



Beth Israel Deaconess  
Medical Center

# 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.

**- Cellulitis Order Set**

To be drawn immediately  Add-on

**Labs**

- CBC + Diff
- Chem-7
- Blood cultures 1 set

**Antibiotic**

- Cephalexin 500 mg PO **\*Allergy**
- Bactrim DS 2 tab PO
- Vancomycin 1000 mg IV
- Doxycycline 100 mg PO ONCE

**+ Other**

**Order**

fections

l precautions

C.

# Real-time phenotype estimation

## Cellulitis

- Specialized order sets, confirm followup care.

The image shows a screenshot of a medical order set interface for Cellulitis. The interface is divided into several sections:

- Cellulitis Order Set**: The main title of the order set.
- To be drawn immediately**: A radio button option for timing.
- Labs**: A section containing checkboxes for:
  - CBC + Diff
  - Chem-7
  - Blood cultures 1
- Antibiotic**: A section containing checkboxes for:
  - Cephalexin 500 (highlighted with a red box)
  - Bactrim DS 2 ta
  - Vancomycin 100
  - Doxycycline 100 mg PO ONCE
- + Other**: A section for additional orders.
- Order**: A button at the bottom left.

A blue box labeled **Followup** is overlaid on the interface, containing the following text:

**All Atrius cellulitis patients require a follow up appointment at Kenmore Urgent Care within 24 hours of discharge.**

Below this, a scrollable text box contains the following message:

Before departure from the Emergency Department, you will have a follow up appointment scheduled at Kenmore Urgent Care (please see below for date and time). Your appointment is time sensitive in relation to your specific prescription so it is extremely important that you keep the

A red arrow points from the follow-up text box to the scrollable message box. On the right side of the interface, there is a button labeled **Followup Options**.

C.

# 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

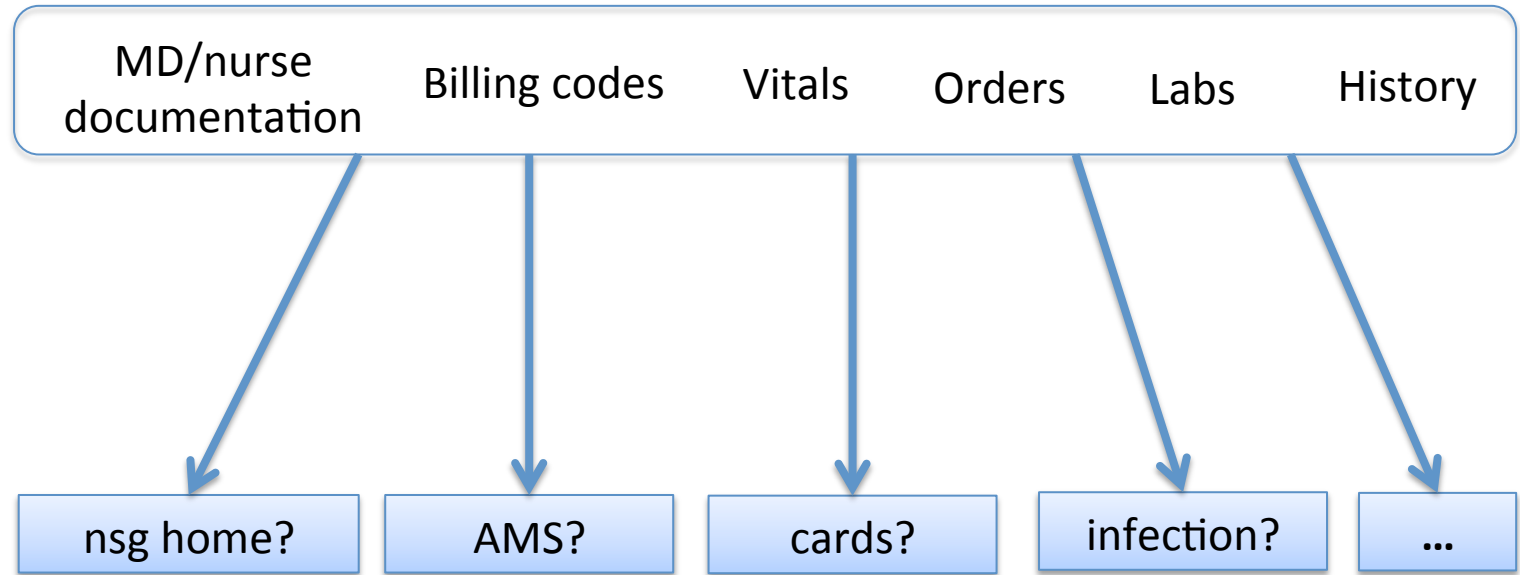
QA review

Contextual  
display

Cohort Selection

# Big Picture

All patient observations



Phenotype variables

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 facilty  
...

text contains:  
*“nursing home”*

Nursing home?

physician response  
(gold standard)

	T	F
T	297	129
F	1,319	34511

PPV  
0.70

Sensitivity  
0.18



# 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 some positive cases. **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

$\mathcal{X}$  repre

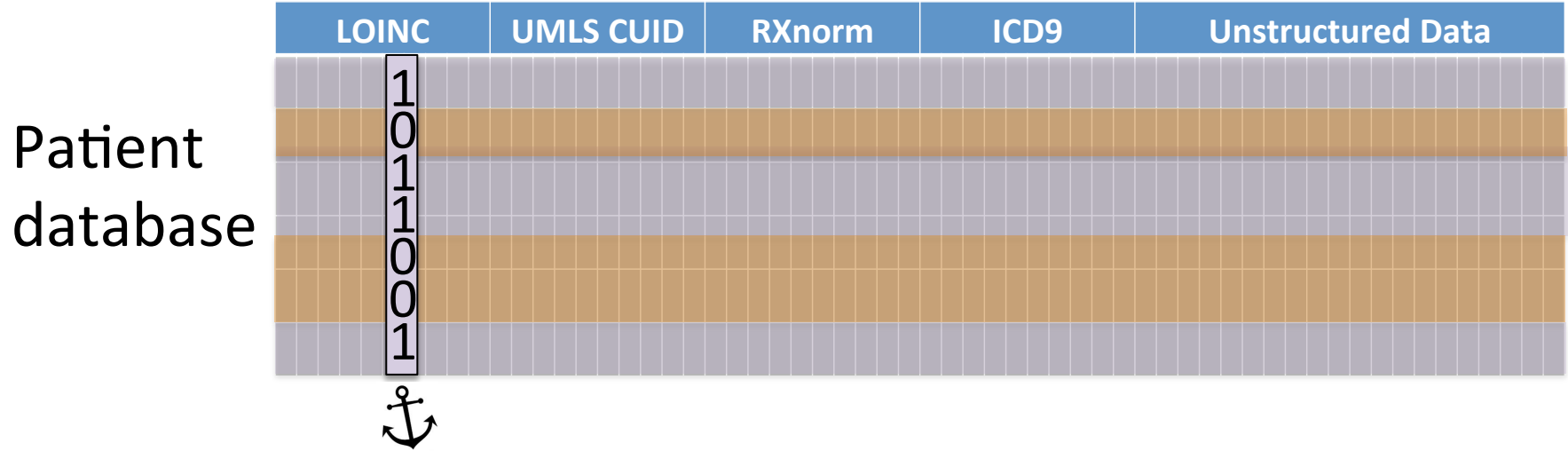
e.g. If we know the patient had **heart failure**, knowing whether the *diagnosis code* appears does not inform us about the rest of the record.

# Theoretical basis for anchors

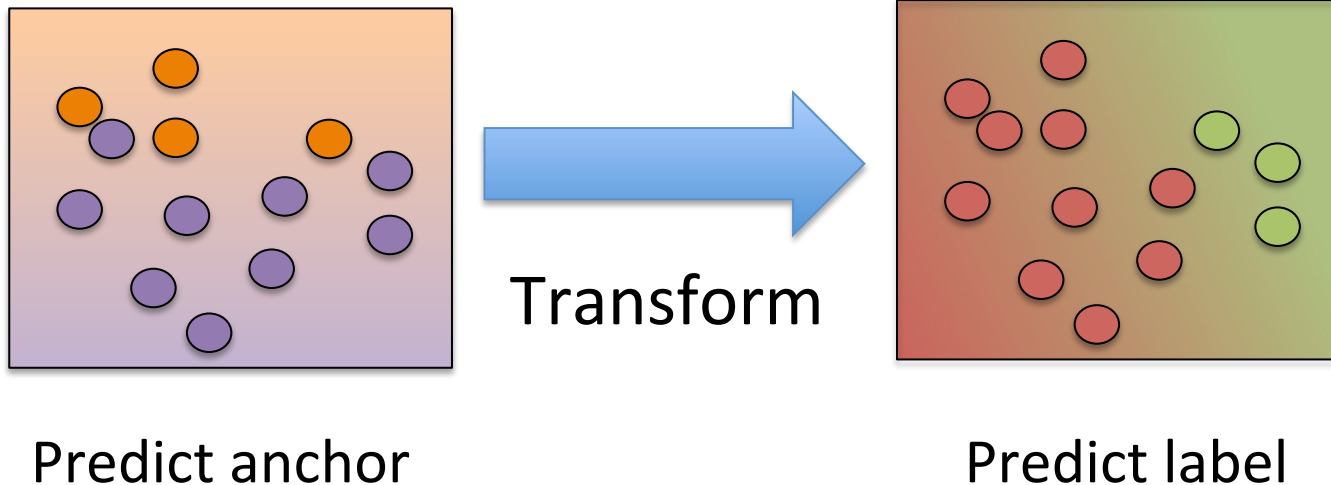
- Unobserved variable:  $Y$ , Observation:  $A$
- $A$  is an **anchor** for  $Y$  if conditioning on  $A=1$  gives uniform samples from the set of *positive cases*.
- Theorem [Elkan & Noto 2008]:

*In the above setting, a function to predict  $A$   
can be transformed to predict  $Y$ !*

# Learning with Anchors



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



# 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  $\mathcal{P}$  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:

$$\frac{1}{C} \Pr(A = 1 \mid \mathcal{X}^{(t)}) \quad \text{if } A^{(t)} = 0$$
$$1 \quad \text{if } A^{(t)} = 1$$

## 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

	LOINC	UMLS CUID	RXnorm	ICD9	Different data types
New institution					

# Generalizability/Portability

	LOINC	UMLS CUID	RXnorm	ICD9	Different data types
New institution					

Data may be very different:

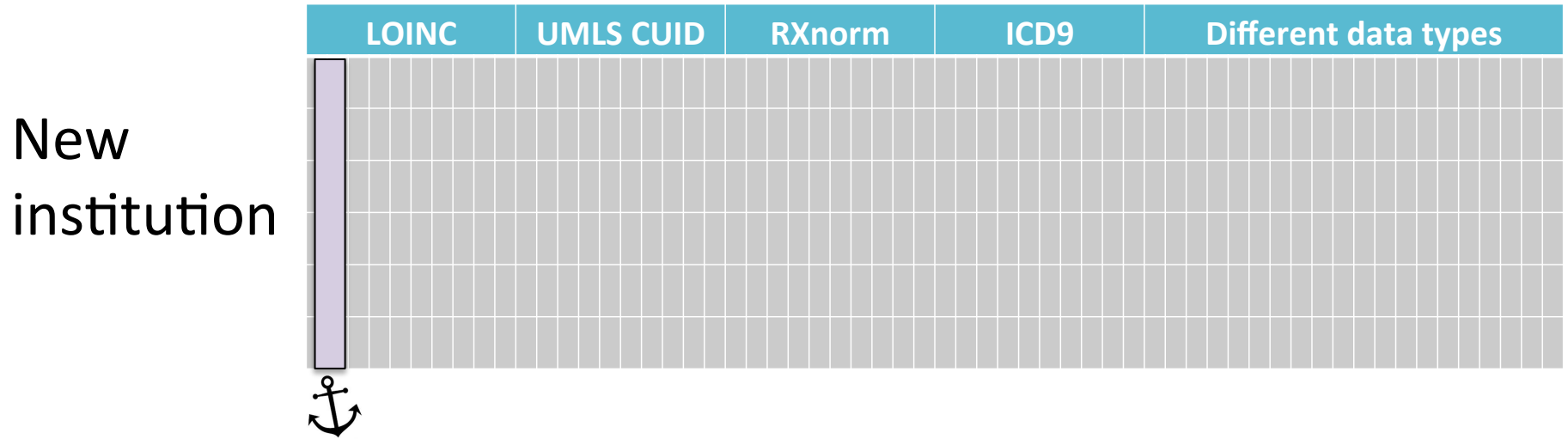
- Language
- Representation
- Population

# Generalizability/Portability

	LOINC	UMLS CUID	RXnorm	ICD9	Different data types
New institution					

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

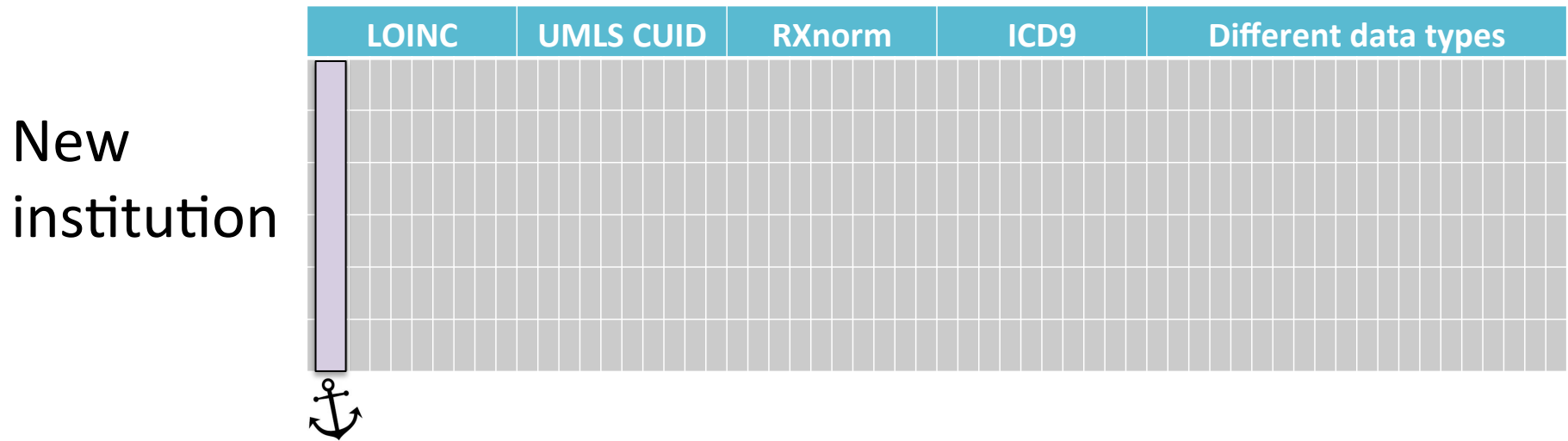
# Generalizability/Portability



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



# Generalizability/Portability



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

anchors

hiv  
hiv+

Specified anchors

suggest code med pyx

cd4  
med\_ATRIPLA  
med\_Truvada  
hep  
id  
med\_Raltegravir  
testing  
test  
160

Automated suggestions

new variable

current var is hiv  
anchored patients: 268  
hand labeled patients: 0  
evaluator patients: 0  
precision@0.8: ?

cd4|

Learn!

ChiefComplaint: r / o flu  
-----  
TriageAssessment: pt with flu like sx ...  
-----  
MDcomments: 44m hiv + ( cd4 400s ) with myalgias ...  
-----  
MedRecon: Alprazolam  
Truvada  
-----  
Diagnosis: FLU W RESP MANIFEST NEC  
DIABETES-NON INSULIN DEP  
LONG-TERM (CURRENT) USE OF INSULIN

Detailed patient display

1.000: 42 M CELLULITUS RT LEG :  
0.999: 51 M DYSPNEA :  
0.999: 49 M SOB :  
0.999: 44 M R/O FLU :  
0.999: 47 M HA WEAKNESS :  
0.999: 53 M SHORTN

Ranked patient list

Patient filters

- do labeling
- view not anchored
- view all anchored
- view selected anchored
- view recently anchored

Code freely available [clinicalml.org](http://clinicalml.org)

# Anchor Explorer V1

The screenshot displays the Anchor Explorer V1 interface with several key components:

- Left Panel:** A list of anchors with 'hiv' selected. Below it, a 'new variable' button and statistics for the current variable 'hiv':  
current var is hiv  
anchored patients: 268  
hand labeled patients: 0  
evaluator patients: 0  
precision@0.8: ?
- Top Center Panel:** 'anchors' section showing 'Specified anchors' with 'hiv' and 'hiv+' listed.
- Top Right Panel:** 'Automated suggestions' section with a list of terms: cd4, med\_ATRIPLA, med\_Truvada, hep, id, med\_Raltegravir, testing, test.
- Bottom Left Panel:** 'Patient filters' section with radio buttons for:  
 do labeling  
 view not anchored  
 view all anchored  
 view selected anchored  
 view recently anchored
- Center Panel:** A 'Learn!' button and a 'Detailed patient display' showing a patient's medical record:  
cd4|  
ChiefComplaint: r / o flu  
TriageAssessment: pt with  
MDcomments: 44m hiv + (  
MedRecon: Alprazolam  
Truvada  
Diagnosis: FLU W RESP MANIFEST NEC  
DIABETES-NON INSULIN DEP  
LONG-TERM (CURRENT) USE OF INSULIN
- Bottom Center Panel:** A 'Ranked patient list' showing a list of patients with their scores and conditions:  
1.000: 42 M CELLULITUS RT LEG :  
0.999: 51 M DYSPNEA :  
0.999: 49 M SOB :  
0.999: 44 M R/O FLU :  
0.999: 47 M HA WEAKNESS :  
0.999: 53 M SHORTN

**Rapid iteration**  
~30 min to add a  
new phenotype

Code freely available [clinicalml.org](http://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

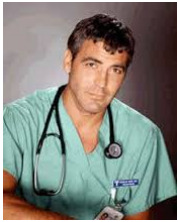


- Active malignancy
- Fall
- Cardiac Etiology
- Infection
- From Nursing Home
- Anticoagulated
- Immunosuppressed
- Septic Shock
- Pneumonia

Does the patient have an active malignancy? <sup>i</sup>

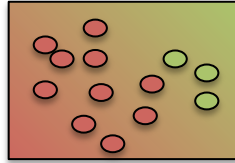
Unlikely                      Unsure                      Likely

# Learned models: Nursing Home



Anchors

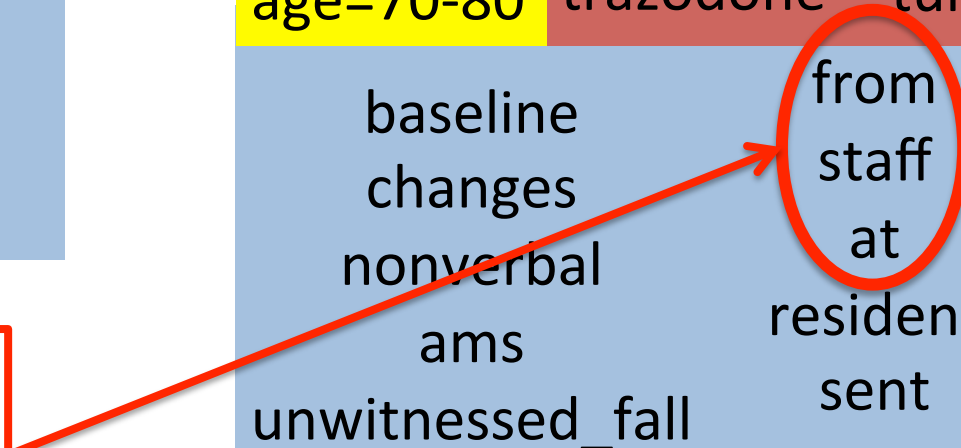
nursing facility  
 nursing home  
 nsg facility  
 nsg home  
 nsg. home



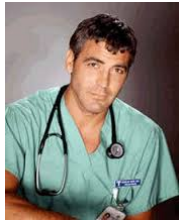
Highly weighted features

Ages	Medications	Pyxis
age=90+ age=80-90 age=70-80 baseline changes nonverbal ams unwitnessed_fall confusion	senna mirtazapine colace maalox trazodone tums from staff at resident sent reported	vancomycin levofloxacin dnr full code g tube foley nh Unstructured text

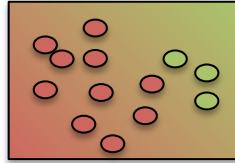
Conditional independence assumption?



# 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+

nstemi  
 stemi  
 ntg  
 lasix  
 nitro

**Medications**  
 lasix  
 furosemide

cp  
 chest pain  
 edema  
 cmed  
 chf exacerbation  
 sob  
 pedal edema

Sex=M

**Pyxis**  
 aspirin  
 clopidogrel  
 Heparin Sodium  
 Metoprolol  
 Tartrate  
 Morphine Sulfate  
 Integrilin  
 Labetalol

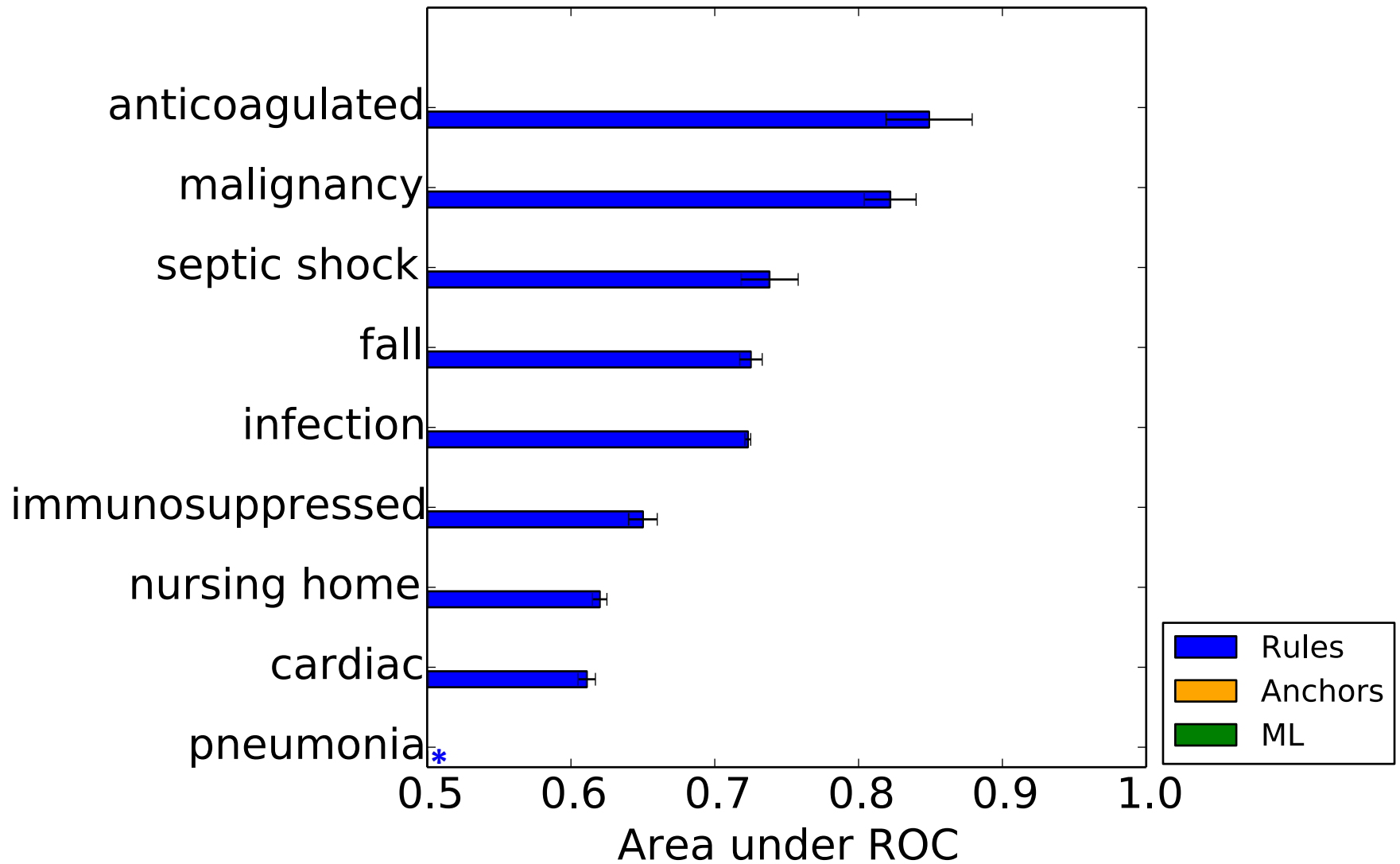
**Unstructured text**

# Comparison to Existing Approaches

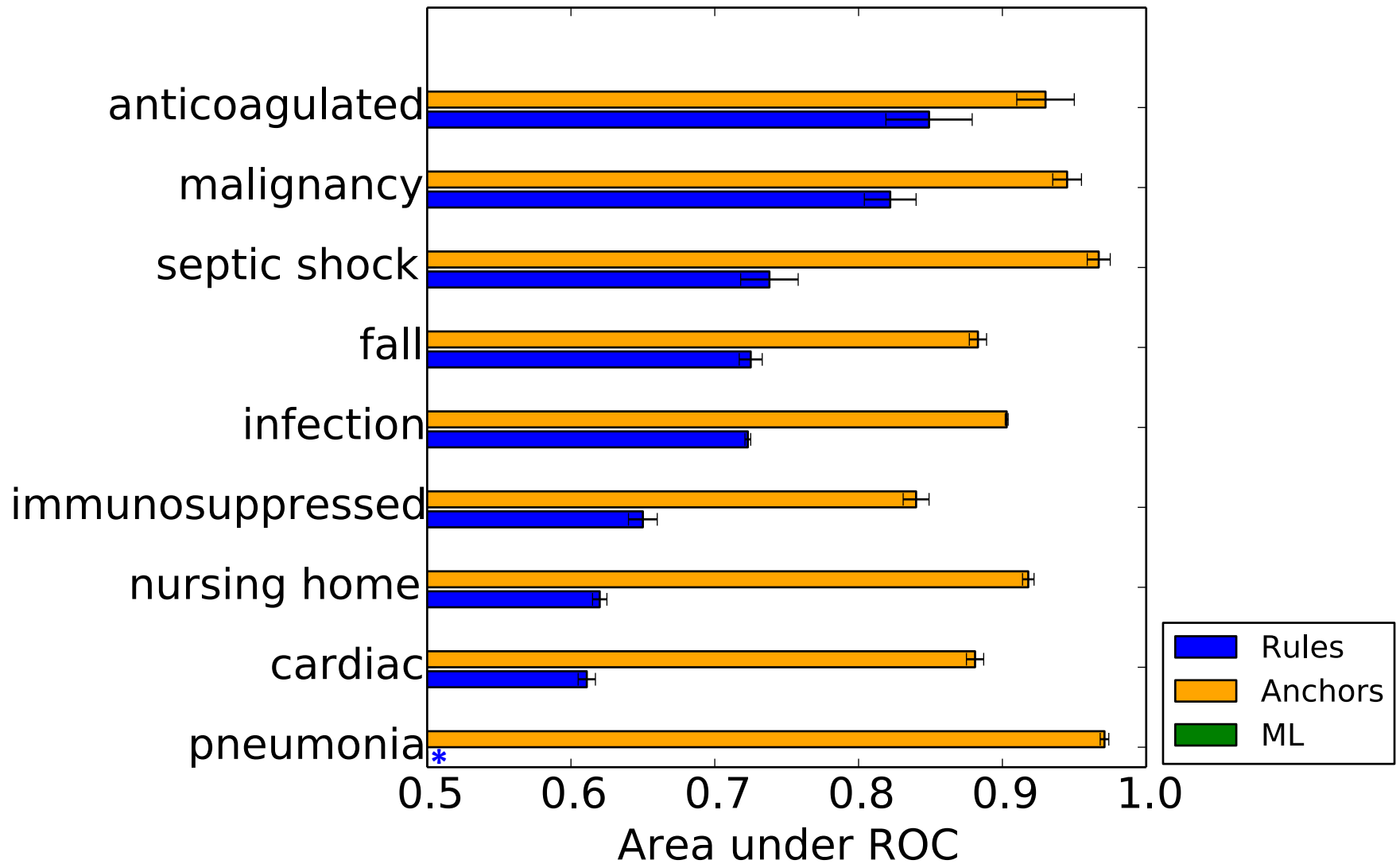
- (Rules) Predict just according to the anchors.
  - 1 if anchor is present, 0 otherwise
- (ML) Machine learning (logistic regression)
  - Using up to 3K labels
  - Improves with more labels, but labels are expensive!



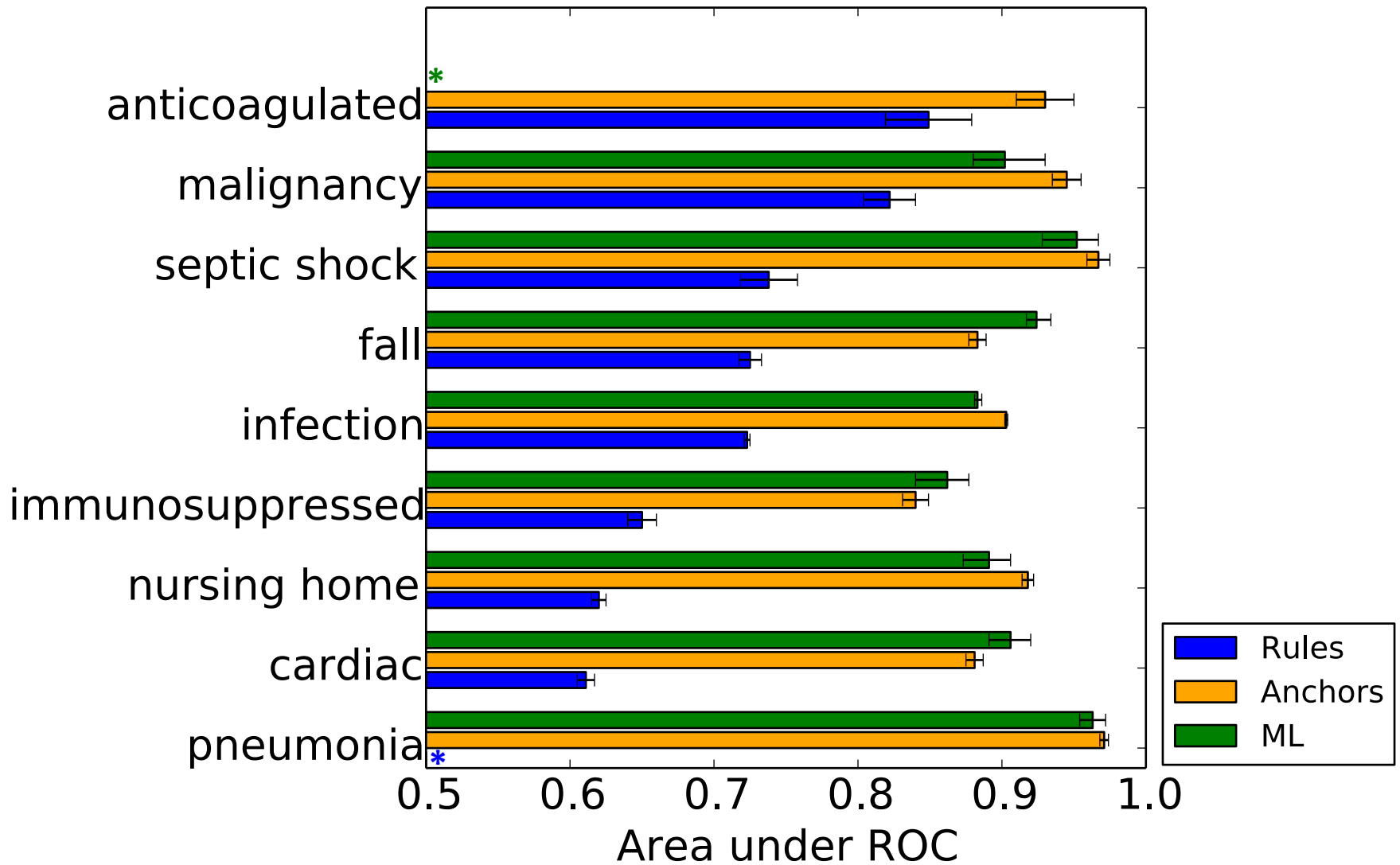
# Phenotype predictions



# 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](http://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  
instituti  
General

Patient [Redacted] (1000000000)

Age / Sex [Redacted]

Chief Complaint **s/p Fall**

Room / Zone **37 / Purple Zone**

Registration **Not Updated**

PCP [Redacted] [Atrius - Kenmore Square]  
Admits to: Dept. of Medicine Hospitalist Group. (HMED) (617-421-8843)

Atrius **Atrius EpicWeb -- 781-292-7272**

Attending **Nathanson, Larry** 😊 [35381] T1

Resident [Redacted] (1000000000)

Nurse [Redacted] (1000000000)

Tech [Redacted] (1000000000)

Referrals [Redacted] (1000000000)

Clinical State **geriatricFall x**

Pathways **Consider Geriatric Falls pathway: (Click Here)**  
[Reset](#)

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 portability of anchors to other institutions
- General cohort selection tool

More info: [clinicalml.org](http://clinicalml.org)