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.

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]

- Labs
  - CBC + Diff
  - Chem-7
  - Blood cultures
- Antibiotic
  - Cephalexin 500
  - Bacitracin DS 2 tabs
  - Vancomycin 10g
  - Doxycycline 100 mg PO ONCE
- Other

Followup

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

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
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?
- ...

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 faclt
- ...

<table>
<thead>
<tr>
<th>Nursing home?</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T</strong></td>
<td>297</td>
<td>129</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>1,319</td>
<td>34511</td>
</tr>
</tbody>
</table>

Text contains: “nursing home”

<table>
<thead>
<tr>
<th></th>
<th>PPV</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.70</td>
<td>0.18</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Possible Anchor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic</td>
<td>gsn:016313 (insulin) in Medications</td>
</tr>
<tr>
<td>Cardiac</td>
<td>ICD9:428.X (heart failure) in Diagnoses</td>
</tr>
<tr>
<td>Nursing home</td>
<td>“from nursing home” in text</td>
</tr>
<tr>
<td>Social work</td>
<td>“social work consulted” in text</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Possible Anchor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic</td>
<td>gsn:016313 (insulin) in Medications</td>
</tr>
<tr>
<td>Cardiac</td>
<td>ICD9:428.X (heart failure) in Diagnoses</td>
</tr>
<tr>
<td>Nursing home</td>
<td>“from nursing home” in text</td>
</tr>
<tr>
<td>Social work</td>
<td>“social work consulted” in text</td>
</tr>
</tbody>
</table>
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:

\[
P(Y = 1 \mid A = 1) = 1 \quad \text{AND} \quad A \perp \chi \mid Y
\]

Positive condition \hspace{2cm} \text{Conditional independence}

\( \chi \) 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:

\[ P(Y = 1|A = 1) = 1 \quad \text{AND} \quad A \perp \chi|Y \]

Positive condition

\begin{tabular}{|l|}
\hline
\textbf{e.g. If patient is taking insulin, the patient is surely diabetic.} \\
\hline
\end{tabular}

Conditional independence

\begin{tabular}{|l|}
\hline
\textbf{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.} \\
\hline
\end{tabular}
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

<table>
<thead>
<tr>
<th>LOINC</th>
<th>UMLS CUID</th>
<th>RXnorm</th>
<th>ICD9</th>
<th>Unstructured Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transform

Predict anchor  

Predict label
# Learning with anchors

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

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
   \]

---

[Elkan & Noto 2008]

**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

<table>
<thead>
<tr>
<th>LOINC</th>
<th>UMLS CUID</th>
<th>RXnorm</th>
<th>ICD9</th>
<th>Different data types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

New institution
Generalizability/Portability

<table>
<thead>
<tr>
<th>LOINC</th>
<th>UMLS CUID</th>
<th>RXnorm</th>
<th>ICD9</th>
<th>Different data types</th>
</tr>
</thead>
</table>

Data may be very different:
- Language
- Representation
- Population

New institution
Generalizability/Portability

<table>
<thead>
<tr>
<th>LOINC</th>
<th>UMLS CUID</th>
<th>RXnorm</th>
<th>ICD9</th>
<th>Different data types</th>
</tr>
</thead>
</table>

New institution

As long as our anchors appear in the new data as well...
## Generalizability/Portability

<table>
<thead>
<tr>
<th>LOINC</th>
<th>UMLS CUID</th>
<th>RXnorm</th>
<th>ICD9</th>
<th>Different data types</th>
</tr>
</thead>
</table>

**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

<table>
<thead>
<tr>
<th>LOINC</th>
<th>UMLS CUID</th>
<th>RXnorm</th>
<th>ICD9</th>
<th>Different data types</th>
</tr>
</thead>
</table>

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

Specified anchors

Automated suggestions

Detailed patient display

Ranked patient list

Code freely available clinicalml.org
Anchor Explorer V1

Specified anchors

Automated suggestions

Rapid iteration
~30 min to add a new phenotype

Detailed patient display

Ranked patient list

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

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

Yes the patient have an active malignancy? 

Unlikely  Unsure  Likely

<-- Previous  Abort  Next -->
Learned models: Nursing Home

Anchors:
- nursing facility
- nursing home
- nsg facility
- nsg home
- nsg. home

Highly weighted features:

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

### Medications
- senna
- mirtazapine
- colace
- maalox
- trazodone
- tums

### Pyxis
- vancomycin
- levofloxacin

### Unstructured text
- from staff
- at resident
- sent
- reported

- baseline changes
- nonverbal
- ams
- unwitnessed_fall
- confusion
- dnr
- full code
- g tube
- foley
- nh
- Unstructured text

Conditional independence assumption?
Learned models: Cardiac Etiology

Anchors

**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
- cp
- chest pain
- edema
- nstemi
- stemi
- ntg
- lasix
- nitro
- cmed
- chf exacerbation
- sob
- pedal edema

**Sex=M**
- aspirin
- clopidogrel
- Heparin Sodium
- Metoprolol Tartrate
- Morphine Sulfate
- Integrilin
- Labetalol

Unstructured text

Highly weighted features
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

- anticoagulated
- malignancy
- septic shock
- fall
- infection
- immunosuppressed
- nursing home
- cardiac
- pneumonia

Area under ROC
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 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 portability of anchors to other institutions
• General cohort selection tool

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