Shaping the Responsible Adoption of AI in Healthcare

Nigam Shah
Chief Data Scientist, Stanford Healthcare
Professor of Medicine, Stanford University

Where I am coming from

Professor of Medicine @ SOM

Research … ways to bring AI into clinical use safely, ethically and cost effectively.

Teach … data science in medicine for the Biomedical Informatics (BMI), Masters in Clinical Information Management (MCIM), the Clinical Informatics, and two Stanford online programs

Consult … the organization in shaping the Stanford Medicine data science ecosystem for clinical and translational research

Chief Data Scientist @ SHC

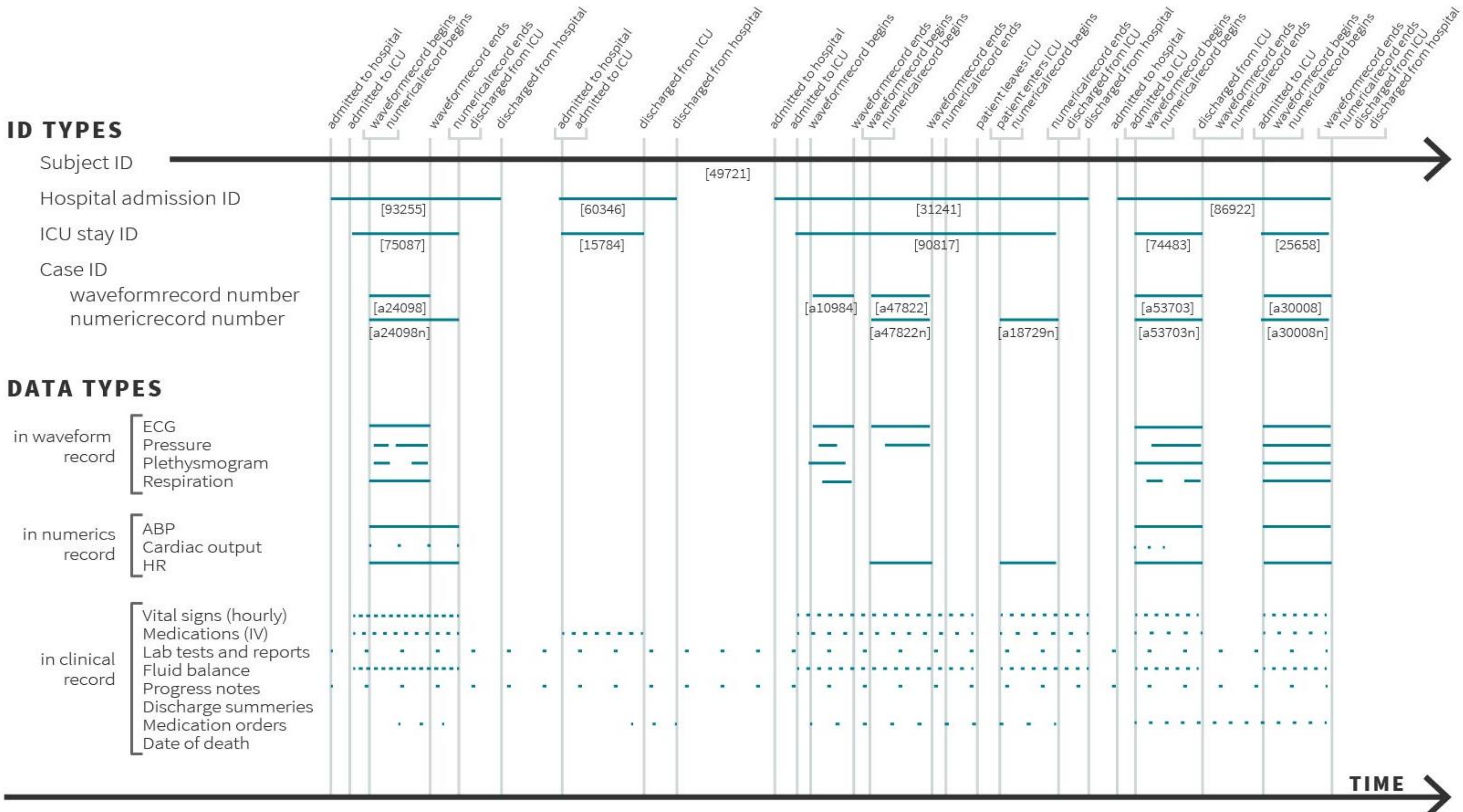
Lead … the team bringing predictive algorithms and AI into the healthcare environment.

Build … the delivery science to assess usefulness, reliability and fairness of AI projects.

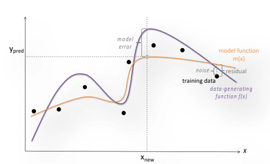
Serve … the organization with cross-functional leadership to effectively use data science.

Represent … Stanford Health Care to foster our reputation as a world leader data science.

We use data from patient timelines to build models



Models classify, predict, or recommend in service of the science, practice or delivery of care



Models of varying complexity:

- Logistic Regression
- Random Forest
- Deep Neural Network

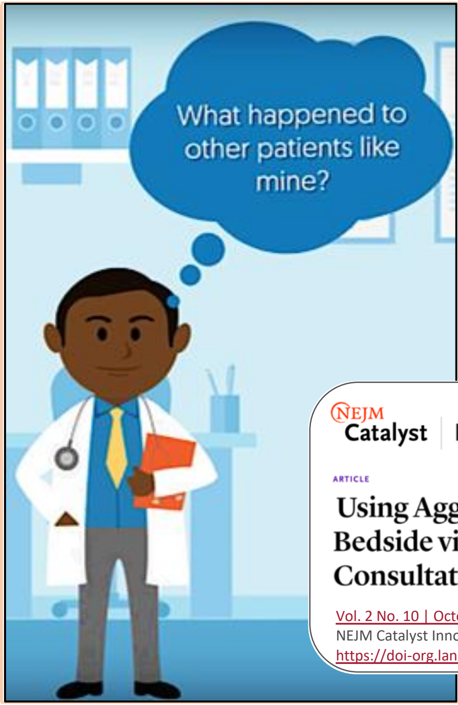
Whether to treat

How to treat

	Science	Practice	Delivery
Classify (Diagnose)	. . .	. . .	. . .
Predict (Prognose)	. . .	. . .	. . .
Recommend (Decide)	. . .	. . .	. . .

The typical consultation request

	Science	Practice	Delivery
Classify (Diagnose)	…	…	…
Predict (Prognose)	…	…	
Recommend (Decide)	…		



NEJM Catalyst | Innovations in Care Delivery

ARTICLE

Using Aggregate Patient Data at the Bedside via an On-Demand Consultation Service

Vol. 2 No. 10 | October 2021
NEJM Catalyst Innovations in Care Delivery 2021; 10
https://doi-org.laneproxy.stanford.edu/10.1056/CAT.21.0224

Why supporting such consultations matters

Deciding without data

Jeffrey R Darst [1], Jane W Newburger, Stephen Resch, Rahul H Rathod, James E Lock

Affiliations + expand

PMID: 20653700 PMCID: PMC4283550 DOI: 10.1111/j.1747-0803.2010.00433.x Paperpile

Free PMC article

"During the 7.5 days, 1188 decisions (158/day) were made. Almost 80% of decisions were deemed by the physicians to have no basis in any prior published data an < 3% of decisions were based on a study specific to the question at hand."

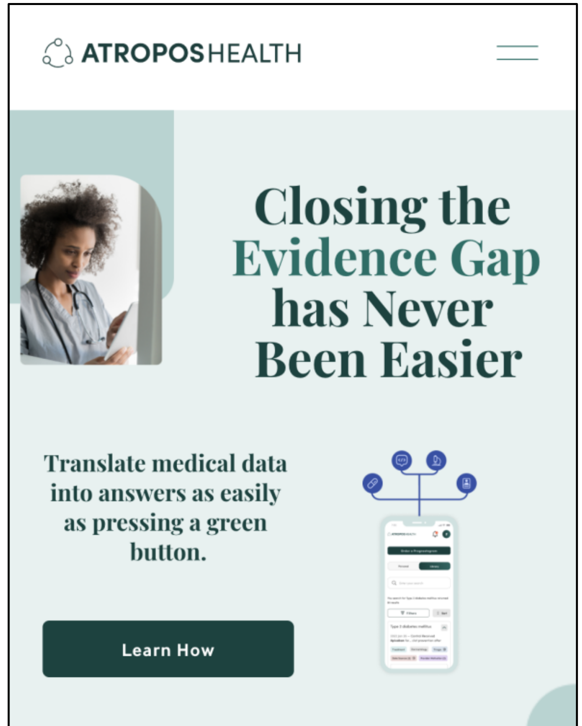
Spin out – Atropos Health, in 2021

NEJM Catalyst | Innovations in Care Delivery

ARTICLE

Using Aggregate Patient Data at the Bedside via an On-Demand Consultation Service

Vol. 2 No. 10 | October 2021
NEJM Catalyst Innovations in Care Delivery 2021; 10
https://doi-org.laneproxy.stanford.edu/10.1056/CAT.21.0224



www.atroposhealth.com

How Medical Records Can Close the Information Gap in Patient Care

by Nigam Shah and John D. Halamka

May 17, 2023

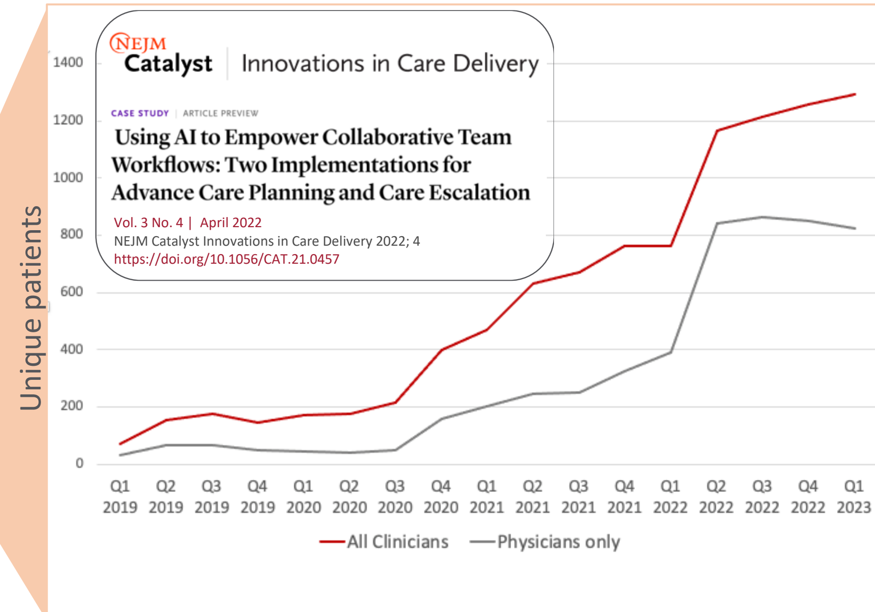
Benjamin Rondel/Getty Images

Summary. In dealing with many cases, doctors lack comparative real-time evidence and are forced to make decisions in spite of unknown variables that can dramatically alter outcomes. Such evidence gaps happen every day, particularly for patients with multiple conditions,... **more**

www.tinyurl.com/HBR-gap

A typical predict-n-act set up

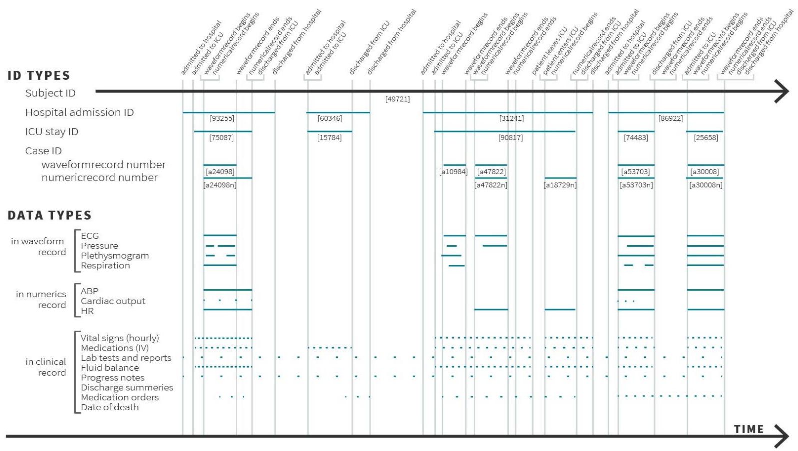
	Science	Practice	Delivery
Classify (Diagnose)	. . .	. . .	. . .
Predict (Prognose)	. . .	. . .	
Recommend (Decide)	. . .	. . .	. . .

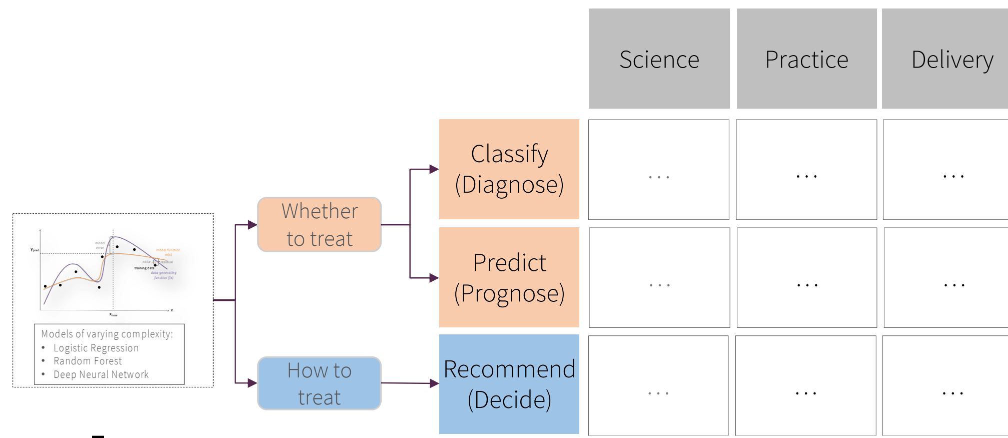


Examples

- Predicting mortality to improve advance care planning
- Classifying ischemic vs. hemorrhagic stroke for prioritizing air ambulance transport
- Predicting long term outcomes after pulmonary embolism using imaging and EHR data
- Multimodal models for recurrence risk in surgically resectable colorectal cancer, to guide adjuvant therapy
- Opportunistic ASCVD risk estimation, using CT images and EMR data
- Predicting no-shows for providing transportation support
- Classifying presence of undiagnosed disease
  - Familial hypercholesterolemia – to order sequencing
  - Peripheral artery disease – to order ABI measurement
- Predicting length of stay, readmissions, bed-demand etc. …

Model stratifies by risk; value comes from taking responsive action





If ($\boxed{\text{Risk}}$ > Th.)

then (do = $\boxed{\text{X}}$)

A framework for making predictive models useful in practice

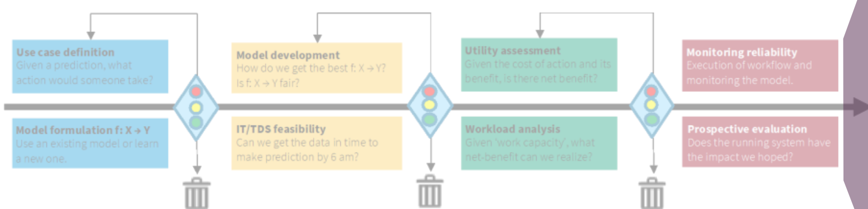
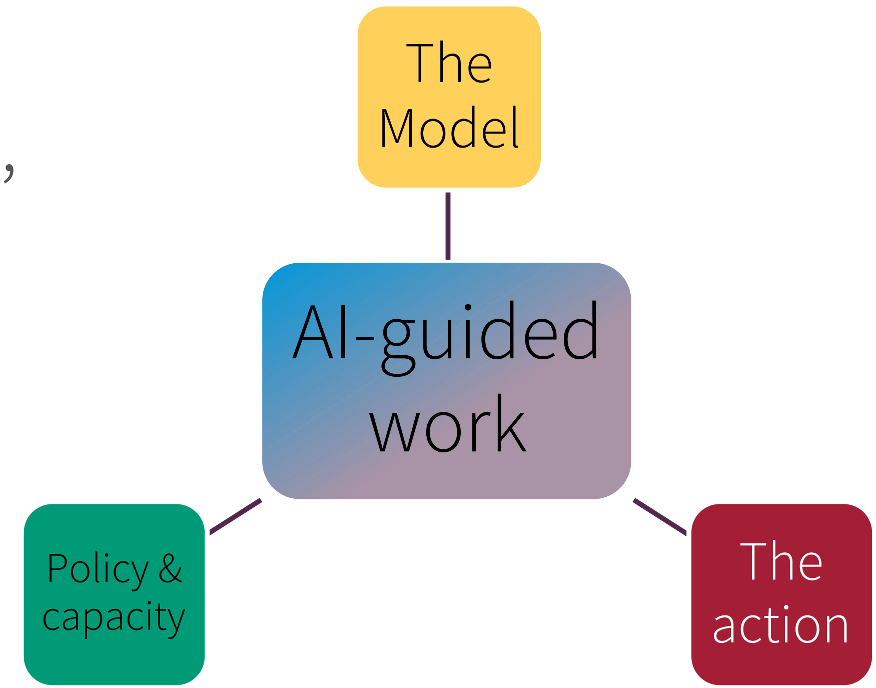


Fig. 4: Development and evaluation of a predictive model throughout its life cycle

1	**Needs Statement** That specifies the problem a model is expected to solve.	Oct 2015
2	**Use case definition** Given a prediction, what action would we take?	Feb 2016
3	**Model formulation f: X → Y** Use an existing model or a new one.	
4	**Model dev: Performance** How do we get the best f: X → Y?	Feb 2017
5	**Model dev: Fairness** Is f: X → Y fair?	
6	**IT feasibility** How do we get the data in time to make predictions?	
7	**[Workflow / Org / App Integration]** How do we get the model output back into care workflow?	Jan 2019
8	**Utility assessment** Given the cost of action, is there net benefit?	Mar 2020
9	**Workload analysis** Given 'work capacity', what net-benefit can we realize?	
10	**Monitoring (workflow, model)** Making predictions and monitoring the model as well as workflow.	
11	**Prospective evaluation** Does the system have the impact we hoped?	Mar 2022
12	**Ethical concerns** Surfaced by stakeholder interviews	Apr 2023
13	**Business case** Enterprise value given the model, the intervention, and patient mix	

Jung et al 2020. doi:10.1093/jamia/ocaa318

There is an interplay among models, capacity, and actions we take



Viewpoint

August 8, 2019

Making Machine Learning Models Clinically Useful

Nigam H. Shah, MD, PhD[1]; Arnold Milstein, MD, MPH[2]; Steven C. Bagley, PhD[3]

» Author Affiliations | Article Information

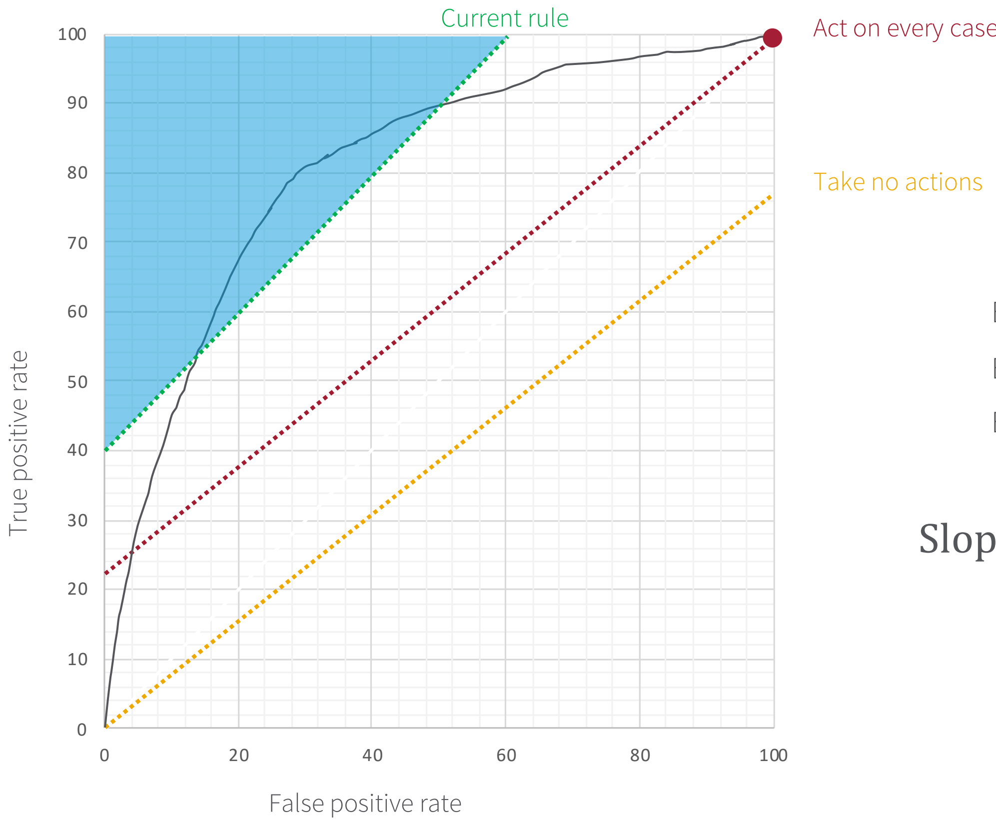
Recommendations for "building good models"

Model Reporting Guideline	Use Case	Model Formulation	Model Dev.	Model Dev: Fairness	Practical Feasibility	Utility Assessment	Deployment Design	Execution of Workflow	Monitoring of model	Prospective Evaluation
Model Cards	8	5	29	**9**	1	0	0	0	0	0
Model Facts Labels	**10**	7	9	0	1	1	0	0	2	1
Guidelines	7	6	31	1	0	1	0	0	1	0
MI-CLAIM	4	3	29	3	0	1	0	0	0	1
MINIMAR	4	4	18	5	0	0	0	0	0	0
TRIPOD	7	**9**	**53**	1	0	**3**	0	0	3	2
CONSORT-AI	**10**	3	23	6	1	0	0	0	2	**19**
SPIRIT-AI	9	3	17	1	2	0	0	0	2	18
Trust and Value	4	0	9	0	2	1	0	0	4	2
ML Test Score	0	0	12	4	1	0	0	**2**	**17**	0
Risk	2	4	24	0	0	1	0	0	2	6
STARD	8	2	37	6	0	1	0	0	0	0
ABCD	1	3	27	0	0	1	0	0	0	0
CHARMS	5	**9**	42	1	2	0	0	0	1	4
PROBAST	4	6	41	0	1	1	0	0	1	0
Total	14	14	104	10	5	4	0	2	19	25

Stanford MEDICINE

Jonathan Lu et al. "Assessment of Adherence to Reporting Guidelines by Commonly Used Clinical Prediction Models."
doi:10.1001/jamanetworkopen.2022.27779

ROC, Utility, and indifference lines



	Positive (r_p=5%)	Negative (r_n=95%)
Positive	$u_{tp} * r_p * TPR$	$u_{fp} * r_n * FPR$
Negative	$u_{fn} * r_p * (1-TPR)$	$u_{tn} * r_n * (1-FPR)$

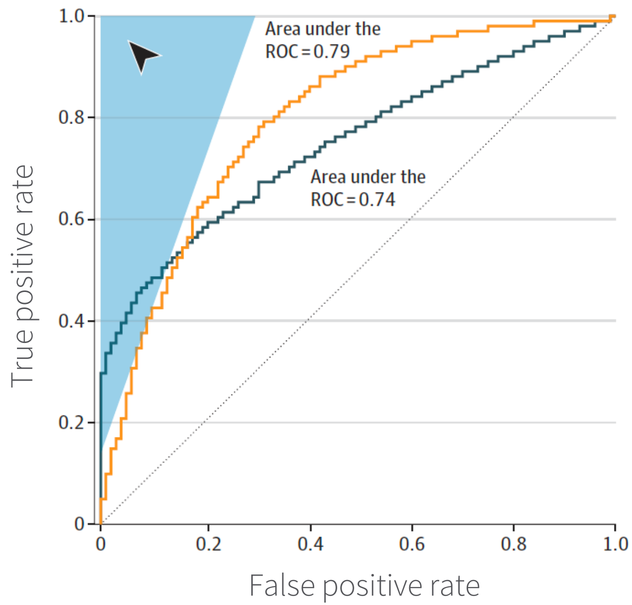
$$E(u) = u_{tp} * r_p * TPR + u_{fn} * r_p * (1-TPR) + u_{fp} * r_n * FPR + u_{tn} * r_n * (1-FPR)$$

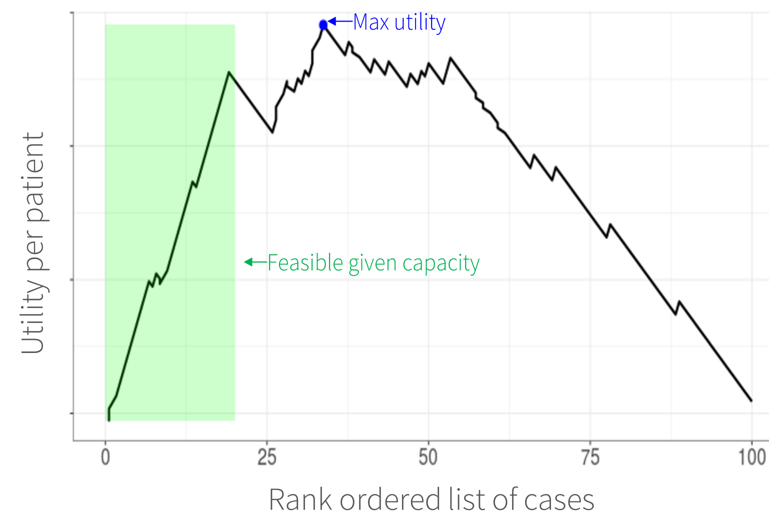
$$E(u) = u_{tp} * r_p * \mathbf{1} + u_{fn} * r_p * (1-\mathbf{1}) + u_{fp} * r_n * \mathbf{1} + u_{tn} * r_n * (1-\mathbf{1})$$

$$E(u) = u_{tp} * r_p + u_{fp} * r_n$$

$$\text{Slope} = \frac{\text{the rate of negatives x the cost of misclassifying a negative}}{\text{the rate of positives x cost of misclassifying a positive}}$$

Focus on achievable utility, given work capacity





Building a model, then separately doing a utility analysis, and later facing work constraints is suboptimal

A model's ROI is often challenging

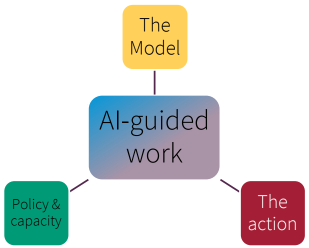
Cost to Build & Deploy the Model		
Proportional cost of shared infrastructure		
Data warehouse cost		45,000.00
Cost to prepare data to train model		
Data pull and label definition	8	19,230.77
Verify clean labels w/SME	26	12,500.00
Cost to learn / validate / evaluate the model		
Hardware cost		5,000.00
Data scientist costs	4	19,230.77
Cost to setup run-time environment		
Hardware cost		10,000.00
Live data procurement costs		10,000.00
Program manager cost	52	24,000.00
Database expert	52	6,000.00
ML engineer	52	10,000.00
Year 1 build cost		160,961.54
Monitoring and maintenance costs		
Yearly maintenance cost per ratio		32,192.31
Cost per prediction		
Model cost over 5 years		289,730.77
Year 1 cost per prediction		32.19
With build cost amortized over 5 years		11.59
Benefit accrued to society or payer		
Yearly benefit accrued		1,739,130.43

Cost for the Model-guided Workflow	
Cost to design the clinical workflow	
One time workflow design cost	20,000.00
Cost for clinical integratioon	
Application integration cost	30,000.00
Execution costs for a flagged case	
Procedure costs	250.00
Laboratory testing	1,500.00
Clinician costs	240.38
Cascade testing (family)	6,000.00
Clinician review	240.38
Intervention cost	
Patient co-pay	50.00
Medication cost	14,000.00
Facility fees	200.00
Execution cost per flagged case	22,480.77
With design cost amortized over 5 years	22,680.77
Healthcare System Profit / Loss	
Cost to find true case using the model	1,158.92
Workflow cost for case found by the model	22,680.77
Passthrough cost per case	14,000.00
Healthcare system revenue per case	8,680.77
Cases found by model	6.96
Year 1 P/L	(100,573.58)
Year 2 P/L	28,195.65
Year 3 P/L	28,195.65
Year 4 P/L	28,195.65
Year 5 P/L	28,195.65

Data Science Team at SHC

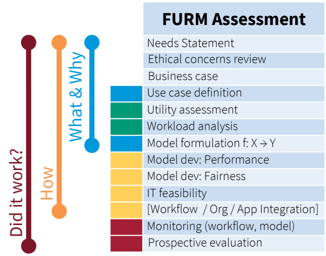
01

Thought leadership for Responsible AI in Healthcare.



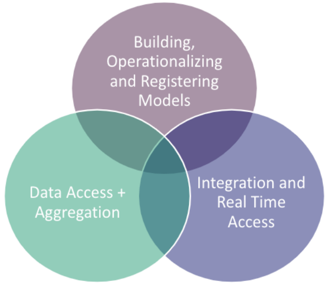
02

Ensure that we create FURM - Fair, Useful, Reliable Models.



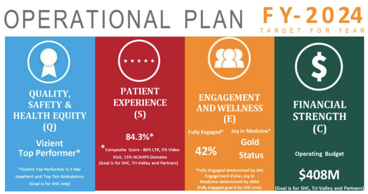
03

Processes and infrastructure for an “AI ready” organization.



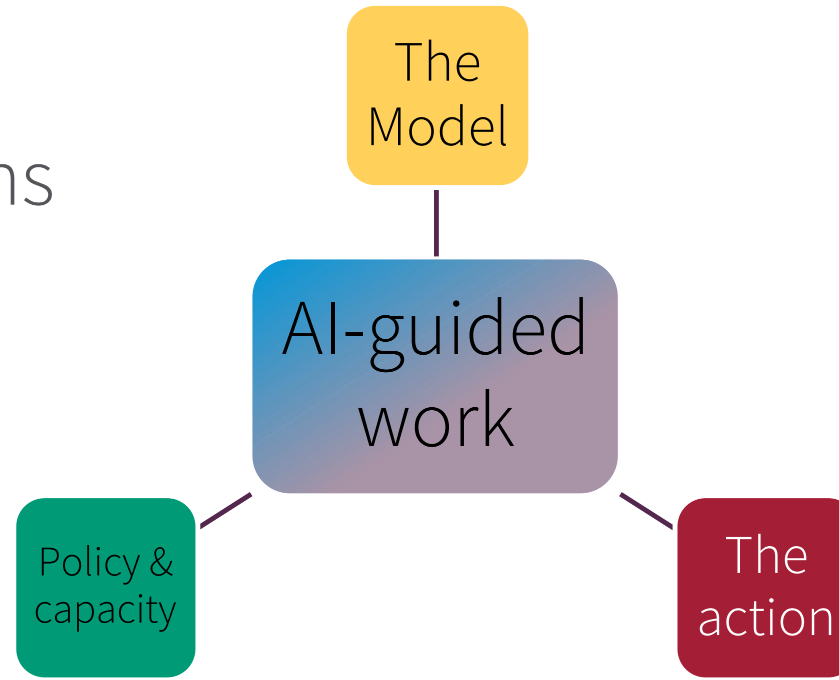
04

Identify and execute, 3-5 projects with enterprise value.

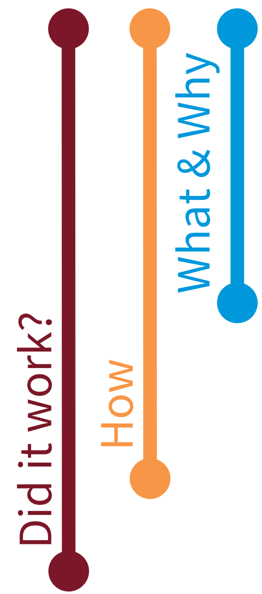


https://dsatshc.stanford.edu/

We continue to study the interplay of models, work capacity, and actions

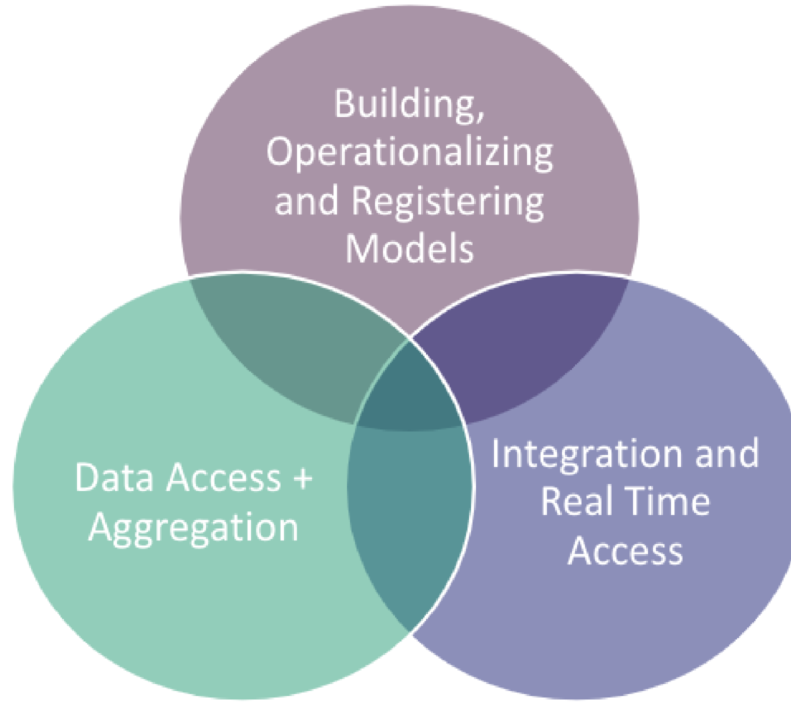


We have developed a way to assess if we are creating Fair, Useful, Reliable Models

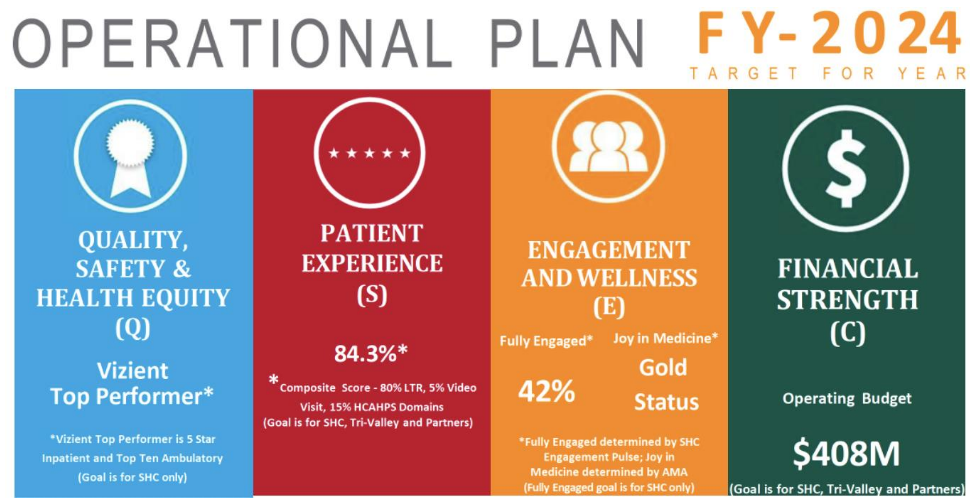


FURM Assessment
Needs Statement
Ethical concerns review
Business case
Use case definition
Utility assessment
Workload analysis
Model formulation f: X → Y
Model dev: Performance
Model dev: Fairness
IT feasibility
[Workflow / Org / App Integration]
Monitoring (workflow, model)
Prospective evaluation

You will need processes as well as infrastructure for being "AI ready"



Governance is crucial for enterprise-wide alignment



The state of AI at Stanford Healthcare

Administrative

Patient Engagement

Quality / Clinical Decision Support

Clinician Efficiency

30+ Vendor Applications in Production Using AI

A Roadmap To Welcoming Health Care Innovation

	Readout	R&D	Clinical	Business
1. Discovery (pilots, explorations)	technical feasibility and user acceptance			
2. Development (deployment, strategic project)	proof of meeting intent of the innovation such as access, quality, or productivity gain			
3. Dissemination (enterprise project, scaling deployment, ROI study)	refine the technology as well as optimize the business model			

Stages per http://goto.stanford.edu/innovation

Scaling beyond Stanford Healthcare

Providing guidelines for the responsible use of AI in healthcare

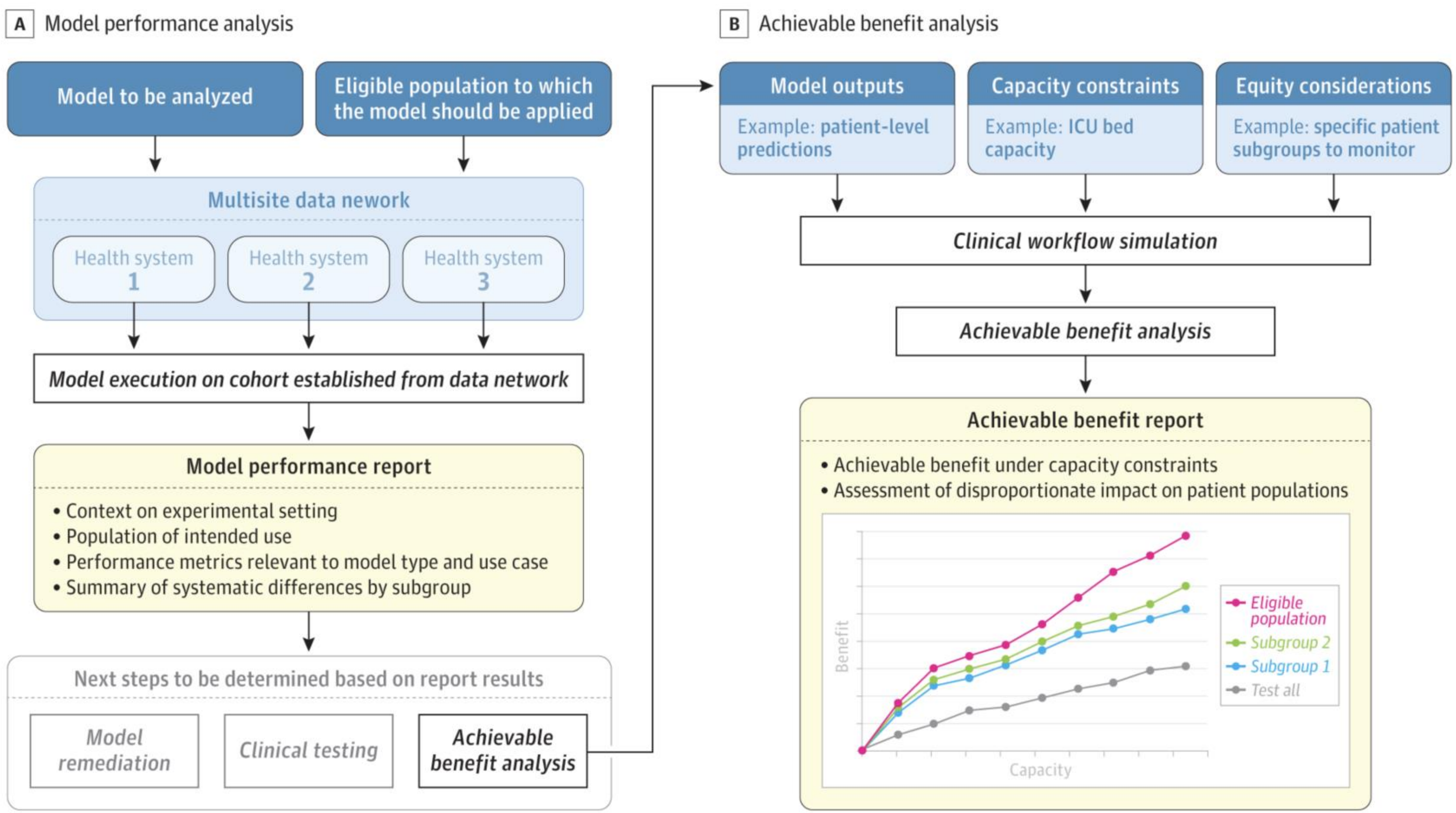
CHAI™
Coalition for Health AI

Learn More Insights News Events Join Us

Our Purpose

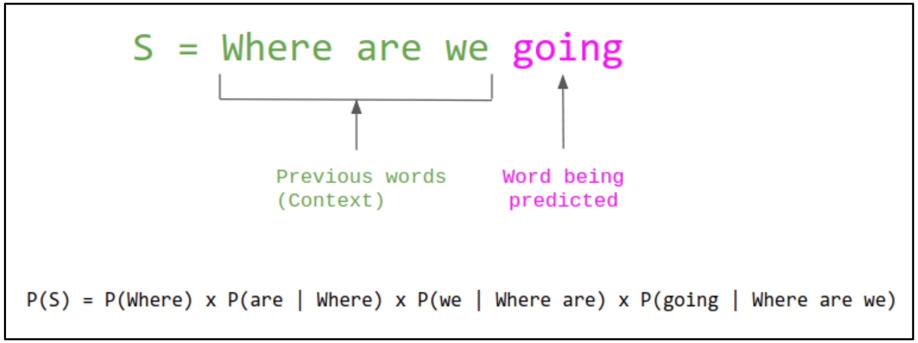
The Coalition for Health AI (CHAI™) is a community of academic health systems, organizations, and expert practitioners of artificial intelligence (AI) and data science. These members have come together to harmonize standards and reporting for health AI and educate end-users on how to evaluate these technologies to drive their adoption. **Our mission is to provide a framework for the landscape of health AI tools to ensure high quality care, increase trust amongst users, and meet health care needs.**

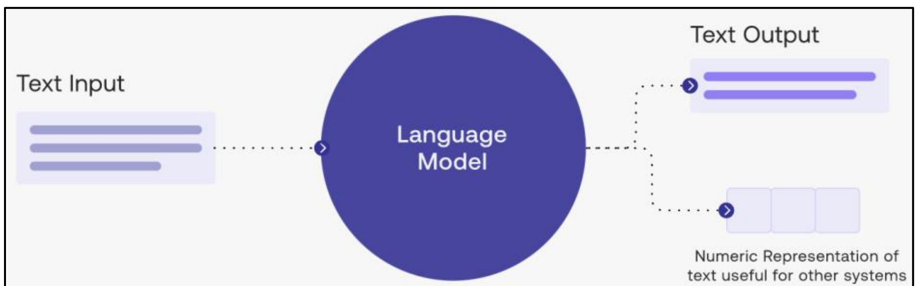
A Nationwide Network of Health AI Assurance Laboratories



Generative AI changes the framework

Language models 101





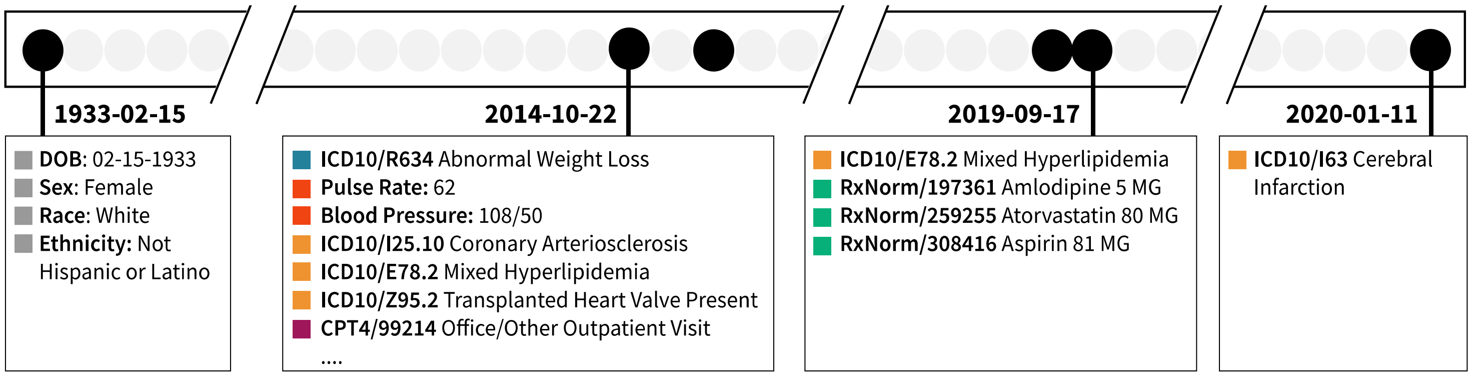
https://docs.cohere.com/docs/introduction-to-large-language-models

Training data

Language model

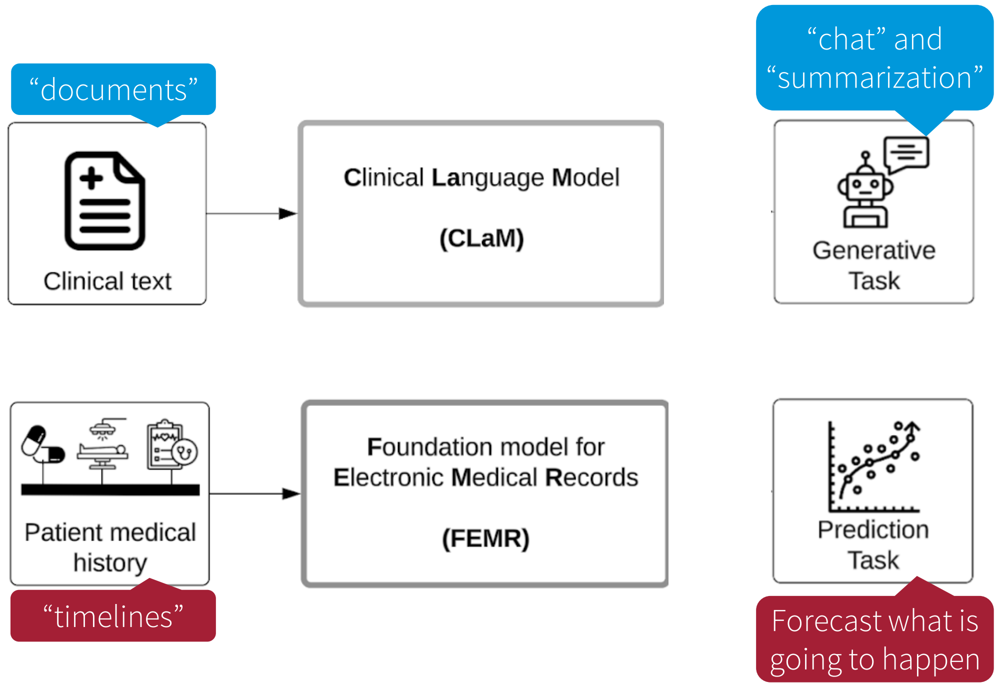
Large language model

Structured EHR data comprise a “language”



EHR “Language”: Visit{R634, 999214} | Rx {308416} | Visit{I63, R69} | …

Two ways to build “language” models using the EHR



https://tinyurl.com/shaky-foundations

Foundation models for **E**lectronic **M**edical **R**ecords

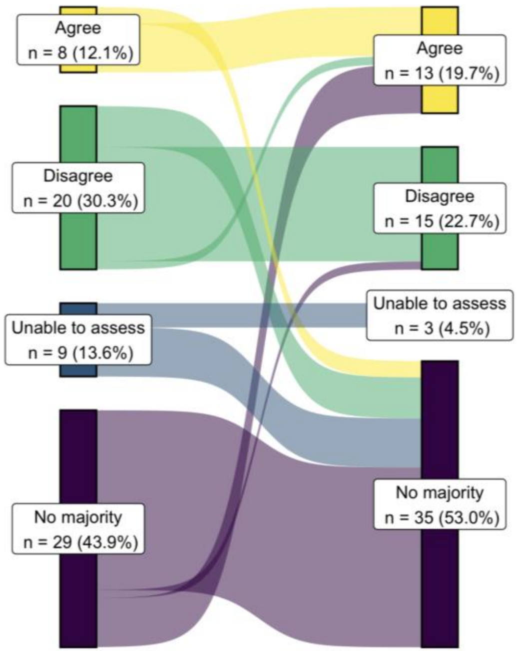
CLMBR: Clinical language modeling-based representations [1]	**2021**	• **3.5 to 19%** increase AUROC of binary tasks • Classifiers **decay less as time passes** [2] • Classifiers **transfer better across subgroups** [3] • Classifiers are **portable across hospitals** [4]
MOTOR: Many Outcome Time Oriented Representations [5]	**2023**	• **First time-to-event foundation model** • Better performance over **long time horizons** • **8x faster training** • **95% less training data**

https://github.com/som-shahlab/femr/

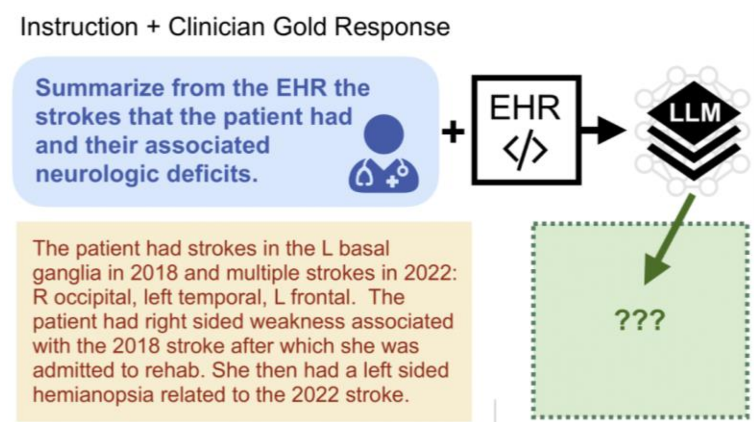
for large-scale, self-supervised learning using electronic health records

Clinical **La**nguage **M**odels

#1: Flashy headlines over-hype memorization



#2: Tuning for medical tasks is limited



Ensuring Useful Adoption of Generative AI in Healthcare

Foundation models transcribe, summarize, or create in service of the science, practice or delivery of care

	Science	Practice	Delivery
Transcribe	. . .	. . .	. . .
Summarize	. . .	. . .	. . .
Create	. . .	. . .	. . .

Contrasting traditional vs. foundation models

	Traditional Models	Foundation Models
Deployment	Top-down	Top-down OR Bottom-up
Cost	Predictable	Unpredictable
Value assessment	Well-understood	Unclear how to measure
Capabilities	Narrow, predefined	Used for tasks the model is never trained for
Output	Well-defined	Emergent, can have 'hallucinations'
Example	Predict which patient with renal injury will progress to dialysis	Write a response to a patient message

Efficiency, effectiveness, and productivity

Doing more with the same

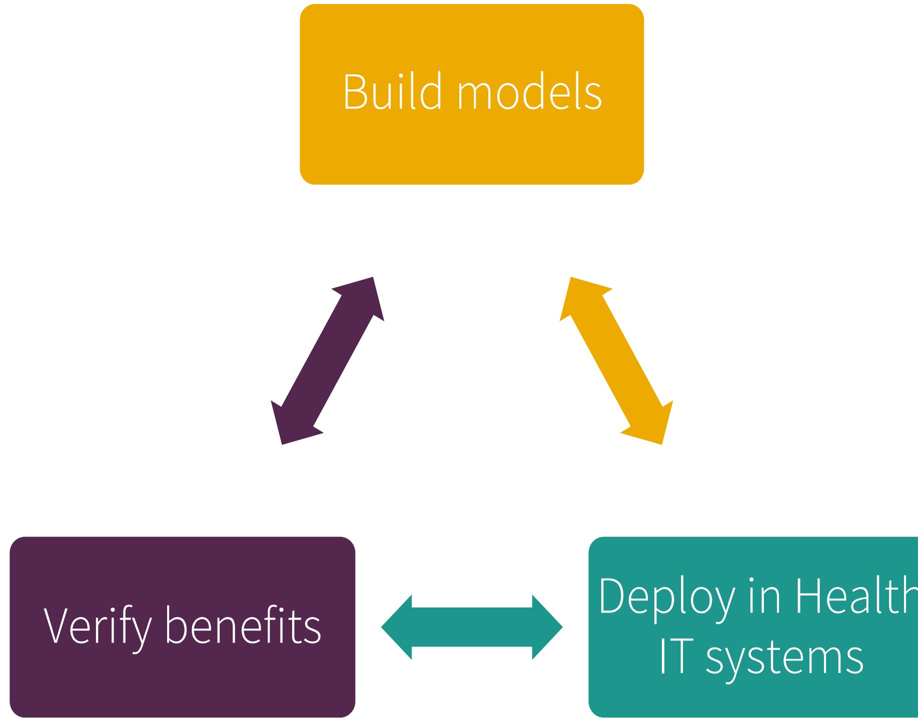
$$\textbf{Productivity} = \frac{\textbf{Output}}{\textbf{Input}}$$

Doing the same with less

$$\text{Effectiveness} = \frac{\text{Actual output}}{\text{Expected output}}$$

$$\text{Efficiency} = \frac{\text{Resource planned}}{\text{Resources used}}$$

We need to focus on defining and verifying benefits



Special Communication | AI in Medicine

August 7, 2023

Creation and Adoption of Large Language Models in Medicine

Nigam H. Shah, MBBS, PhD[1,2,3]; David Entwistle, BS, MHSA[1]; Michael A. Pfeffer, MD[1,2]

Acknowledgements

Funding:

- Federal – NLM, NHLBI (Past: NIGMS, NHGRI, NINDS, NCI, FDA)
- Institutional – Dept. of Medicine, Dean's office, Stanford Hospital
- Fellowships – Med Scholars, Siebel Scholars, Stanford Graduate Fellowship, NSF, DoD
- Industry – Healogics, Janssen R&D, Oracle, Baidu, Amgen, Google, Apixio, CollabRx, Curai
- Emerson Collective, Mark & Debra Leslie

Questioning conventional wisdom https://tinyurl.com/hai-blogs