Machine Learning for Healthcare

6.871x

Learning with noisy labels

David Sontag

Labels may be noisy

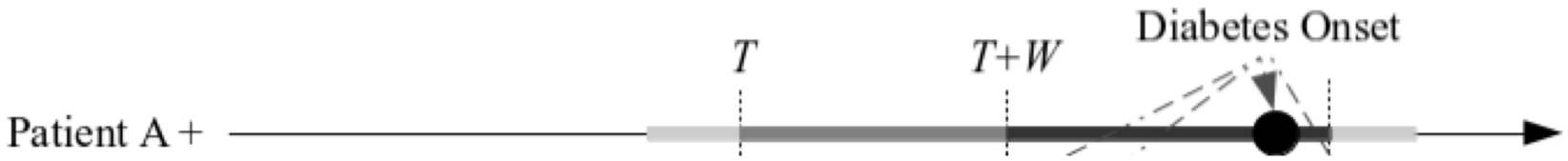
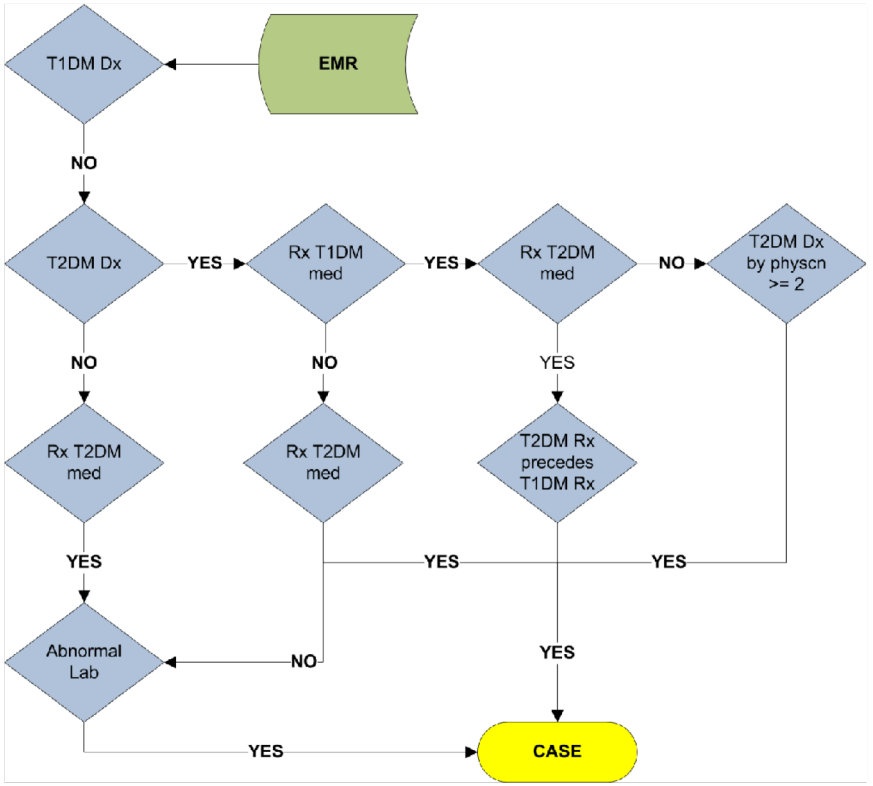


Figure 1: Algorithm for identifying T2DM cases in the EMR.

If the derived label is noisy, how does it affect learning?



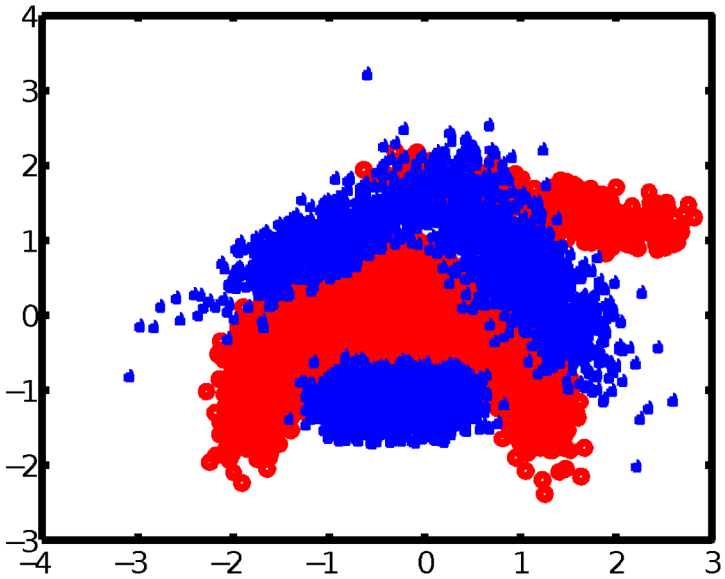
Source: https://phekb.org/sites/phenotype/files/T2DM-algorithm.pdf

Labels may be noisy

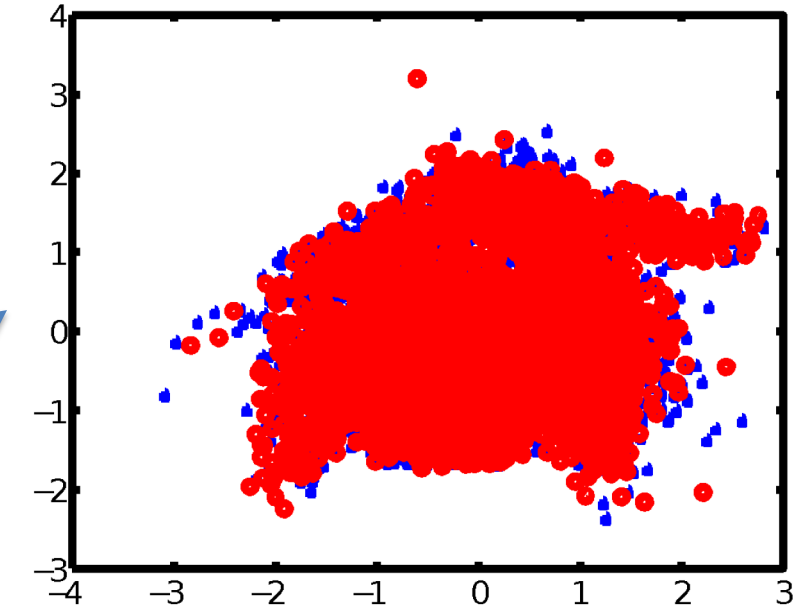
Fibrosis

red = mislabeled
orange = maybe mislabeled

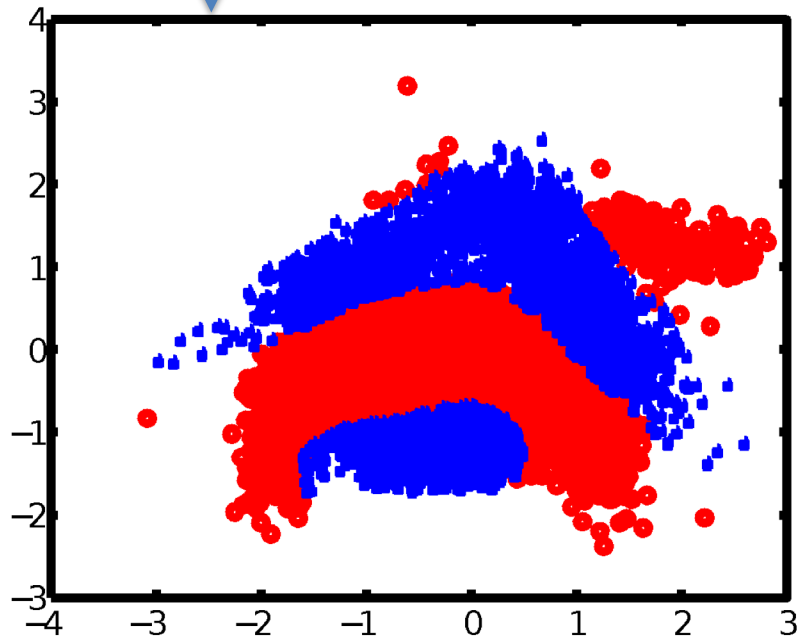
[Wang et al., “Chest X-ray8”]
figure credit: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/



40% label noise



Machine learning



[Natarajan et al., NeurIPS ’13. Figure 2]

Learning with noisy labels

We will show that if we have

a) ***class-conditional*** label noise and

b) ***lots*** of training data,

learning as usual, substituting noisy labels, works!

This opens the door to using noisy labels for training, and coming up with clever ways of deriving these for free

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Natarajan et al: Introduction

- Features X
- True unobserved labels $Y \in \{-1, 1\}$
- Noisy observed labels $\tilde{Y} \in \{-1, 1\}$
- True distribution $P(X, Y, \tilde{Y})$

X (age)	Y (diabetic)	$\tilde{Y}$ (noisy version)
30	-1	-1
64	1	1
75	1	-1

- Data sampled from $P(X, \tilde{Y}) = \sum_y P(X, Y = y, \tilde{Y})$

Y *exists*, but it is hidden during training

X (age)	$\tilde{Y}$ (noisy version)
30	-1
64	1
75	-1

Assumption: class-conditional label noise

- Assume that $\tilde{Y} \perp X|Y$:

$$P(X, Y, \tilde{Y}) = P(X, Y)P(\tilde{Y}|Y)$$

$\tilde{Y}$ only depends on Y: label noise is independent of input features

- Since Y is binary, need two parameters to fully define $P(\tilde{Y}|Y)$: $\rho_+ = P(\tilde{Y} = -1|Y = 1)$ & $\rho_- = P(\tilde{Y} = 1|Y = -1)$

- Assume that $\rho_+ + \rho- < 1$ and that ρ_+, ρ_- are known

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Learning an unbiased estimator

- If we could learn $\eta(X) = P(Y|X)$, then we would be able to predict optimally.

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Learning an unbiased estimator

- If we could learn $\eta(X) = P(Y|X)$, then we would be able to predict optimally.

$$\tilde{\eta}(X) = P(\tilde{Y} = 1|X)$$

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Learning an unbiased estimator

- If we could learn $\eta(X) = P(Y|X)$, then we would be able to predict optimally.

$$\begin{aligned}\tilde{\eta}(X) &= P(\tilde{Y} = 1|X) \\ &= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X)\end{aligned}$$

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Learning an unbiased estimator

- If we could learn $\eta(X) = P(Y|X)$, then we would be able to predict optimally.

$$\begin{aligned}\tilde{\eta}(X) &= P(\tilde{Y} = 1|X) \\ &= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X) \\ &= P(Y = 1|X)P(\tilde{Y} = 1|Y = 1) \\ &\quad + P(Y = -1|X)P(\tilde{Y} = 1|Y = -1)\end{aligned}$$

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Learning an unbiased estimator

- If we could learn $\eta(X) = P(Y|X)$, then we would be able to predict optimally.

$$
\begin{aligned}
\tilde{\eta}(X) &= P(\tilde{Y} = 1|X) \\
&= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X) \\
&= P(Y = 1|X)P(\tilde{Y} = 1|Y = 1) \\
&\quad + P(Y = -1|X)P(\tilde{Y} = 1|Y = -1) \\
&= \eta(X)(1 - \rho_+) + (1 - \eta(X))\rho_-
\end{aligned}
$$

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Learning an unbiased estimator

- If we could learn $\eta(X) = P(Y|X)$, then we would be able to predict optimally.

$$
\begin{aligned}
\tilde{\eta}(X) &= P(\tilde{Y} = 1|X) \\
&= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X) \\
&= P(Y = 1|X)P(\tilde{Y} = 1|Y = 1) \\
&\quad + P(Y = -1|X)P(\tilde{Y} = 1|Y = -1) \\
&= \eta(X)(1 - \rho_+) + (1 - \eta(X))\rho_- \\
&= \eta(X)(1 - \rho_+ - \rho_-) + \rho_-
\end{aligned}
$$

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Learning an unbiased estimator

$$\rightarrow \eta(X) = \frac{\tilde{\eta}(X) - \rho_-}{1 - \rho_+ - \rho_-}$$

- Learn $\tilde{\eta}(X)$ using any ML algorithm which returns calibrated classifiers. Substitute $\tilde{\eta}(X)$ in the above equation to get an unbiased estimate of $\eta(X)$!

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Application to electronic phenotyping

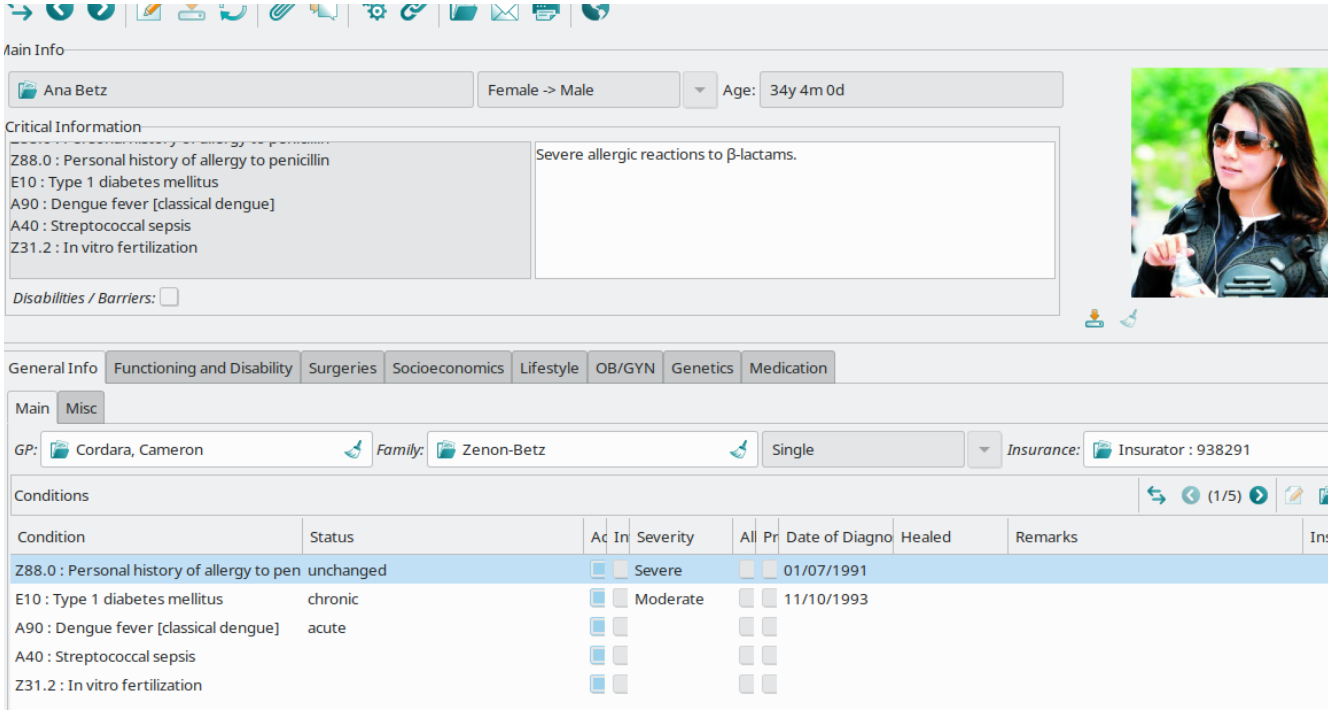


Figure source:
https://commons.wikimedia.org/wiki/File:GNU_Health_patient_main_screen.png

Hundreds of relevant clinical variables

Abdominal pain
Active malignancy
Altered mental status
Cardiac etiology
Renal failure
Infection
Urinary tract infection
Shock
Smoker
Pregnant
Lower back pain
Motor Vehicle accident
Psychosis
Anticoagulated
Type II diabetes
…

[Halpern, Horng, Choi, Sontag, AMIA ’14; Halpern, Horng, Choi, Sontag, JAMIA ‘16]

Simplest approach: rules

- We would like to estimate, for every patient, which phenotypes apply to them (at some point in time)
- Common practice is to derive manual rules:

Need to include:
nursing facility
nursing care
facility nursing / rehab
nsg facility
nsg faclty
…

text contains: *"nursing home"*

Nursing home? (physician response (gold standard))

	T	F	
T	297	129	PPV 0.70
F	1,319	34511	
	Sensitivity 0.18		

Slow, expensive, poor sensitivity.

Often we can find noisy labels WITHIN the data!

Phenotype	Example of noisy label ("anchor")
Diabetic (type I)	gsn:016313 (insulin) in **Medications**
Strep Throat	Positive strep test in **Lab results**
Nursing home	"from nursing home" in **Text**
Pneumonia	"pna" in **Text**
Heart attack	ICD10 I21 in **Billing codes**

How can we use these for machine learning?

Often we can find noisy labels WITHIN the data!

Phenotype	Example of noisy label (anchor)
Heart attack	ICD10 I21 in **Billing codes**

- Suppose we want to know, was a patient admitted to the emergency department for a heart attack?
- Billing codes not available at prediction time, but can be used for labels
- Reasonable to assume that $\rho_- = P(\tilde{Y} = 1 | Y = -1) \approx 0$, but because of noisy nature of billing codes, $\rho_+ = P(\tilde{Y} = -1 | Y = 1)$ likely non-zero

Called "positive only" noise since it implies $P(Y = 1 | \tilde{Y} = 1) = 1$

Anchor & Learn Algorithm

(special case for anchors derived from future data)

Training

1. Treat the anchors as "true" labels
2. Learn a classifier to predict whether the ***anchor*** $\tilde{Y}$ appears
3. Calibration step: divide by $\frac{1}{|P|}\sum_P P(\tilde{Y}=1|X)$

P = data points with $\tilde{Y}=1$

Test time

1. Predict using the learned classifier (with calibration)

Often we can find noisy labels WITHIN the data!

Phenotype	Example of noisy label (anchor)
Nursing home	"from nursing home" in **Text**

- We again assume that $\rho_- = P(\tilde{Y} = 1 | Y = -1) \approx 0$, but because many ways to write "from nursing home" in text, we have $\rho_+ = P(\tilde{Y} = -1 | Y = 1)$ likely non-zero
- If we simply learn to predict $\tilde{Y}$ using the notes, we will learn a trivial classifier! It will simply extract mentions of this phrase!
- This is a clear violation of the assumption $\tilde{Y} \perp X | Y$, since $\tilde{Y}$ is derived from X

Anchor & Learn Algorithm

Training

1. Treat the anchors as "true" labels
2. Learn a classifier to predict whether the ***anchor*** appears based on ***all other features***
3. Calibration step: divide by $\frac{1}{|P|}\sum_P P(\tilde{Y}=1|X)$

P = data points with $\tilde{Y}=1$

Test time

1. If the anchor is present: Predict 1
2. Else: Predict using the learned classifier (with calibration)

Evaluating phenotypes

- Derived anchors and learned phenotypes using 270,000 patients' emergency department medical records

History
Alcoholism
Anticoagulated
Asthma/COPD
Cancer
Congestive heart failure
Diabetes
HIV+
Immunosuppressed
Liver malfunction

Acute		
Abdominal pain	Deep vein thrombosis	Laceration
Allergic reaction	Employee exposure	Motor vehicle accident
Ankle fracture	Epistaxis	Pancreatitis
Back pain	Gastroenteritis	Pneumonia
Bicycle accident	Gastrointestinal bleed	Psych
Cardiac etiology	Geriatric fall	Obstruction
Cellulitis	Headache	Septic shock
Chest pain	Hematuria	Severe sepsis
Cholecystitis	Intracerebral hemorrhage	Sexual assault
Cerebrovascular accident	Infection	Suicidal ideation
	Kidney stone	Syncope
		Urinary tract infection

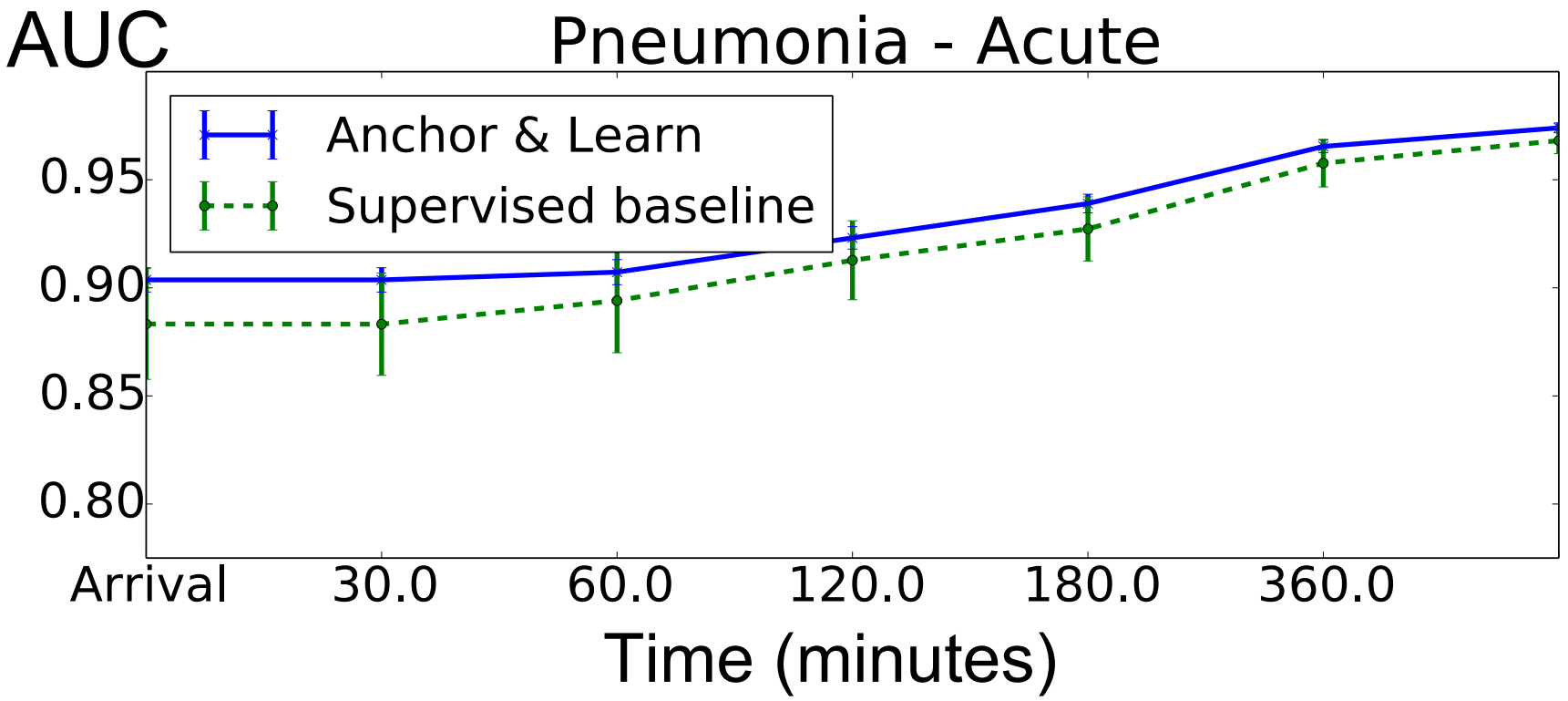
[Halpern, Horng, Choi, Sontag, AMIA '14]
[Halpern, Horng, Choi, Sontag, JAMIA '16]

Evaluating phenotypes

- Derived anchors and learned phenotypes using 270,000 patients' emergency department medical records
- To obtain ground truth, added a small number of questions to patient discharge procedure, rotated randomly

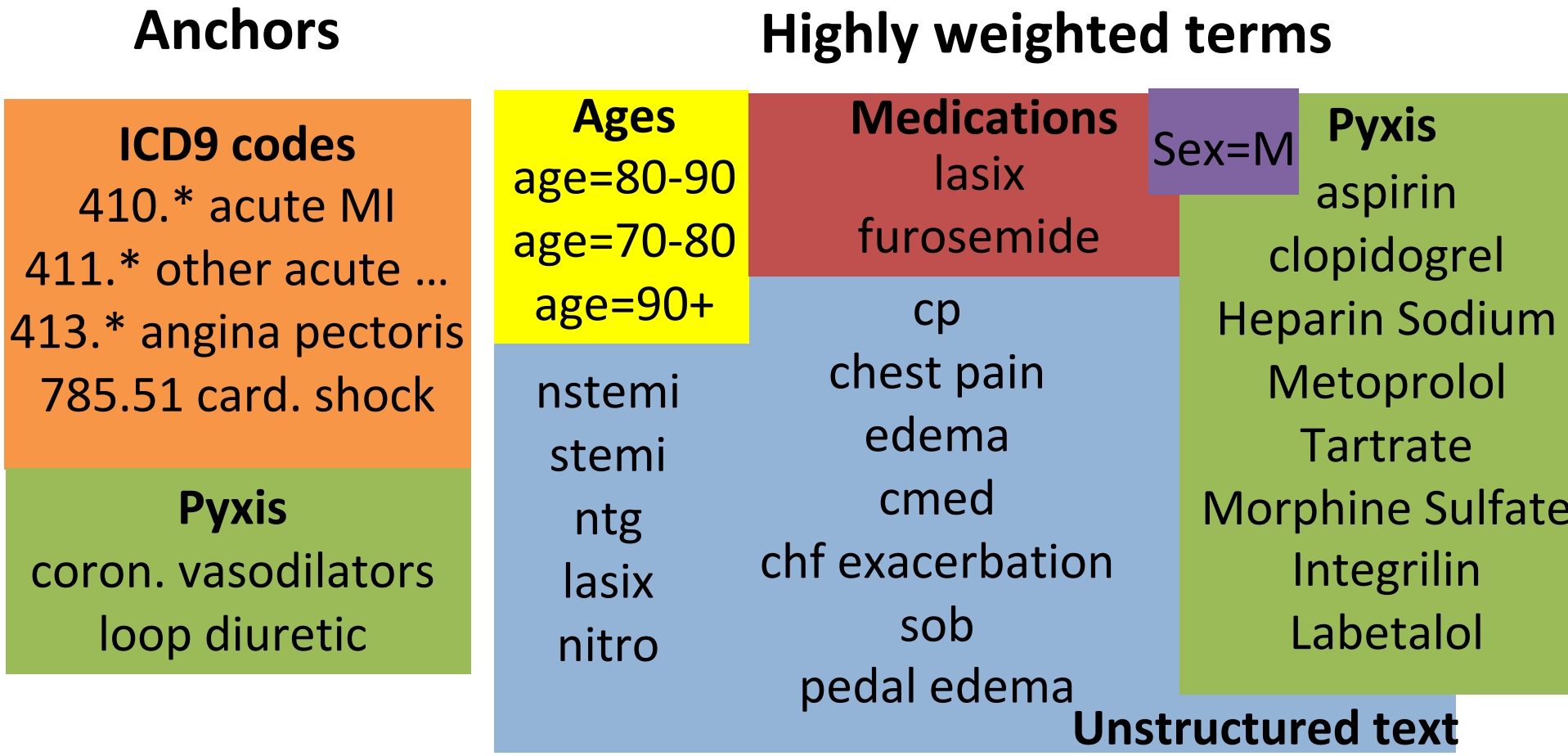
[Halpern, Horng, Choi, Sontag, AMIA '14]
[Halpern, Horng, Choi, Sontag, JAMIA '16]

Evaluating phenotypes



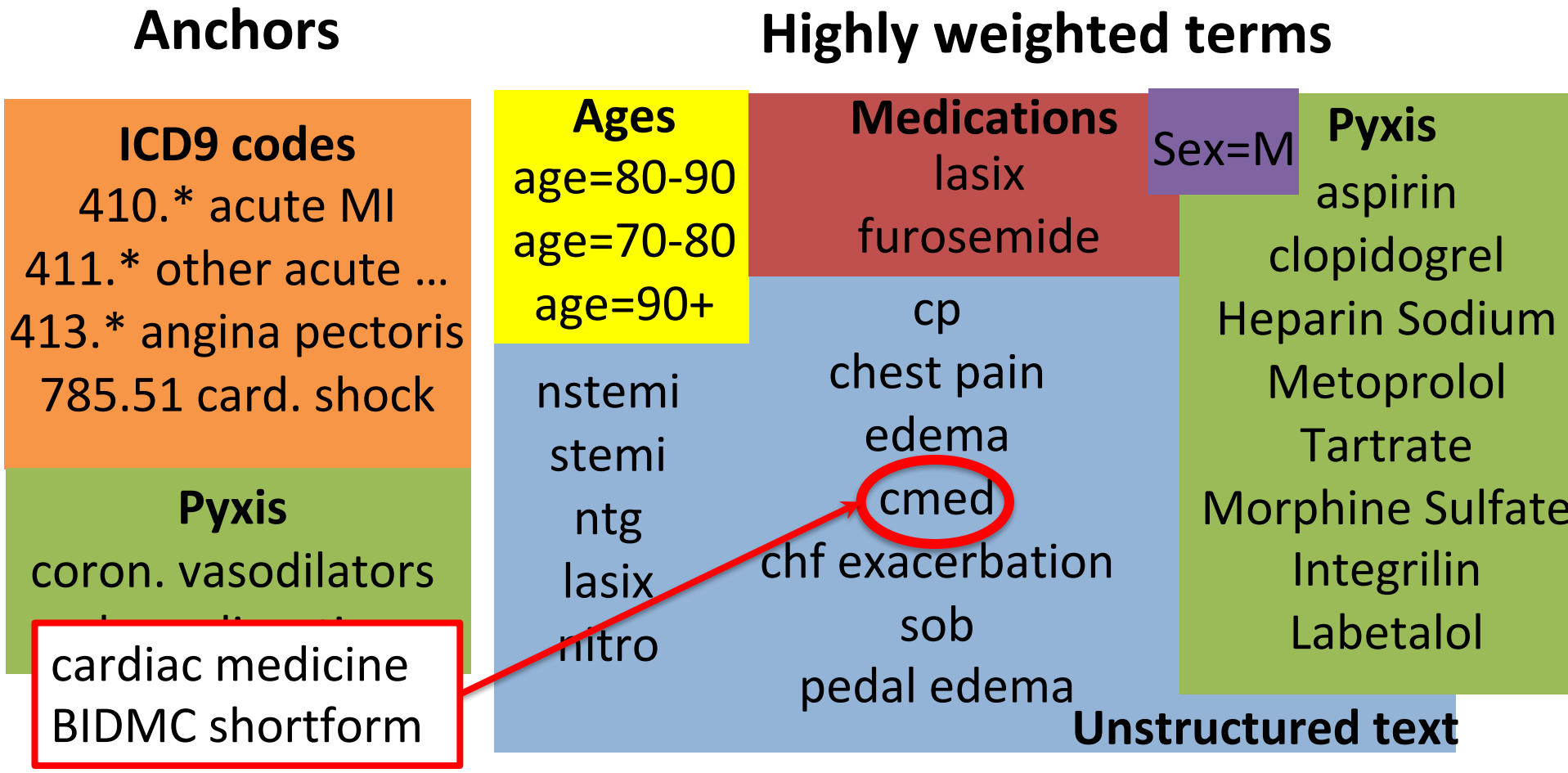
Comparison to supervised learning using labels for 5000 patients

Evaluating phenotypes – example model (cardiac etiology)



[Halpern, Horng, Choi, Sontag, AMIA '14]
[Halpern, Horng, Choi, Sontag, JAMIA '16]

Evaluating phenotypes – example model (cardiac etiology)



[Halpern, Horng, Choi, Sontag, AMIA '14]
[Halpern, Horng, Choi, Sontag, JAMIA '16]