Estimating Clinical State Variables without Labeled Data

Yoni Halpern

Joint work with:

Youngduck Choi
Steven Horng MD, MMSc
David Sontag, PhD

Disclosure

The authors have no relationships with commercial interests.

Emergency Department:

- **Limited resources**
- **Time sensitive**
- **Critical decisions**

Outline

- Use case: Real-time phenotype estimation
- Current Approaches
  - manual rules, machine learning
- Anchor-based learning
  - defining anchors
  - learning framework
  - interactive anchor specification
- Evaluation
- Conclusion and Next Steps

Real-time phenotype estimation

Cellulitis

– Specialized order sets, confirm followup care.

From nursing home

– Higher risk for acquired infections

Geriatric fall

– Alert transport staff for fall precautions

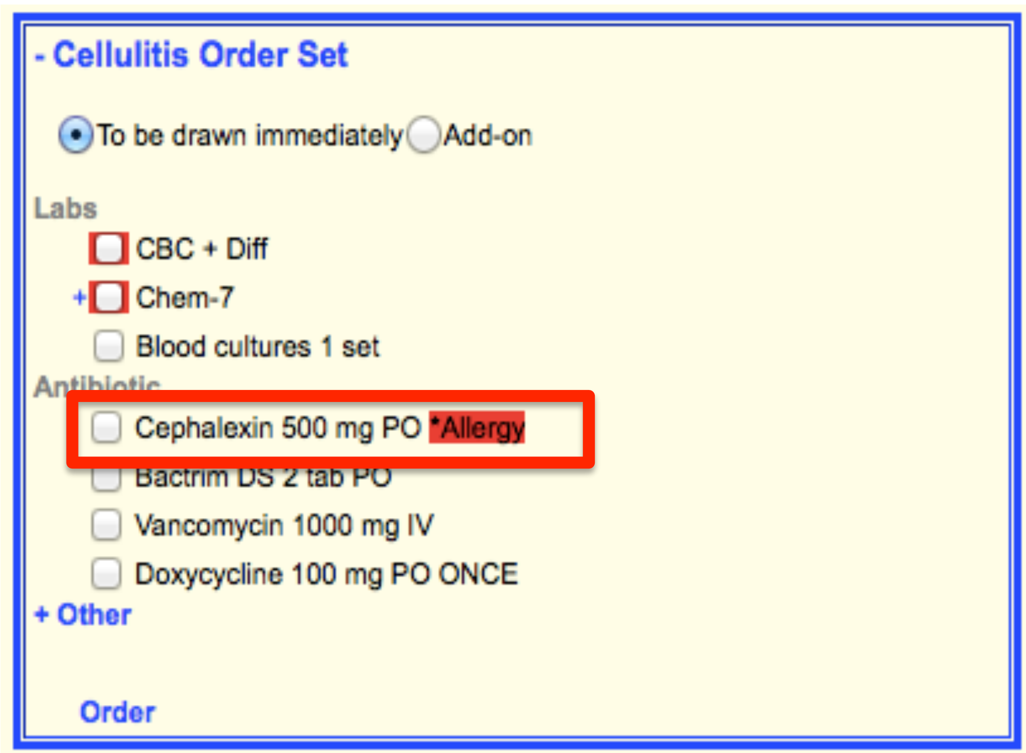
Many more

– GI bleed, syncope, DKA, etc.

Real-time phenotype estimation

Cellulitis

– Specialized order sets, confirm followup care.



ections

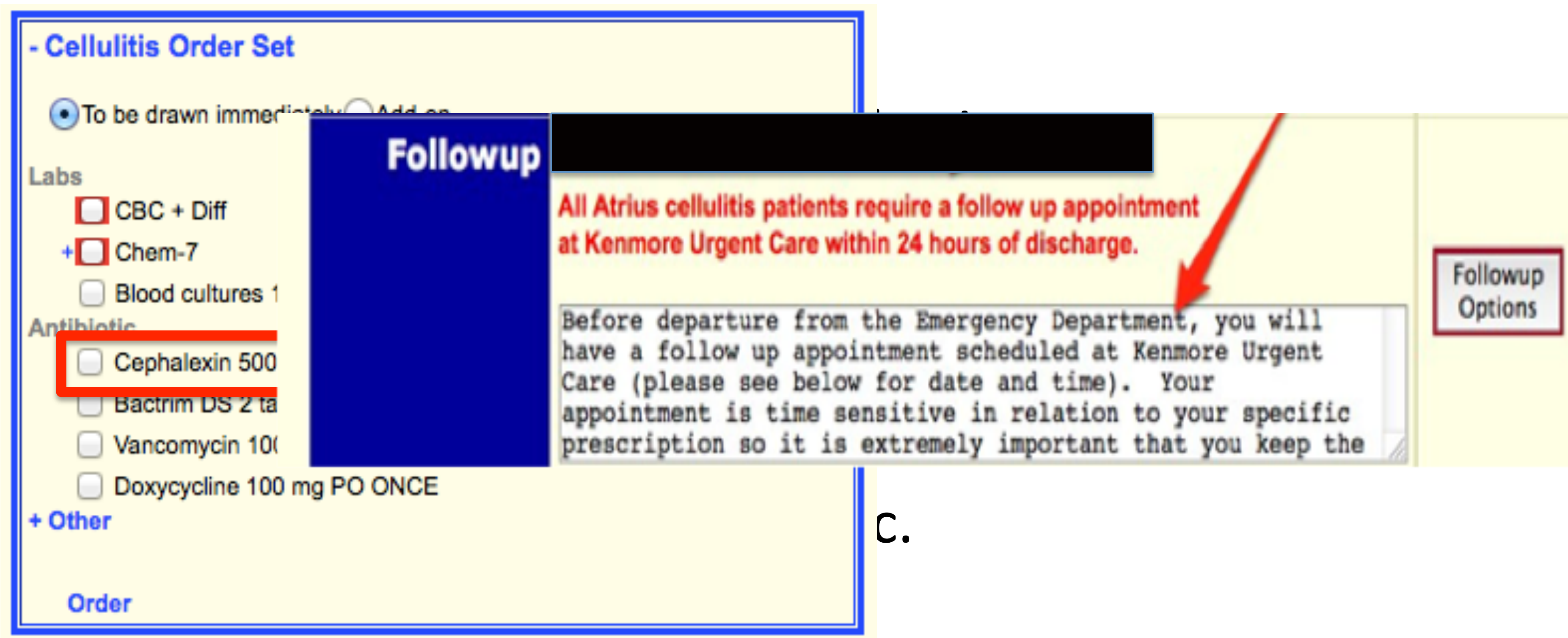
l precautions

c.

Real-time phenotype estimation

Cellulitis

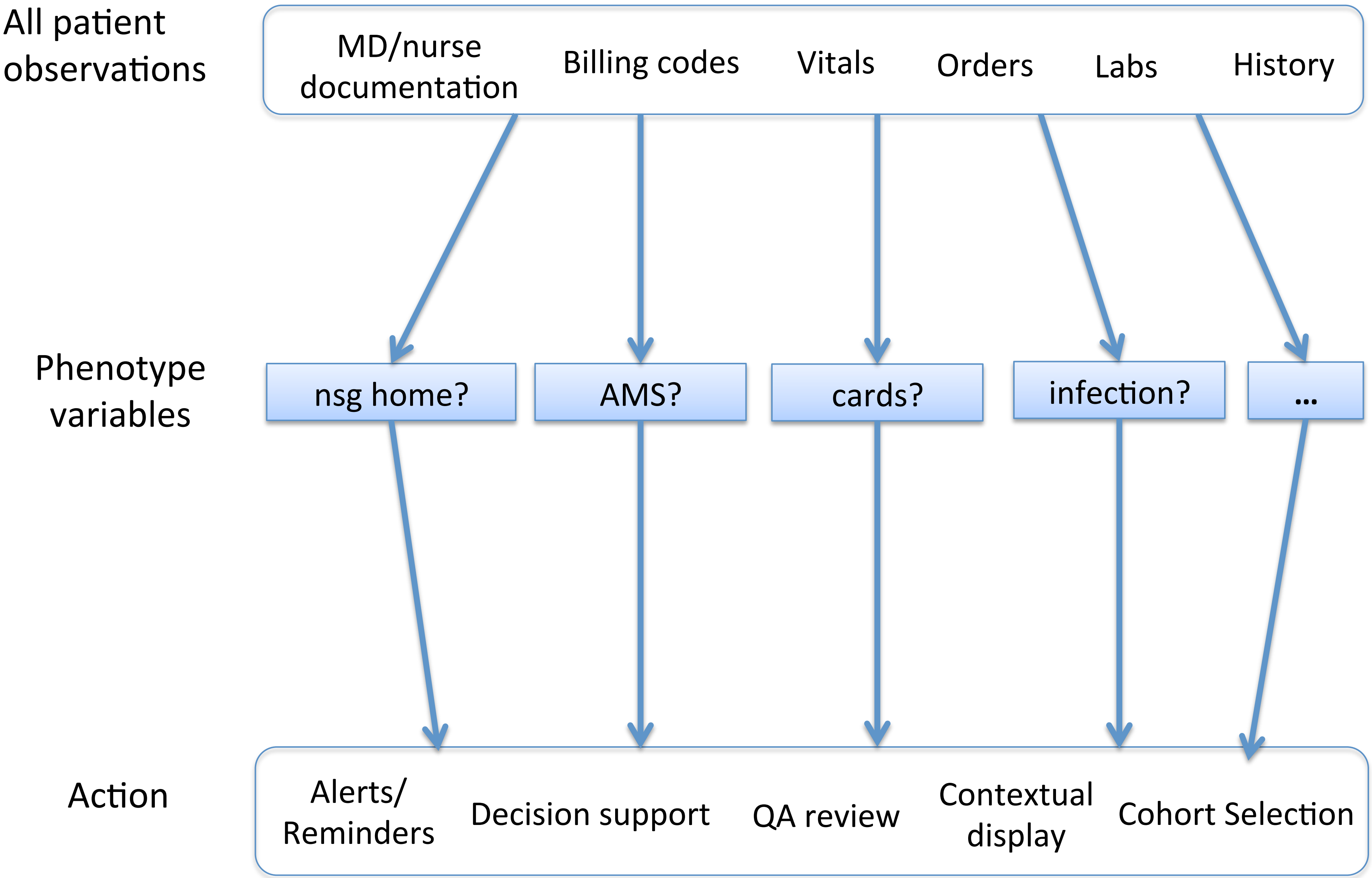
– Specialized order sets, confirm followup care.



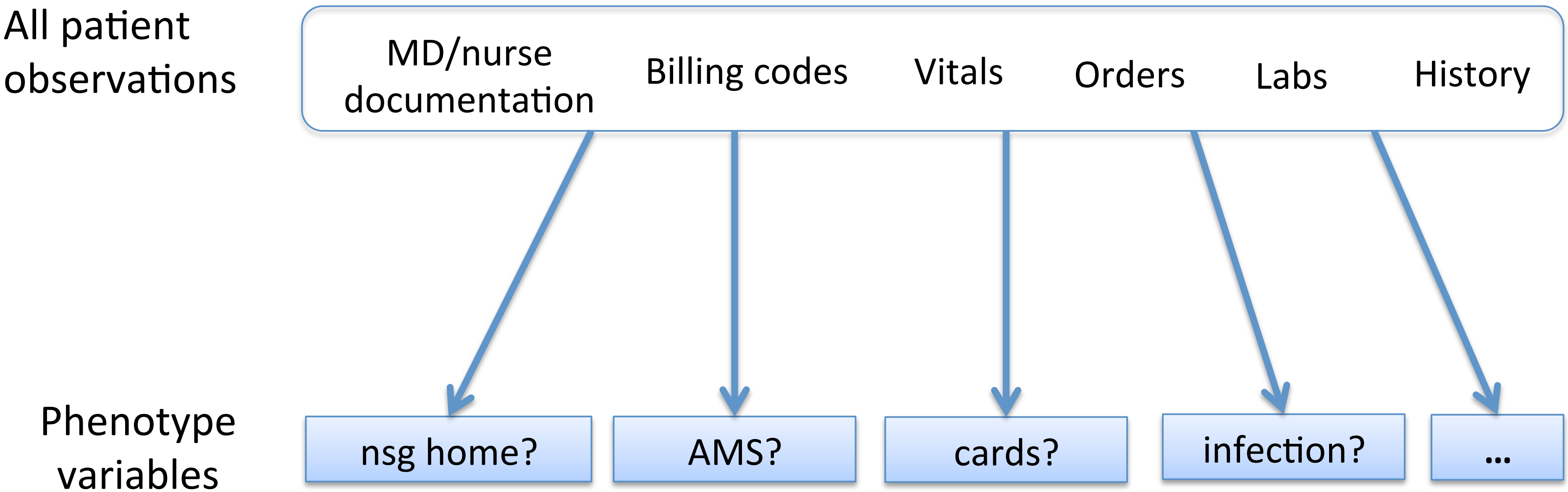
Representation

- To trigger effective decision support, the computer needs to know:
  - Is the patient from a nursing home?
  - Does the patient have an infection, altered mental status, require a cardiology consult, …?
- Hundreds of such phenotype variables that would be valuable for decision support
  - Entering this information in structured form would be a nightmare!

Big Picture


All patient observations
MD/nurse documentation
Billing codes
Vitals
Orders
Labs
History
Phenotype variables
nsg home?
AMS?
cards?
infection?
...
Action
Alerts/ Reminders
Decision support
QA review
Contextual display
Cohort Selection

Big Picture



How do we extract this representation from patient records?

Easy – Structured: allergies, medications, vital signs

Harder – Free text: smokes, domiciled, infection, histories

Hardest – Inferred: DDx, risk assessment, missing info

Current Approaches I: Manually created rules

- **Nursing home:** is the phrase "nursing home" in the patient's notes? Regional list of names, addresses.
- **Active Malignancy:** Diagnosis codes, key phrases in radiology reports,…
- Time consuming, often have low sensitivity

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

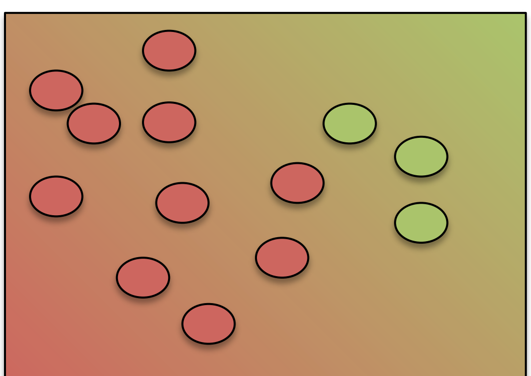
Nursing home?

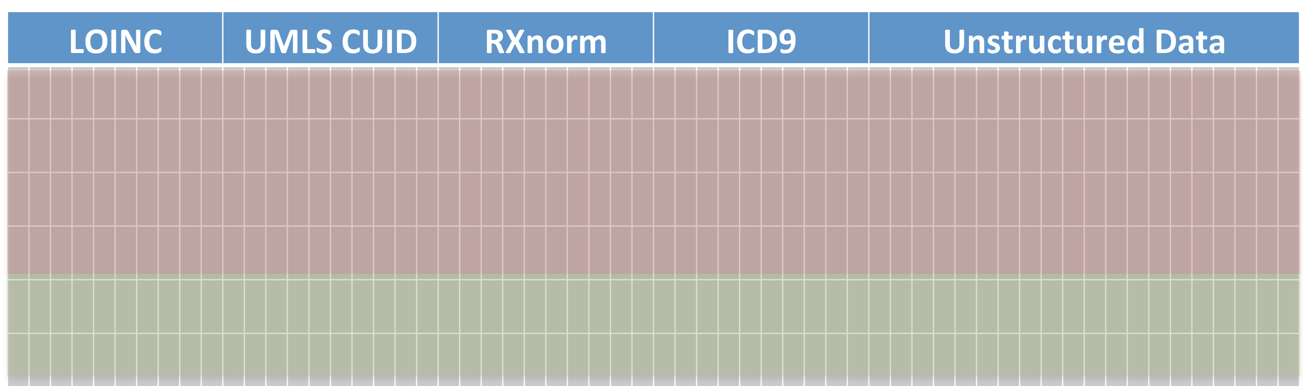
physician response (gold standard)

text contains: *"nursing home"*	T	F	
T	297	129	PPV 0.70
F	1,319	34511	
	Sensitivity 0.18		

Current Approaches II: Machine Learning

- Leverage large clinical databases to learn predictive rules.
- Need labeled data
- Classifiers often don't generalize across institutions





Our contribution: Learning with Anchors

- Use a combination of domain expertise (simple rules) and vast amounts of data (machine learning).
- Method does not require any manual labeling.
- Anchors are highly transferable between institutions.

What are anchors?

- Rather than provide gold-standard labels, construct a simple rule that can catch some positive cases.
- Examples:

Phenotype	Possible Anchor
Diabetic	gsn:016313 (insulin) in Medications
Cardiac	ICD9:428.X (heart failure) in Diagnoses
Nursing home	“from nursing home” in text
Social work	“social work consulted” in text

What are anchors?

- Rather than provide gold-standard labels, construct a simple rule that can catch <u>some positive cases</u>. **Low sensitivity here is ok!**
- Examples:

Phenotype	Possible Anchor
Diabetic	gsn:016313 (insulin) in Medications
Cardiac	ICD9:428.X (heart failure) in Diagnoses
Nursing home	"from nursing home" in text
Social work	"social work consulted" in text

Theoretical basis for anchors

- Unobserved variable: Y, Observation: A
- *A* is an **anchor** for *Y* if conditioning on *A*=1 gives uniform samples from the set of *positive cases*.
- Alternative formulation – two necessary conditions:

$$P(Y = 1|A = 1) = 1 \quad \text{AND} \quad A \perp \mathcal{X}|Y$$

Positive condition **Conditional independence**

$\mathcal{X}$ represents all *other* observations.

Theoretical basis for anchors

- Unobserved variable: Y, Observation: A
- *A* is an **anchor** for *Y* if conditioning on *A*=1 gives uniform samples from the set of *positive cases*.
- Alternative formulation – two necessary conditions:

$$P(Y = 1 | A = 1) = 1 \quad \text{AND} \quad A \perp \mathcal{X} | Y$$

Positive condition

e.g. If patient is taking *insulin*, the patient is surely **diabetic**.

Conditional independence

e.g. If we know the patient had **heart failure**, knowing whether the *diagnosis code* appears does inform us about the rest of the record.

$\mathcal{X}$ repre… s.

Theoretical basis for anchors

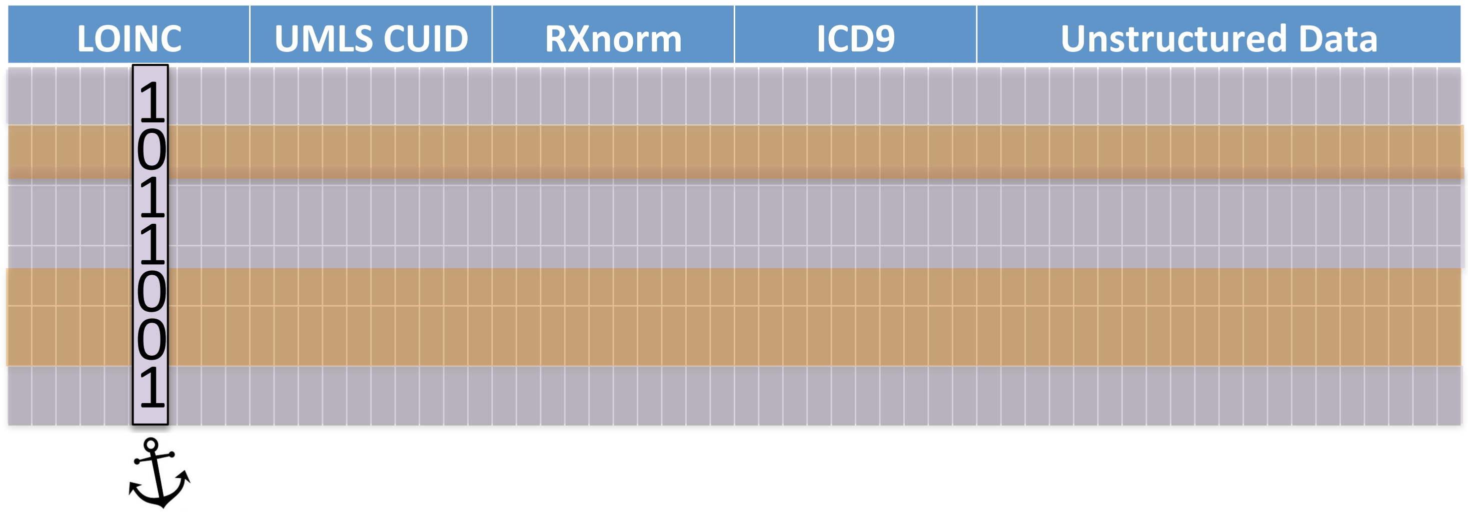
- Unobserved variable: Y, Observation: A
- *A* is an **anchor** for *Y* if conditioning on *A*=1 gives uniform samples from the set of *positive cases*.
- Theorem [Elkan & Noto 2008]:

In the above setting, a function to predict A

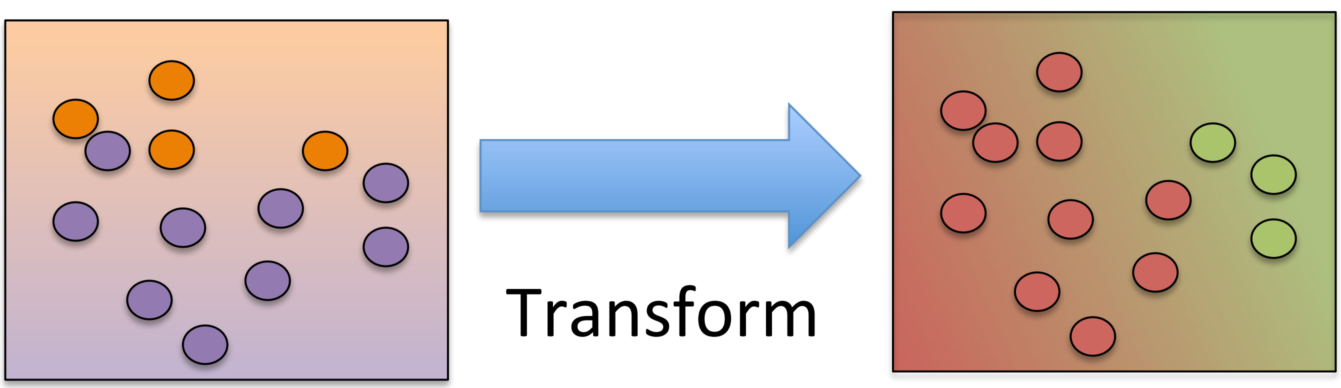
can be transformed to predict Y!

Learning with Anchors

Patient database



- Identify anchors
- Learn to predict the anchors (anchor as pseudo-labels)
- Account for the difference between anchors and labels



Predict anchor

Predict label

Learning with anchors

[Elkan & Noto 2008]

Input: anchor A

unlabeled patients

Output: prediction rule

1. Learn a calibrated classifier (e.g. logistic regression) to predict:

$$\Pr(A = 1 \mid \mathcal{X})$$

2. Using a validate set, let *P* be the patients with A=1. Compute:

$$C = \frac{1}{|\mathcal{P}|} \sum_{k \in \mathcal{P}} \Pr(A = 1 \mid \mathcal{X}^{(k)})$$

3. For a previously unseen patient *t*, predict:

$$\begin{array}{ll} \frac{1}{C}\Pr(A = 1|\mathcal{X}^{(t)}) & \text{if } A^{(t)} = 0 \\ 1 & \text{if } A^{(t)} = 1 \end{array}$$

Learning
Learn to predict *A* from the other variables.

Calibration
C is the average model prediction for patients with anchors.

Transformation
If no anchor present, according to a scaled version of the anchor-prediction model.

Generalizability/Portability

New institution

LOINC	UMLS CUID	RXnorm	ICD9	Different data types

Generalizability/Portability

New institution

LOINC	UMLS CUID	RXnorm	ICD9	Different data types

Data may be very different:

- Language
- Representation
- Population

Generalizability/Portability

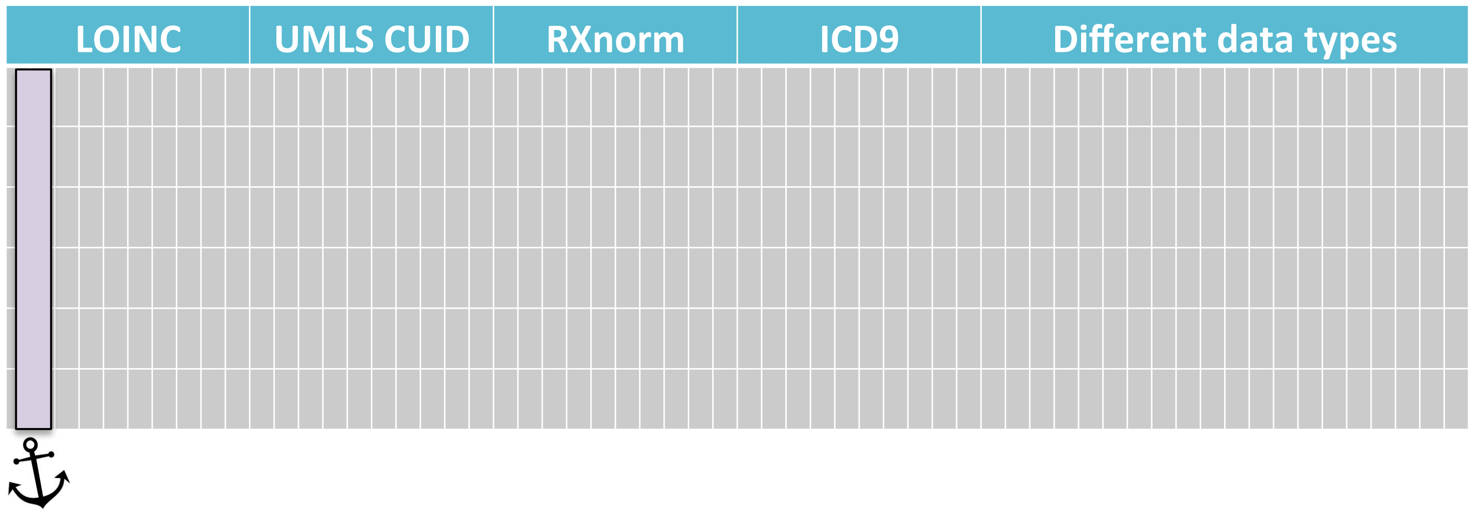
New institution

LOINC	UMLS CUID	RXnorm	ICD9	Different data types

As long as our anchors appear in the new data as well…

Generalizability/Portability

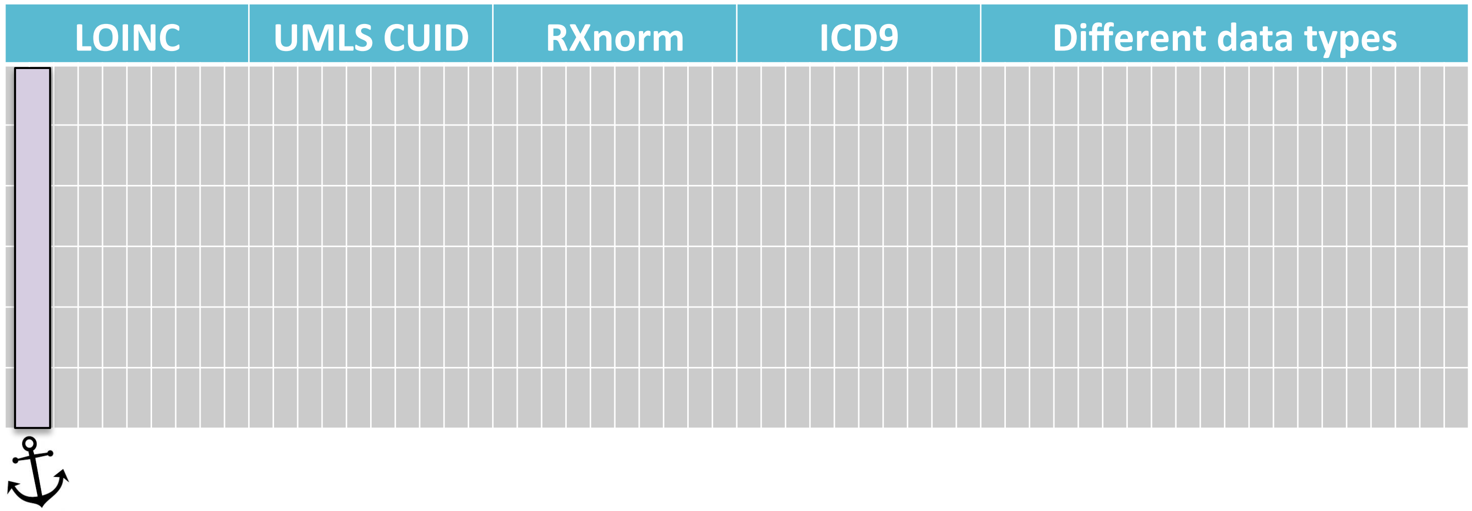
New
institution



As long as our anchors appear in the new data as well…
Can learn a new model, specific to the new institution.

Generalizability/Portability

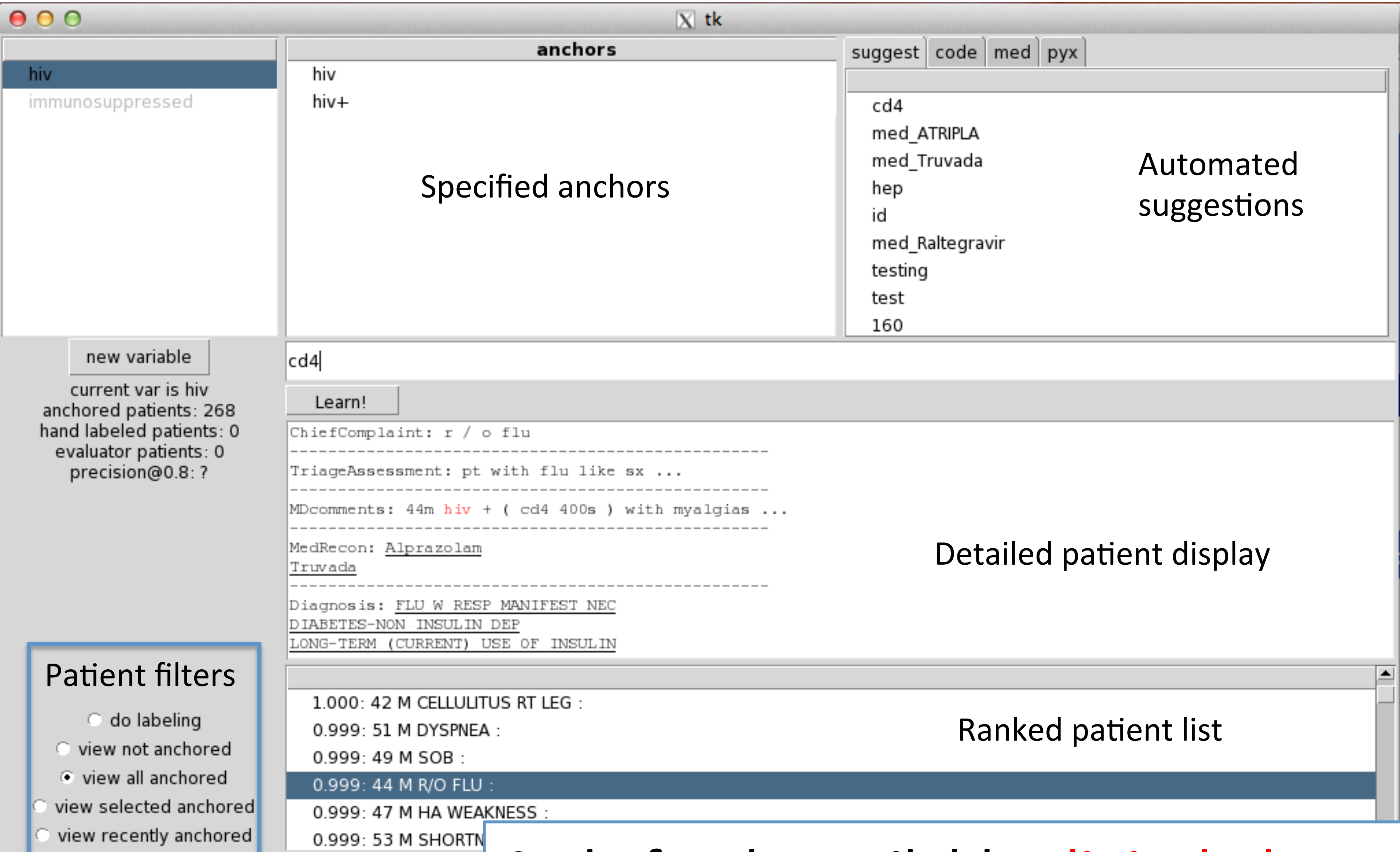
New institution



As long as our anchors appear in the new data as well…
Can learn a new model, specific to the new institution.

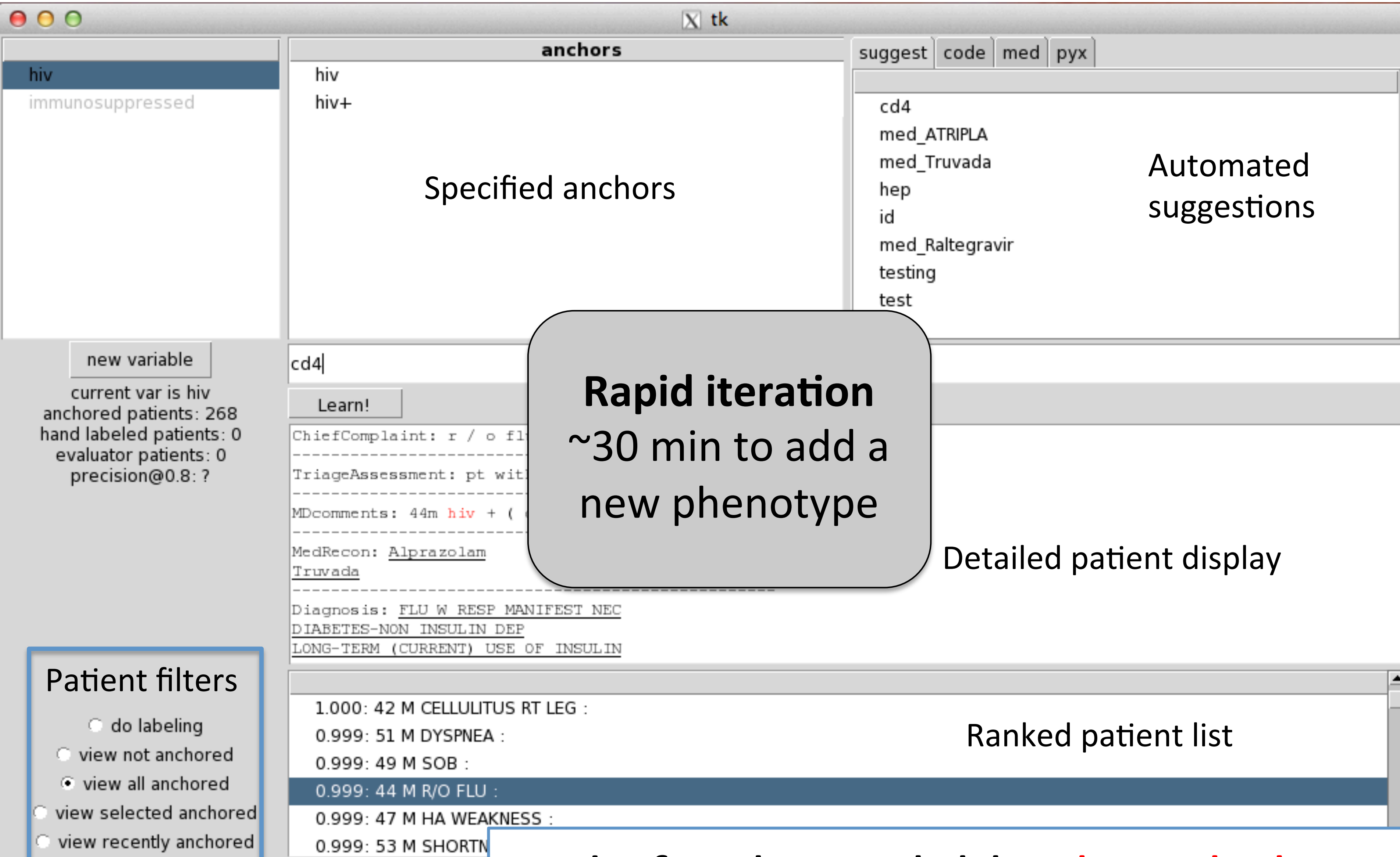
Only need to share **anchor definitions**,
Each site trains models on its own data.

Anchor Explorer V1



Code freely available clinicalml.org

Anchor Explorer V1



Rapid iteration

~30 min to add a new phenotype

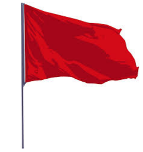
Code freely available clinicalml.org

Evaluation

- 273,174 emergency department patients from Beth Israel Deaconess ED
- Observations:
  - Structured: age, sex, ICD9 codes*, medRecon, pyxis
  - Preprocessed free text: chief complaint, triage assessment, and physician's comments
- Anchors specified by a single ED physician using our interface.

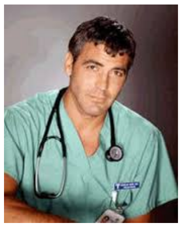
*Diagnosis codes available for *training* but not at test time (real time decision support setting).

Test variables: ED red flags



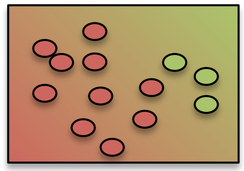
- Active malignancy
- Fall
- Cardiac Etiology
- Infection
- From Nursing Home
- Anticoagulated
- Immunosuppressed
- Septic Shock
- Pneumonia

Learned models: Nursing Home



Anchors

- nursing facility
- nursing home
- nsg facility
- nsg home
- nsg. home



Highly weighted features

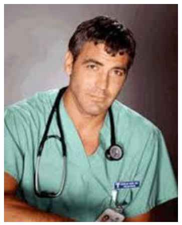
Ages	Medications		Pyxis
age=90+	senna	mirtazapine	vancomycin
age=80-90	colace	maalox	levofloxacin
age=70-80	trazodone	tums	

Unstructured text

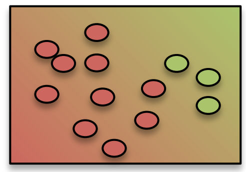
baseline	from	dnr
changes	staff	full code
nonverbal	at	g tube
ams	resident	foley
unwitnessed_fall	sent	nh
confusion	reported	

Conditional independence assumption?

Learned models: Cardiac Etiology



Anchors



Highly weighted features

ICD9 codes
410.* acute MI
411.* other acute …
413.* angina pectoris
785.51 card. shock

Pyxis
coron. vasodilators

cardiac medicine
BIDMC shortform

Ages
age=80-90
age=70-80
age=90+

Medications
lasix
furosemide

Sex=M

Pyxis
aspirin
clopidogrel
Heparin Sodium
Metoprolol
Tartrate
Morphine Sulfate
Integrilin
Labetalol

nstemi
stemi
ntg
lasix
nitro

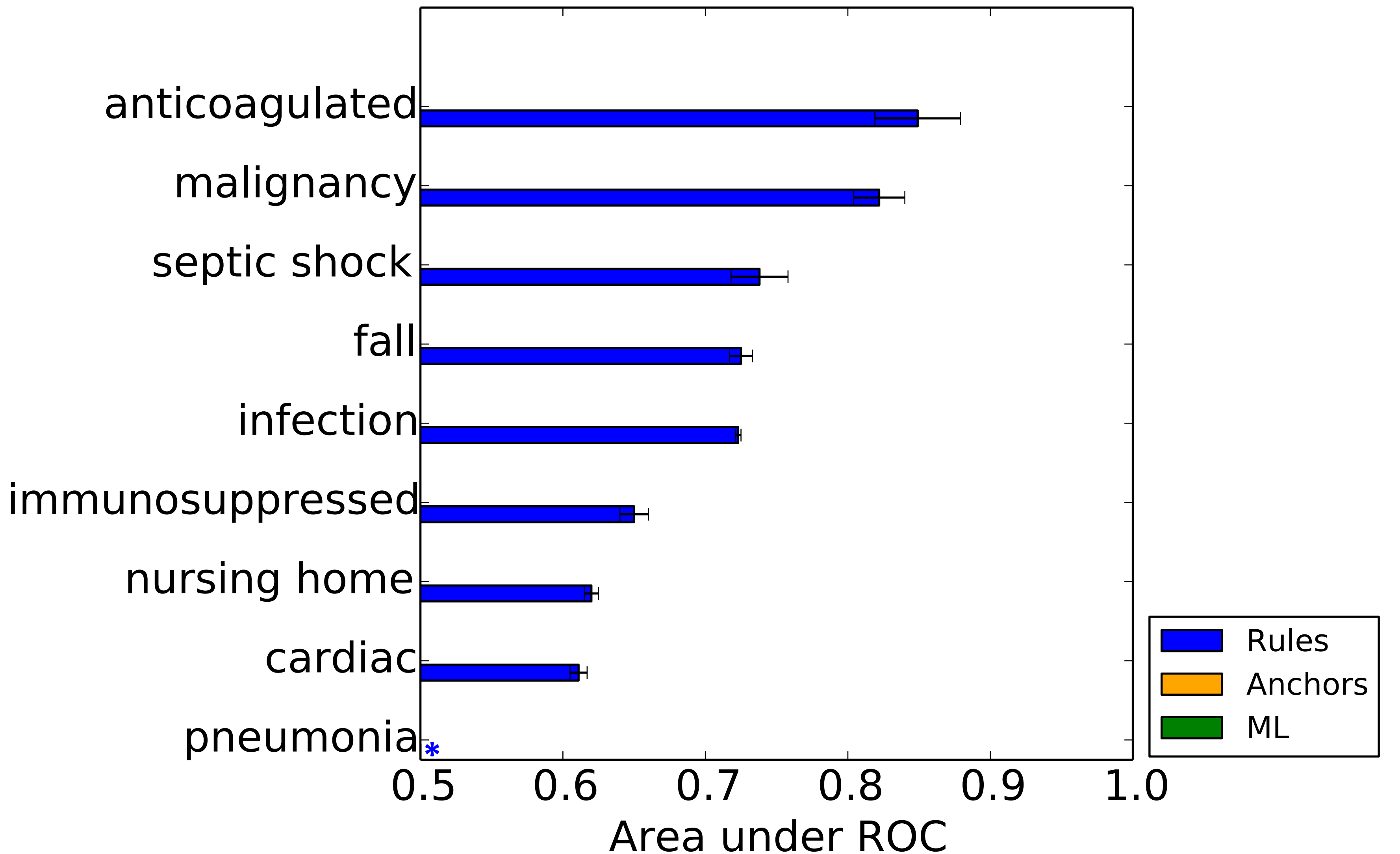
cp
chest pain
edema
cmed
chf exacerbation
sob
pedal edema

Unstructured text

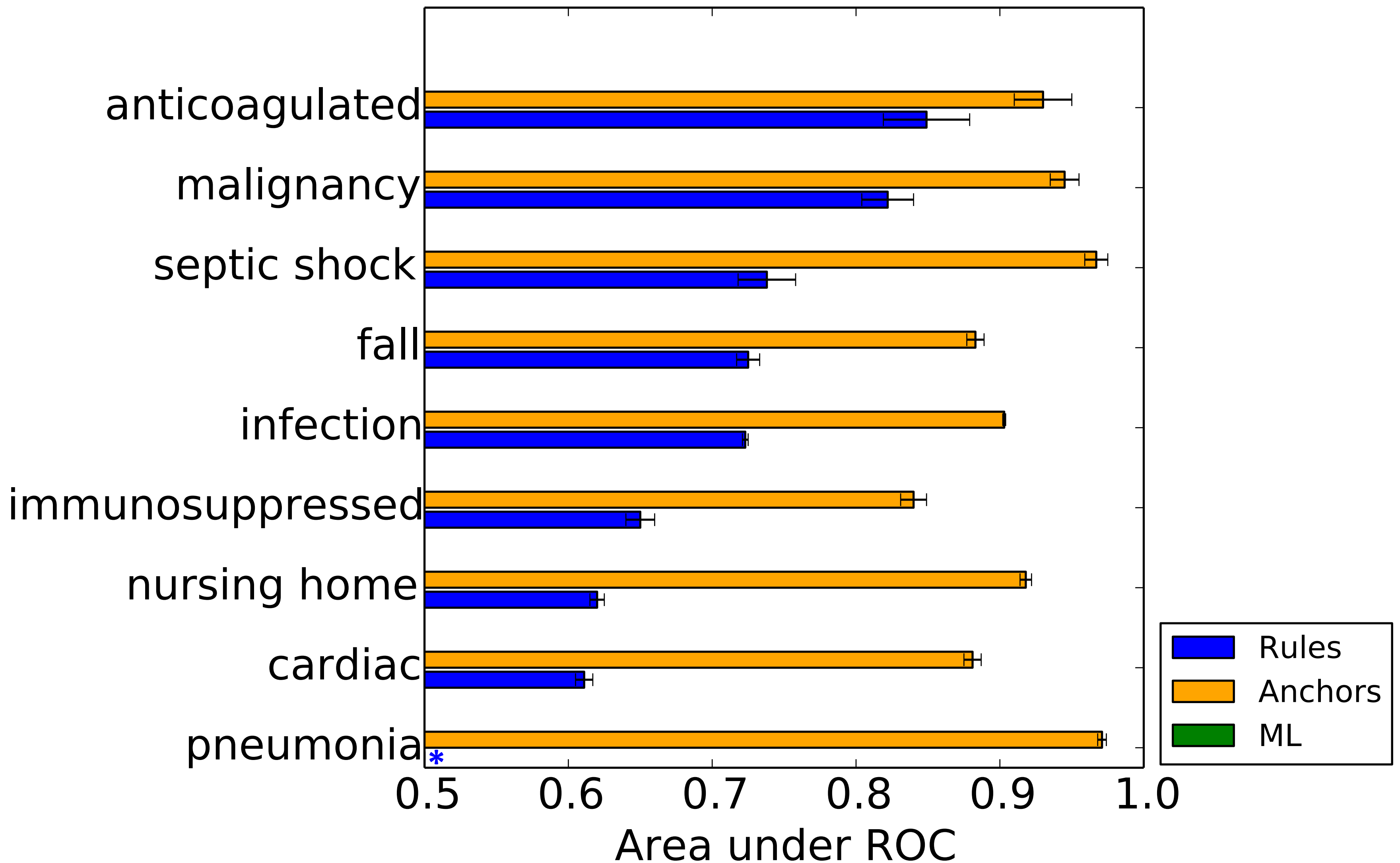
Comparison to Existing Approaches

- (Rules) Predict just according to the anchors.
  - 1 if anchor is present, 0 otherwise
- (ML) Machine learning (logistic regression)
  - Using up to 3K labels
  - Improves with more labels, but labels are expensive!

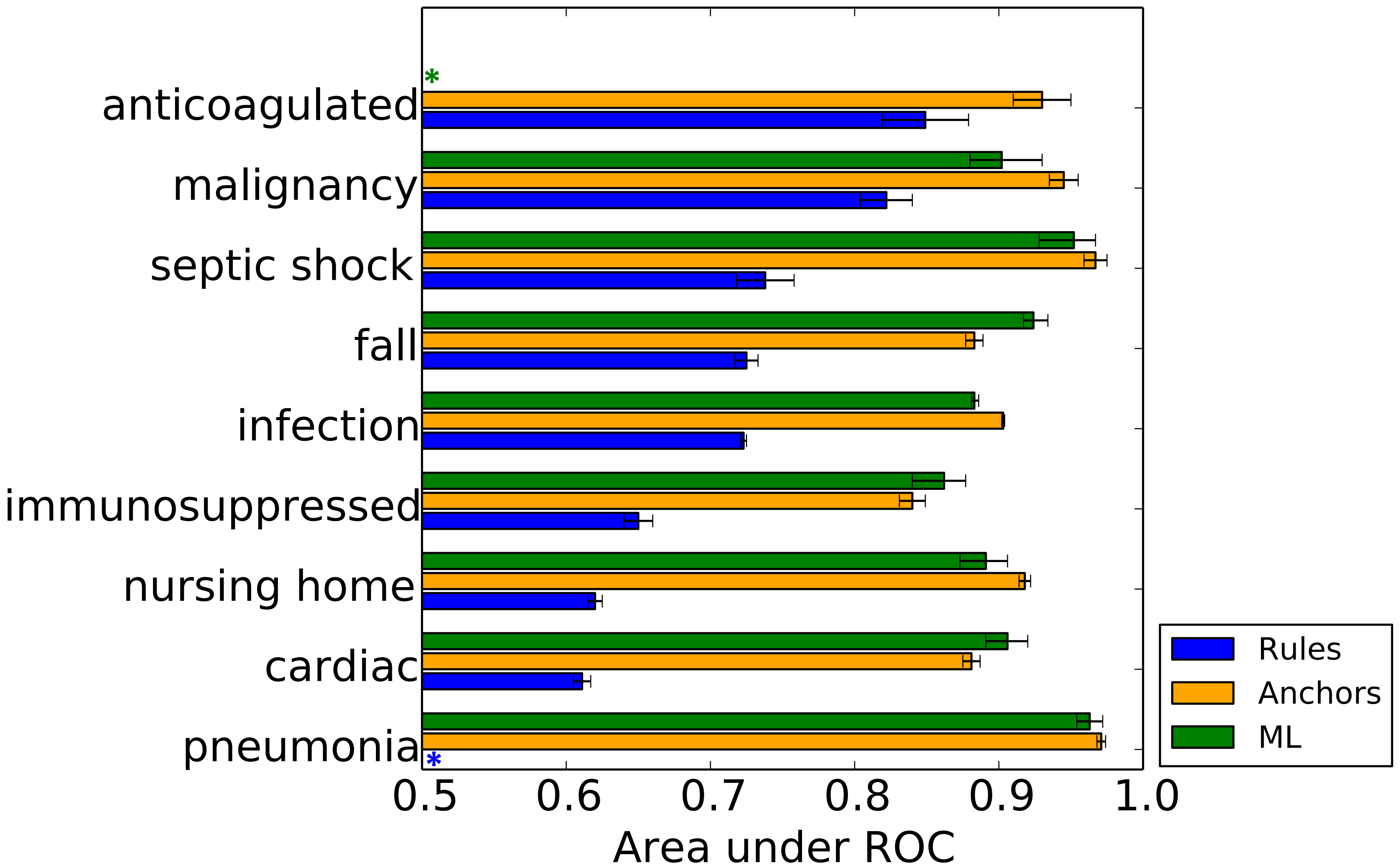
Phenotype predictions



Phenotype predictions



Phenotype predictions



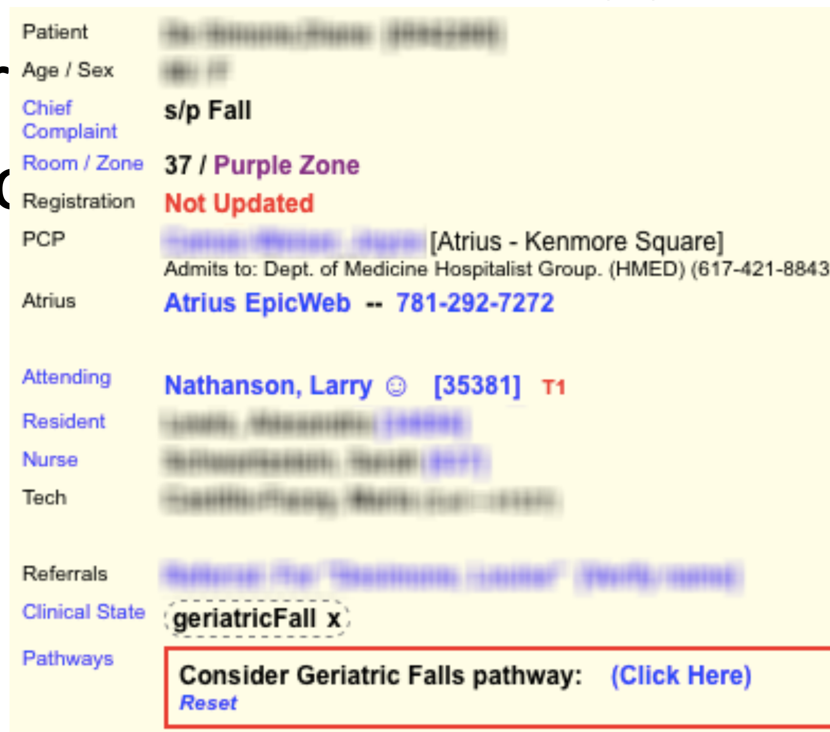
Our next steps

- Shared library of anchored phenotypes
- Real-time estimation of clinical states and actual use for decision support within ED
- Test portability of anchors to other institutions
- General cohort selection tool

More info: clinicalml.org

Our next steps

- Shared library of anchored phenotypes
- Real-time estimation of clinical states and actual use for decision support within ED
- Test por ner instituti
- General



inicalml.org

Our next steps

- Shared library of anchored phenotypes
- Real-time estimation of clinical states and actual use for decision support within ED
- Test portability of anchors to other institutions
- General cohort selection tool

More info: clinicalml.org