Unsupervised approaches for phenotyping using electronic health record data

July 29, 2020 | Sentinel Innovation Center Seminar

Katherine P. Liao, MD, MPH

Associate Professor of Medicine | Assistant Professor of Biomedical Informatics
Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital
Massachusetts Veterans Epidemiology Research and Information Center, VA Boston Healthcare System

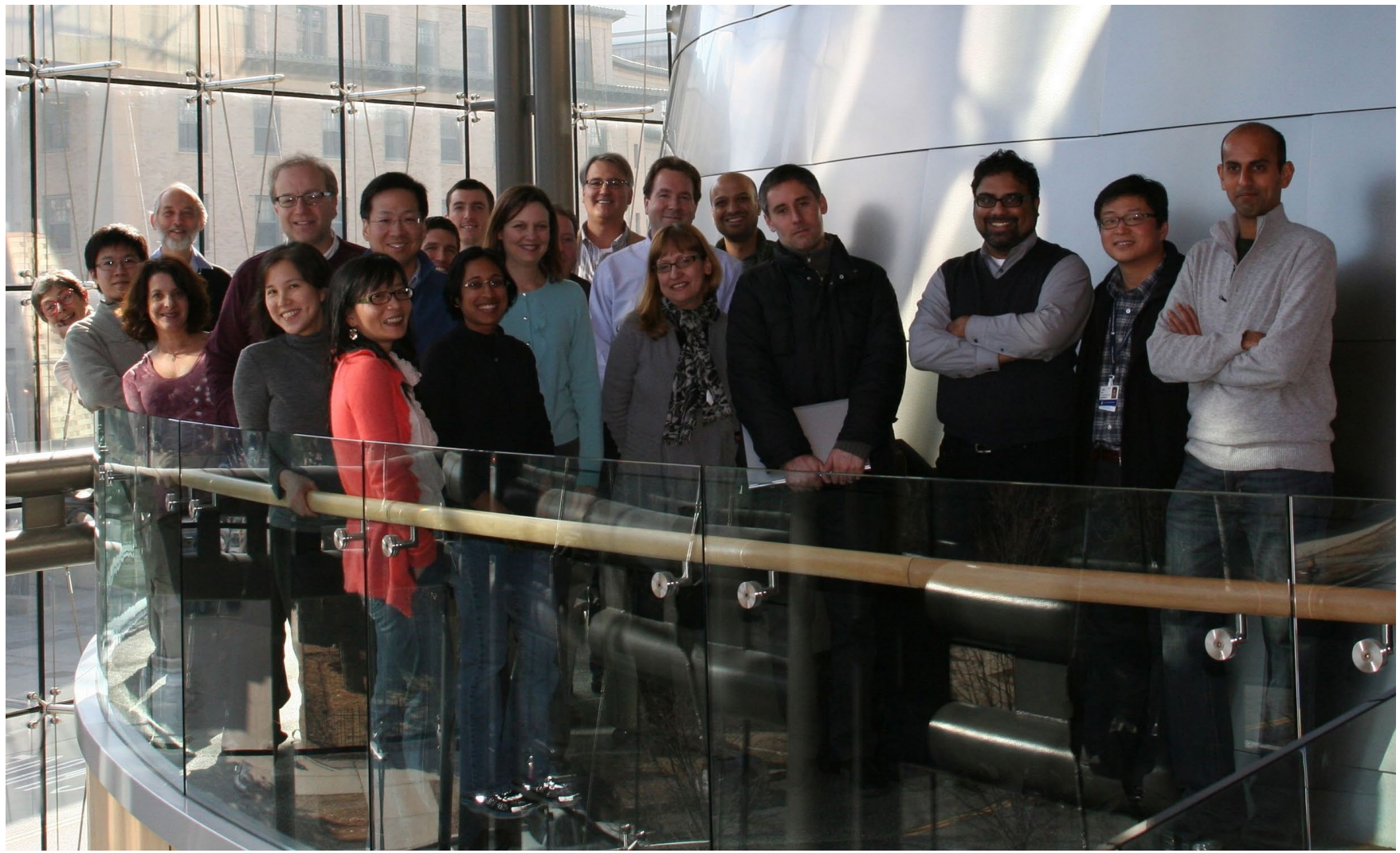
Outline

- Rationale for development of phenotyping approaches using EHR
- Brief background of ML for phenotyping
  - Supervised vs unsupervised
- Unsupervised approaches for phenotyping w/ EHR data
  - Strengths and limitations

i2b2

Informatics for Integrating Biology & the Bedside

A National Center for Biomedical Computing



Who has rheumatoid arthritis (RA) in the EHR?

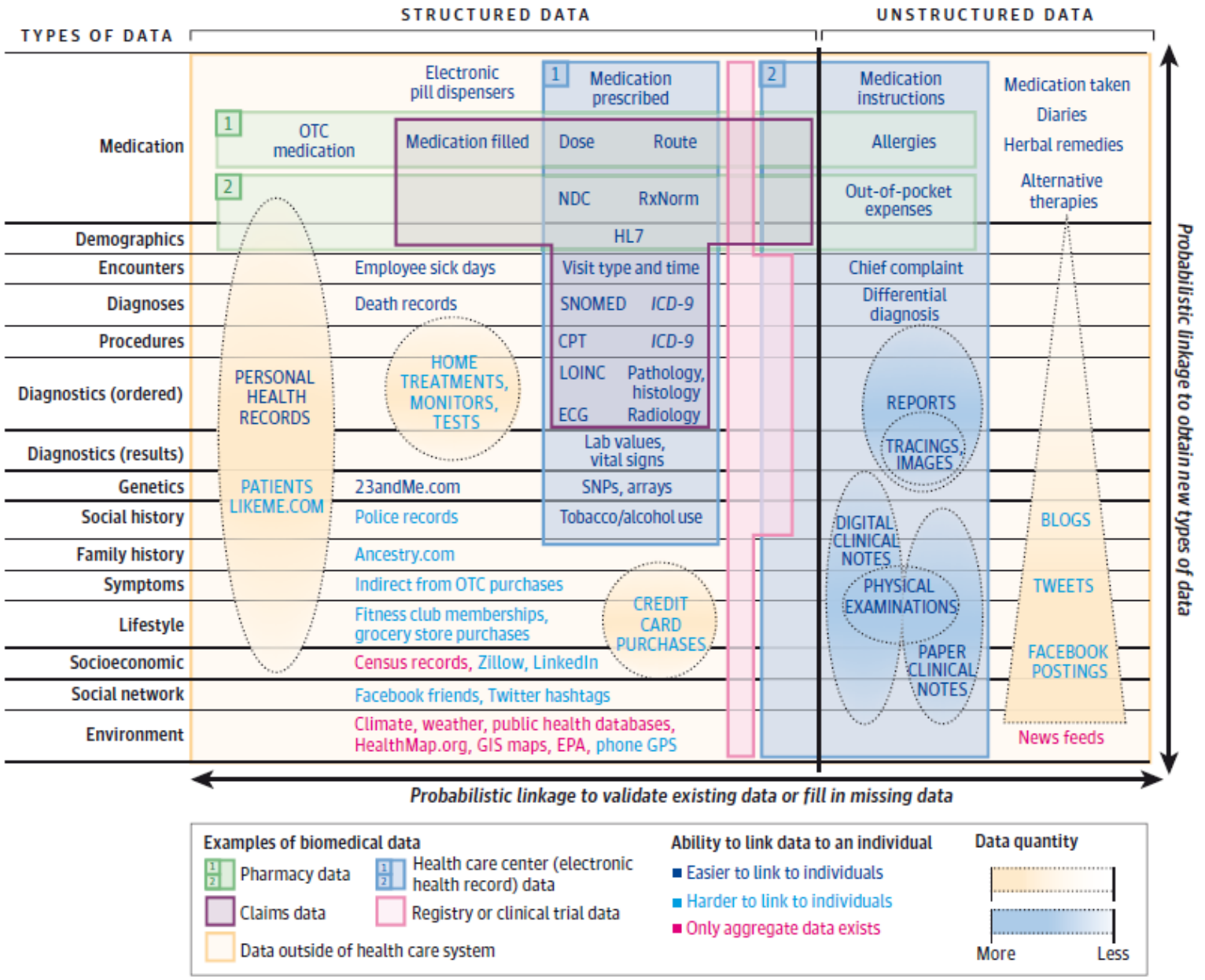
Table 4. Comparison of performance characteristics from validation of the complete classification algorithm (narrative and codified) with algorithms containing codified-only and narrative-only data*

Model	RA by algorithm or criteria, no.	PPV (95% CI), %	Sensitivity (95% CI), %	Difference in PPV (95% CI), %†
Algorithms				
Narrative and codified (complete)	3,585	94 (91–96)	63 (51–75)	Reference
Codified only	3,046	88 (84–92)	51 (42–60)	6 (2–9)‡
NLP only	3,341	89 (86–93)	56 (46–66)	5 (1–8)‡
Published administrative codified criteria				
≥3 ICD-9 RA codes	7,960	56 (47–64)	80 (72–88)	38 (29–47)‡
≥1 ICD-9 RA codes plus ≥1 DMARD	7,799	45 (37–53)	66 (57–76)	49 (40–57)‡

* The complete classification algorithm was also compared with criteria for RA used in published administrative database studies. RA = rheumatoid arthritis; PPV = positive predictive value; 95% CI = 95% confidence interval; NLP = natural language processing; ICD-9 = International Classification of Diseases, Ninth Revision; DMARD = disease-modifying antirheumatic drug.
† Difference in PPV = PPV of complete algorithm − comparison algorithm or criteria.
‡ Significant difference in PPV compared with the complete algorithm.

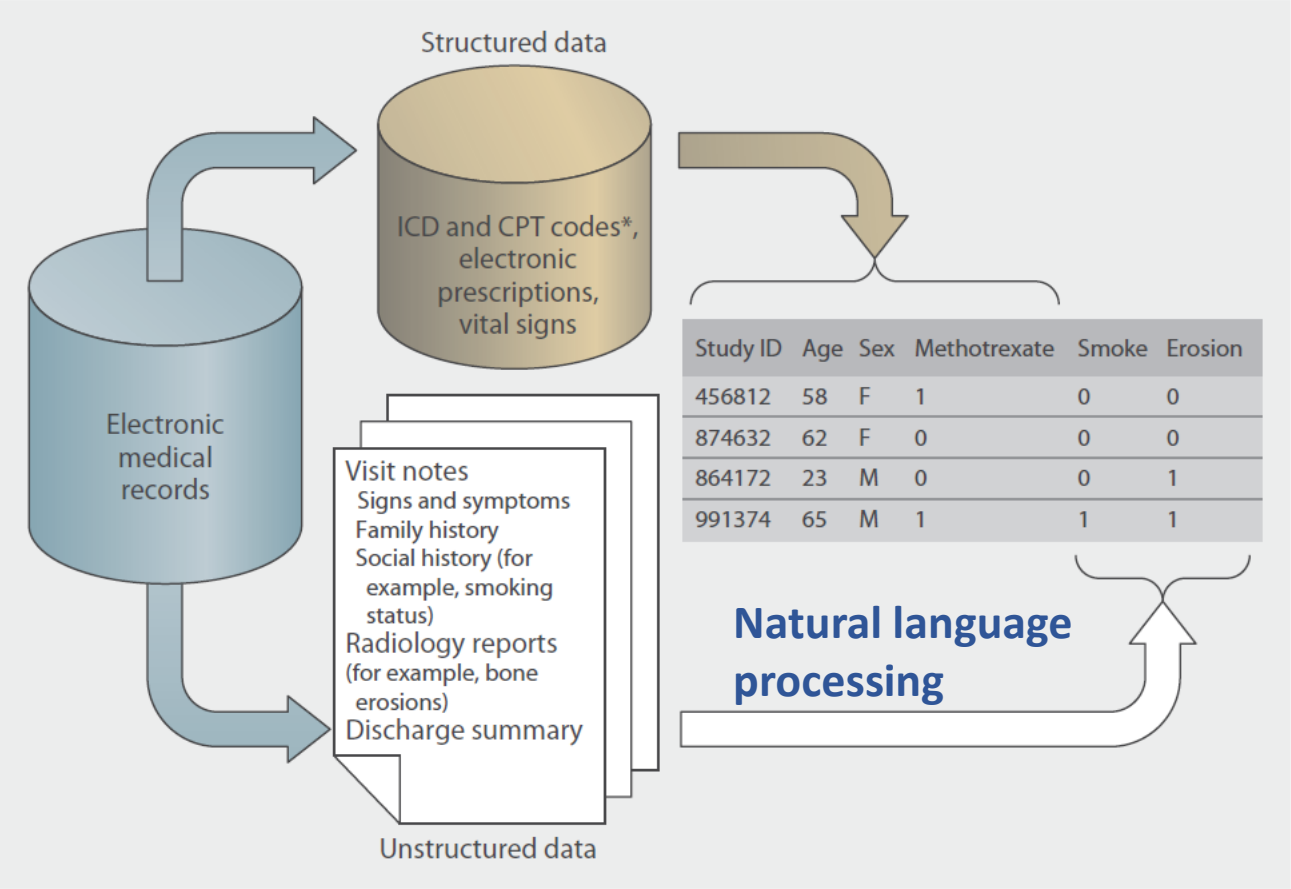
Liao, et al., Arth Care Res 2010

Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care



Weber et al, JAMA 2014

Types of EMR data



Liao, Cai, et al., BMJ 2015

Natural language processing (NLP)

Computational method for text processing based on the rules of linguistics

NLP

I	saw	the	girl	with	the	ophthalmoscope.
w1	w2	w3	w4	w5	w6	w7
pronoun	verb	article	noun	prep	article	noun

Slide adapted from Drs. Guergana Savova and Wendy Chapman

NLP ≠ "find" command in Word

- Negation
  - The patient has no erosions in the MCPs.
- Inverted syntax
  - Colon, ascending and descending, biopsy
- Relation
  - Tamoxifen is used in the treatment of breast cancer
- Morphologic variations
  - Tobacco, 30 pack years, past smoker, +tob → smoking

Illustrative dataset

ID	Age	Sex	Dx code	Lab	Dis+
9	22	M	0	-	0
10	45	F	1	31	1
11	75	F	1	40	1
12	67	M	0	-	0
13	56	M	0	56	1
14	54	F	0	11	0
15	81	F	1	42	1
16	48	F	0	5	0

Training set

Pattern recognition

ID	Age	Sex	Dx code	Lab	Dis+
9	22	M	0	-	0
10	45	F	1	31	1
11	75	F	1	40	1
12	67	M	0	-	0
13	56	M	0	56	1
14	54	F	0	11	0
15	81	F	1	42	1
16	48	F	0	5	0

+200 subjects

Training set

+1000 features

Pattern recognition

- More potential "features" *may* enable more accurate algorithms
  - Features can also add noise
- Challenge to identify the important features and their patterns

+200 subjects

ID	Age	Sex	Dx code	Lab	Dis+
9	22	M	0	-	0
10	45	F	1	31	1
11	75	F	1	40	1
12	67	M	0	-	0
13	56	M	0	56	1
14	54	F	0	11	0
15	81	F	1	42	1
16	48	F	0	5	0

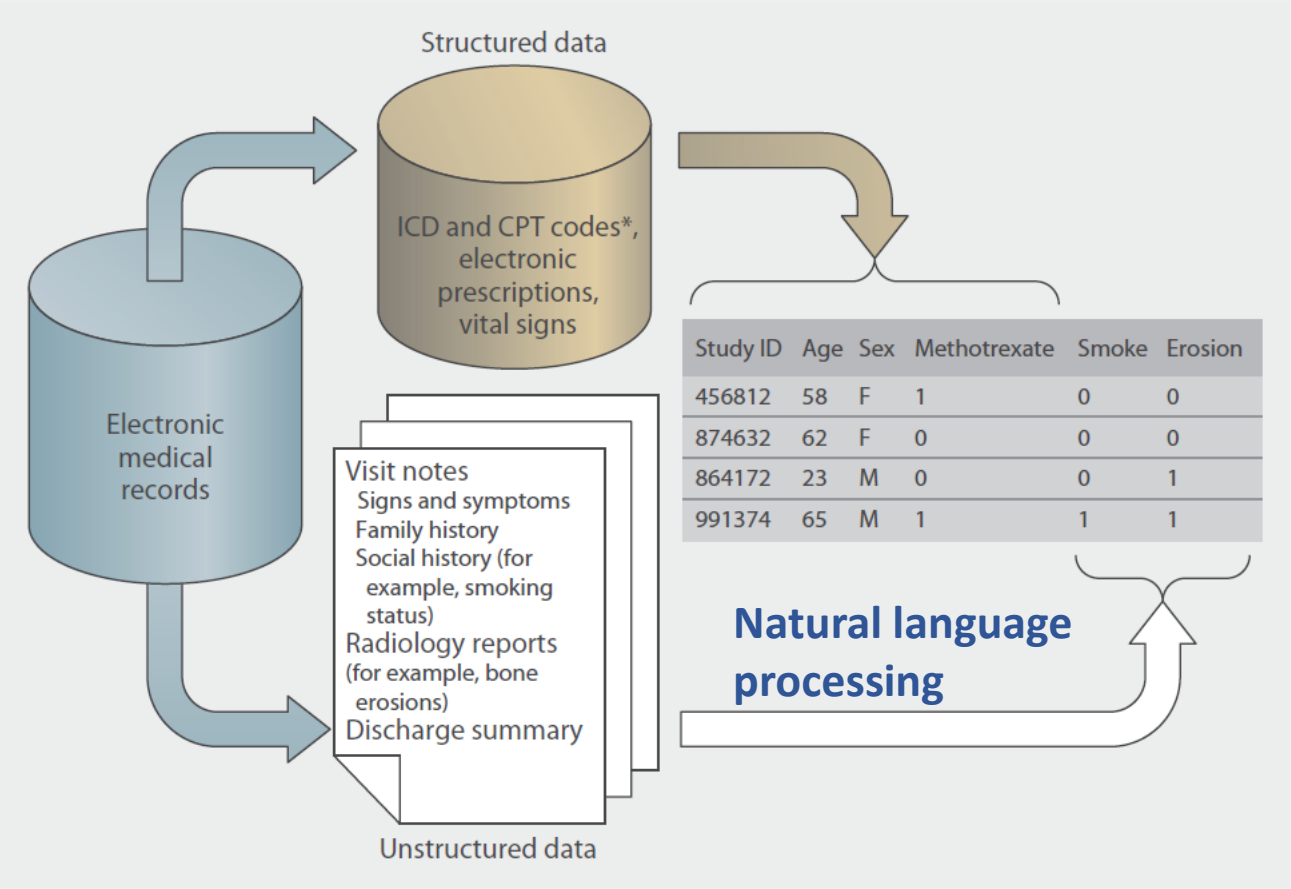
Training set

+1000 variables

Artificial Intelligence & Machine Learning

- Artificial intelligence (AI)
  - Intelligence demonstrated by machines
    - Contrast to human intelligence
- Machine learning (ML) → subset of AI
  - Requires training set
  - Focus on prediction (vs causality)
    - Does not address why or how to change outcomes
  - Learning structure from data
    - Pattern recognition
  - Examples
    - Least absolute shrinkage and selection operator (LASSO) regression
    - Support vector machine (SVM)

Types of EHR data



Liao, Cai, et al., BMJ 2015

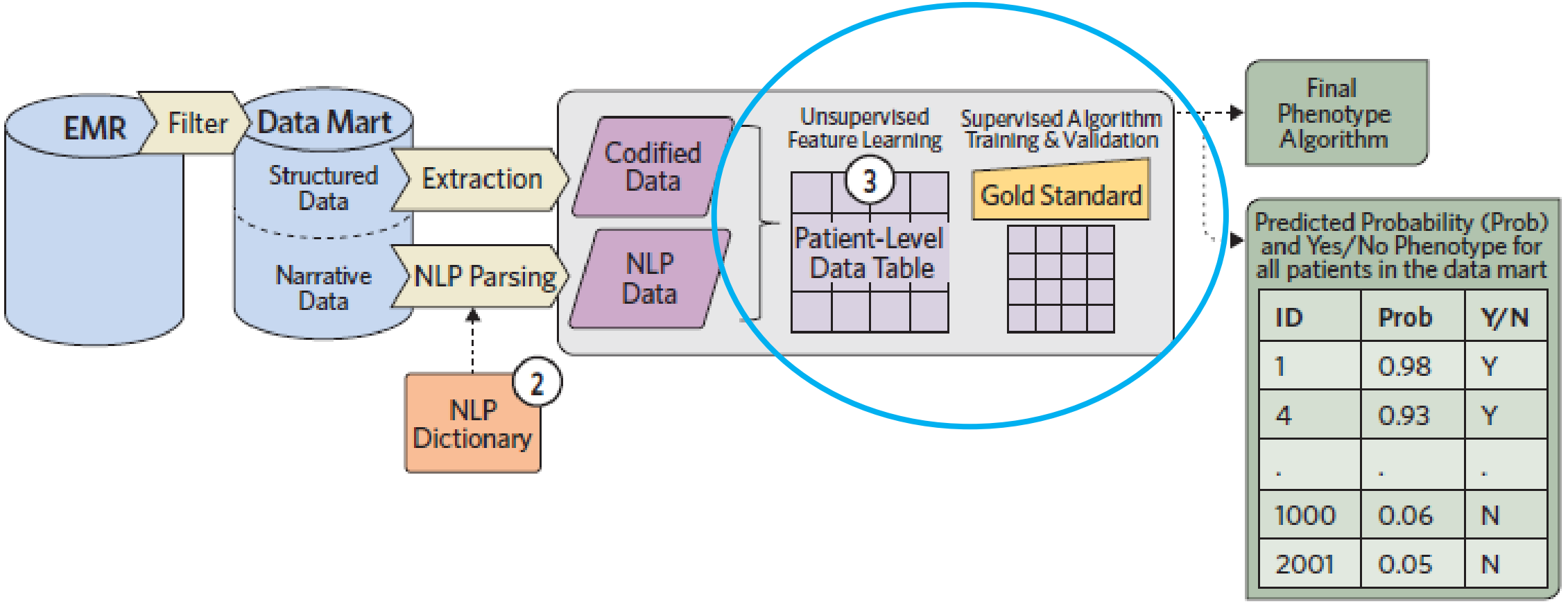
Approach to developing phenotype algorithms using EHR data

- Chart review- not feasible

- Rule-based
  - Relies on human expertise to identify important features
  - Algorithm is a combination of AND, NOT, OR

- Machine learning
  - Data driven method to select features and develop algorithm

Machine learning, NLP, and EHR
Pipeline for phenotyping



Zhang, Cai, Yu, et al., Nat Protocols 2019

Limitations of supervised ML approaches for phenotyping

- Require gold standard labels through manual chart review
  - Notes not always available
  - Time and resource intensive
  - Not scalable

- Inefficient
  - Large amount of unlabeled data contains ”noisy labels”

Comparison of EHR phenotype algorithm approaches

Characteristics	Supervised or semi-supervised	Unsupervised
Manual chart review for labels	Y	N
Feature selection	Manual or automated	Automated
Rule-based, e.g. 2 ICD + 1 Rx	Option	N
Machine learning	Option	Y
Efficiency	Varies	High
Accuracy	Data available	Needs validation

Unsupervised approaches for phenotyping w/ EHR data

Unsupervised approaches

- Anchor, Halpern et al., 2014
- XPRESS, Agarwal et al., 2016
- APHRODITE, Banda et al., 2017
- **PheNorm, Yu et al…Cai, 2017**
- MAP, Liao, Sun et al…Cai, 2019

Journal of the American Medical Informatics Association, 0(0), 2017, 1–7
doi: 10.1093/jamia/ocx111
Research and Applications

Research and Applications

Enabling phenotypic big data with PheNorm

Sheng Yu,[1,2] Yumeng Ma,[3] Jessica Gronsbell,[4] Tianrun Cai,[5] Ashwin N Ananthakrishnan,[6] Vivian S Gainer,[7] Susanne E Churchill,[8] Peter Szolovits,[9] Shawn N Murphy,[7,10] Isaac S Kohane,[8] Katherine P Liao,[11] and Tianxi Cai[4]

[1]Center for Statistical Science, Tsinghua University, Beijing, China, [2]Department of Industrial Engineering, Tsinghua University, Beijing, China, [3]Department of Mathematical Sciences, Tsinghua University, Beijing, China, [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [5]Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA, [6]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA, [7]Research Information Science and Computing, Partners HealthCare, Charlestown, MA, USA, [8]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, [9]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA, USA, [10]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA and [11]Department of Medicine, Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, MA, USA

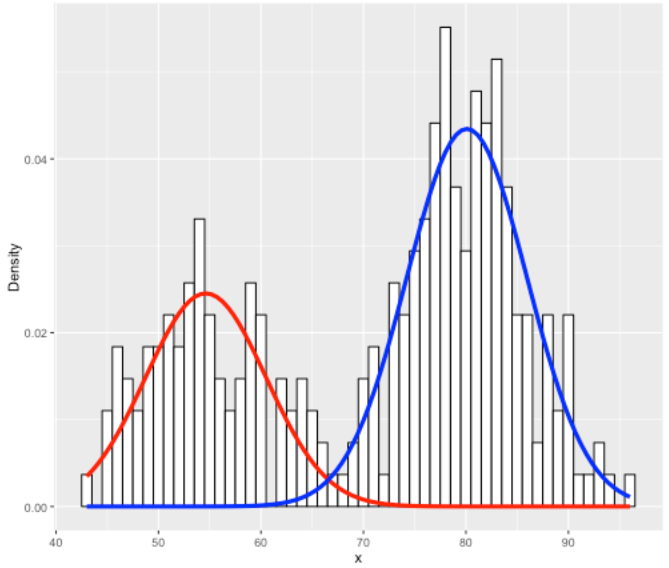
Corresponding Author: Sheng Yu, Center for Statistical Science, Tsinghua University, Weiqinglou Rm 209, Beijing, 100084, China. E-mail: syu@tsinghua.edu.cn. Tel: +86-10-62783842

Received 19 June 2017; Revised 5 August 2017; Editorial Decision 9 September 2017; Accepted 14 September 2017

PheNorm: Assumption

Surrogate disease labels S_i (i.e. ICD-9 codes) normalized by a patient's healthcare utilization U_i (i.e. count of patient notes) are log-normally distributed with mean μ_Y dependent on the patient's true disease status Y_i

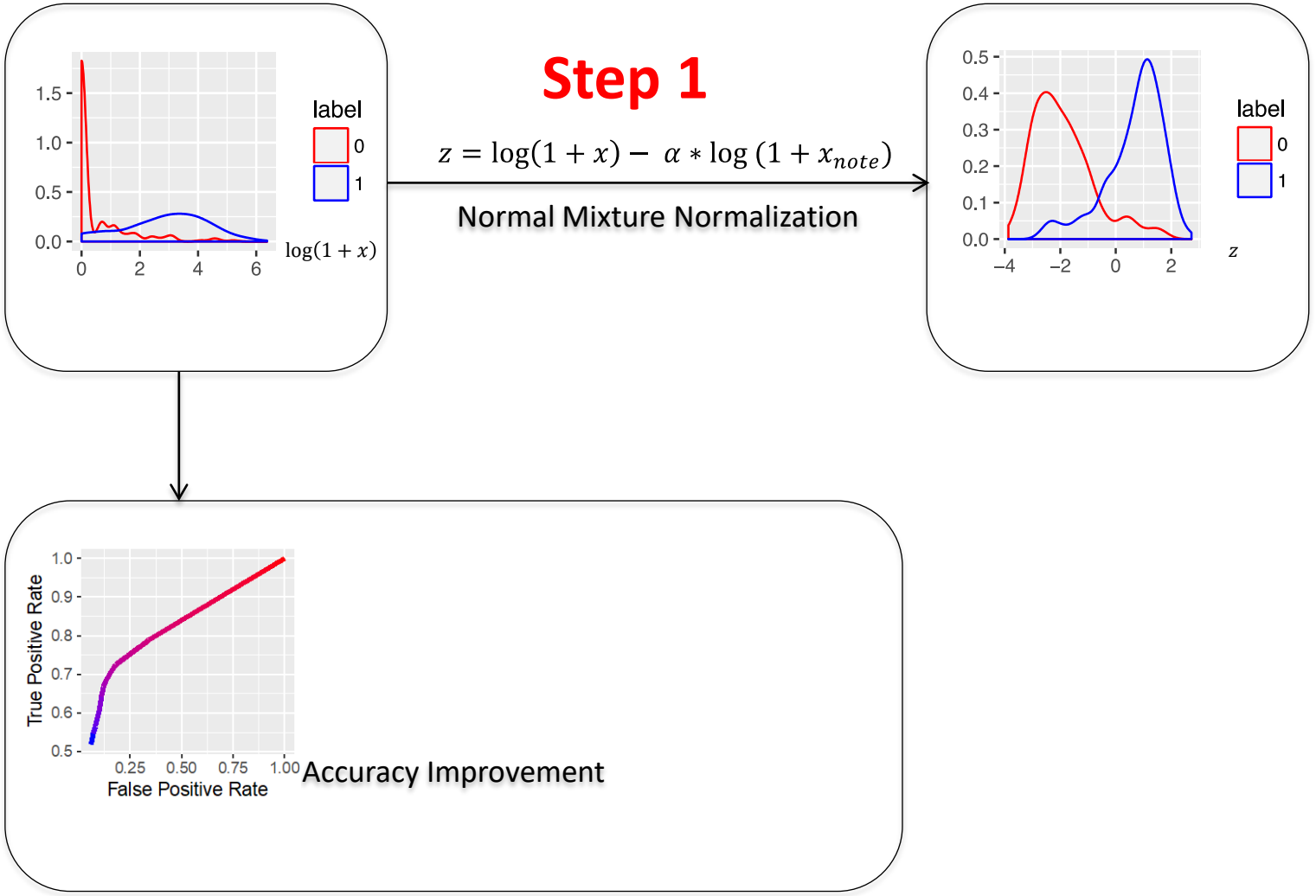
$$\log(S_i) \sim Norm(\mu_Y + c\log(U_i))$$



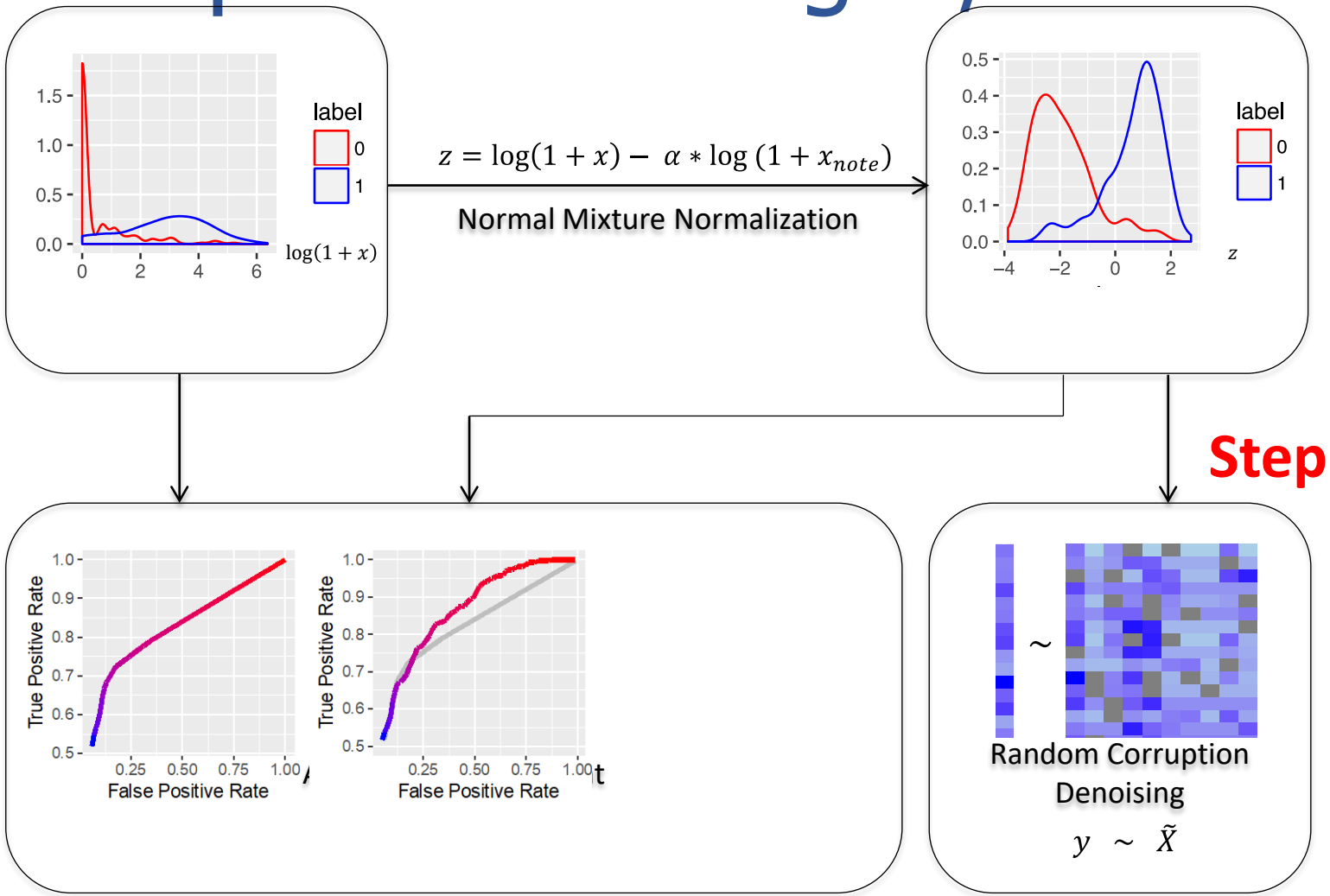
Abbreviations

- Main Features
  - $\boldsymbol{x_{ICD}}$: # ICD-9 codes of target phenotype for each patient
  - $\boldsymbol{x_{NLP}}$: # positive NLP mentions only, e.g. not negated, remove mention from family hx, of target phenotype from all notes for a given patient
  - $\boldsymbol{x_{ICDNLP}} = \boldsymbol{x_{ICD}} + \boldsymbol{x_{NLP}}$
- Healthcare utilization: $\boldsymbol{x_{note}}$ = # notes for each patient
- Additional potential features: $\boldsymbol{x_1} \ldots \boldsymbol{x_p}$
  - Counts of medication, mentions of signs and symptoms in the notes, etc
  - Can be curated through prior knowledge or via data-driven approaches

PheNorm Step 1: Normalization



PheNorm Step 2: Denoising w/ other features



PheNorm workflow

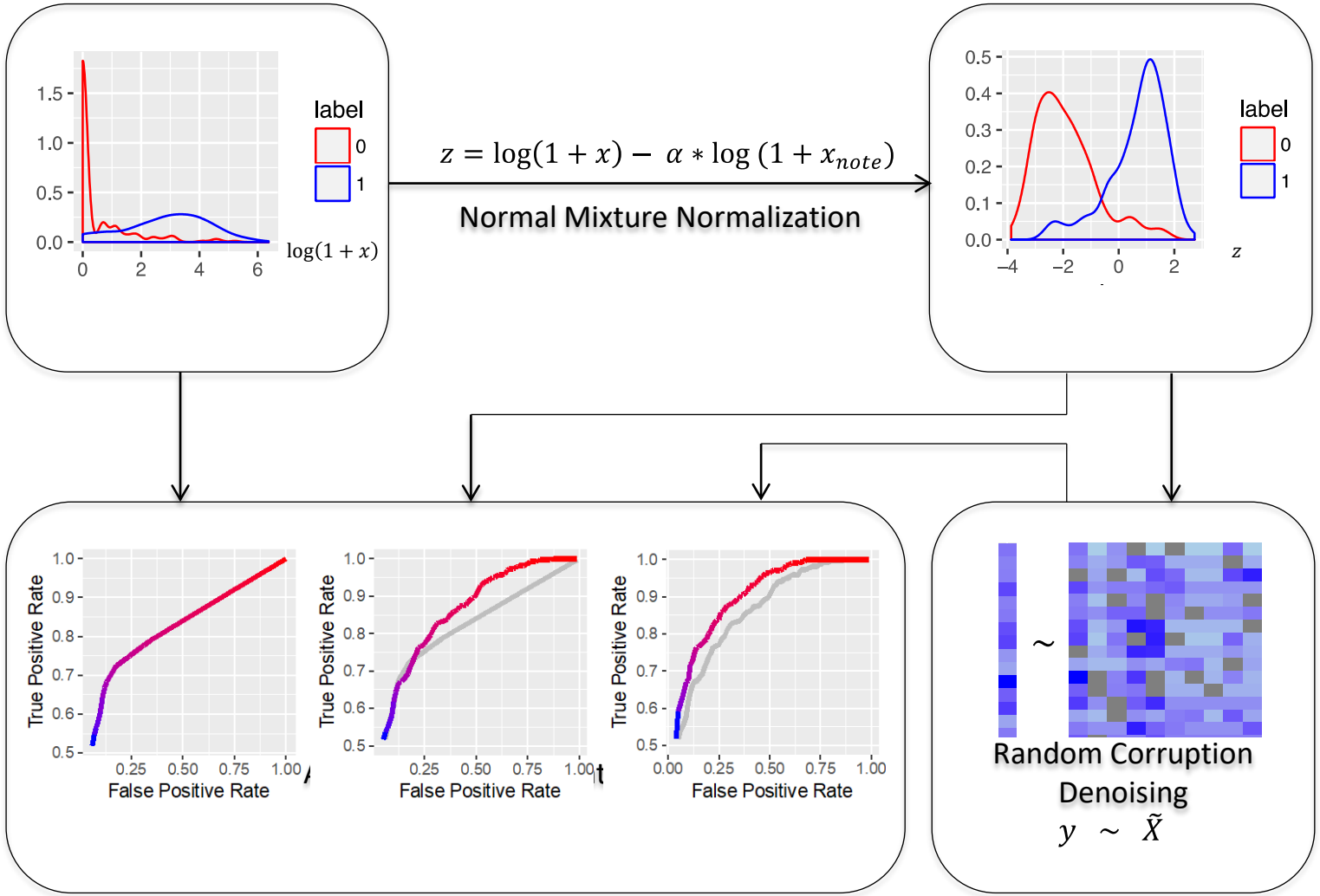


Table 1. AUCs of the raw feature x, the normalized feature z, the PheNorm scores using SAFE feature for denoising with a dropout rate of 0.3, $\text{PheNorm}_{\text{vote}}$, the supervised algorithms trained with SAFE features with $N = 100$, 200, or 300 labels, as well as the XPRESS and Anchor algorithms.

	CAD	RA	CD	UC
x_{ICD}	0.844	0.868	0.824	0.812
z_{ICD}	$0.875^{0.031*}_{0.010}$	$0.901^{0.033*}_{0.008}$	$0.877^{0.053*}_{0.013}$	$0.859^{0.047*}_{0.012}$
$PheNorm_{ICD}$	$0.899^{0.024*}_{0.004}$	$0.929^{0.028*}_{0.009}$	$0.911^{0.033*}_{0.005}$	$0.900^{0.041*}_{0.005}$
x_{NLP}	0.840	0.898	0.906	0.904
z_{NLP}	$0.864^{0.025*}_{0.011}$	$0.923^{0.025*}_{0.011}$	$0.947^{0.041*}_{0.007}$	$0.931^{0.026*}_{0.006}$
$PheNorm_{NLP}$	$0.884^{0.019*}_{0.003}$	$0.937^{0.014*}_{0.005}$	$0.948^{0.001}_{0.004}$	$0.935^{0.004*}_{0.002}$
x_{ICDNLP}	0.865	0.903	0.902	0.901
z_{ICDNLP}	$0.895^{0.030*}_{0.008}$	$0.935^{0.032*}_{0.009}$	$0.944^{0.042*}_{0.008}$	$0.933^{0.032*}_{0.007}$
$PheNorm_{ICDNLP}$	$0.899^{0.004*}_{0.002}$	$0.936^{0.001}_{0.002}$	$0.945^{0.001}_{0.002}$	$0.935^{0.002}_{0.002}$
$PheNorm_{vote}$	0.899	0.937	0.945	0.933

Comparison is with the previous step; asterisk indicates positive increment at the significance level of 0.05.

MAP: a refinement of PheNorm

- Limitations of PheNorm
  - Output linear score vs predicted probability of disease
  - Does not identify threshold value for classifying subjects as cases

- MAP (multi-modal automated phenotyping)
  - Fit a sequence of mixture models → predicted probabilities for all patients & estimates of disease prevalence from each fitting
  - Synthesize information via model averaging
  - Classifying as a case if predicted probabilities exceed threshold

Step 1: Assemble NLP & ICD data for each PheWAS group

- Mappings
  - ICD9 codes in a Phecode group → UMLS CUIs
  - ICD9 code → UMLS CUI
  - ICD9 string → UMLS CUI
  - PheWAS string → UMLS CUI

UMLS= Unified Medical Language System
CUI= concept unique identifier

phenotype group: rheumatoid arthritis

Code String	Code	ICD_9	ICD9_Str	CUI ICD9	CUI ICD9_String	CUI Code_String
Rheumatoid arthritis	714.1	714.0	rheumatoid arthritis	C0003873	C0003873	C0003873
		714.1	Felty's syndrome	C0015773	C0015773	C0003873
		714.2	Other rheumatoid arthritis with visceral or systemic involvement	C0157914	C0157914	C0003873
Rheumatoid arthritis and other inflammatory polyarthropathies	**714**	714.4	Chronic postrheumatic arthropathy	C0152084	C0152084	C0157913
		714.8	Other specified inflammatory polyarthropathies	C0157919	C0157919	C0157913
		714.89	Other specified inflammatory polyarthropathies	C0157919	C0157919	C0157913
		714	Rheumatoid arthritis and other inflammatory polyarthropathies	C0157913	C0157913	C0157913

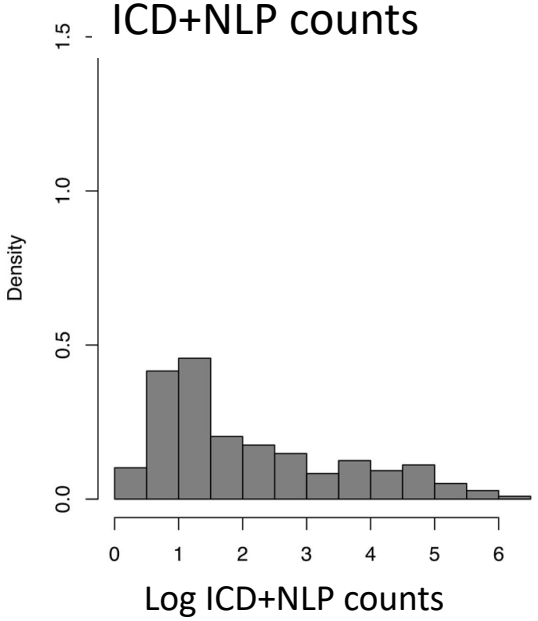
ICD9 Counts: ICD_RA

NLP Counts: NLP_RA

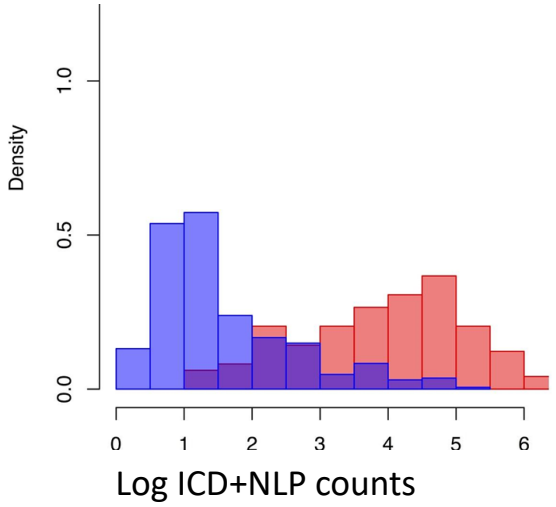
Step 2: Joint Analysis of NLP & ICD

- Fit multiple Poisson and log-normal mixture models to {NLP,ICD} counts → probabilities of phenotype(+)
- Adjust for healthcare utilization

Distribution of log ICD+NLP counts



Distribution of log ICD+NLP counts by RA status

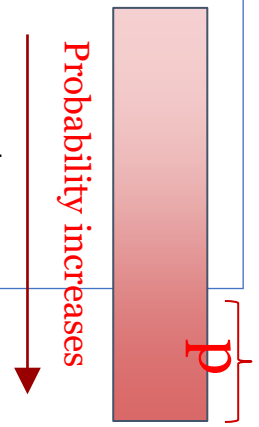


Step 3: Synthesize information from all model fittings

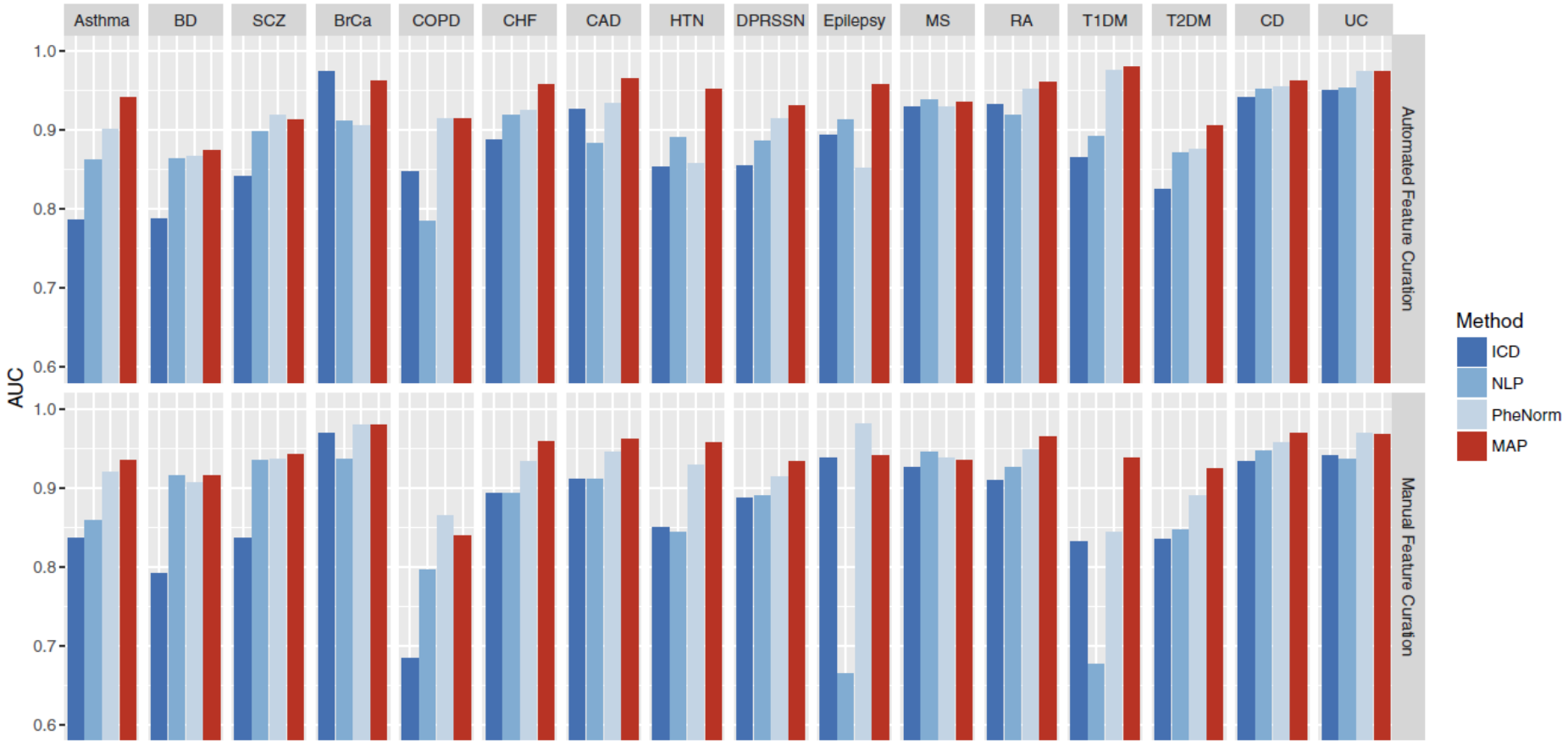
- Each fitted model provides a predicted probability of phenotype for each patient
- The final predicted probability of phenotype(+) is the average predicted probabilities from all fitted models

Step 4: Cut-off estimate based on population prevalence p

- Fitted mixture models → estimated phenotype prevalence
- Classify p% patients with highest predicted probabilities as phenotype(+) (as opposed to the standard method based on ICD code thresholding)

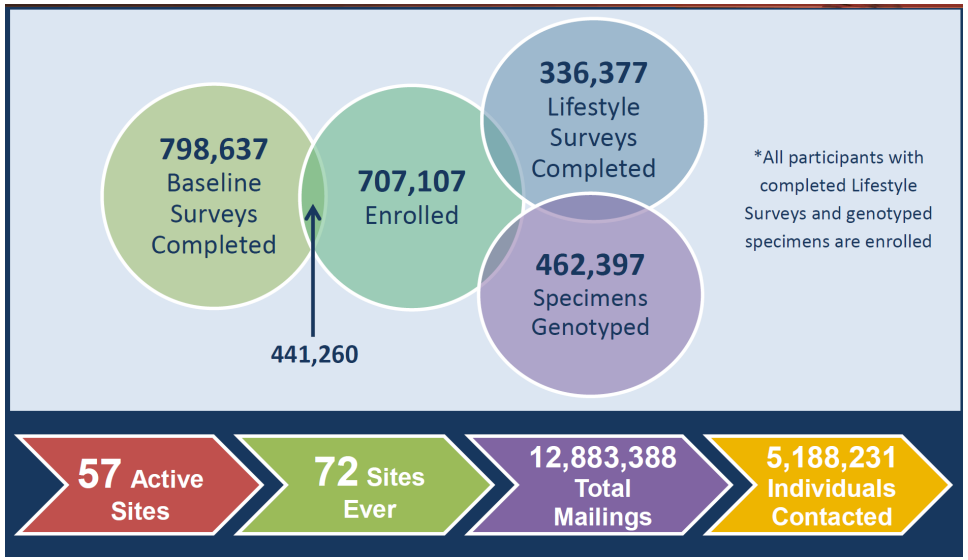


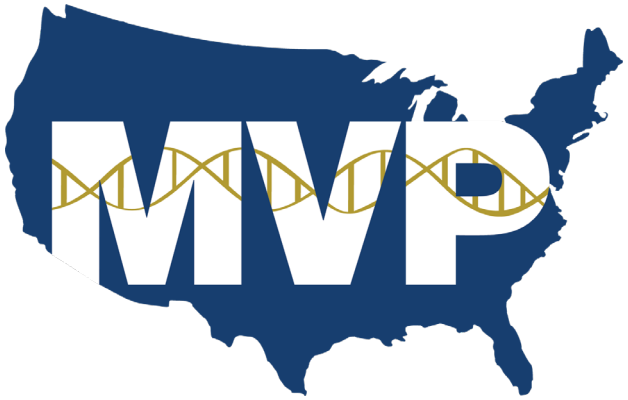
Performance of phenotype algorithms across conditions



Applications: Phenomics Library

- Veterans Affairs Health Centers
  - ~22 million veterans nationwide
    - Million Veteran Program (MVP)
  - Ported and validated supervised and unsupervised approaches





EHR research platform for translational studies

VA EHR data

Summary

- Phenotyping approaches designed for prevalent conditions
- Optimized for EHR data
- Robust and portable
- Supervised vs unsupervised based on downstream use
  - Cohort creation
  - Phenotype screens, e.g. PheWAS
  - Association studies

- Future directions
  - Algorithms for incident or recurrent conditions
  - Can existing algorithms catch incident conditions within a time window?

Thank you

BWH

Selena Huang
Brittany Weber
Austin Cai
Zeling He
Nicholas Link (VA)
Kumar Dahal
Dana Weisenfeld
Charlotte Golnik
Thany Seyok
Andrew Cagan
Jackie Stratton

DBMI, Harvard Medical School

Tianxi Cai
Chuan Hong
Hajime Uno (DFCI)
Junwei Liu (HSPH)
Amanda King
Isaac Kohane
Susanne Churchill

Boston VA Healthcare System/MAVERIC

J. Michael Gaziano
Christopher O'Donnell
Kelly Cho

Paul Monach
Anne Ho
Jennifer Huffman
Lauren Costa
Petra Schubert
Laura Tarko
Ashley Galloway

National Heart Lung and Blood Institute

MVP

NIAMS