Shaping the Future of AI in Healthcare

Nigam Shah
Acknowledgements

Funding:
- Federal – NLM, NHLBI (Past: NIGMS, NHGRI, NINDS, NCI, FDA)
- Institutional – Dept. of Medicine, Population Health Sciences, Dean’s office, Stanford Hospital
- Fellowships – Med Scholars, Siebel Scholars Foundation, Stanford Graduate Fellowship, NSF, DoD
- Industry – Healogics, Janssen R&D, Oracle, Baidu USA, Amgen, Google, Apixio, CollabRx, Curai
- Philanthropic Gifts
Where I am coming from:

**Institutional**

- Associate CIO for Data Science
- Associate Dean of Research (Informatics, and our CTSA)

**Faculty**

1. The Green Button project
2. Stanford Medicine Program for AI in Healthcare
3. COVID-19 (last 90 days)
Let’s meet Laura

A teenager with systemic lupus erythematous, proteinuria, pancreatitis and positive for antiphospholipid antibodies

Let’s meet Vera

A 70 year old Asian woman with a history of hypertension and asthma. She is on metformin but has uncontrolled diabetes.
If \( \text{Risk} > \text{Th.} \) then \( \text{do} = \text{X} \)

Decide whether to act

Guide choice of how to act

A 70 year old Asian woman with a history of hypertension and asthma. She is on metformin but has uncontrolled diabetes.

A teenager with systemic lupus erythematosus, proteinuria, pancreatitis and positive for antiphospholipid antibodies.
If \((\text{Risk} > \text{Th.})\) then \((\text{do} = X)\)

- **Decide whether to act**
- **Guide choice of how to act**

**Digital Drive:**
Advancing Precision Health Takes Real Smarts—Artificially Speaking
The Stanford Program for AI Health Care

AI identifies risk of cholesterol-raising genetic disease
Stanford scientists and their collaborators have devised an algorithm to predict the risk of a disease that, untreated, can lead to heart attack or stroke.

[http://greenbutton.stanford.edu](http://greenbutton.stanford.edu)
Lessons from 200 million patient timelines

Take proactive action

Automate a tedious task
Lessons in converting timelines to datasets

Decisions made:
- About source and choice of features
- About how much to agonize over textual data
- About handling of time
- About defining an electronic phenotype
- About building a cohort
# Lessons in finding the right problems

<table>
<thead>
<tr>
<th>Science</th>
<th>Practice</th>
<th>Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classify</strong>&lt;br&gt;Finding subtypes of heart failure with preserved ejection fraction</td>
<td><strong>Who might be at high risk for a thromboembolism?</strong></td>
<td><strong>Who is burnt out?</strong></td>
</tr>
<tr>
<td><strong>Predict</strong>&lt;br&gt;Increased Monocyte Count is marker for bad prognosis in fibrotic diseases</td>
<td><strong>Which patients are likely to die in the next 3-12 months?</strong></td>
<td><strong>Who will be a no show?</strong></td>
</tr>
<tr>
<td><strong>Act/Treat</strong>&lt;br&gt;Colon tumors can be treated by allogeneic chimeric antigen receptor T-cell Rx</td>
<td><strong>What is a good second line drug to manage diabetes after metformin?</strong></td>
<td><strong>Request four back up nurses on Wed, for the Ortho OR.</strong></td>
</tr>
</tbody>
</table>
The Green Button project

Given a specific case, provides a report summarizing similar patients in Stanford’s clinical data warehouse, the common treatment choices made, and the observed outcomes.

An institutional review board approved study (IRB # 39709).

http://greenbutton.stanford.edu
Pilot phase completed, August 2019
patients with cryptogenic stroke

\[\text{var st} = \text{Intersect(OR(icd9=436, icd9=434), NOT(OR(icd9=393, icd9=394, icd9=397.1, icd9=397.9, icd9=398, icd9=246, icd9=424.9, icd9=V43, icd9=433.1, icd9=431, icd9=434.11, icd9=434.01)), \text{AGE (40 years, 90 years), VISIT TYPE="INPATIENT"}, \text{NOT(TEXT="thyroid diseases"}, \text{NOT(TEXT="heart valve prosthesis"}, \text{NOT(TEXT="disease of mitral valve"}, \text{NOT(TEXT="rheumatic heart disease"}))}\]

those that got diagnosed with \text{Afib}

\[\text{var afib} = \text{FIRST_MENTION(icd9=427.31)}\]

those with a cryptogenic stroke, and then \text{Afib} in 1 to 5 years

\[\text{SEQUENCE ($st, afib$)} + (\text{-5 years, -1 year})\]

1. Phenotype definition
2. Knowledge graph use
3. Cohort generation
4. Searching timelines

www.tinyurl.com/search-ehr
How ‘reliable’ are the results?

1. Comparing with two reference sets
   - Applies to the treatment effect estimation consults
   - 13-22% were “false discoveries”

2. Comparing across datasets (Truven, Optum)
   - Agreed 68-74% of the time
   - About the same rate as how often RCTs agree with each other

3. Comparing patient matching strategies
   - Agreed 79% of the time
1. Implementation
2. Rethinking utility
3. Safety, ethics, and system effects
4. Training and partnerships

Compassionate intelligence
Can machine learning bring more humanity to health care?

AI identifies risk of cholesterol-raising genetic disease
Stanford scientists and their collaborators have devised an algorithm to predict the risk of a disease that, untreated, can lead to heart attack or stroke.
Example research and perspectives

1. What is the individual level “cost” of group level algorithmic fairness?
2. Can we learn accurate ASCVD risk models for populations not covered by the current cohorts?
3. Can we learn generically useful and reusable patient representations?

1. The ‘best’ model isn’t always the most useful. (JAMA)
2. Machine-learning systems should reflect the ethical standards that guide other actors in health care. (NEJM)
3. Deployment cost—or the organizational effort required to integrate the output of a model into clinical workflow—should be a metric of evaluation. (Nature Medicine)
Palliative care and ACP: too little, too late

- 3.5 - 8% of inpatients are estimated to benefit from palliative care and advance care planning.
  - less than 50% are offered these options.

- Almost none (0.08%) are offered these options > 6 months before death.
  - most ACP notes written within one month of death

# days before death
ACP Workflow: 21 steps, 7 handoffs, 48 hrs
Label choice: Predicting a surrogate event

We built models to predict:

- 3-12 month mortality.
- Probabilistic forecasts of time to event.

Evaluation using held out test-sets

- AUC = 0.85 | AUPRC = 0.41
- AUC = 0.81 | AUPRC = 0.39
Before deploying

• Validity of the surrogate label
  • 235 patients in a blinded prospective study.

• Model’s predictions agree with experts’ prognosis judgments for both 0-3, and 0-12 months.

Ensure that the increased workload is manageable

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>Future</th>
<th>% increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Medicine</td>
<td>343</td>
<td>583</td>
<td>69%</td>
</tr>
<tr>
<td>Total</td>
<td>1272</td>
<td>1512</td>
<td>19%</td>
</tr>
</tbody>
</table>
Before deploying

Establishing a baseline

A heuristic of “3 or more admissions”, flags 21% of cases that are in need for advanced care planning at a cost of screening 2.46 cases to find one true case.

Quantifying improvement

• At 21% recall, the model prompts for screening of 1.08 admissions (cuts work into less than half).

• Fixing the number needed to screen at 2 admissions, the model has 85% recall (i.e. finds 4x cases).

• The model finds cases 58 days earlier than the “3 or more admissions” heuristic.
Is there utility, given cost and benefit of actions?

<table>
<thead>
<tr>
<th>Utility</th>
<th>Desc</th>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{tp}$</td>
<td>Utility for true positives (ACP is appropriate and provided)</td>
<td>-28,613</td>
<td>Gade et al. Net savings of 4855 * inflation multiplier, subtracted from $U_{fn}$</td>
</tr>
<tr>
<td>$U_{fn}$</td>
<td>Utility for false negatives (ACP is appropriate but not provided)</td>
<td>-37,085</td>
<td>Gade et al. Original value of 21252 * inflation multiplier of 1.745</td>
</tr>
<tr>
<td>$U_{fp}$</td>
<td>Utility for false positives (ACP is not appropriate but provided)</td>
<td>-14,970</td>
<td>$U_{tn}$ plus inflation adjusted cost of intervention.</td>
</tr>
<tr>
<td>$U_{tn}$</td>
<td>Utility for true negatives (ACP is not appropriate and not provided)</td>
<td>-11,646</td>
<td>Per capita spend in US, 2018, Peterson-Kaiser</td>
</tr>
</tbody>
</table>
Realized utility, given work capacity constraints

$2^4 = 16$ possible combinations

Best case trade off

Worst case trade off
Bottom line: is my model useful?

- Impact of rejecting recommended ACP
- Impact of capacity constraints
- Impact of loss to discharge
- Impact of “outpatient rescue”
We need a “delivery science” for AI/ML solutions

2: Model Dev: How do we get the best $f: X \rightarrow Y$?
- Does using textual content help?
- How do we train fair models?
- Can we use $f: X \rightarrow Y$ in the real world?
  - Can we get the data by 5 am, to make prediction by 6 am?

4: Running system = model applied to each case + execution of workflow.
- Evaluate the impact of the running system
- Maintenance is a liability – who will carry the pager?
- Monitoring is unexplored

1: Use case
- What clinical outcome(s) are we trying to affect?
- What action would someone take?
- Who will take that action?
- Technical formulation $f: X \rightarrow Y$ subject to...
  - use an existing equation vs. learn a new equation.

3: Utility assessment
- Given the actions and its benefit, is there net utility?
- Deployment design
- Given ‘work capacity’, what net-benefit can we realize?
  - Do we require new workflows?

Technical formulation

Use case

Model development

Technical validation

Deployment design

Utility assessment

Running system

Maintain, monitor

Prospective study
**Weak Supervision**

Use cheaper label sources to build training sets

No hand-labeled training data

more weak supervision info

**snorkel**
https://www.snorkel.org/

Transform off-the-shelf ontologies, etc. into *labeling functions*

---

**She reports that she had contact with +COVID patient on Feb 8**

I am testing for COVID-19.

---

**Rules**

Weakly Supervised

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.6</td>
<td>69.1</td>
<td>75.2</td>
</tr>
</tbody>
</table>

**Inkfish: Weakly Supervised Biomedical Entity Tagging**

<table>
<thead>
<tr>
<th>Entity</th>
<th>Domain</th>
<th>Rule F1</th>
<th>Inkfish F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorder</td>
<td>EHR</td>
<td>72.4</td>
<td>76.6</td>
</tr>
<tr>
<td>Drug</td>
<td>EHR</td>
<td>82.8</td>
<td>86.9</td>
</tr>
<tr>
<td>Disease</td>
<td>Literature</td>
<td>75.7</td>
<td>79.7</td>
</tr>
<tr>
<td>Chemical</td>
<td>Literature</td>
<td>79.8</td>
<td>89.4</td>
</tr>
</tbody>
</table>

**+4.1 to 9.6 F1 Improvement**
Profiling presenting symptoms of patients screened for SARS-CoV-2

Alison Callahan*, Jason A. Fries*, Saurabh Gombar, Birju Patel, and Nigam H. Shah (*equal contributors)

There is high interest in characterizing the presenting symptoms of individuals with COVID-19 to inform diagnosis and triage decisions as well as identify patients at risk of serious complications. As one of the many efforts in Stanford Medicine’s data science response to the current pandemic, we developed a text processing system to identify clinical observations in the notes written by care providers when screening patients for COVID-19.

We’d need about 20 symptoms to get $P(+ve \mid \text{symptoms}) > 0.8$
Viral RNA detected for up to 30 days

Gombar et al, Persistent detection of SARS-CoV-2 RNA in patients and healthcare workers with COVID-19 accepted in the Journal of Clinical Virology
More at


email: nigam@stanford.edu