Stanford HEALTH CARE

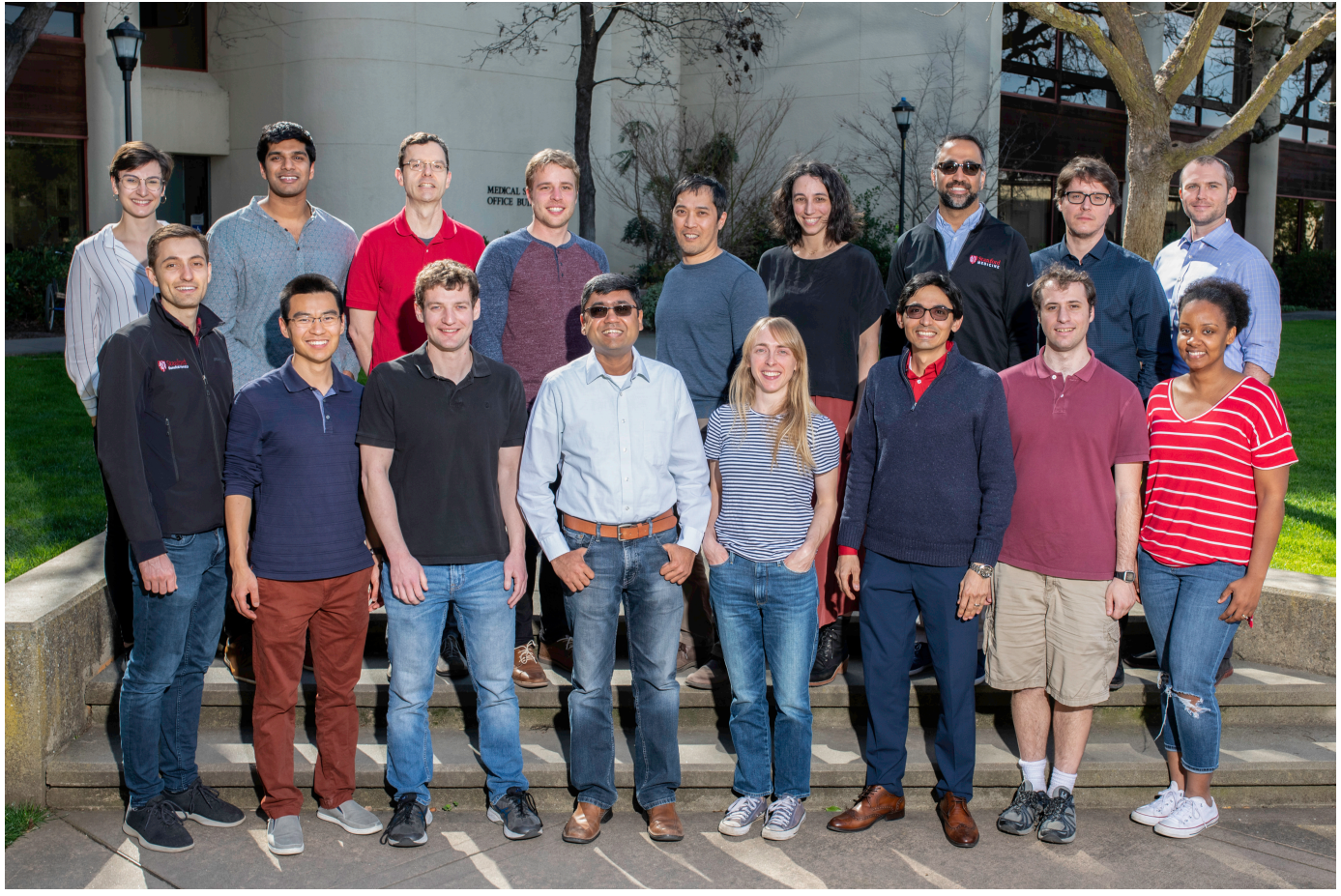
Shaping the Future of AI in Healthcare

Nigam Shah

SHC DIGITAL SOLUTIONS

Stanford | MEDICINE

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 erythematosus, proteinuria, pancreatitis and positive for antiphospholipid antibodies

www.webmd.com/lupus/picture-of-acute-systemic-lupus-erythematosus

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.

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

www.webmd.com/lupus/picture-of-acute-systemic-lupus-erythematosus

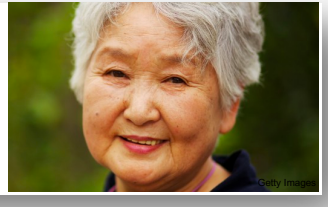
Decide whether to act

If (Risk > Th.)

then (do = X)

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.



DIGITALLY DRIVEN

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.

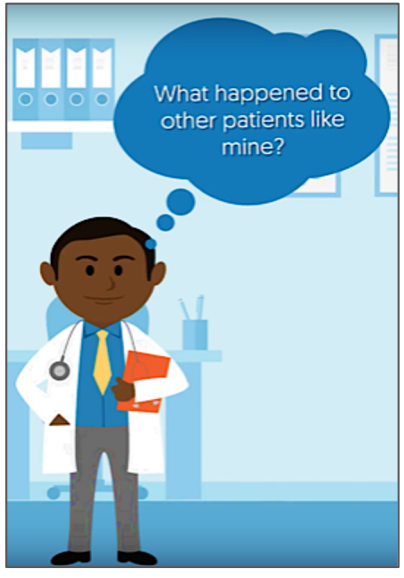
Decide whether to act

If (Risk > Th.)

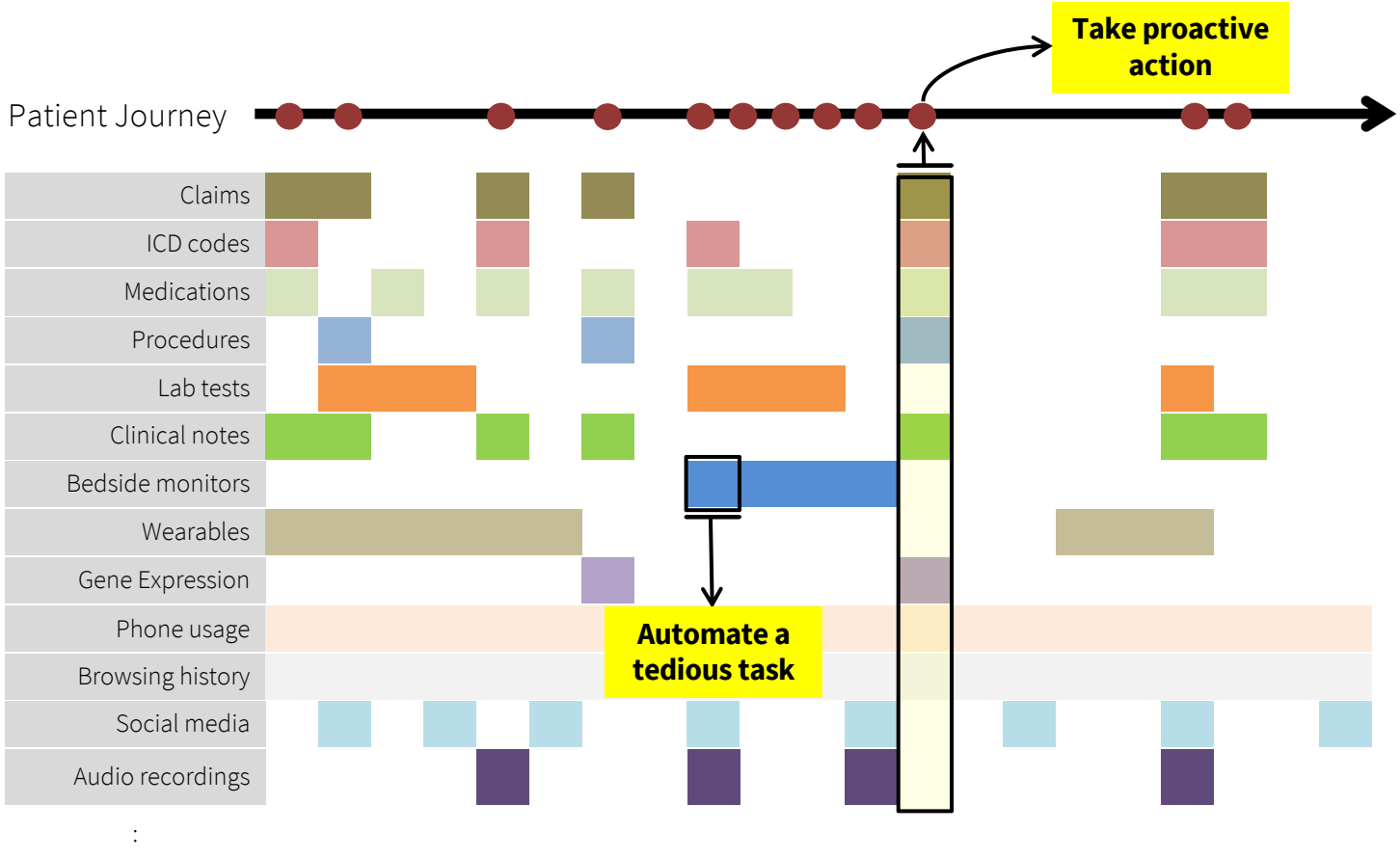
then (do = X)

Guide choice of how to act

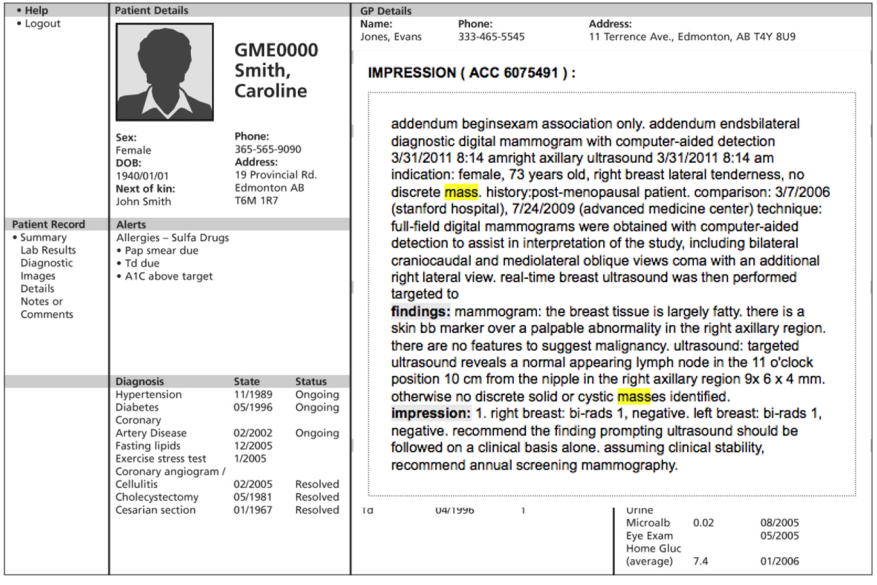
http://greenbutton.stanford.edu

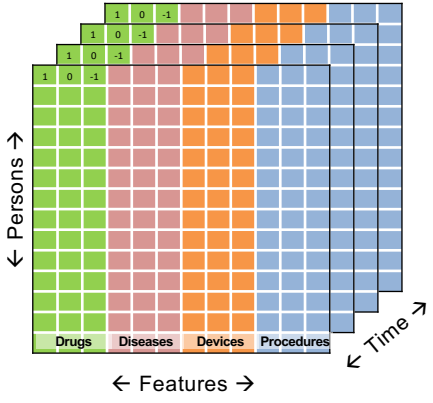


Lessons from 200 million patient timelines



Lessons in converting timelines to datasets





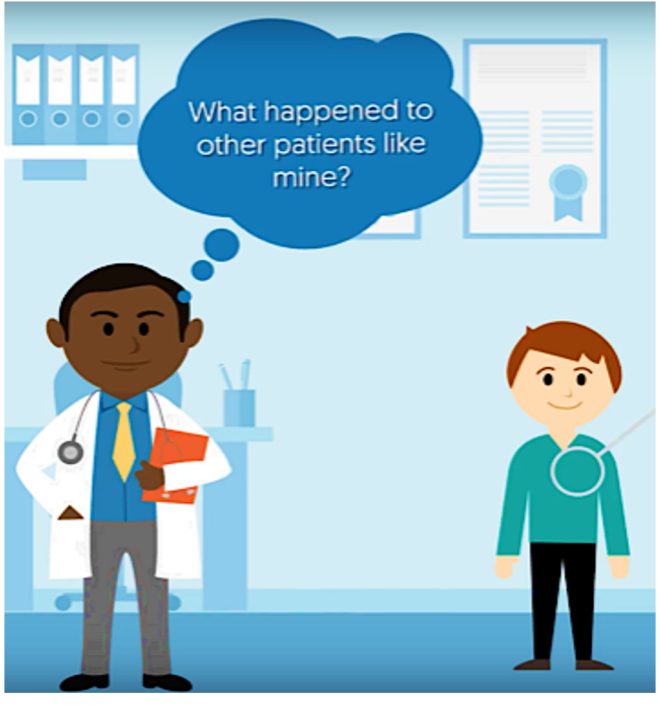
Decisions made:

- 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

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

The Green Button project

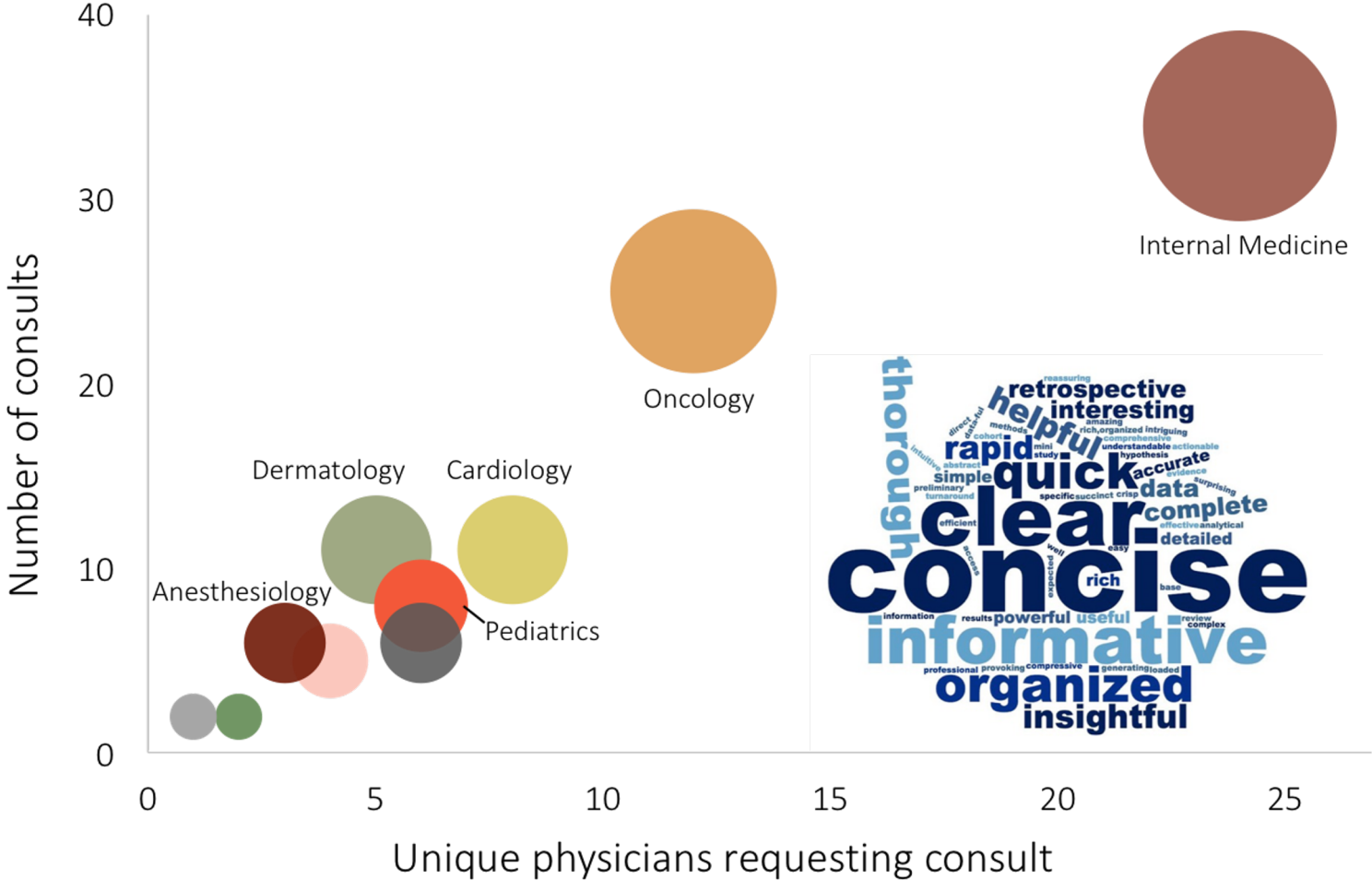


Given a specific case, provides a report summarizing similar patients in Stanford's clinical data warehouse, the common treatment choices made, and the observed outcomes.

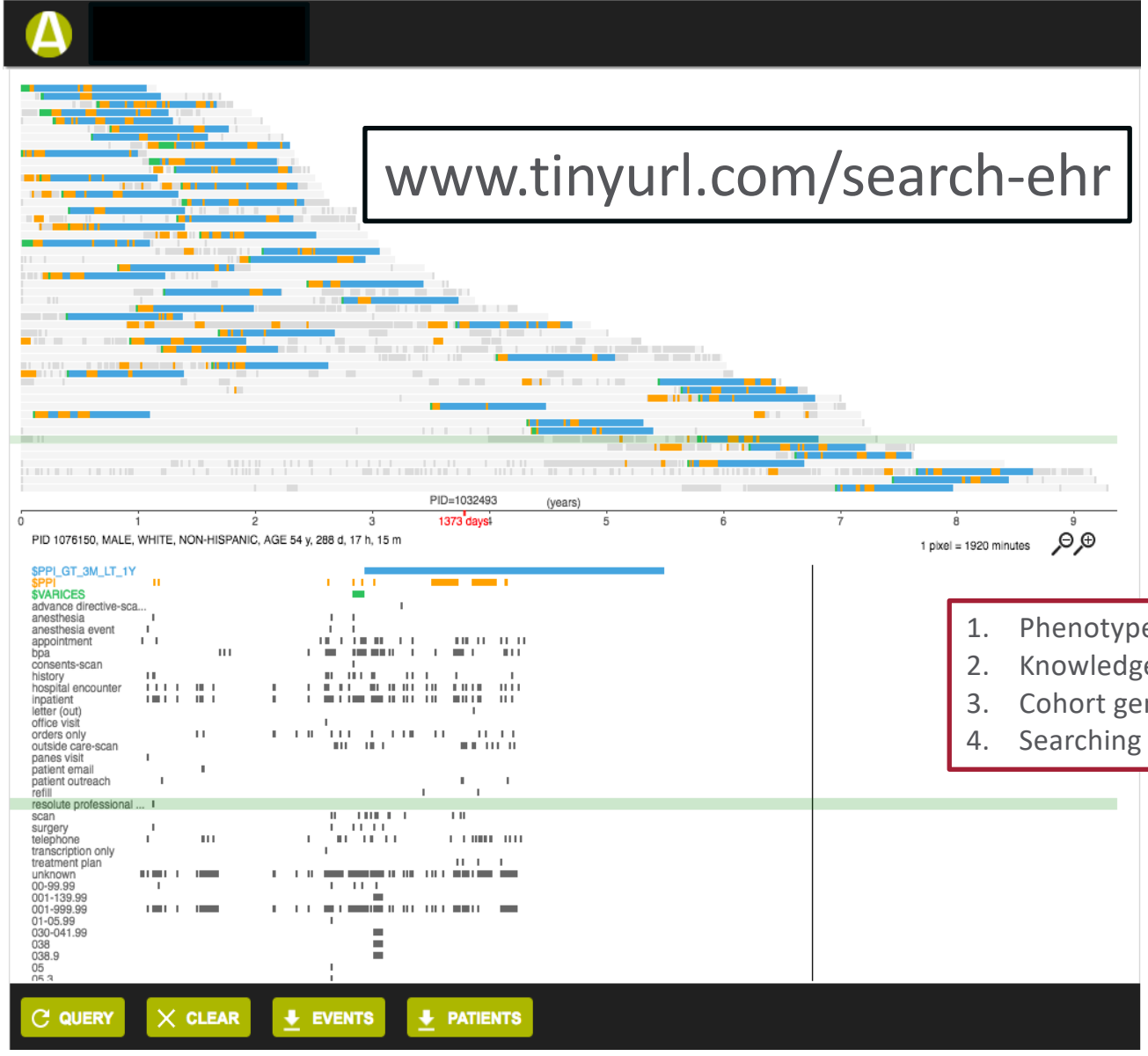
An institutional review board approved study (IRB # 39709).

http://greenbutton.stanford.edu

Pilot phase completed, August 2019



SHC



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

SHC DIC

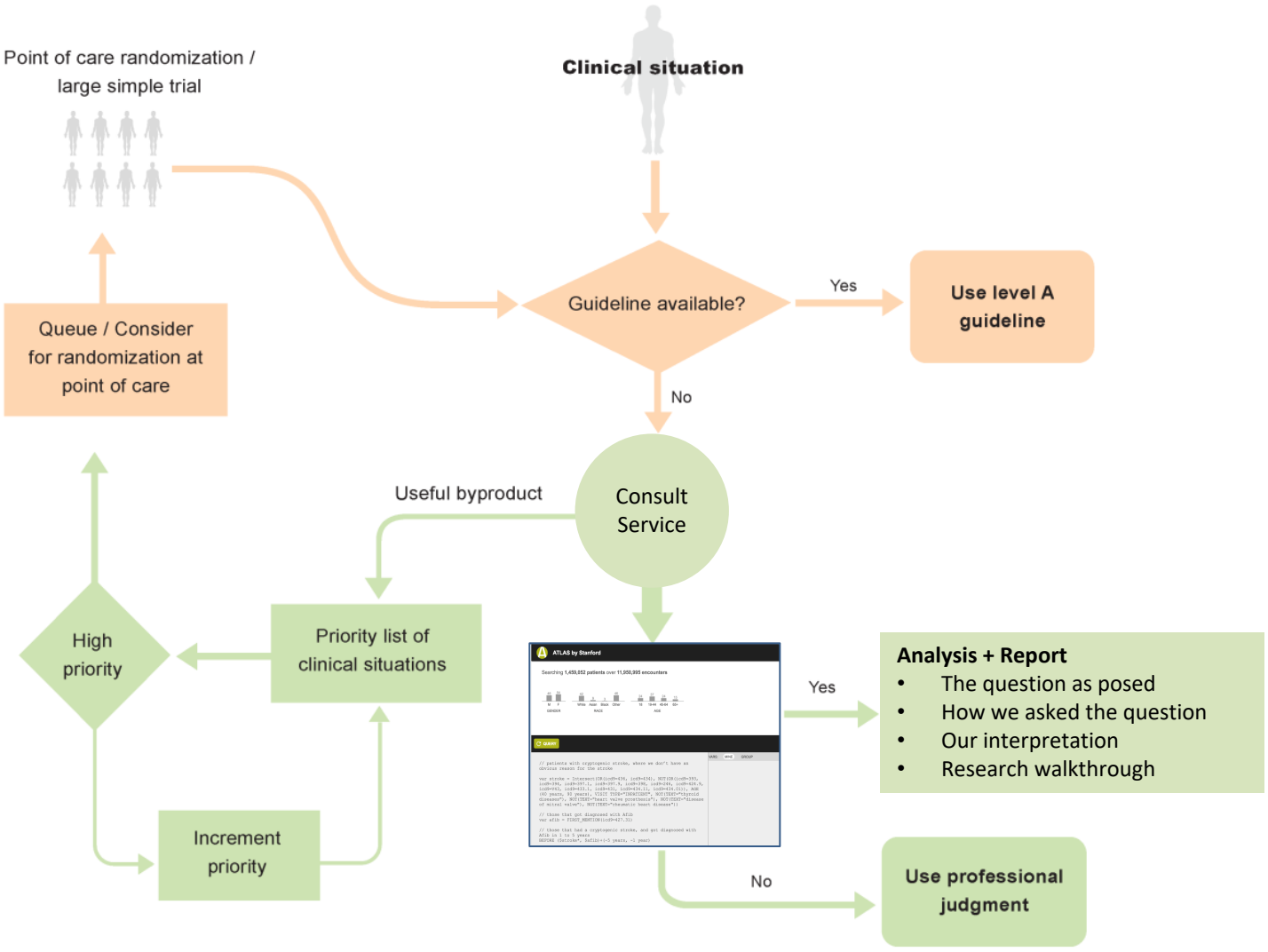
How ‘reliable’ are the results?

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

Green button → Informatics Consult



Stanford Medicine Program for AI in Healthcare

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

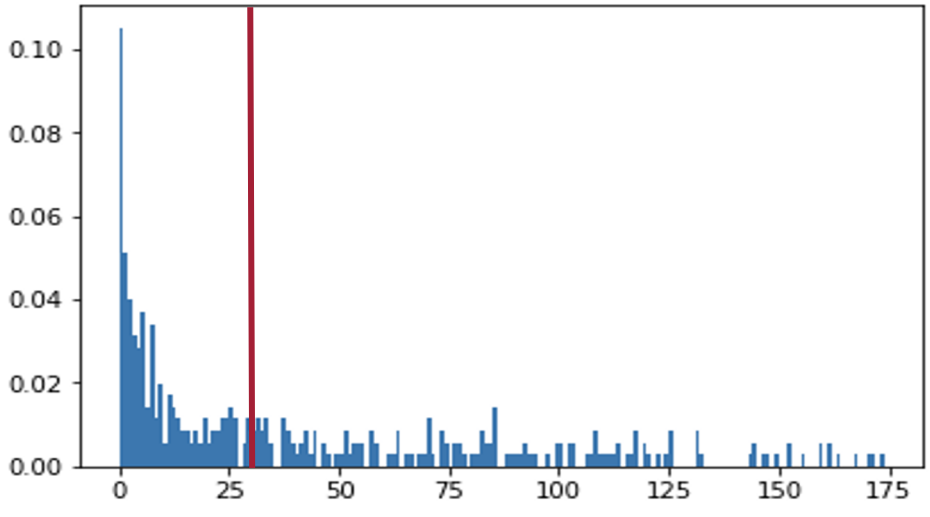
1. What is the individual level **“cost” of group level algorithmic fairness**?
2. Can we **learn accurate ASCVD risk models** for populations not covered by the current cohorts?
3. Can we learn generically useful and **reusable patient representations**?

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**)

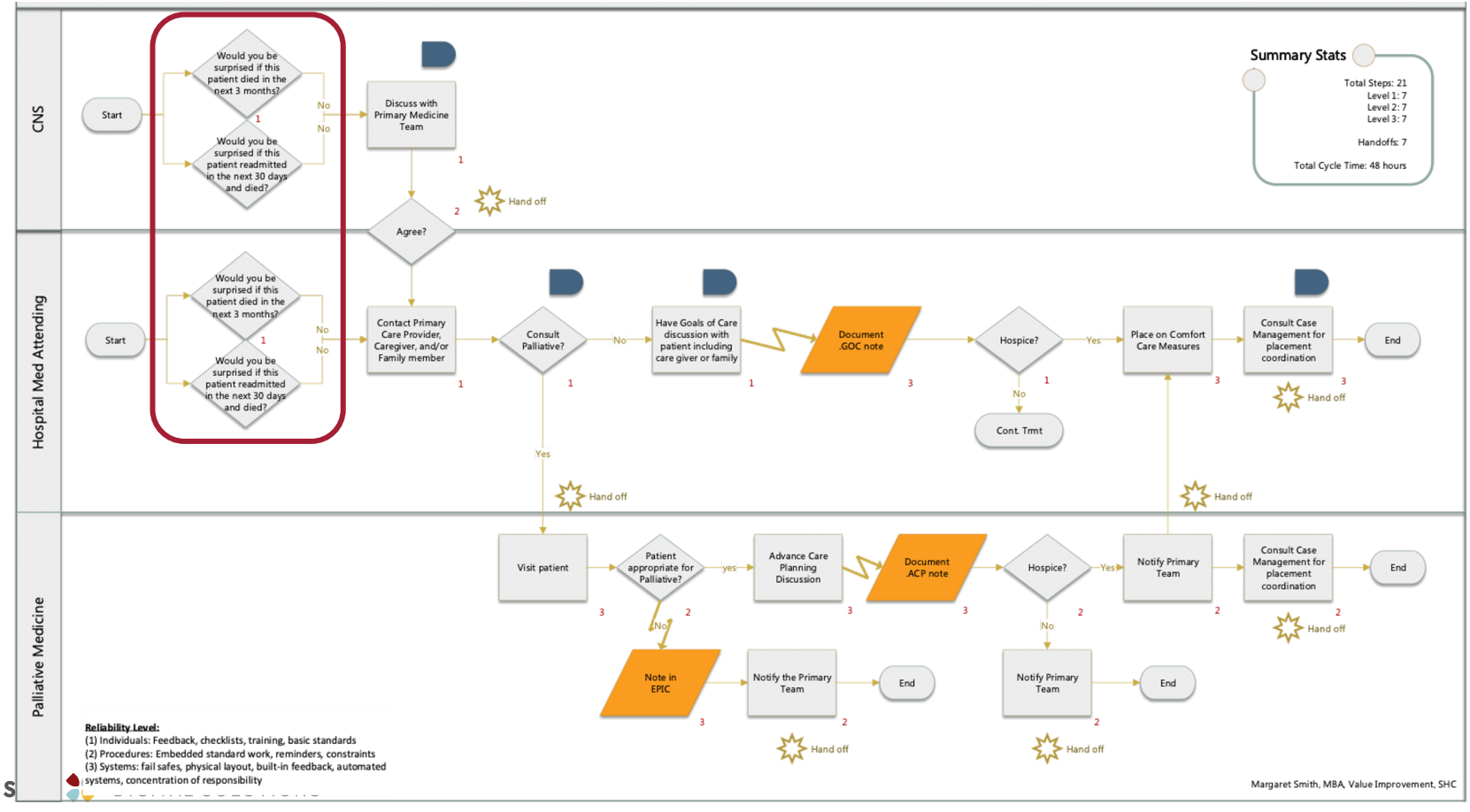
Palliative care and ACP: too little, too late

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

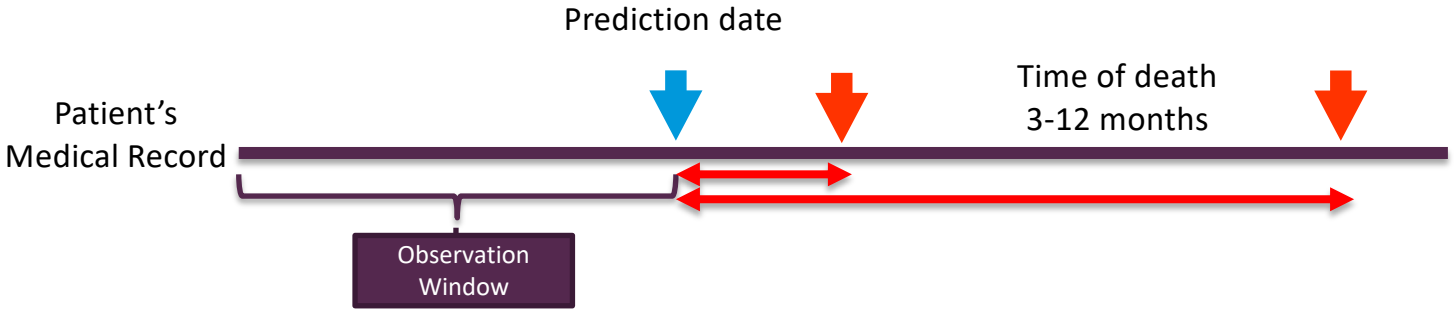
- Almost none (0.08%) are offered these options > 6 months before death.
  - most ACP notes written within one month of death



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

	Current	Future	% increase
General Medicine	343	583	69%
Total	1272	1512	19%

Before deploying

Establishing a baseline

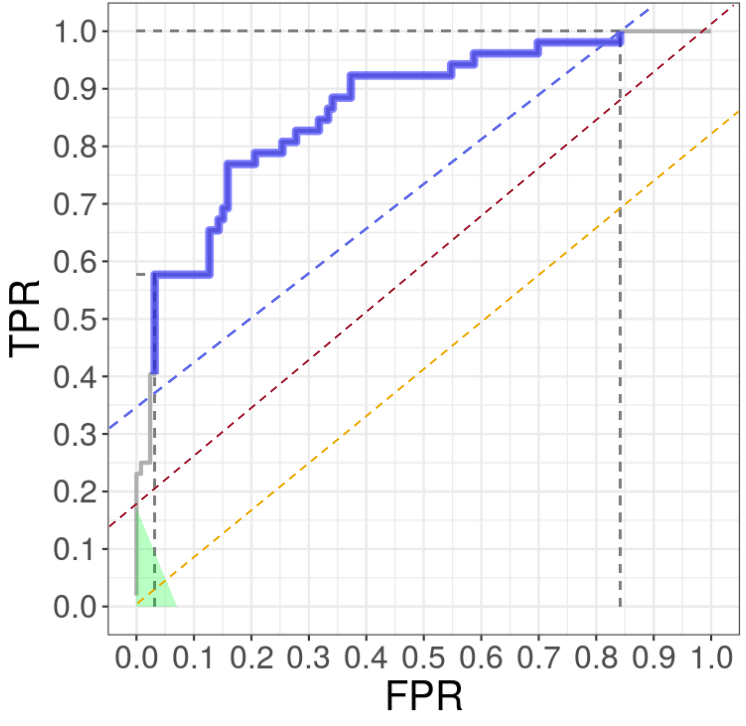
A heuristic of “3 or more admissions”, flags 21% of cases that are in need for advanced care planning at a cost of screening 2.46 cases to find one true case.

Quantifying improvement

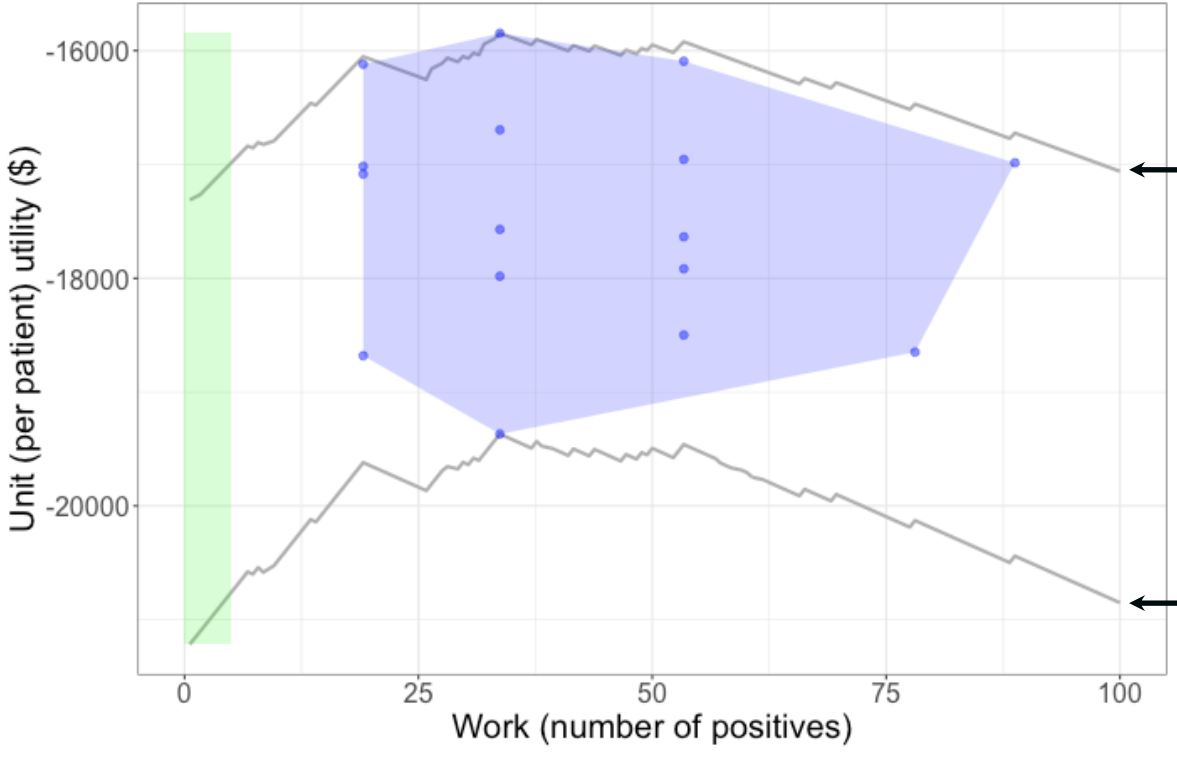
- At 21% recall, the model prompts for screening of 1.08 admissions (cuts work into less than half).
- Fixing the number needed to screen at 2 admissions, the model has 85% recall (i.e. finds 4x cases).
- The model finds cases 58 days earlier than the “3 or more admissions” heuristic.

Is there utility, given cost and benefit of actions?

Utility	Desc	Value	Source
U^{tp}	Utility for true positives (ACP is appropriate and provided)	-28,613	Gade et al. Net savings of 4855 * inflation multipler, subtracted from U^{fn}
U^{fn}	Utility for false negatives (ACP is appropriate but not provided)	-37,085	Gade et al. original value of 21252 * inflation multiplier of 1.745
U^{fp}	Utility for false positives (ACP is not appropriate but provided)	-14,970	U^{tn} plus inflation adjusted cost of intervention.
U^{tn}	Utility for true negatives (ACP is not appropriate and not provided)	-11,646	Per capita spend in US, 2018, Peterson-Kaiser



Realized utility, given work capacity constraints



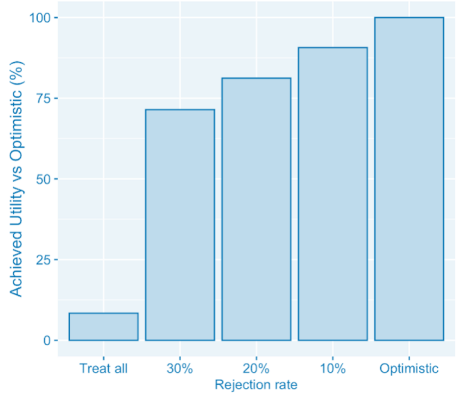
Best case trade off

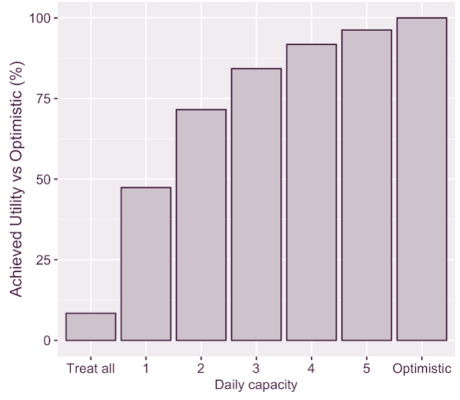
2^4 = 16 possible combinations

Worst case trade off

Bottom line: is my model useful?

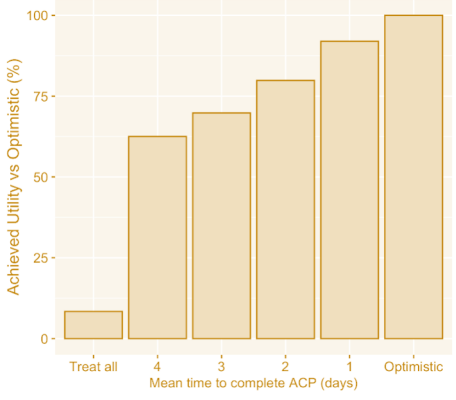
Impact of rejecting recommended ACP

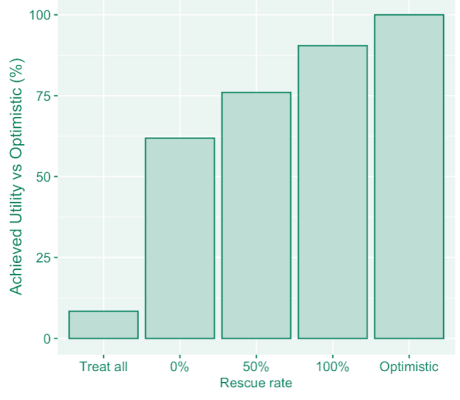




Impact of capacity constraints

Impact of loss to discharge



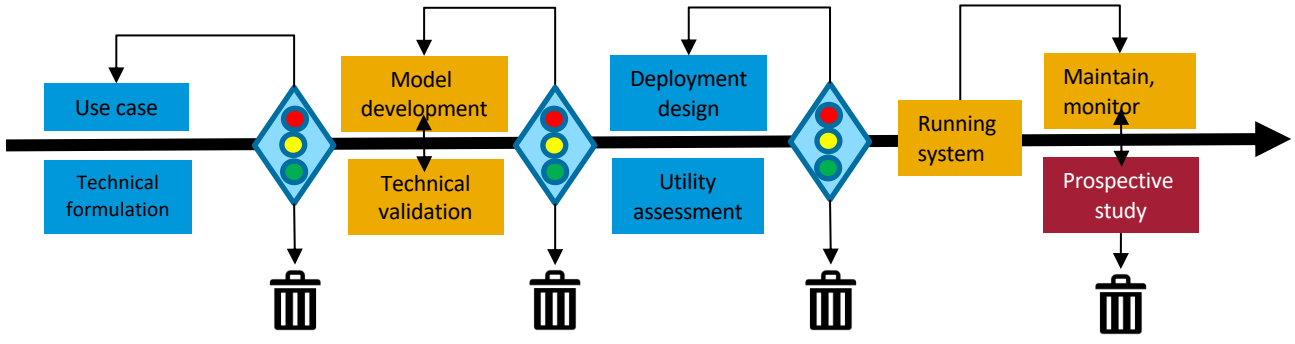


Impact of "outpatient rescue"

We need a “delivery science” for AI/ML solutions

2: Model Dev: How do we get the best f: X -> Y?
Does using textual content help?
How do we train fair models?
Can we use f: X -> 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 -> Y subject to…
use an existing equation vs. learn a new equation.

3: Utility assessment
Given the 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?

Methods development + COVID-19

Weak Supervision

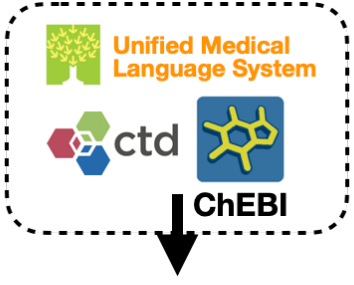
Use cheaper label sources to build training sets

No hand-labeled training data

more weak supervision info

snorkel

https://www.snorkel.org/



Transform off-the-shelf ontologies, etc. into

labeling functions

Entity	Domain	Rule F1	Inkfish F1
Disorder	EHR	72.4	**76.6**
Drug	EHR	82.8	**86.9**
Disease	Literature	75.7	**79.7**
Chemical	Literature	79.8	**89.4**

+4.1 to 9.6 F1 Improvement

Inkfish: Weakly Supervised Biomedical Entity Tagging

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

I am testing for COVID-19.
```

Rules

Precision **82.6** Recall **69.1** F1 **75.2**

Weakly Supervised

Precision **87.2** Recall **74.5** F1 **80.4**

www.tinyurl.com/symptom-profile

Profiling presenting symptoms of patients screened for SARS-CoV-2

Nigam Shah
Apr 3 · 2 min read

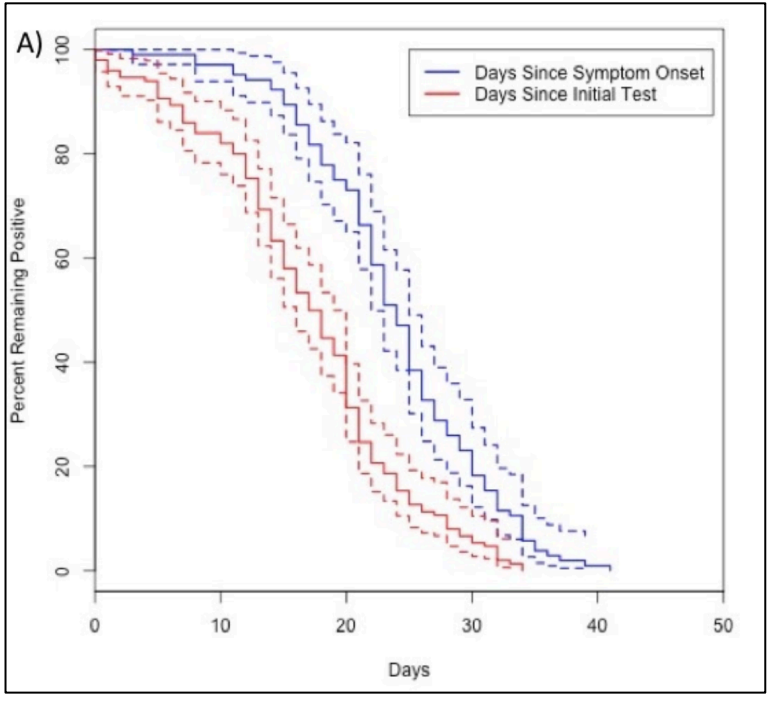
Alison Callahan, Jason A. Fries*, Saurabh Gombar, Birju Patel, and Nigam H. Shah (*equal contributors)*

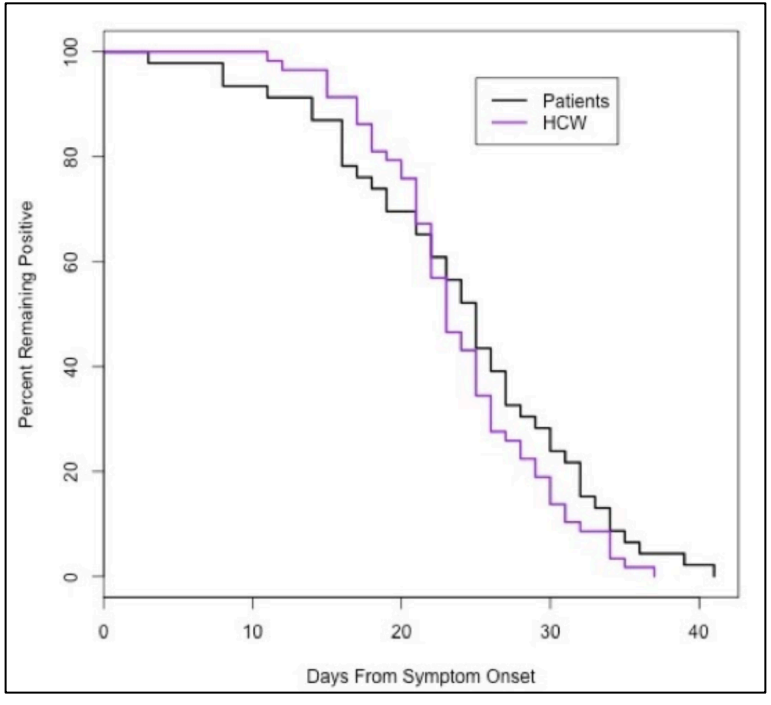
There is high interest in characterizing the presenting symptoms of individuals with COVID-19 to inform diagnosis and triage decisions as well as identify patients at risk of serious complications. As one of the many efforts in Stanford Medicine's data science response to the current pandemic, we developed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

Clinical observation	Count (observation)	Count (observation & +ve)	Count (observation & -ve)	P(observation)	P(observation | +ve)	P(observation | -ve)	P(+ve | observation)	P(-ve | observation)
cough	577	51	526	0.645	0.797	0.633	0.088	0.912
dyspnea	526	41	485	0.588	0.641	0.584	0.078	0.922
febrile	396	44	352	0.442	0.688	0.424	0.111	0.889
sore throat	244	13	231	0.273	0.203	0.278	0.053	0.947
chest pain	129	11	118	0.144	0.172	0.142	0.085	0.915
congestion	124	7	117	0.139	0.109	0.141	0.056	0.944
rash	109	6	103	0.122	0.094	0.124	0.055	0.945
nausea and vomiting	101	8	93	0.113	0.125	0.112	0.079	0.921
fatigue	99	12	87	0.111	0.188	0.105	0.121	0.879
myalgia	98	10	88	0.109	0.156	0.106	0.102	0.898
influenza	92	7	85	0.103	0.109	0.102	0.076	0.924
tachycardia	91	8	83	0.102	0.125	0.100	0.088	0.912
acetaminophen	81	10	71	0.091	0.156	0.085	0.123	0.877
pain	81	5	76	0.091	0.078	0.091	0.062	0.938
chills	80	14	66	0.089	0.219	0.079	0.175	0.825
hypertension	77	5	72	0.086	0.078	0.087	0.065	0.935
malaise	77	12	65	0.086	0.188	0.078	0.156	0.844
headache	76	9	67	0.085	0.141	0.081	0.118	0.882

We'd need about 20 symptoms to get P(+ve | 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

1. http://shahlab.stanford.edu/**greenbutton**
2. http://shahlab.stanford.edu/**paihc**
3. http://shahlab.stanford.edu/**covid19**

email: nigam@stanford.edu