Healthcare and Medicine for AI Experts LMP2392H - Lecture 1

By Anna Goldenberg and Susan Poutanen

Winter 2026

Syllabus - important points

- Course webpage: https://goldenberglab.ca/LMP2392H_winter_2026_course
- My office hours: Mon 11am-12pm (hybrid)
- Co-coordinator: Susan Poutanen
- TA: Tina Behrouzi Office Hours: Fri 11am-12pm (hybrid)
- Reflection Quizzes: available on Quercus by the end of Wednesday, submit on Monday. Format: 2-3 questions, a paragraph length in response.
- Presentations: 15min each. Wed 12-1pm. Please sign up after class *today*
- Projects
  - Groups of 2
  - Pick from the list (still waiting) or choose your own
  - Requirements: relevance to healthcare/medicine, a problem that can be helped by AI

Grade breakdown

20% Weekly take-home Reflection Quizzes
Note: late submissions are 20% penalty per day (no credit after 5 days)

10% Presentations
Two 15 min presentations of the papers recommended by the experts per student per semester

70% Project
10% Proposal (due Feb 4)
10% Project Presentation (in addition to the class presentation above)
10% Code
40% Final write up (due Apr 1)
Motivation - 5%, Related work - 5%, Methods - 15%, Results - 10%, Conclusion 5%

Team

Anna Goldenberg
PhD
coordinator

Susan Poutanen
MD, MPH, FRCPC
co-coordinator

Tina Behrouzi
Current PhD student
TA

Schedule

Week	Date	Topic	Assignment
1	January 7, 2026	Introduction to AI in medicine and the hidden dragons	
2	January 14, 2026	**Emergency Department** (Jason Fischer)	Announcement of potential projects
3	January 21, 2026	**Psychiatry** (Venkat Bhat)	
4	January 28, 2026	**Liver Transplant/Hepatology** (Mamatha Bhat)	
5	February 4, 2026	**Surgery** (Amin Madani)	Project proposal due date
6	February 11, 2026	**General Internal Medicine** (Amol Verma)	
	February 18, 2026	**Reading week (no class)**	
7	February 25, 2026	**Pathology** (Phedias Diamandis)	
8	March 4, 2026	**Cancer** (Alejandro Berlin)	
9	March 11, 2026	**Infectious Diseases** (Susan Poutanen)	
10	March 18, 2026	**ICU** (Asad Siddiqui)	
11	March 25, 2026	Term project in-class	
12	April 1, 2026	Term project in-class presentation	Project write up due

tech

AlexNet

2012

DL Pathology Pioneer

2014

Hospital Bedside decision support (e.g. kidney) AI in the NHS UK

2015

GANS

2017

Transformers

2018

First FDA-cleared autonomous AI diagnostic (diabetic retinopathy)

BERT

2019

BioBERT/ ClinBERT

2020

GPT-3

/AlphaFold2

2021

CLIP (multimodal)

2022

First sepsis model deployed at scale

2023

Biological foundation models at scale

2024

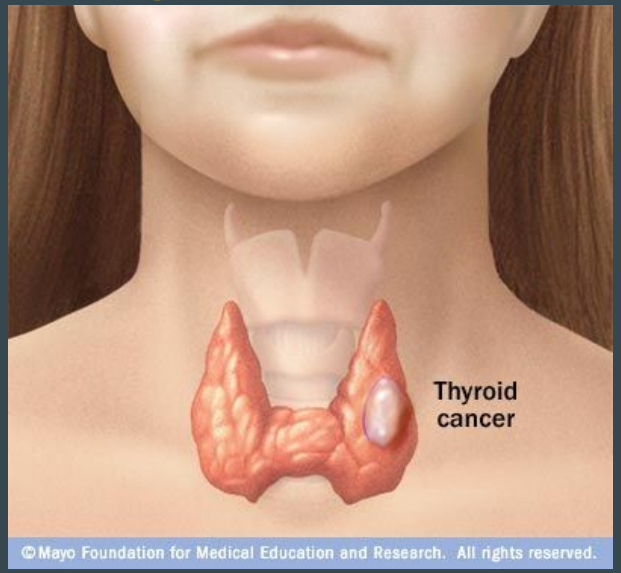
First AI discovered drug in trial

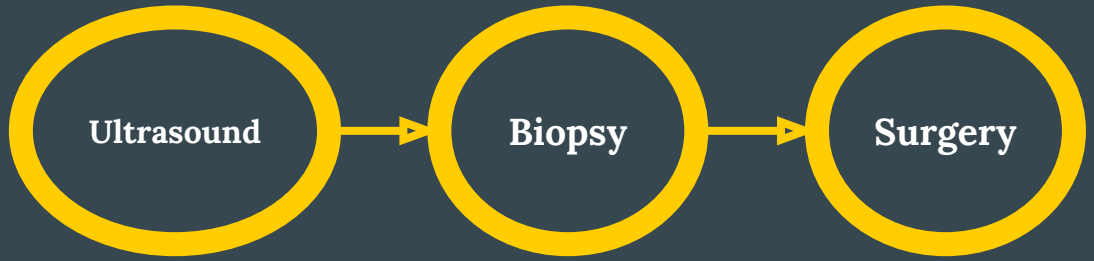
healthcare

Many successes in AI

From the Goldenberg lab

Thyroid Cancer





67% of surgeries find nodule to be benign ⇨ 30% using Random-Forest derived model using image statistics

Radebe et al, 2021

Early cancer detection

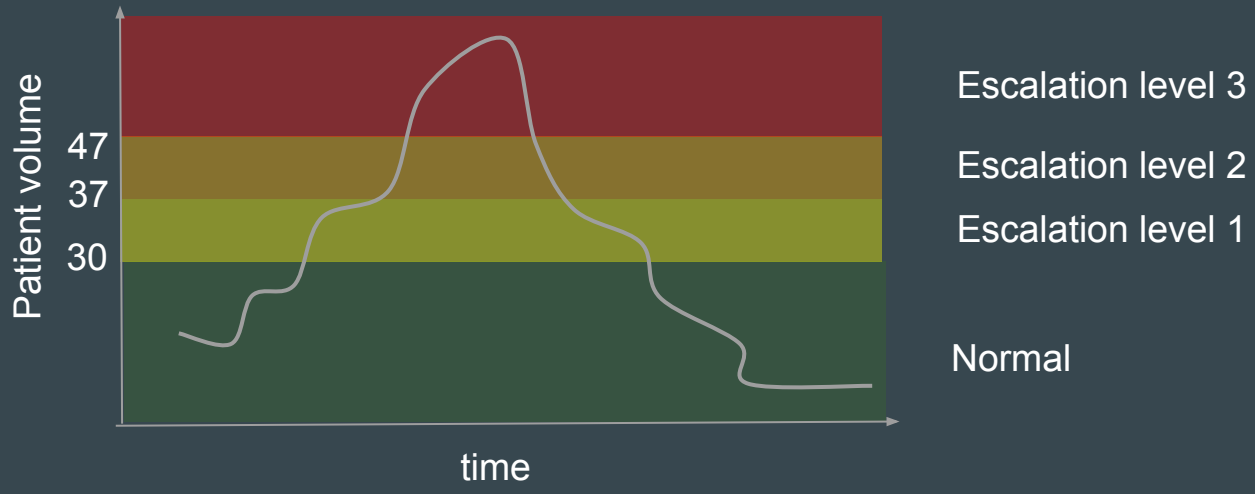
Conditional VAEs

Chang et al, 2021

Critical Care Unit

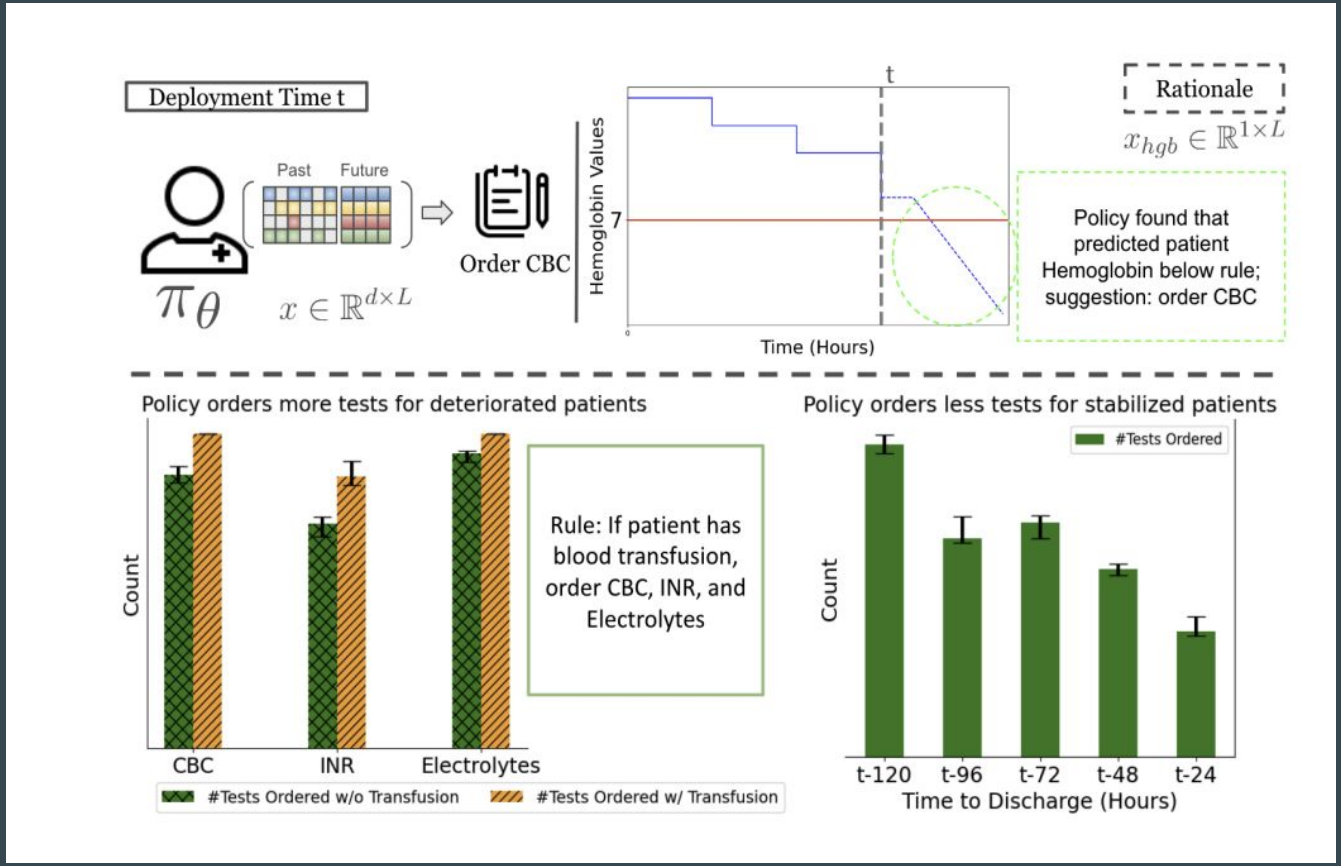
Detecting between 5 - 45 min ahead of cardiac arrest for 70% of the cases (Tonekaboni et al, 2018)

Staff planning in the Emergency Department



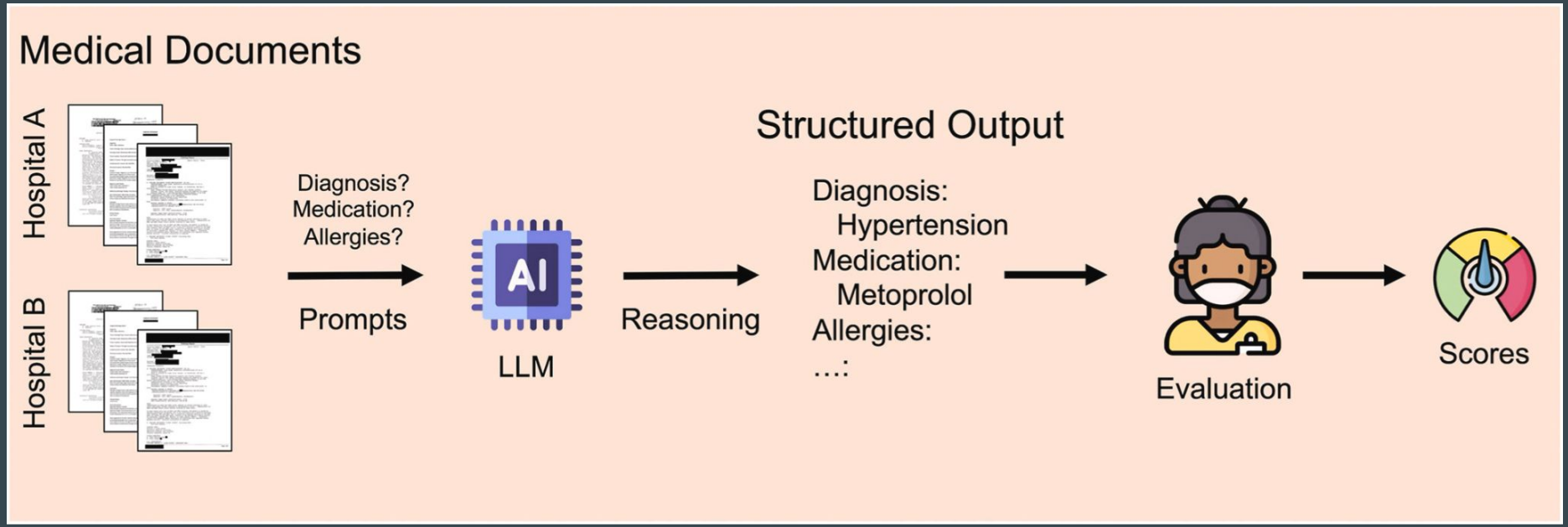
ML Goal: Make a 24-hour forecast of expected number and range of patients

Optimal policy for test ordering

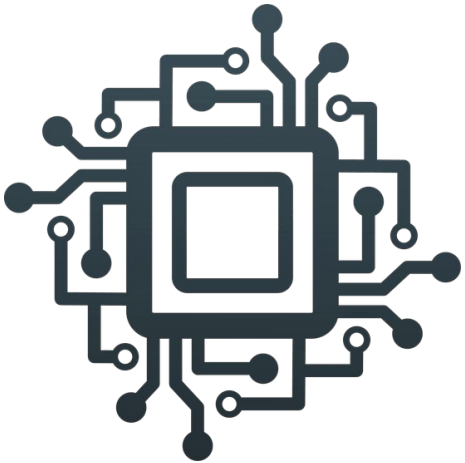


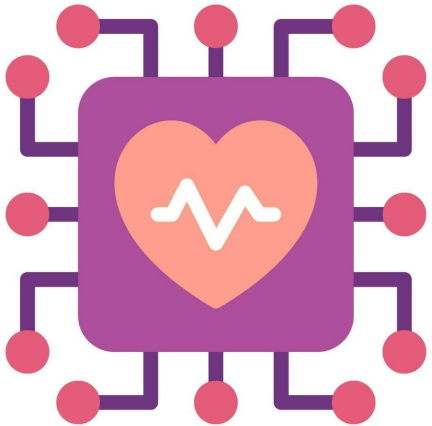
Ji et al, 2025

Language-based interaction with tools

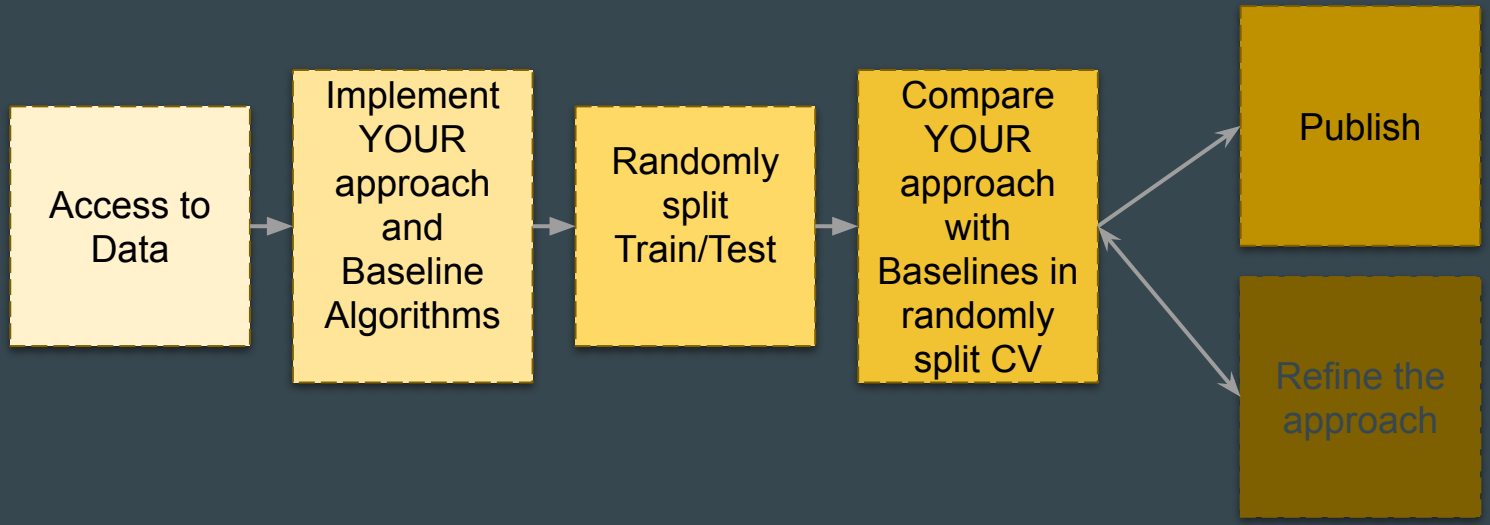


Siepmann et al, 2025

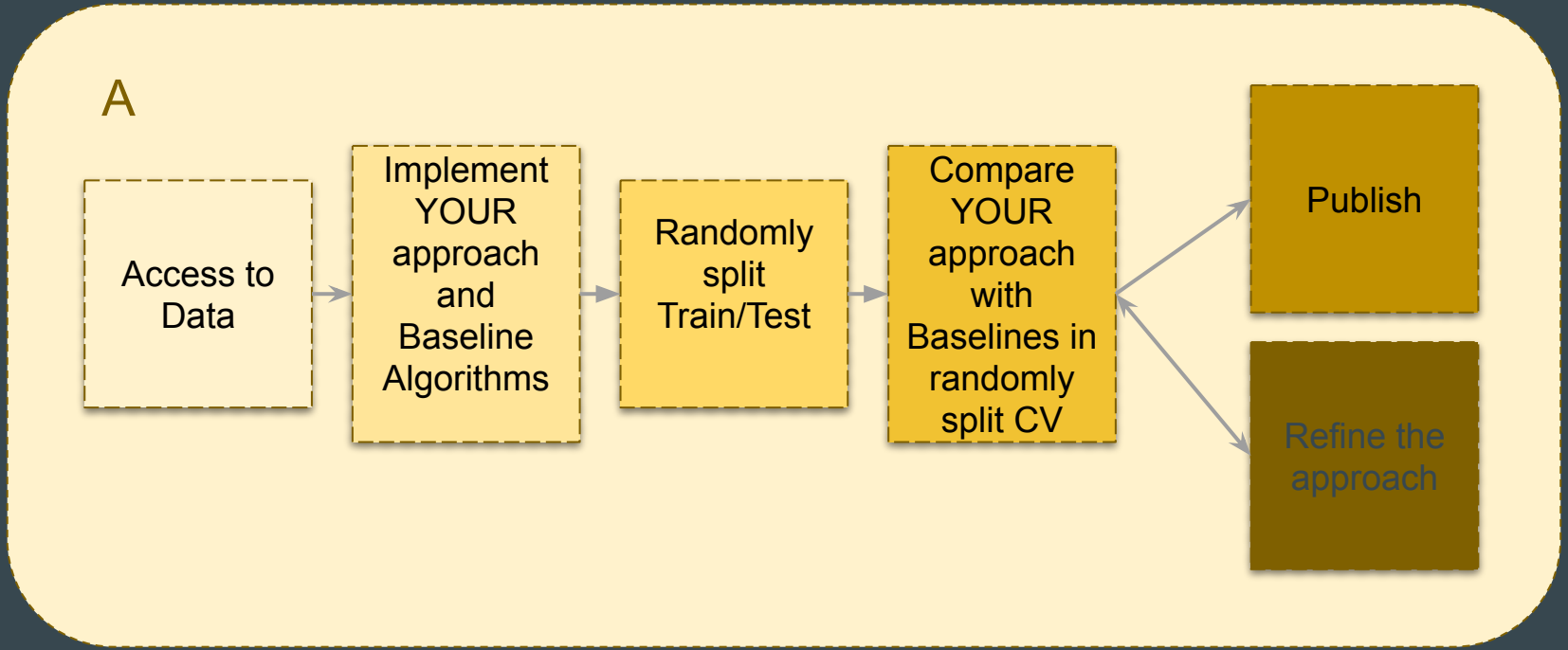




Typical cycle of ML development (in healthcare)



Typical cycle of ML development (in healthcare)



Typical cycle of ML deployment in healthcare

Capture data at point of use

Set up train and validation to imitate prospective deployment

Refine the model

Set up the complete silent mode - testing in real time

Conceptualize and implement UI/Front End

Get all the permissions from the place of deployment (eg hospital)

DEPLOYMENT

Deployment Example

Problem

Research

Predict impending cardiac arrest in ICU, more specifically the need for life saving interventions.

Deployment

Cardiac arrest is very infrequent ~100/yr

42 beds x 365 days/yr x 24 hr/day x 120 times/hr= 44,140,400 evaluations per year

Redefine the task - Risk of Deterioration towards ANY critical event

Caveat: there are no labels… still using proxies for training

Cohort

Research:

Limited retrospective set:

- From 2013 to 2019
- 67 resuscitations
- 60 patients
- 0.5 Hz signals

Missing data points:

- < 20 % of missing physiological signals
- <15 % of missing data points per signal

Deployment:

Low frequency real-time signals:

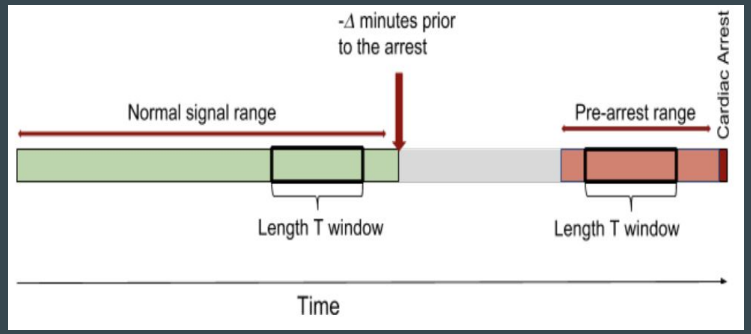
- From 2020 up to early 2023
- From 42 beds
- 0.5 Hz signals (up to 1 Hz)
- 360 patients:
  - 88 required a life saving intervention.
  - 272 “control”

Missing data points:

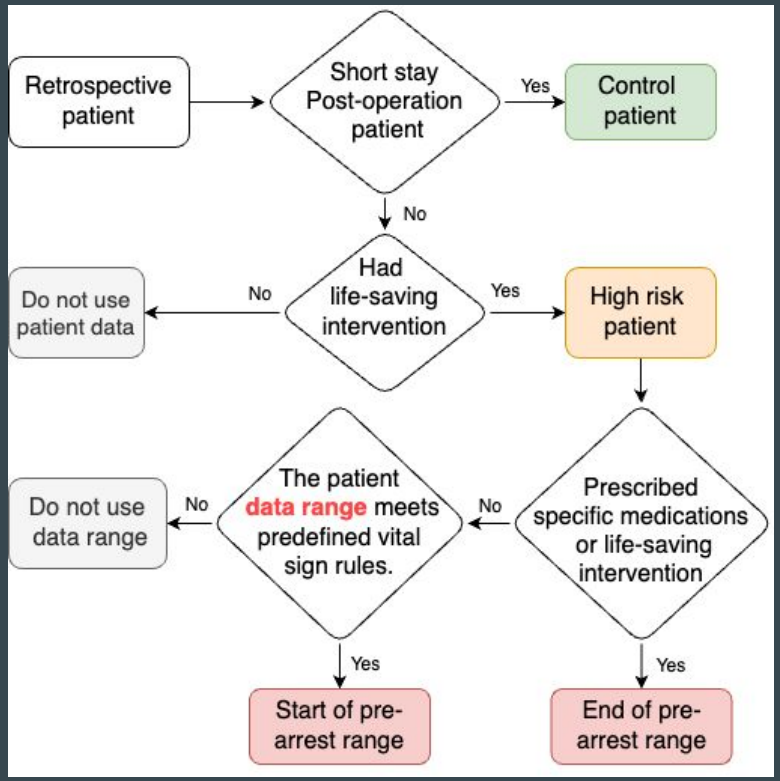
- 40–60 % of missing physiological signals
- 25–40 % of missing data points per signal

Labels

Research



Deployment



Results

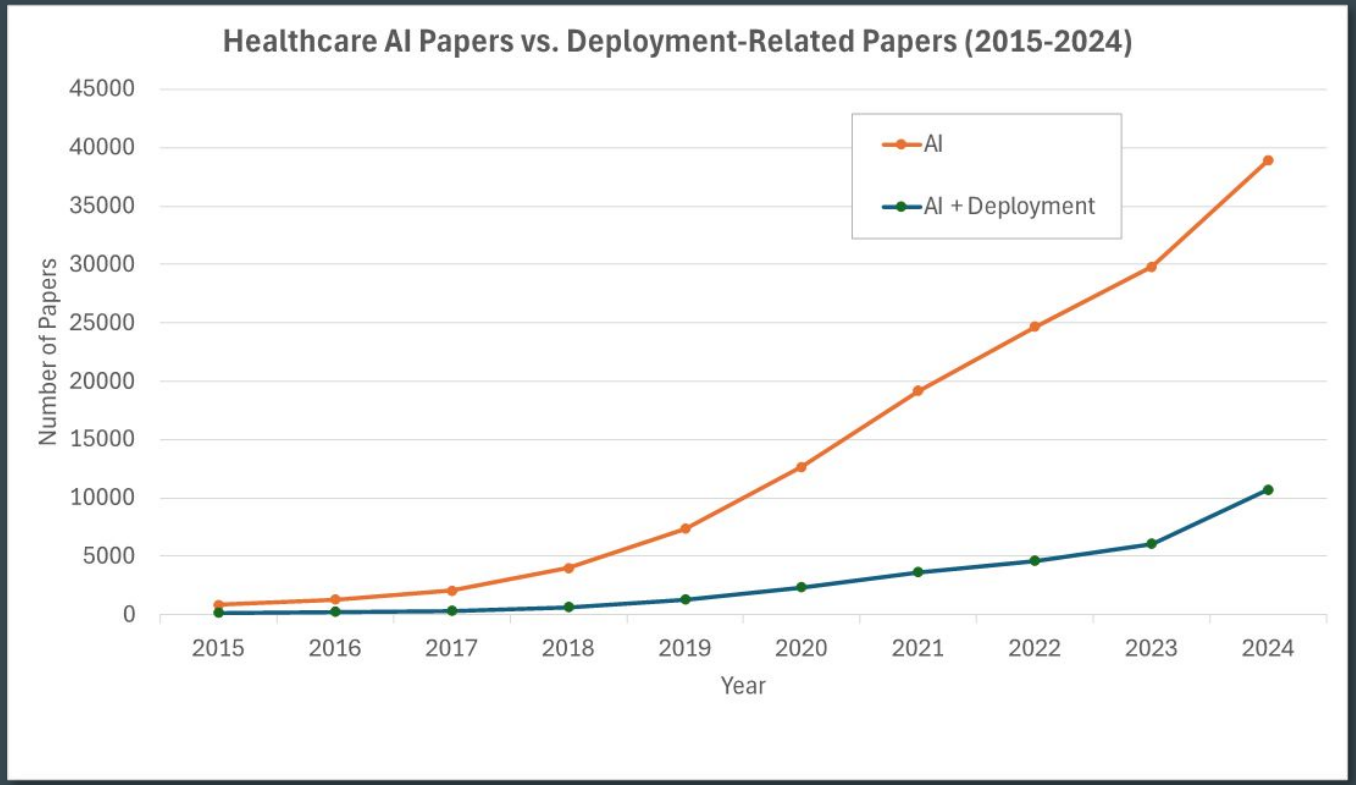
Research

AUROC	AUPRC	Precision	Recall
0.91 ± 0.03	0.54 ± 0.22	0.22 ± 0.08	0.88 ± 0.01

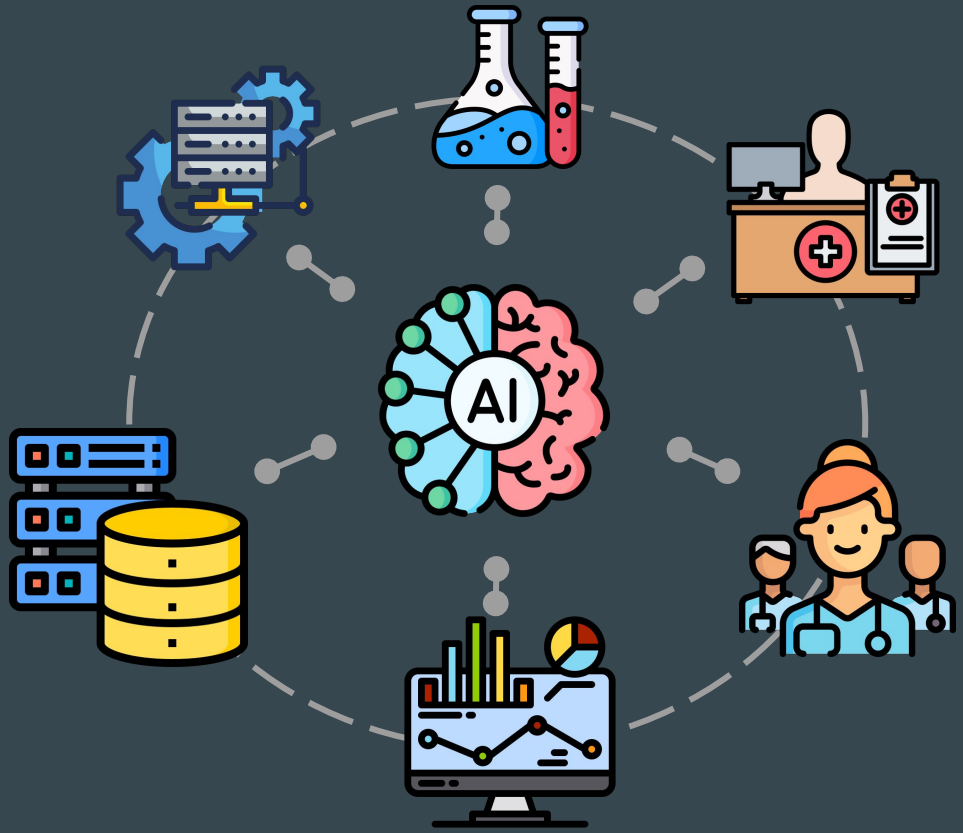
“Deployment”

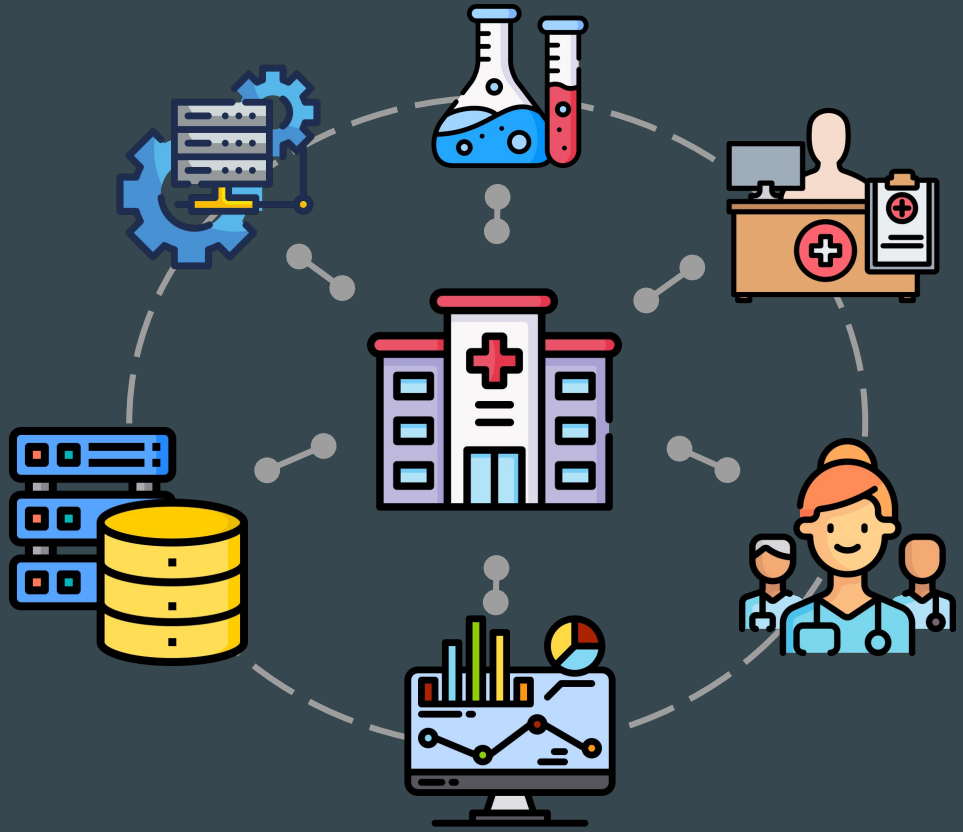
AUROC	AUPRC	Precision	Recall
0.85 ± 0.05	0.43 ± 0.28	0.10 ± 0.03	0.67 ± 0.18

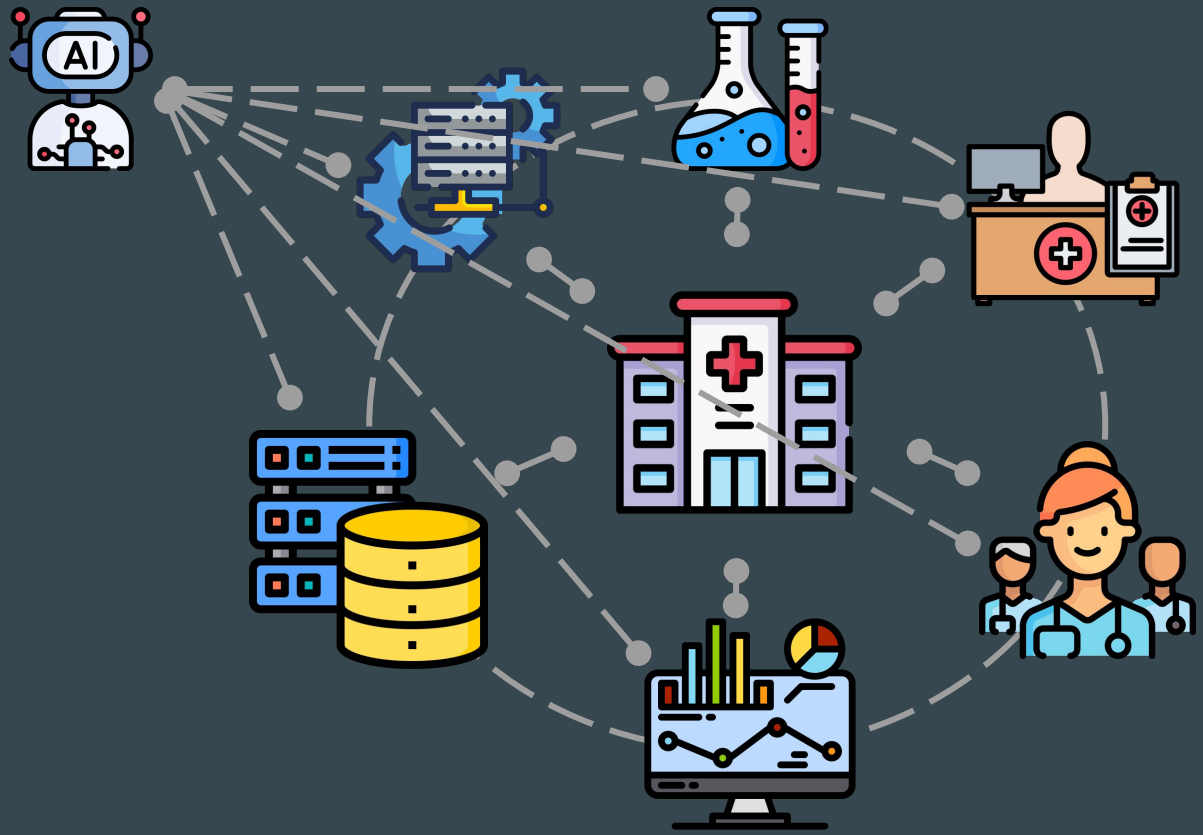
Idea vs practice



1,016
AI/ML-enabled medical devices FDA authorized

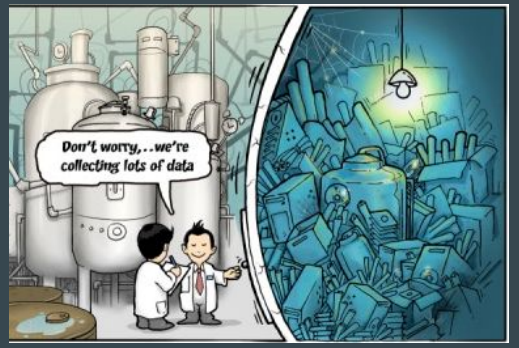

AI




AI

What are the main impediments to integration?

- Lack of integrative pipelines
- No approvals/governance frameworks
- Lack of clear prioritization
- AI is not ready...



AI Challenges

1 Lack of context 1: Policy Creep

Reality:

Patient with asthma has pneumonia and is treated more aggressively

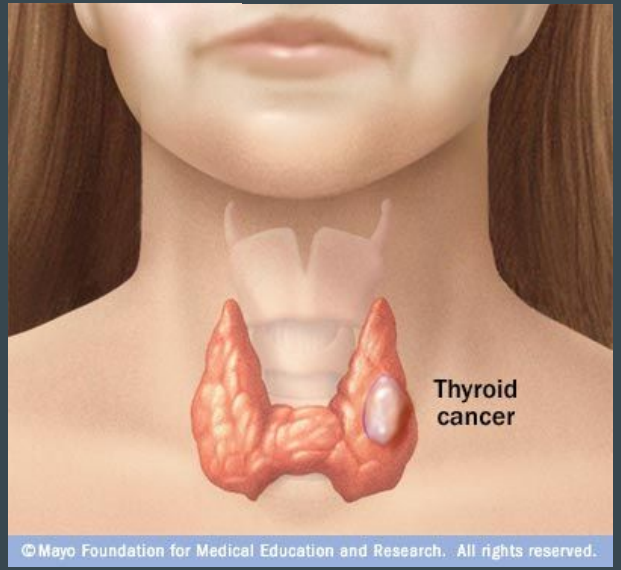
Fewer patients with asthma die of pneumonia

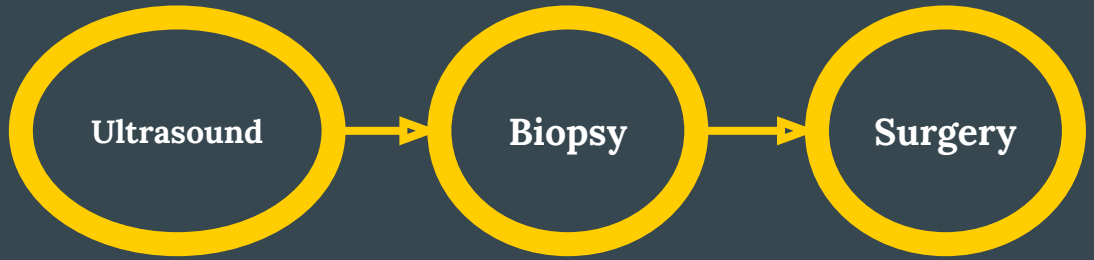
Learned:

If you get pneumonia, it's better if you already have asthma too!

Caruana

1 Lack of context 2: features not available at training





Unnecessary: 67% of surgeries ⇨ ~30%

Small validation set of 10 patients:
7 benign patients, had surgery -> predict 2/7
3 malignant patients had surgery -> predict 2/3

2 Encoding bias in the data

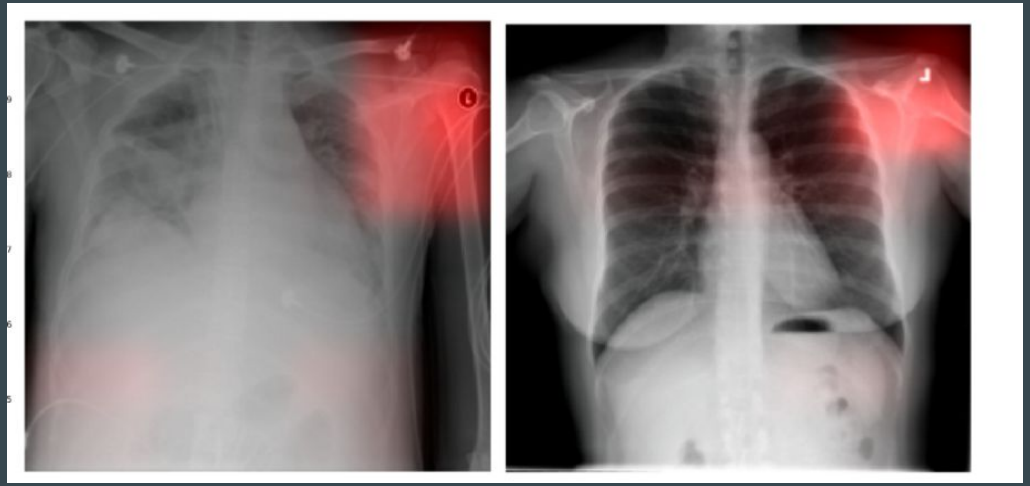
Classifier of no-show for appointments

Learned to discriminate based on race and SES

Result: overbooked appointments only for African American and poor people

From personal communication

3 Are we learning about disease or artifacts?



Pneumonia==metal token??

Source: Zech et al (2018)

4

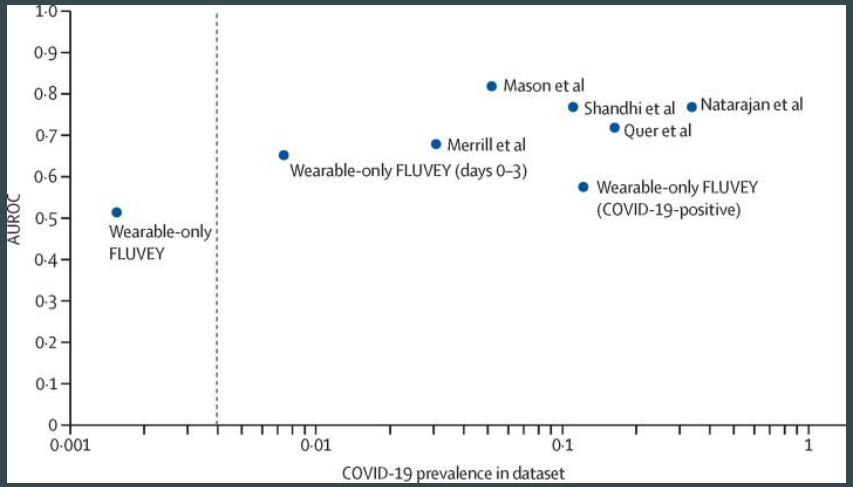
High false positive rate

1% error = 30,000+ false positives

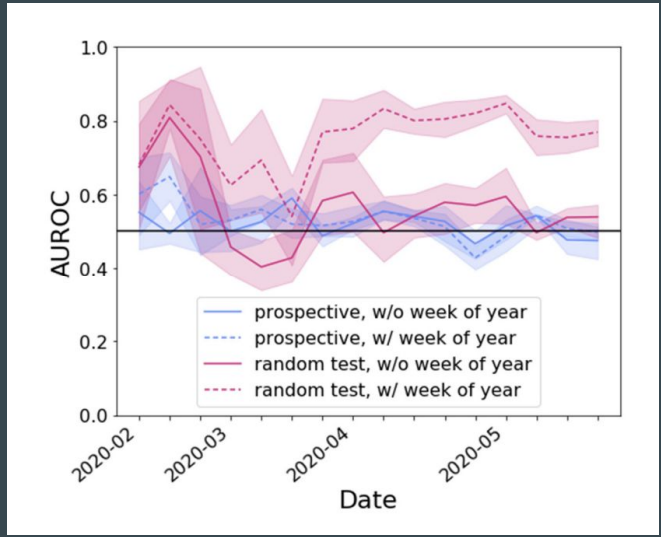
5

Inappropriate testing of the tool (have to use prospective validation)

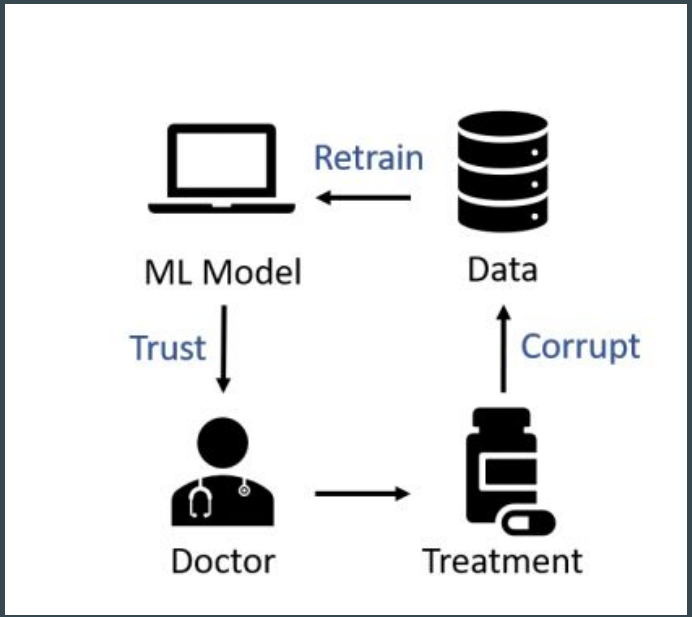
We can predict/detect Covid using wearables



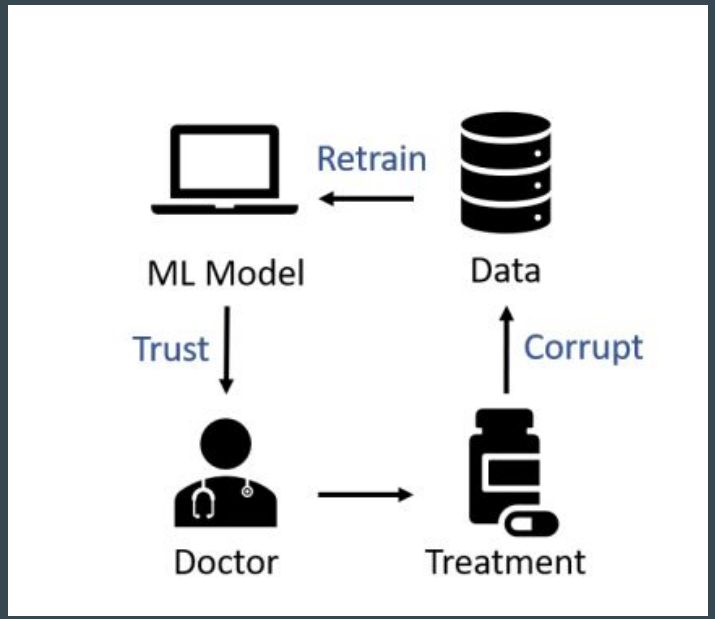
We CANNOT predict/detect Covid using wearables

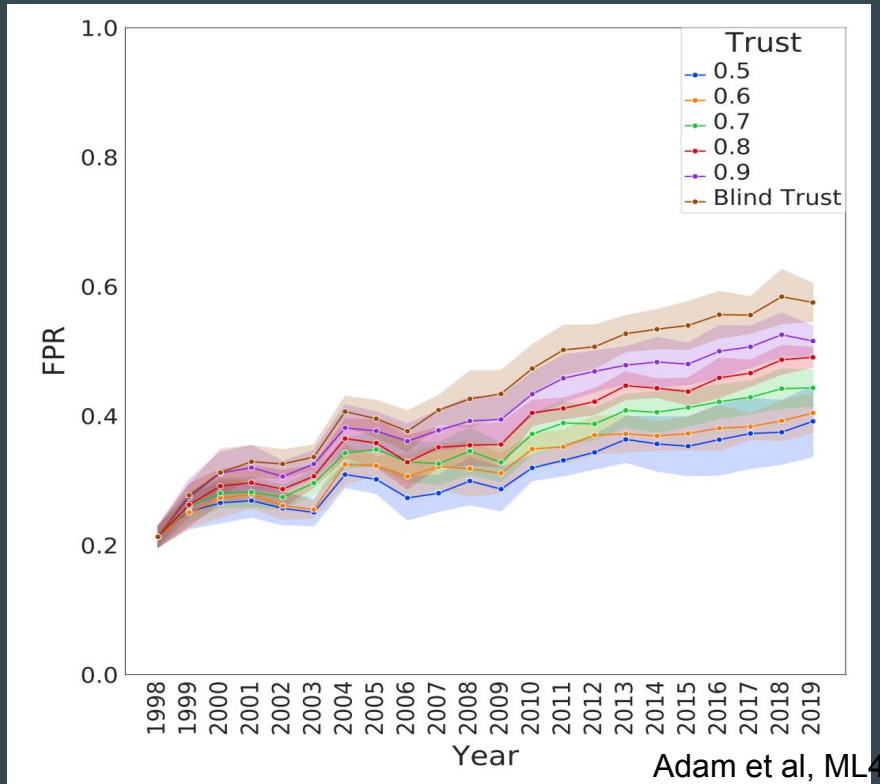


6 Feedback loop problem



6 Feedback loop problem





Adam et al, ML4HC, 2019

Many other issues

- Not optimizing the right objective – not actionable
- Not determining conditions under which the model is valid
- The model is using too many resources for production
- Evaluation is not appropriate
  - Dependent on data that is recorded after the decision has to be made
  - Tested in environment that amplifies signal

Ethics (w/ help of M McCradden)

Morals: values concerning right and wrong; individual - can be subjective and varying

Ethics: concerning the moral parameters of particular activities; requires defensibility, evidence, publicity, (some) generalizability

ETHICS

"...the most important thing for computer scientists to consider when developing AI content for medicine is that medicine has long established ethical and scientific norms that have developed over many decades to ensure that medical interventions are safe and effective. If AI is to be effectively integrated into medicine, it's development must be attentive to these norms. This process will entail some adjustment of these norms as well, but their wholesale rejection will almost certainly lead to distrust, safety and efficacy concerns and, ultimately, slow uptake of AI...."

James Anderson, bioethicist, SickKids

Bias

Ethical Machine Learning in Health Care

Irene Y. Chen,[1] Emma Pierson,[2] Sherri Rose,[3] Shalmali Joshi,[4] Kadija Ferryman,[5] and Marzyeh Ghassemi[4,6]

[1]Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA; email: iychen@mit.edu
[2]Microsoft Research, Cambridge, MA, 02143, USA
[3]Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA, 94305, USA
[4]Vector Institute, Toronto, ON, Canada
[5]Department of Technology, Culture, and Society, Tandon School of Engineering, New York University, Brooklyn, NY, 11201, USA
[6]Department of Computer Science, University of Toronto, Toronto, ON, Canada

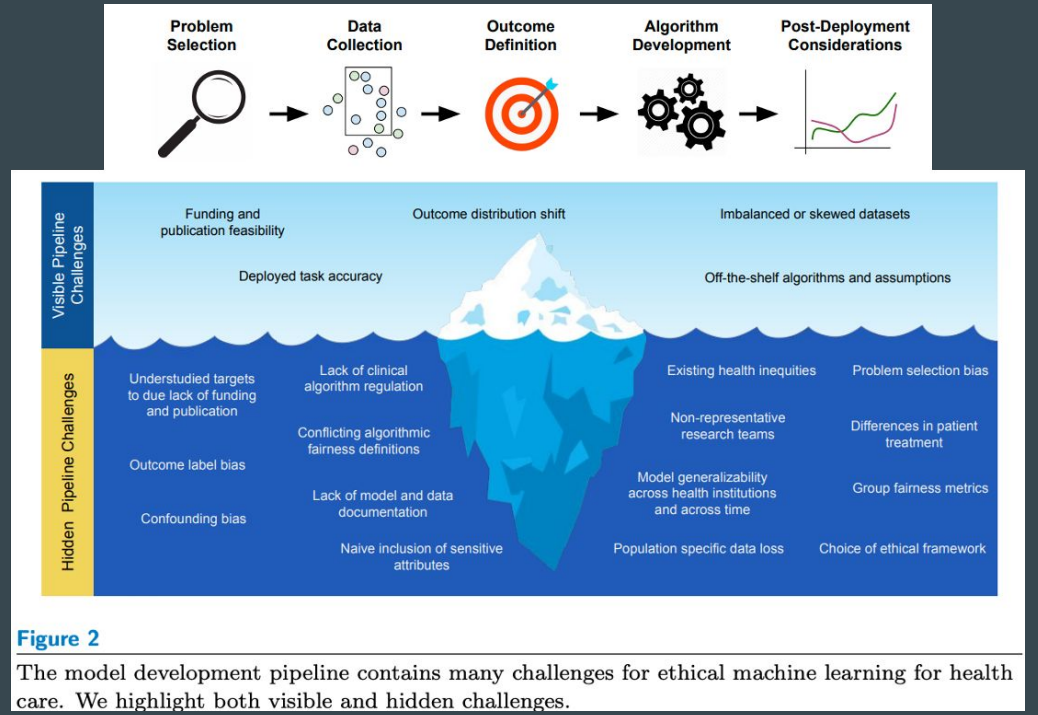


Figure 2

The model development pipeline contains many challenges for ethical machine learning for health care. We highlight both visible and hidden challenges.

Problem selection

BEHAVIOR & SOCIETY | OPINION

Yes, Science Is Political

Scientists need to acknowledge that fact—and to act on it in these most dire of times

By Alyssa Shearer, Ingrid Joylyn Paredes, Tiara Ahmad, Christopher Jackson on October 8, 2020

SCIENTIFIC AMERICAN®

Worldwide disparities in the study and funding of health research

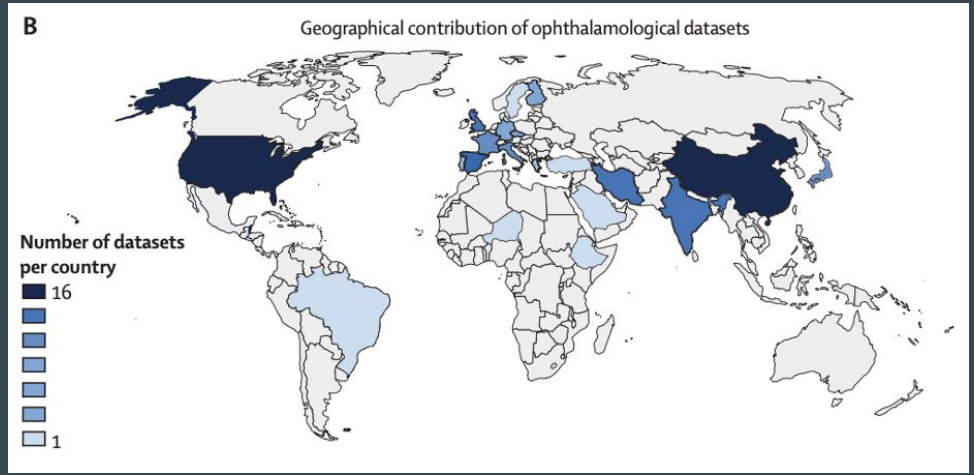
Gap for: health among citizens of the Global South, poverty-related diseases, racialized groups, women's health, mental health

Health data poverty

Chen et al., "Ethical machine learning in health care" 2020 *arXiv preprint*; Ibrahim et al., "Health data poverty: an assailable barrier to equitable digital health care" *Lancet Digital Health* 2021

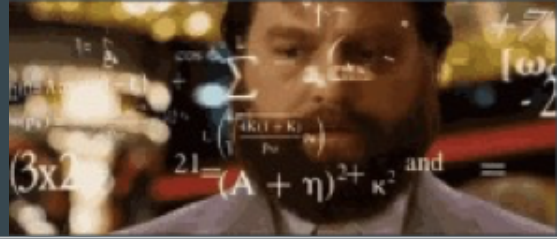
Data Collection

- Representation
- Data quality
  - Interventional trial data, social media data, health records
- Diversity in scientific workforce



Outcome definition

- Labels reflect the current state of knowledge on a topic and are affected by societal attitudes, policy, law
- Label noise: access, incentives, inconsistencies, structural biases
- Both under- and over-diagnosis2
- Risk of preserving our axioms of the past



Algorithm Development

- Algorithms are not neutral
- Risks of particular model choices put some patients more at risk than others
- Confounding: finding non-causal, associationist patterns in data
- Performance metrics
- Algorithmic fairness

Post deployment considerations

- Most important: auditing and oversight
- Continuous quality improvement
- Promote accountability, identify targets for improvement
- Real-world evaluation through evidence-gathering paradigms
- Requires considerations of scale and actionability

Questions to answer for a robust pipeline to deployment

- How is the data being accessed?
- What is the engine that brings data to the inference part? (e.g. real time streaming physiological data requires A LOT of bandwidth)
- How often to retrain/update the model?
- What is being displayed to the clinician (user) – is it *actionable?*
- What data is collected to ensure that model is being impactful in practice?

Summary

- Many successes
- Much still remains to be done
  - More robust ML
  - Guidelines for when the computational method is ready for translation
    - Rate of false alarms is not prohibitive?
    - Ensured the right objectives were optimized?
    - Checked for bias?
    - Right team is onboard to make the transition?
  - Seamless pipelines from data access to implementation in the clinic

Questions