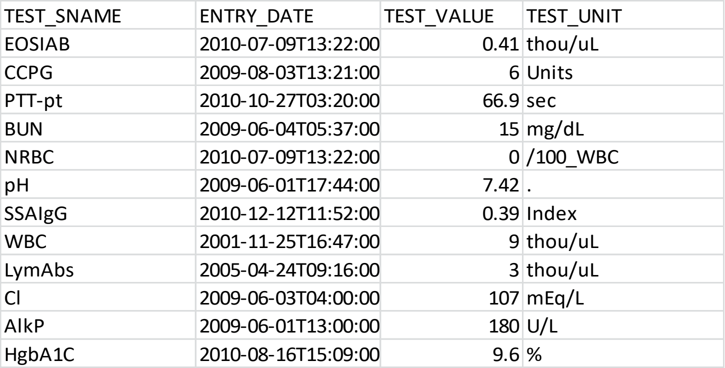
Manual chart review is slow

Automated methods are preferred

Statistical and machine learning approaches can meet this need

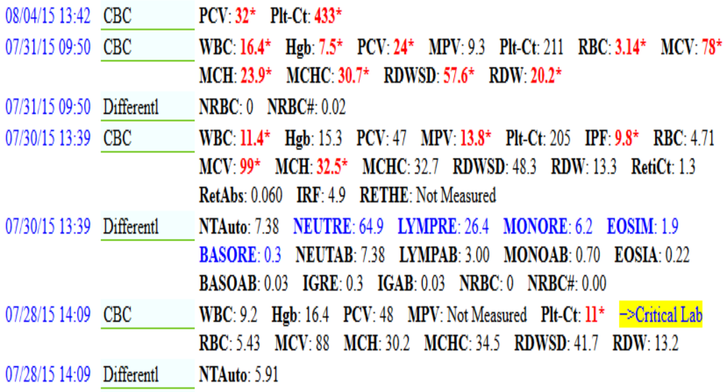
Name (Last, First MI)	ICD-9	Glucose (mg/dL)
VanHouten, Jacob P	47.01	82

EHR data are not (inherently) suited for research



TEST_SNAME	ENTRY_DATE	TEST_VALUE	TEST_UNIT
EOSIAB	2010-07-09T13:22:00	0.41	thou/uL
CCPG	2009-08-03T13:21:00	6	Units
PTT-pt	2010-10-27T03:20:00	66.9	sec
BUN	2009-06-04T05:37:00	15	mg/dL
NRBC	2010-07-09T13:22:00	0	/100_WBC
pH	2009-06-01T17:44:00	7.42	.
SSAIgG	2010-12-12T11:52:00	0.39	Index
WBC	2001-11-25T16:47:00	9	thou/uL
LymAbs	2005-04-24T09:16:00	3	thou/uL
Cl	2009-06-03T04:00:00	107	mEq/L
AlkP	2009-06-01T13:00:00	180	U/L
HgbA1C	2010-08-16T15:09:00	9.6	%

???



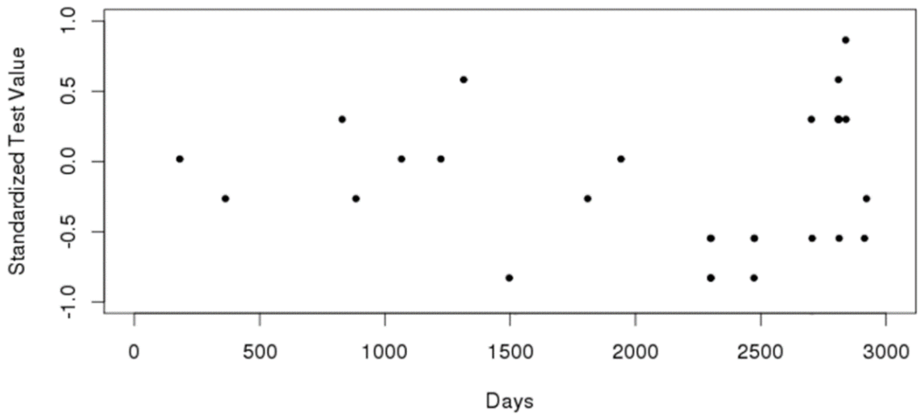
08/04/15 13:42	CBC	PCV: 32* Plt-Ct: 433*
07/31/15 09:50	CBC	WBC: 16.4* Hgb: 7.5* PCV: 24* MPV: 9.3 Plt-Ct: 211 RBC: 3.14* MCV: 78* MCH: 23.9* MCHC: 30.7* RDWSD: 57.6* RDW: 20.2*
07/31/15 09:50	Differentl	NRBC: 0 NRBC#: 0.02
07/30/15 13:39	CBC	WBC: 11.4* Hgb: 15.3 PCV: 47 MPV: 13.8* Plt-Ct: 205 IPF: 9.8* RBC: 4.71 MCV: 99* MCH: 32.5* MCHC: 32.7 RDWSD: 48.3 RDW: 13.3 RetiCt: 1.3 RetAbs: 0.060 IRF: 4.9 RETHE: Not Measured
07/30/15 13:39	Differentl	NTAuto: 7.38 NEUTRE: 64.9 LYMPRE: 26.4 MONORE: 6.2 EOSIM: 1.9 BASORE: 0.3 NEUTAB: 7.38 LYMPAB: 3.00 MONOAB: 0.70 EOSIA: 0.22 BASOAB: 0.03 IGRE: 0.3 IGAB: 0.03 NRBC: 0 NRBC#: 0.00
07/28/15 14:09	CBC	WBC: 9.2 Hgb: 16.4 PCV: 48 MPV: Not Measured Plt-Ct: 11* ->Critical Lab RBC: 5.43 MCV: 88 MCH: 30.2 MCHC: 34.5 RDWSD: 41.7 RDW: 13.2
07/28/15 14:09	Differentl	NTAuto: 5.91

	Var 1	Var 2	Var 3	…	Var *p*	Outcome *y*
Patient 1	*val*	*val*	*val*	…	*val*	0
Patient 2	*val*	*val*	*val*	…	*val*	1
Patient 3	*val*	*val*	*val*	…	*val*	0
…	…	…	…	…	…	…
Patient *n*	*val*	*val*	*val*	*val*	*val*	1

Orientation to Examples

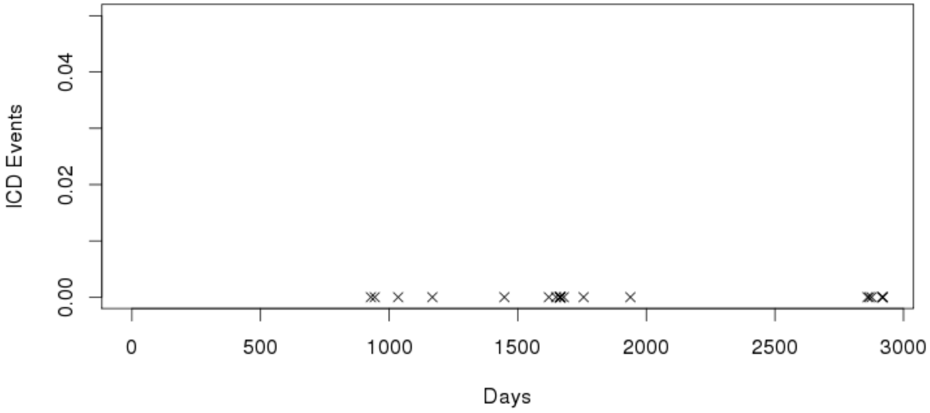
TEST_SNAME	ENTRY_DATE	TEST_VALUE	TEST_UNIT
EOSIAB	2010-07-09T13:22:00	0.41	thou/uL
CCPG	2009-08-03T13:21:00	6	Units
PTT-pt	2010-10-27T03:20:00	66.9	sec
BUN	2009-06-04T05:37:00	15	mg/dL
NRBC	2010-07-09T13:22:00	0	/100_WBC
pH	2009-06-01T17:44:00	7.42	.
SSAIgG	2010-12-12T11:52:00	0.39	Index
WBC	2001-11-25T16:47:00	9	thou/uL
LymAbs	2005-04-24T09:16:00	3	thou/uL
Cl	2009-06-03T04:00:00	107	mEq/L
AlkP	2009-06-01T13:00:00	180	U/L
HgbA1C	2010-08-16T15:09:00	9.6	%

→



ENTRY_DATE	CODE
2005-10-10T00:00:00.0000000	786.5
2000-02-15T00:00:00.0000000	786.5
2000-06-12T00:00:00.0000000	786.5
2009-11-24T00:00:00.0000000	786.5
2009-11-22T00:00:00.0000000	786.5
2007-04-15T00:00:00.0000000	786.5
2009-11-24T00:00:00.0000000	786.5
2006-09-17T00:00:00.0000000	786.5
2009-11-22T00:00:00.0000000	786.5
2002-07-29T00:00:00.0000000	786.5

→

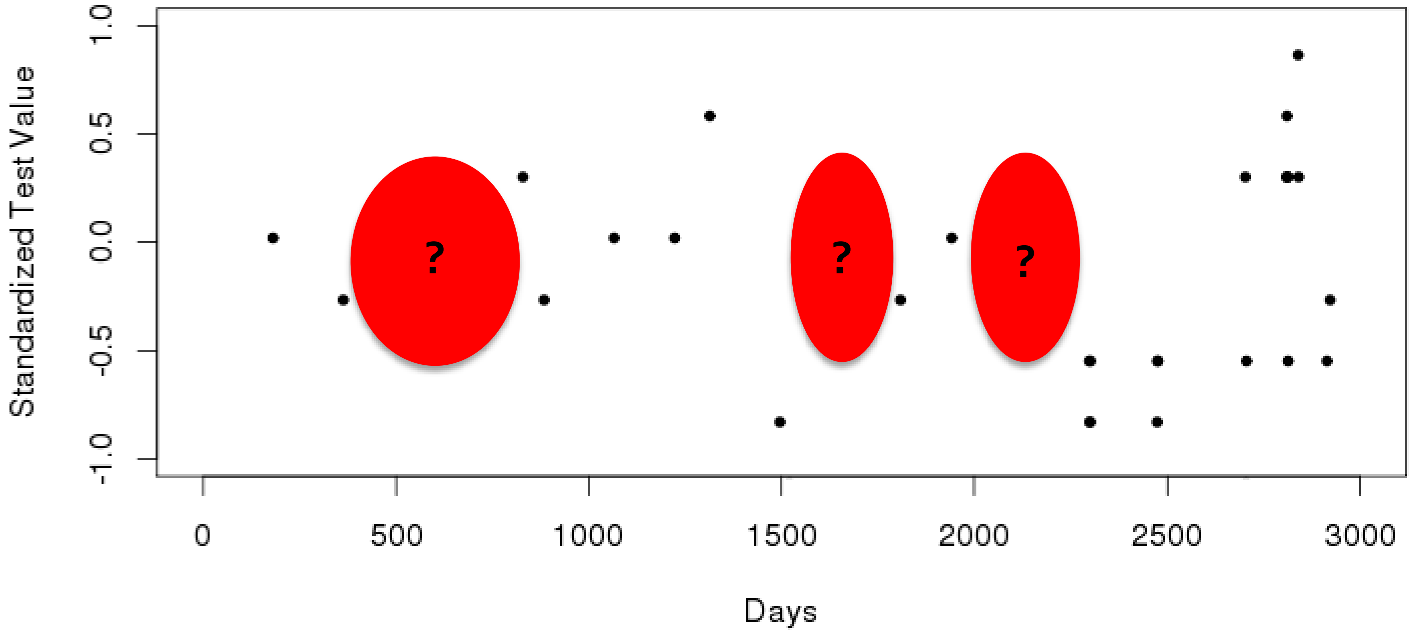


Problems with structured data

Sparsity

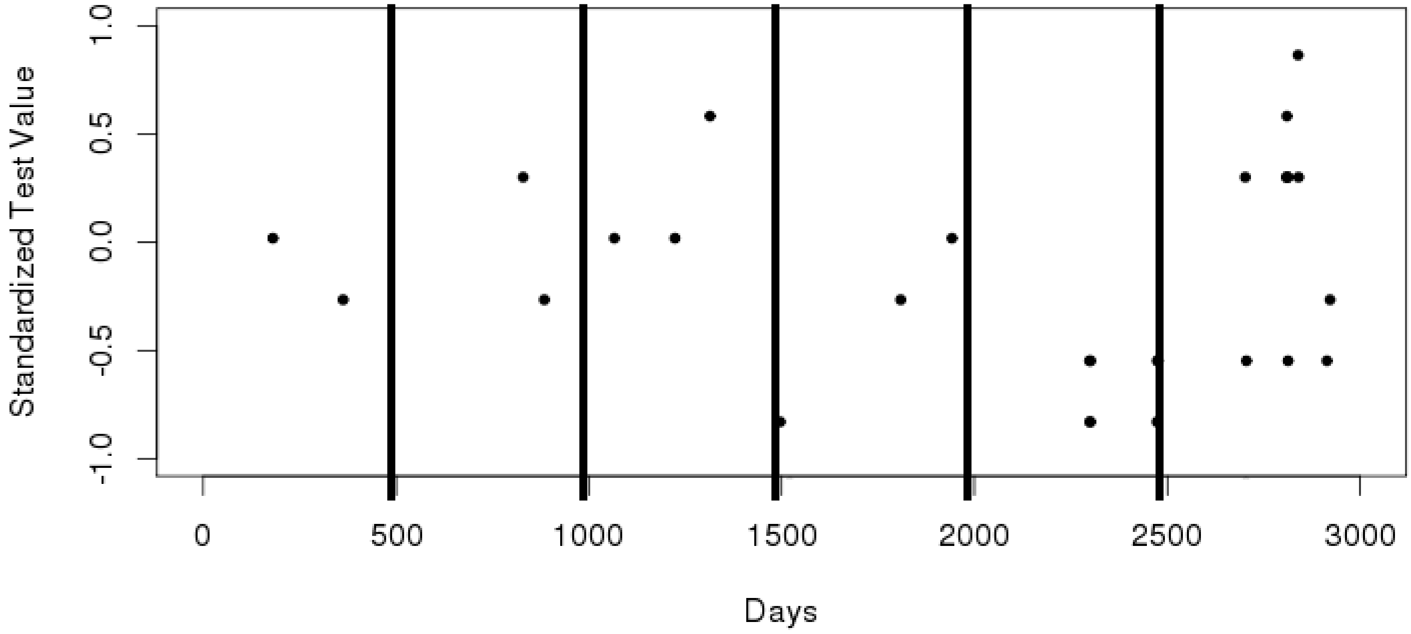
Irregularity

Asynchrony



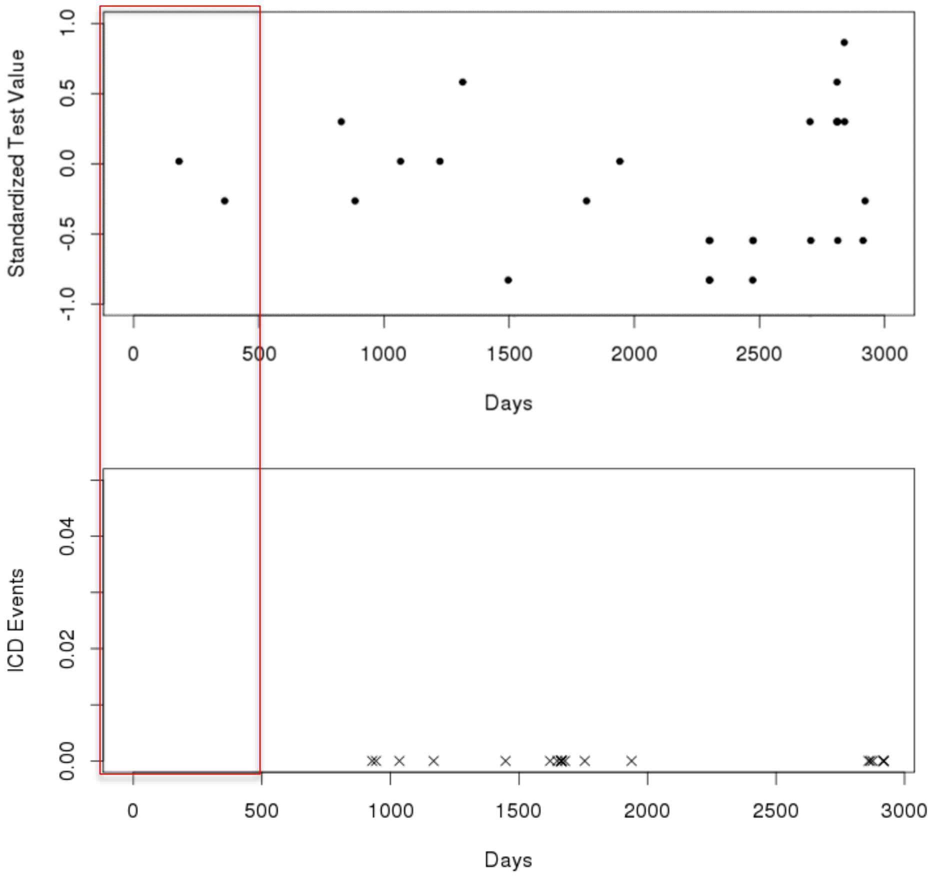
Problems with structured data

Sparsity Irregularity Asynchrony



Problems with structured data

Sparsity Irregularity Asynchrony



Data abstraction models allow us to overcome these problems

We are interested in other characteristics of these abstractions

- How do they affect model performance?
- Do they afford us any new modeling capabilities?
- Can we use this information when making decisions about how to model clinical data?

Classification Tasks

Goal: Predict outcome of interest based on a number of input variables

Typically, computational methods of classification limit these variables

- Small number
- Highly specific

We are interested in using non-specific data

Anderson J Biomed Info 2015
Yu Jamia 2015

Specific vs. Non-specific

Specific vs Non-specific Evidence for Diabetes Mellitus

Specific	Non-specific
Elevated HgbA1C	Coronary artery disease
Metformin	Increased serum creatinine
ICD9 250	Lisinopril

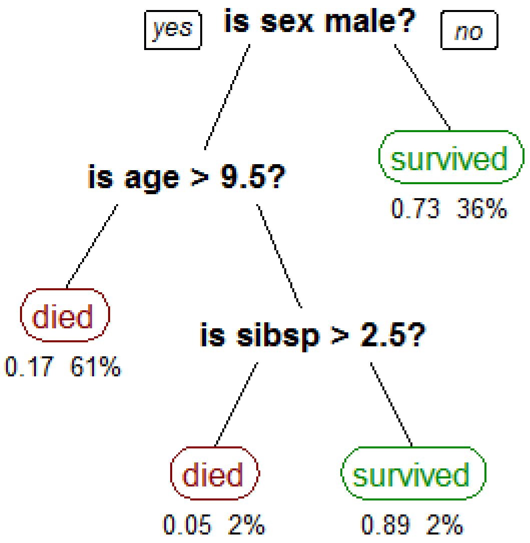
In aggregate, such non-specific information may also be useful in indicating the presence or absence of an outcome of interest.

Overview of Methods

We produced several abstraction models of non-specific laboratory data

We explored the performance of predictive algorithms that used these models as inputs for binary classification tasks

Random Forests



Machine learning ensemble built from many decision trees

- Each tree is randomly different

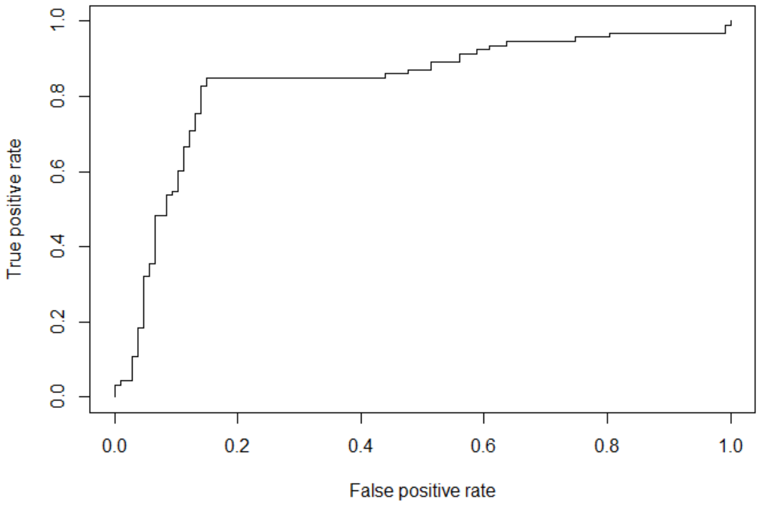
Many desirable properties

- Typically scale well
- Robust to output noise
- Can learn non-linear combinations of variables
- Provide an internal estimate of generalization error
- By permutation tests, can determine variable importance

www.wikipedia.org

Area Under the Curve (AUC)

Example ROC Curve



Single number derived from ROC curve

Used to evaluate the discrimination accuracy of models

Closer to 1 = better

Lasko J Biomed Inform 2005

Methods - Data

Synthetic Derivative Data

Top ~~150~~ 143 lab tests

Most recent 8 years of data

Records needed

- 10/143 labs
- One lab with 3 entries
- No missing sex or race

325,461 records

Methods - Abstraction Models

	Glucose	Na	Cl	TRPI
Binary	[1]	[1]	[1]	[1]
Counts	[20]	[5]	[5]	[1]
Counts/yr	[0, 2, 0, 1, 4, 5, 4, 4]	[0, 0, 0, 0, 1, 2, 1, 1]	[0, 0, 0, 0, 1, 2, 1, 1]	[0, 0, 0, 0, 0, 1, 0, 0]
Cumulative	[0, 2, 2, 3, 7, 12, 16, 20]	[0, 0, 0, 0, 1, 3, 4, 5]	[0, 0, 0, 0, 1, 3, 4, 5]	[0, 0, 0, 0, 0, 1, 1, 1]
Mean	[-0.10]	[0.32]	[-0.42]	[0.35]
Quintiles	[2, 5, 8, 5, 0]	[0, 0, 3, 2, 0]	[0, 3, 1, 1, 0]	[0, 0, 0, 0, 1]
Combo	[(20, -0.10)]	[(5, 0.32)]	[(5, -0.42)]	[(1, 0.35)]

Methods – Classification Tasks

Used these abstraction models as input to random forests

We selected 13 classification tasks

- Could be posed as binary
- Represented varying degrees of difficulty

Methods – Classification Tasks

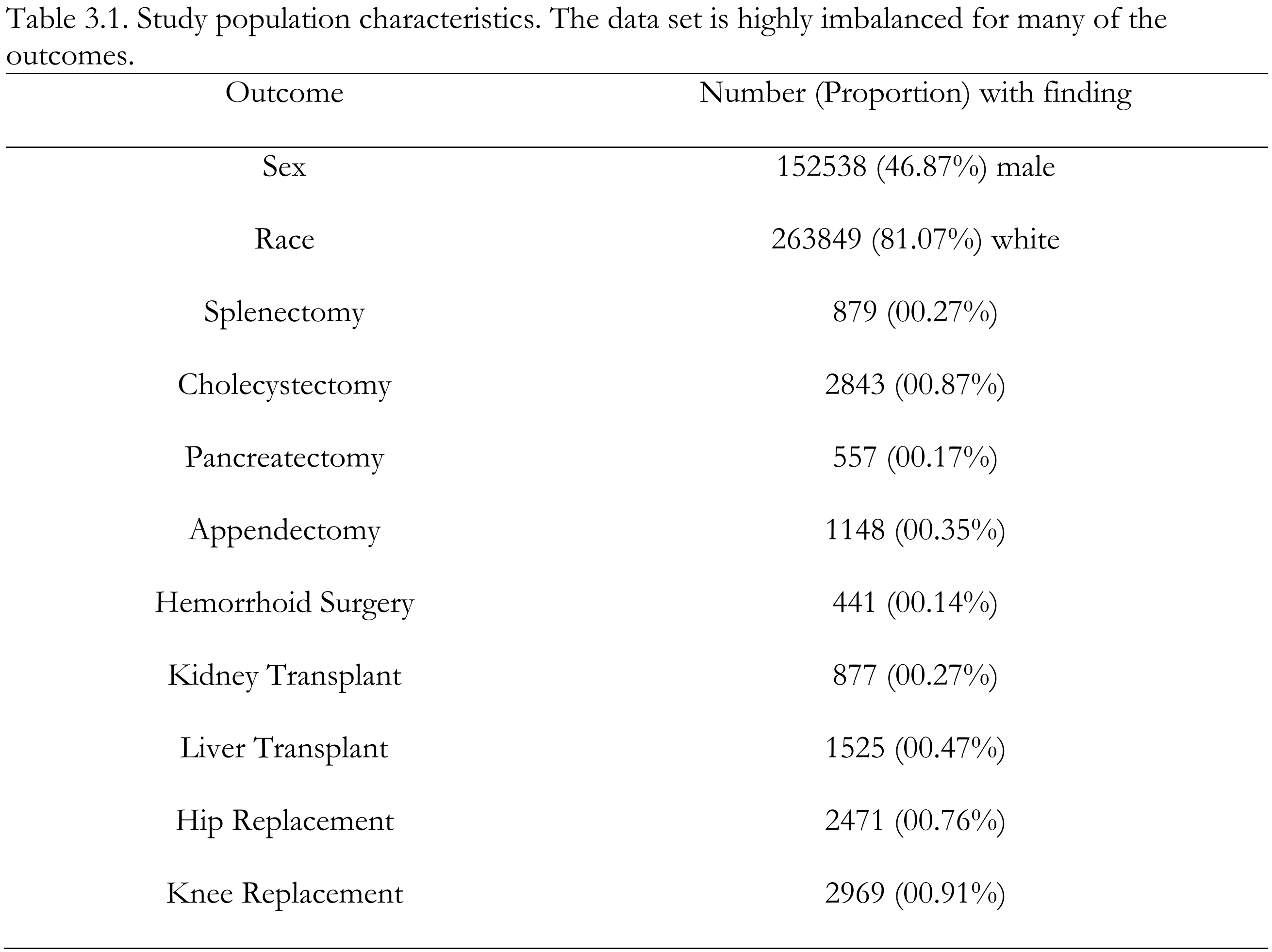


Table 3.1. Study population characteristics. The data set is highly imbalanced for many of the outcomes.

Outcome	Number (Proportion) with finding
Sex	152538 (46.87%) male
Race	263849 (81.07%) white
Splenectomy	879 (00.27%)
Cholecystectomy	2843 (00.87%)
Pancreatectomy	557 (00.17%)
Appendectomy	1148 (00.35%)
Hemorrhoid Surgery	441 (00.14%)
Kidney Transplant	877 (00.27%)
Liver Transplant	1525 (00.47%)
Hip Replacement	2471 (00.76%)
Knee Replacement	2969 (00.91%)

Results - AUC

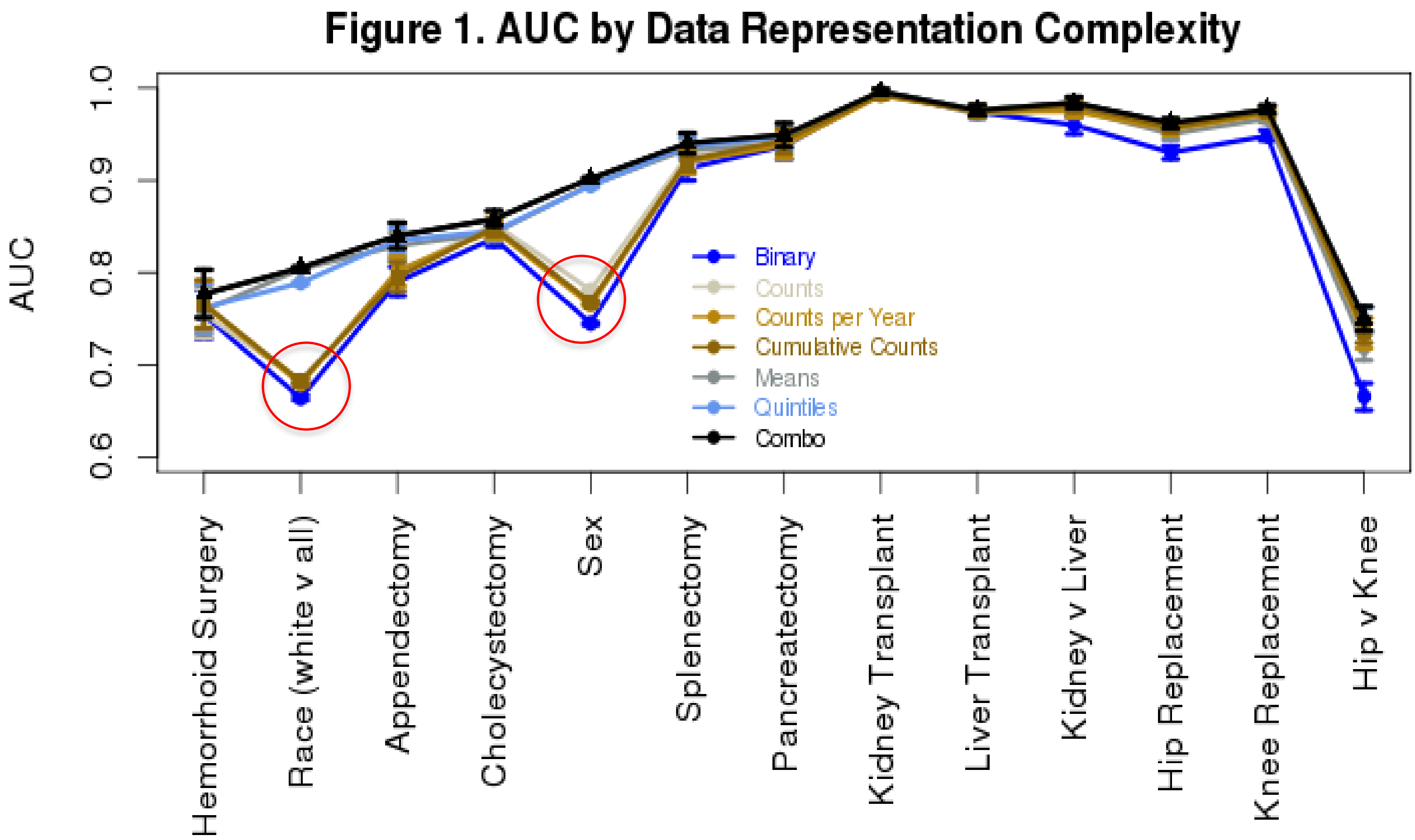
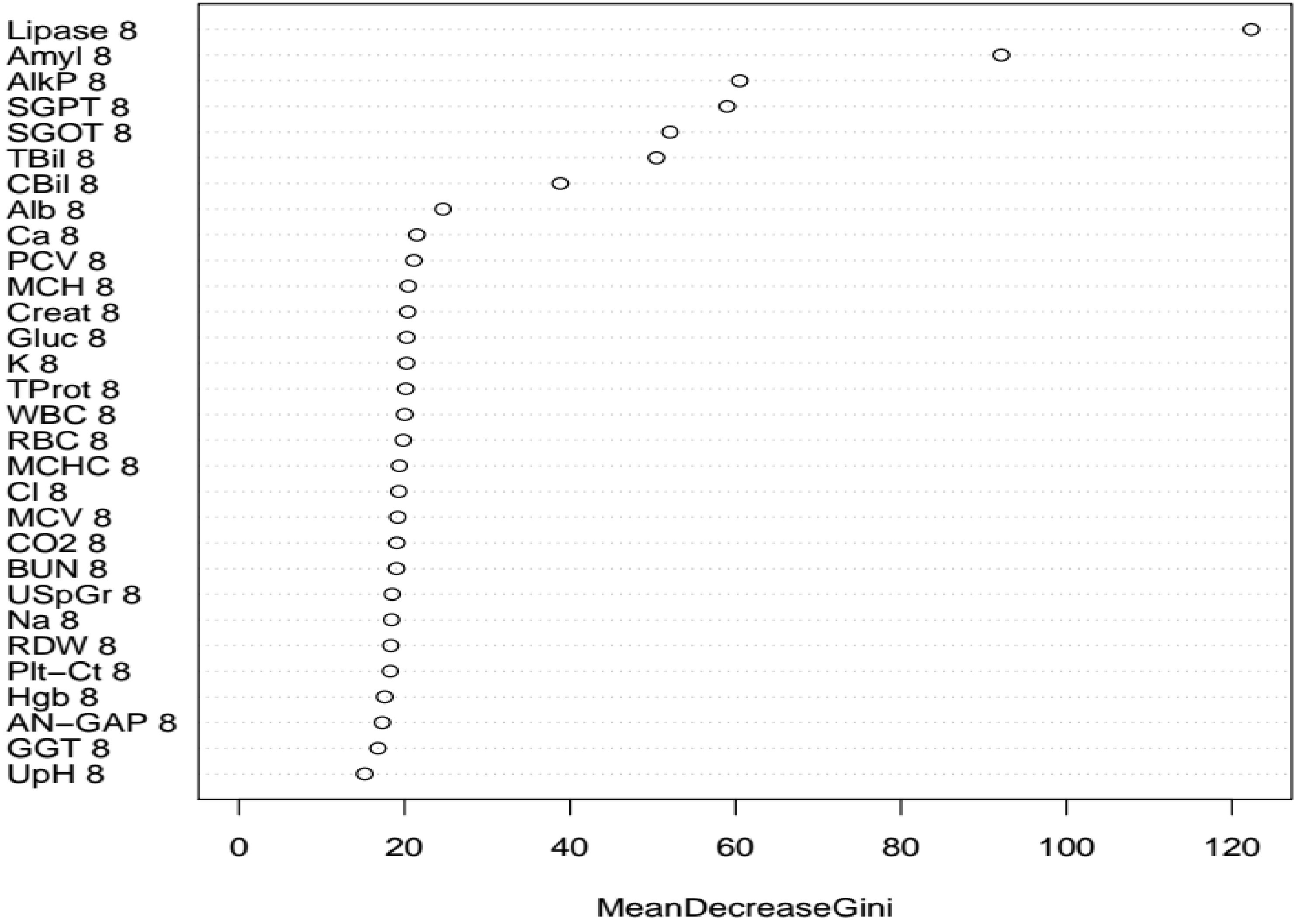


Figure 1. AUC by Data Representation Complexity
AUC
0.6
0.7
0.8
0.9
1.0
Binary
Counts
Counts per Year
Cumulative Counts
Means
Quintiles
Combo
Hemorrhoid Surgery
Race (white v all)
Appendectomy
Cholecystectomy
Sex
Splenectomy
Pancreatectomy
Kidney Transplant
Liver Transplant
Kidney v Liver
Hip Replacement
Knee Replacement
Hip v Knee

Discussion


Lipase 8
Amyl 8
AlkP 8
SGPT 8
SGOT 8
TBil 8
CBil 8
Alb 8
Ca 8
PCV 8
MCH 8
Creat 8
Gluc 8
K 8
TProt 8
WBC 8
RBC 8
MCHC 8
Cl 8
MCV 8
CO2 8
BUN 8
USpGr 8
Na 8
RDW 8
Plt–Ct 8
Hgb 8
AN–GAP 8
GGT 8
UpH 8
0
20
40
60
80
100
120
MeanDecreaseGini

Conclusion

Provides a proof of concept that non-selected features can be used as good inputs

Abstraction models allow for the use of such information

Low-complexity, information-rich representations were the best for these surgical phenotypes

Querying clinical data using continuous abstraction models

Goal: Explore the uses of longitudinal data abstractions

Allow querying of specific laboratory combinations and values

Uncover known (and unknown) associations

Overview of Methods

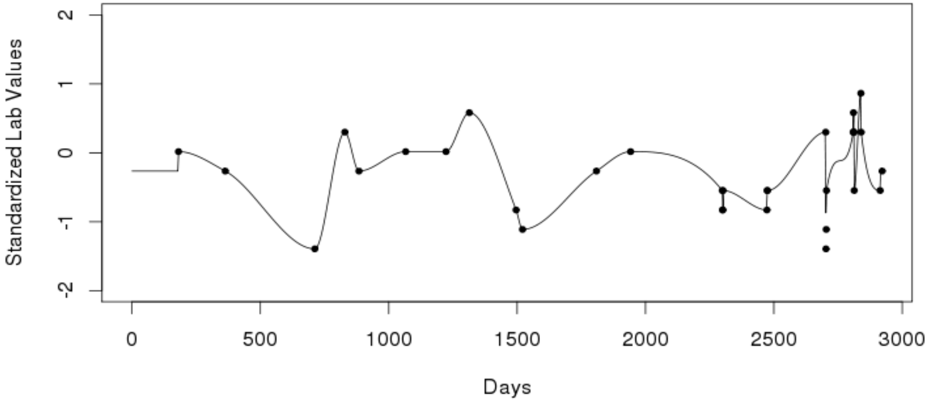
We modeled laboratory values and discrete billing codes as continuous abstractions

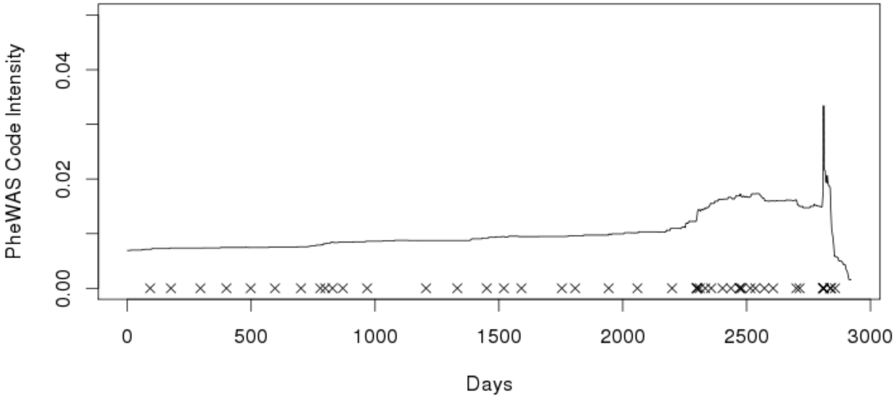
We calculated similarity measure between target lab combinations and each record

We identified associations via correlation

We explored these associations in the context of other associated variables using linear regression

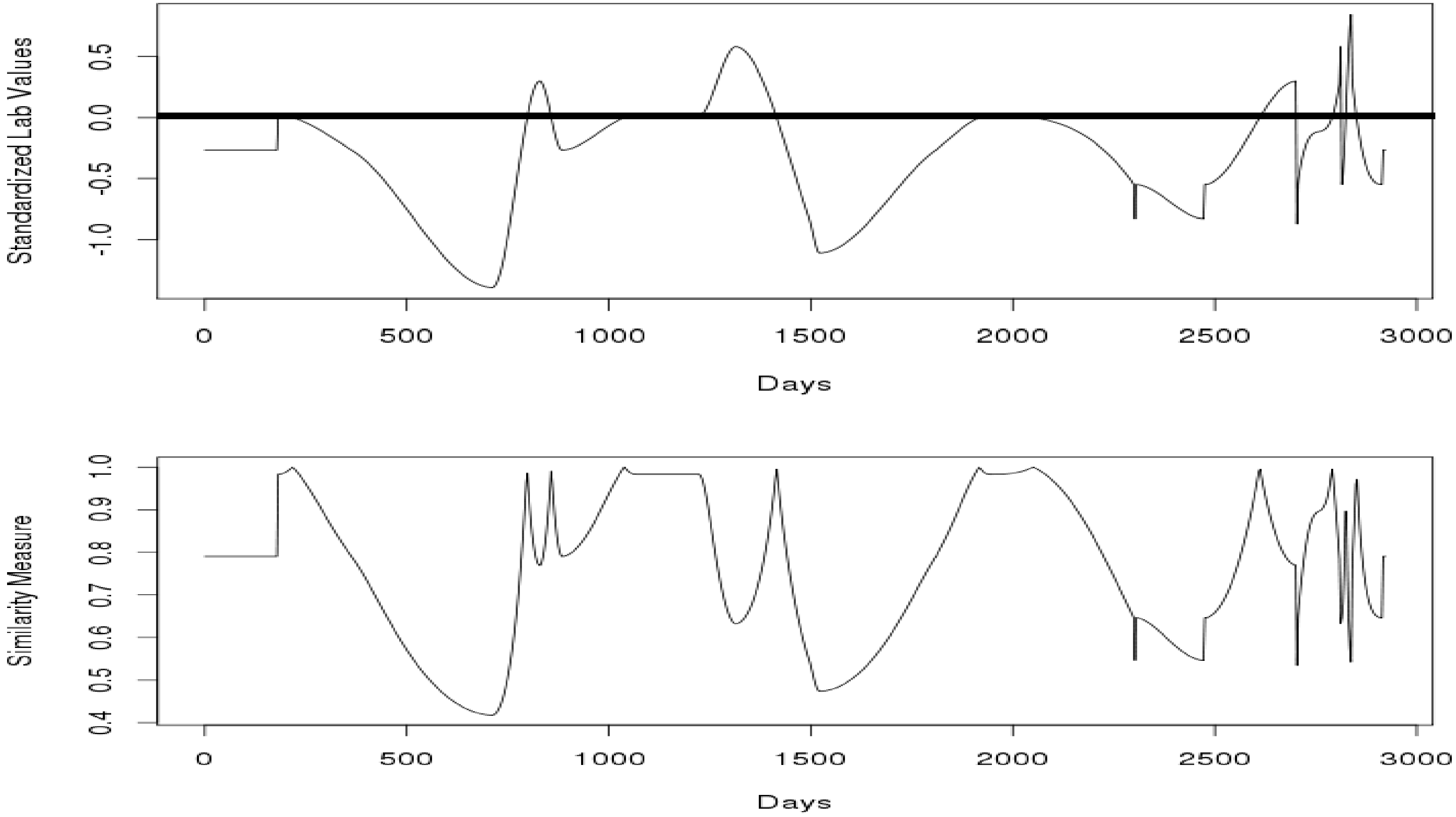
Continuous interpolation abstractions



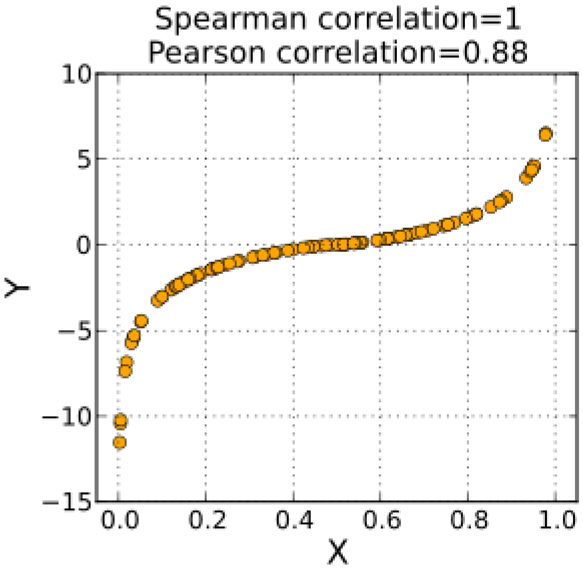


Moler SIAM 2008
Lasko & Bajor (submitted)

Similarity Measure



Correlation



Measures the strength of association between two variables

We chose Spearman's ρ

Range [-1, 1]

www.Wikipedia.org

Linear regression

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i \, l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha) ||\beta||_2^2 / 2 + \alpha ||\beta||_1 \right]$$

Regression can adjust for other relationships

Penalized regression can provide better performance AND perform automatic feature selection (elastic net)

Penalty balance parameter α

Typically uses cross validation to select penalty

Zou & Hastie JR Stat Soc 2005

Methods - Data

Same cohort as previously described

Generated longitudinal representations for lab values and billing codes

Took a single cross-section from each record

288, 966 records

Methods - Testing the Approach

Looked for known univariate associations using correlations

This means that “what codes have higher intensity as we get closer to this laboratory value(s)?”

Selected correlation cutoff of 0.1 (or top 3 values)

Multiple distinct values of labs across some clinical range

- Use case of anti-rejection drugs

Methods - Regression

Look to see if these associations were affected by simultaneously considering additional results

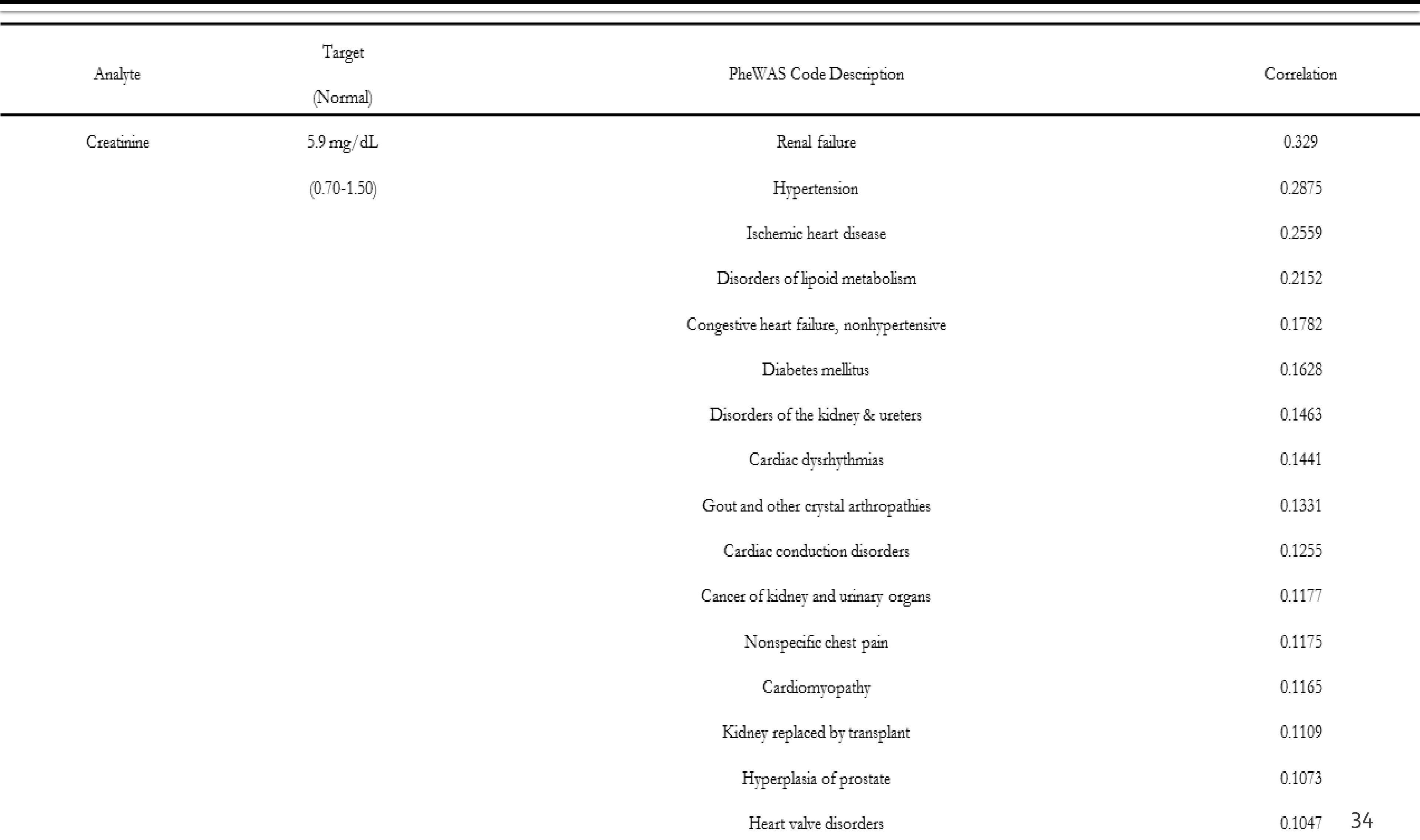
We selected α= 0.5

10-fold cross-validation to set lambda

Results – Correlations

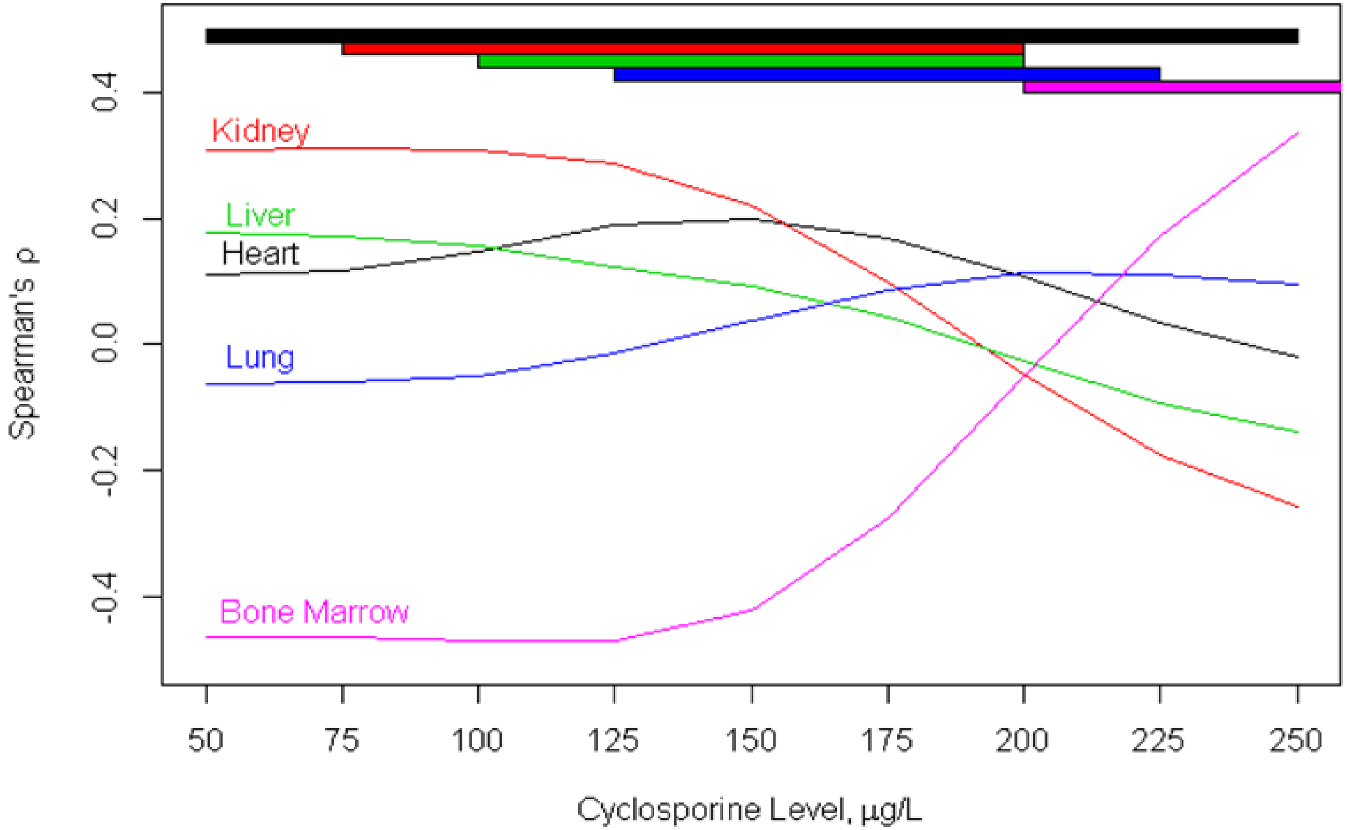
Analyte	Target (Normal Range)	Phewas Code	Correlation
Glucose	450 mg/dL (70-100)	Diabetes mellitus	0.3573
		Hypertension	0.1309
		Ischemic heart disease	0.1301
Troponin	50 ng/mL (<=0.03)	Ischemic heart disease	0.2760
		Congestive heart failure, nonhypertensive	0.1658
		Respiratory failure; insufficiency; arrest	0.1371
		Renal failure	0.1174
		Cardiomyopathy	0.1098
		Shock	0.1080
		Pleurisy	0.1063
		Cardiomegaly	0.1007
		Abnormal serum enzyme levels	0.1004
Lipase	1200 U/L (10-60)	Diseases of pancreas	0.1311
		Chronic liver disease and cirrhosis	0.0827
		Alcohol-related disorders	0.0766
Vitamin B12	1500 pg/mL (180-1000)	Chronic liver disease and cirrhosis	0.0803
		Fluid, electrolyte, & acid-base balance disorders	0.0792
		Other anemias	0.0785

Results - Correlations



Analyte	Target (Normal)	PheWAS Code Description	Correlation
Creatinine	5.9 mg/dL	Renal failure	0.329
	(0.70-1.50)	Hypertension	0.2875
		Ischemic heart disease	0.2559
		Disorders of lipoid metabolism	0.2152
		Congestive heart failure, nonhypertensive	0.1782
		Diabetes mellitus	0.1628
		Disorders of the kidney & ureters	0.1463
		Cardiac dysrhythmias	0.1441
		Gout and other crystal arthropathies	0.1331
		Cardiac conduction disorders	0.1255
		Cancer of kidney and urinary organs	0.1177
		Nonspecific chest pain	0.1175
		Cardiomyopathy	0.1165
		Kidney replaced by transplant	0.1109
		Hyperplasia of prostate	0.1073
		Heart valve disorders	0.1047

Curves allow us to approximately recover clinical guidelines



Curves allow querying against combinations of labs

Analyte	Target (Normal Range)	Phewas Code	Correlation
Glucose	450 mg/dL (70-100)	Diabetes mellitus	0.3573
		Hypertension	0.1309
		Ischemic heart disease	0.1301
Glucose, HbA1C	450 mg/dL; 5.5% (70-100; 4.0-6.5)	Abnormal glucose	0.1366
		Hypertension	0.1272
		Ischemic heart disease	0.1018

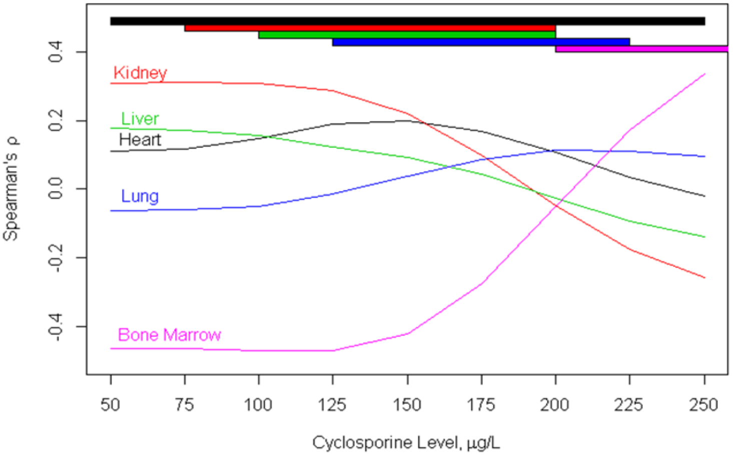
Including normal HbA1C changes the correlation structure

- Diabetes becomes less correlated
- Abnormal glucose becomes the most correlated

Regression coefficients

PheWAS Codes	Coefficient
Gestational diabetes	0.2115
Abnormal glucose	0.2057
Disorders of lipoid metabolism	0.1759
Heart valve disorders	0.1532
Sleep disorders	0.1071
Overweight	0.1051
Known or suspected fetal abnormality	0.0986
Lung transplant	0.097
Other conditions of the mother complicating pregnancy	0.0764
Allergic rhinitis	0.0758
Other and unspecified complications of birth; puerperium affecting management of mother	0.0713
Heart transplant/surgery	0.0712
Back pain	0.0711
Tobacco use disorder	0.0666
Abnormality of organs & soft tissues of pelvis complicating pregnancy, childbirth, or the puerperium	0.0663
Pulmonary collapse; interstitial/compensatory emphysema	0.0649
Pain in joint	0.0596
Liver replaced by transplant	0.0574
Vitamin deficiency	0.0573

Discussion



How would you perform this kind of analysis without longitudinal representations and similarity measures?

Discussion

Temporality is a problem in this method

Principled way to decide on the threshold for the correlations

Granularity of PheWAS codes were too low

Conclusions

Using continuous curve representations helped us

- Overcome data challenges
- Target specific lab values
- Recover known associations, guidelines
- Identify known (but uncommon) associations

Overall Discussion

For “ever” phenotypes, the most compressed, data-rich representations are best

For phenotyping by time, continuous curves were better

GPs as a better way to get uncertainty around the estimate

Conceptual framework for including other data types (images, free text)

Overall Conclusion

Abstraction can be used to overcome sparsity, irregularity and asynchrony

Different abstractions are suited for particular tasks

Applying these methods judiciously could allow wider use of machine learning on clinical data