UCSF
University of California
San Francisco

Quality assurance and improvement for Machine Learning-based medical devices

Jean Feng

University of California, San Francisco

Disclaimer: The views presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of the FDA/HHS, PCORI, or the U.S. Government.

Develop new AI methods

+

Train and Implement AI algorithms

- Develop new methods and frameworks for assessing and improving the safety, effectiveness, and equity of AI algorithms

FDA

CERSI
UCSF-Stanford

- Train, evaluate, and deploy clinical AI algorithms for hospital quality improvement efforts

ZUCKERBERG SAN FRANCISCO GENERAL
Hospital and Trauma Center

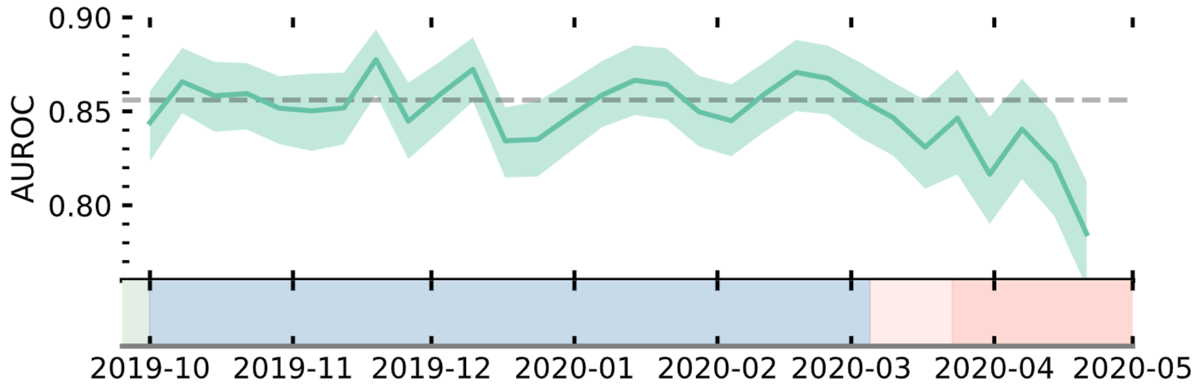
www.jeanfeng.com

Why do we need to monitor and update ML algorithms?

ML algorithms can deteriorate in performance

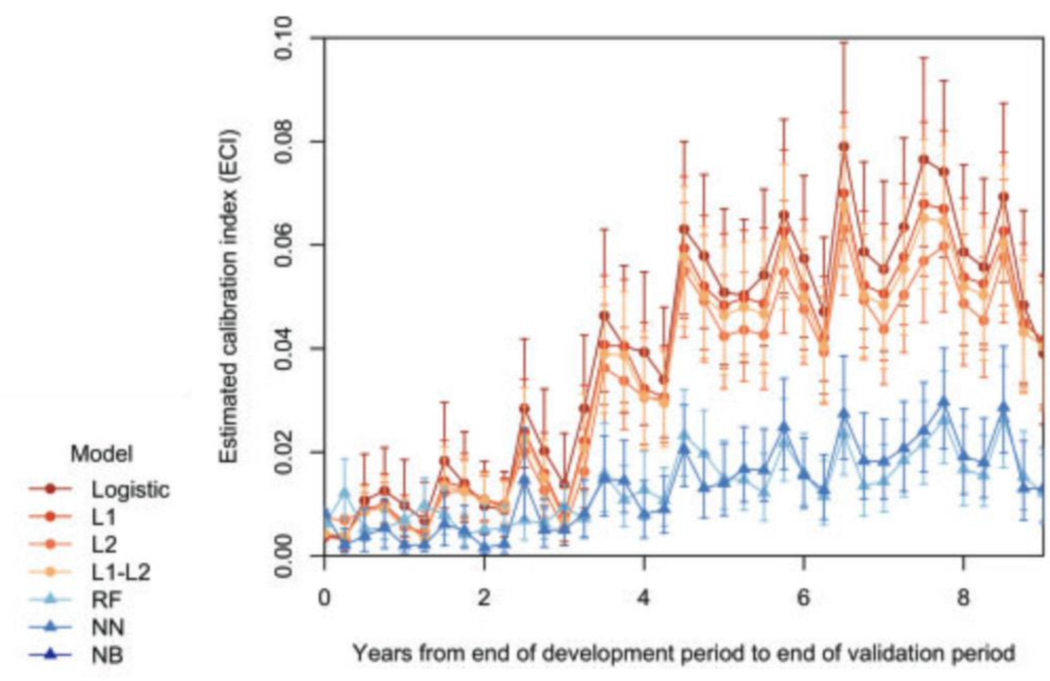
Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19

Christopher Duckworth[1✉], Francis P. Chmiel[1], Dan K. Burns[1], Zlatko D. Zlatev[1], Neil M. White[1], Thomas W. V. Daniels[2,3], Michael Kiuber[4] & Michael J. Boniface[1]



Calibration drift in regression and machine learning models for acute kidney injury

Sharon E Davis,[1] Thomas A Lasko,[1] Guanhua Chen,[2] Edward D Siew,[3,4] Michael E Matheny[1,2,3,5]



Evolving ML algorithms can deteriorate in performance

How Is ChatGPT's Behavior Changing over Time?

Lingjiao Chen[†], Matei Zaharia[‡], James Zou[†]

[†]Stanford University [‡]UC Berkeley

Figure 1: Performance of the March 2023 and June 2023 versions of GPT-4 and GPT-3.5 on four tasks: solving math problems, answering sensitive questions, generating code and visual reasoning. The performances of GPT-4 and GPT-3.5 can vary substantially over time, and for the worse in some tasks.

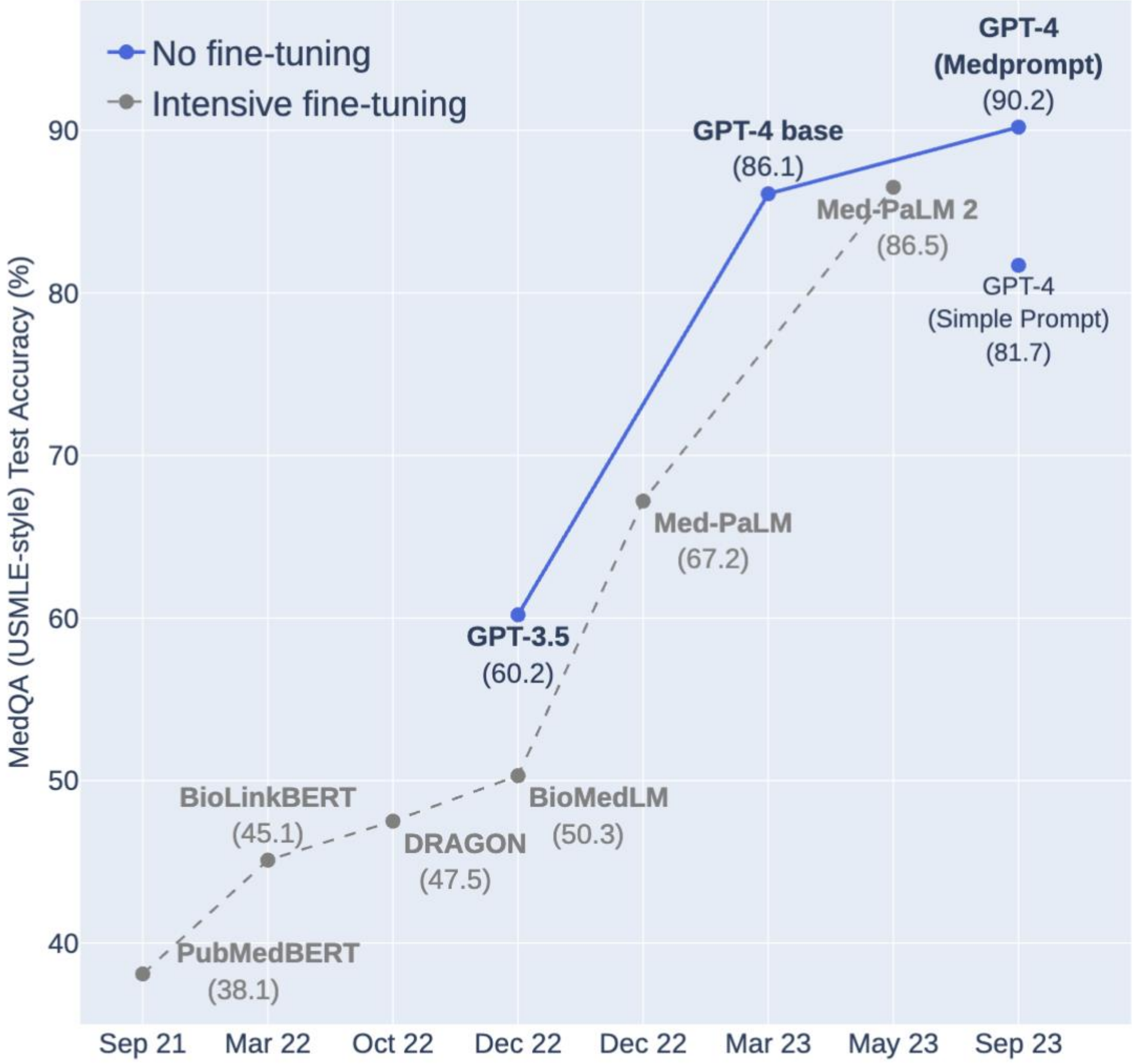
Evolving ML algorithms can also improve in performance

Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine

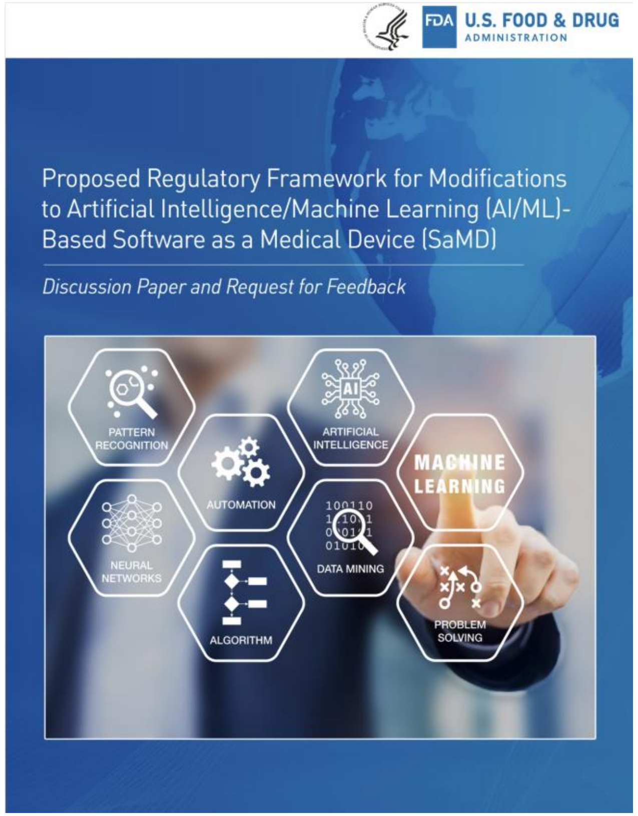
Harsha Nori[*‡], Yin Tat Lee[*], Sheng Zhang[*], Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney[†], Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz[‡]

Microsoft

November 2023



There is a need for model monitoring and updating…



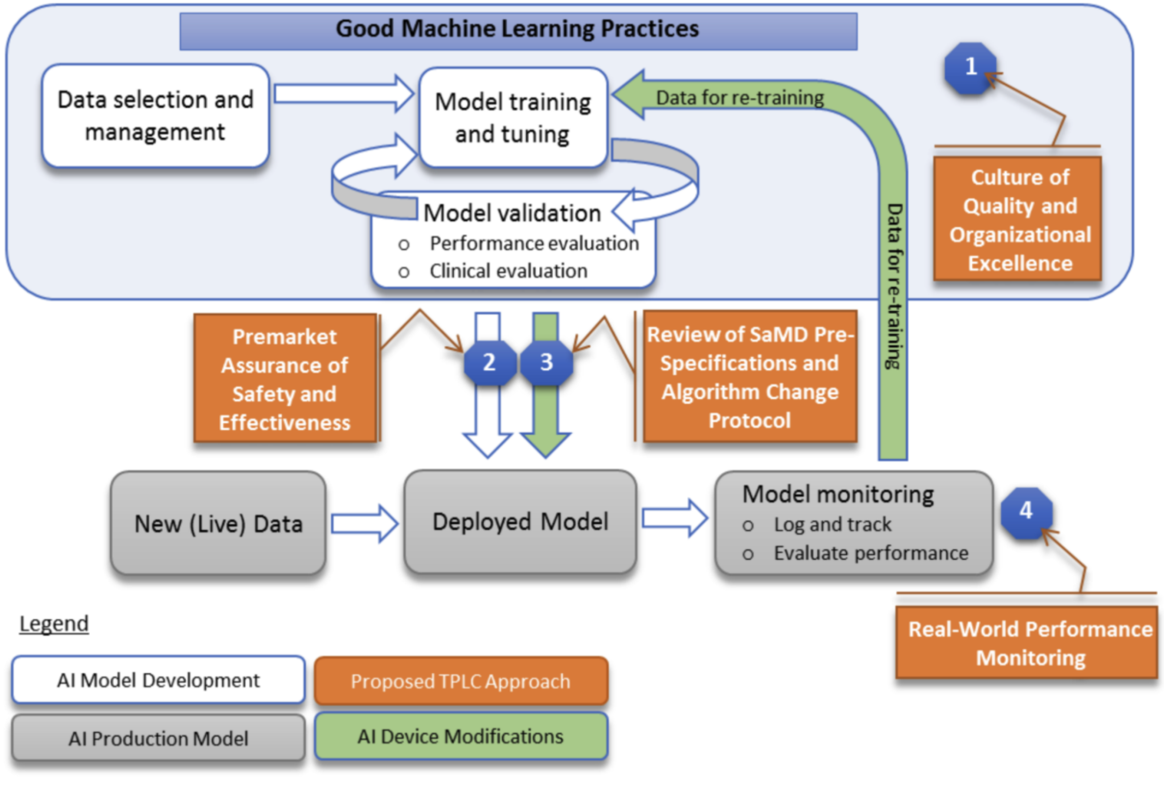


Figure 2: Overlay of FDA's TPLC approach on AI/ML workflow

There is a need for model monitoring and updating…

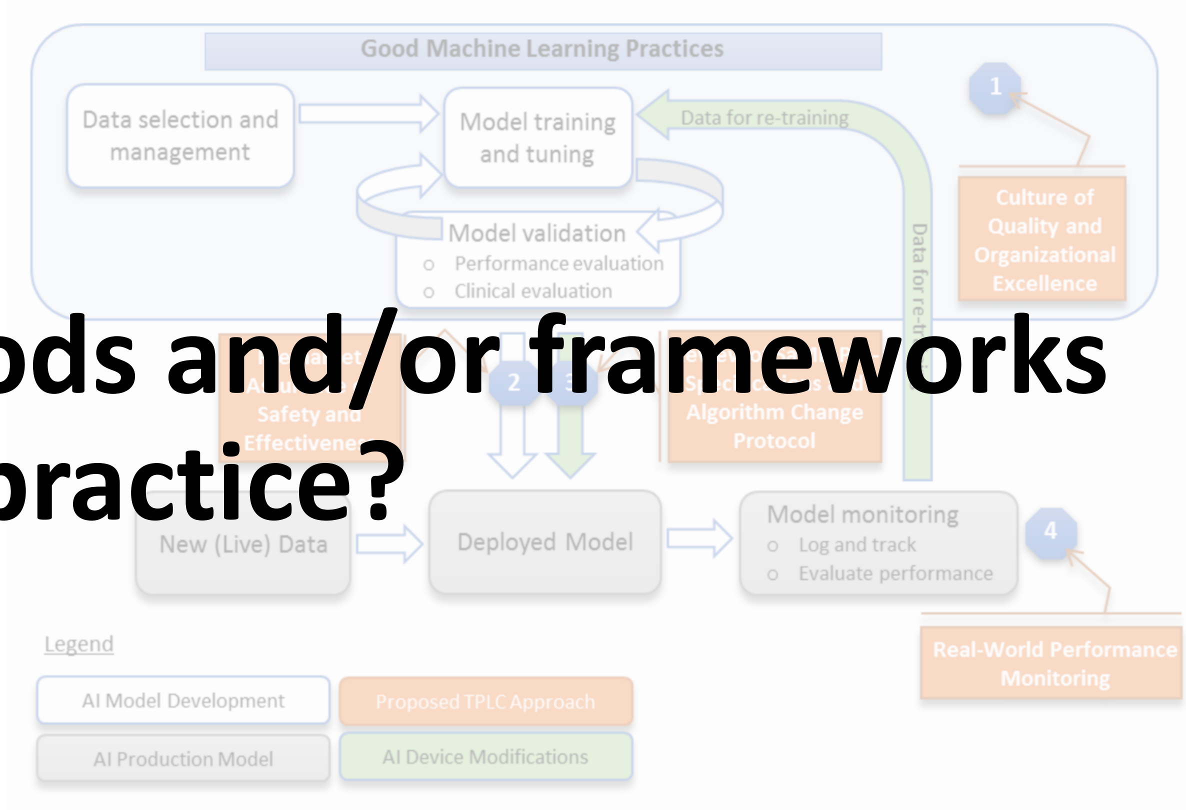


Figure 2: Overlay of FDA's TPLC approach on AI/ML workflow

… But what methods and/or frameworks should we use in practice?

Do we need something entirely new?

npj | digital medicine

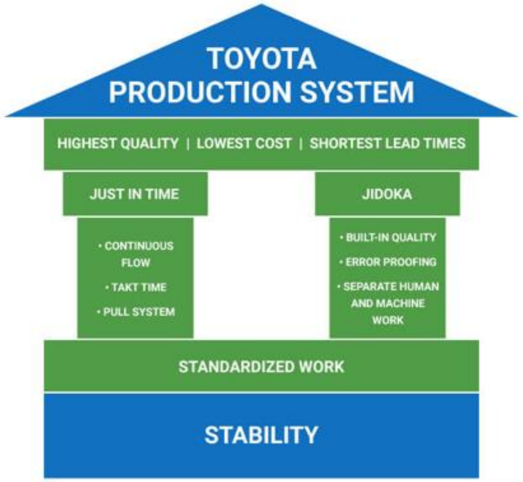
Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare

Jean Feng[1,2] ✉, Rachael V. Phillips[3], Ivana Malenica[3], Andrew Bishara[2,4], Alan E. Hubbard[3], Leo A. Celi[5] and Romain Pirracchio[2,4]

Model monitoring ⟷ Quality assurance (QA)

Model updating ⟷ Quality improvement (QI)

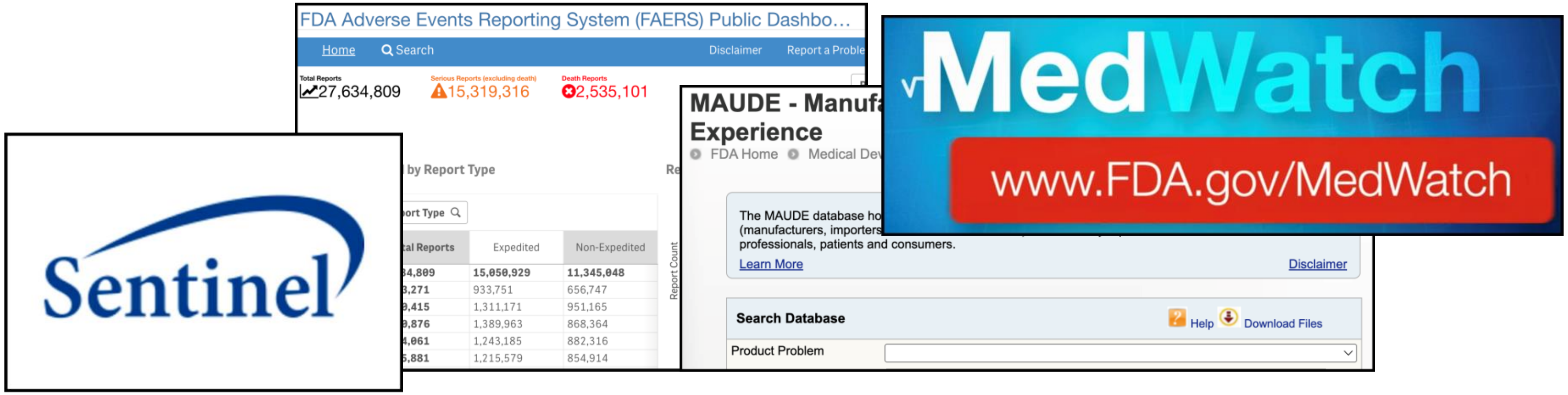
Quality assurance (QA) & Quality improvement (QI)



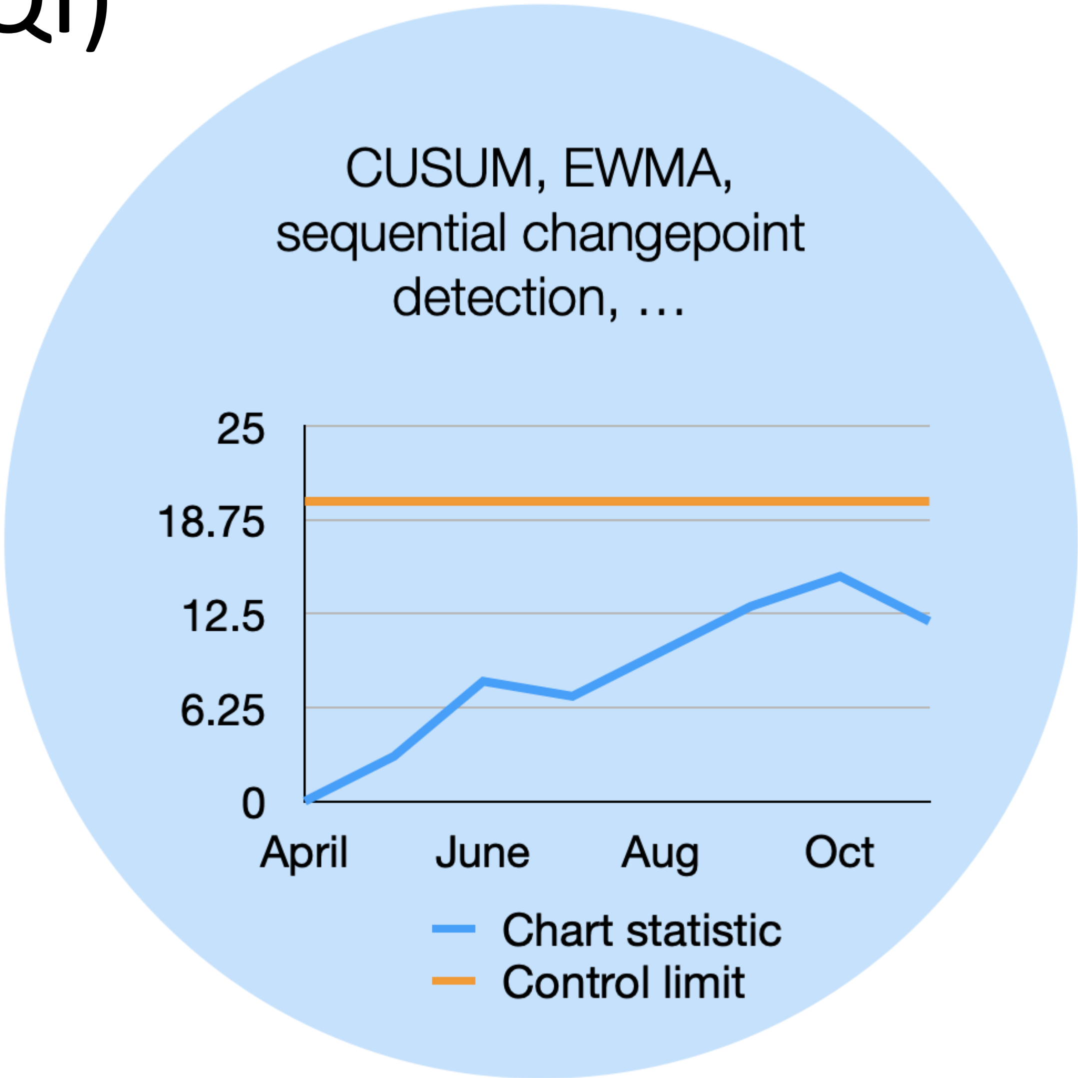
Manufacturing

Healthcare

Public Health

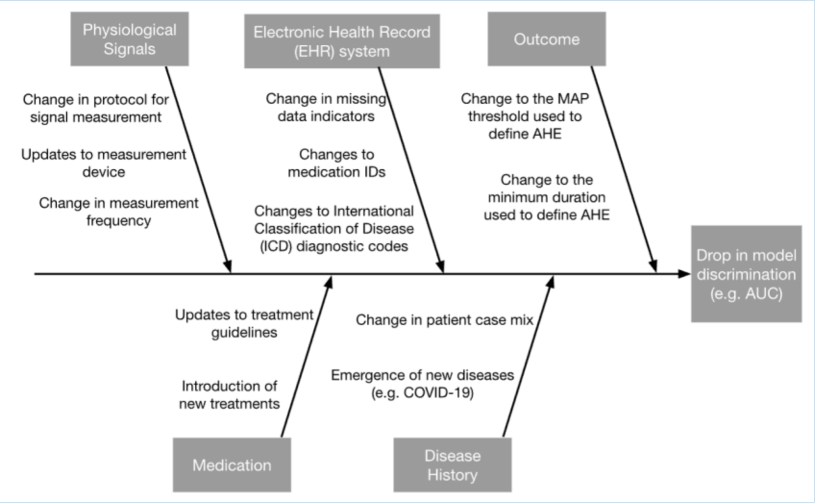


Quality assurance (QA) & Quality improvement (QI)

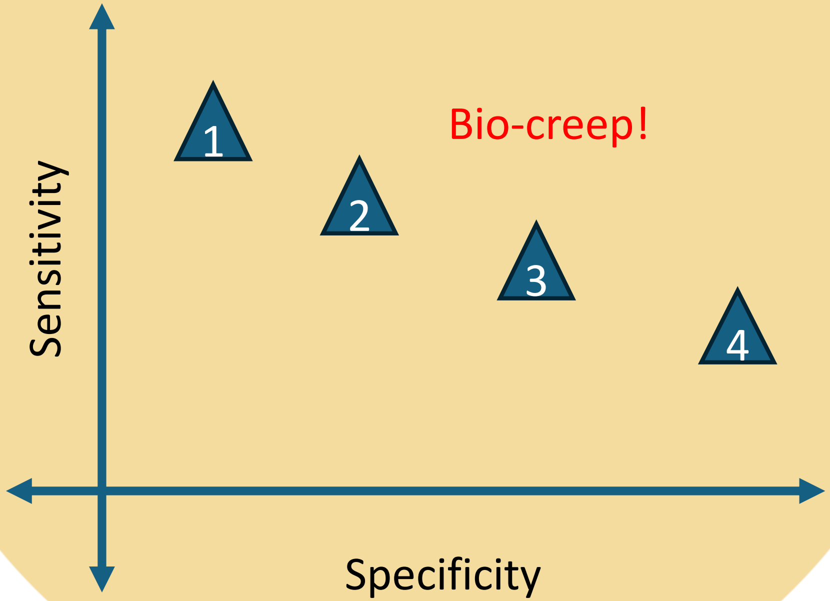


Quality assurance (QA) & Quality improvement (QI)

Cause-and-effect plots



Online hypothesis testing



Feng et. al. 2020

Cool, there are tools for ML QA/QI!

some

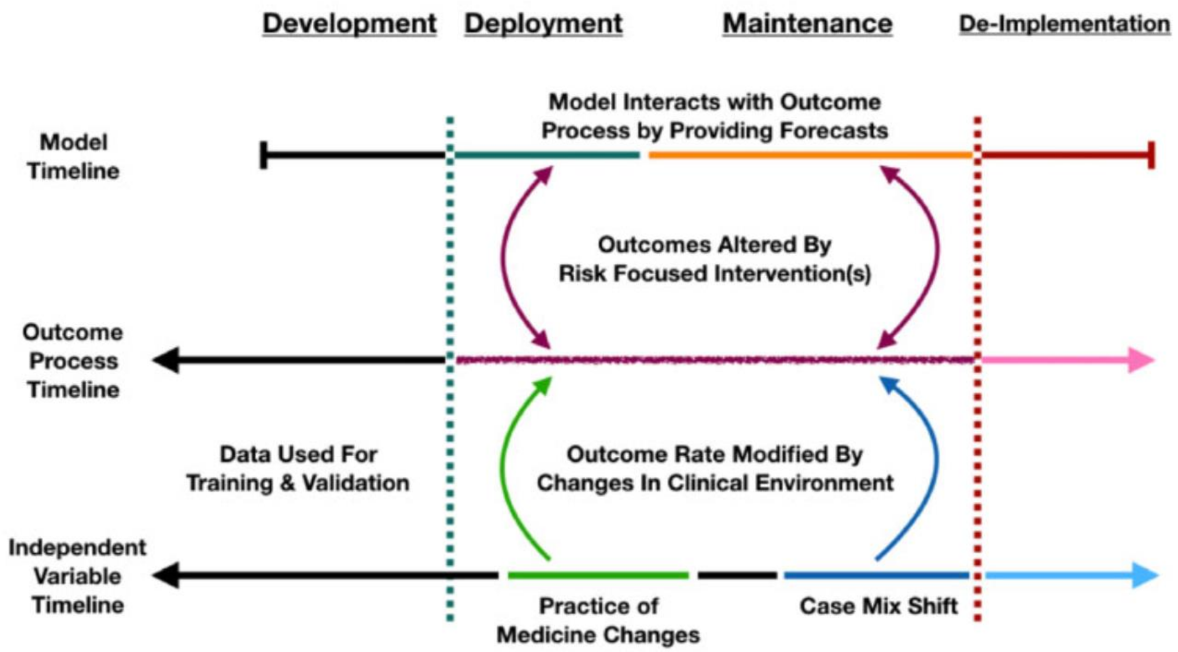
So is model monitoring and updating easy?

No!

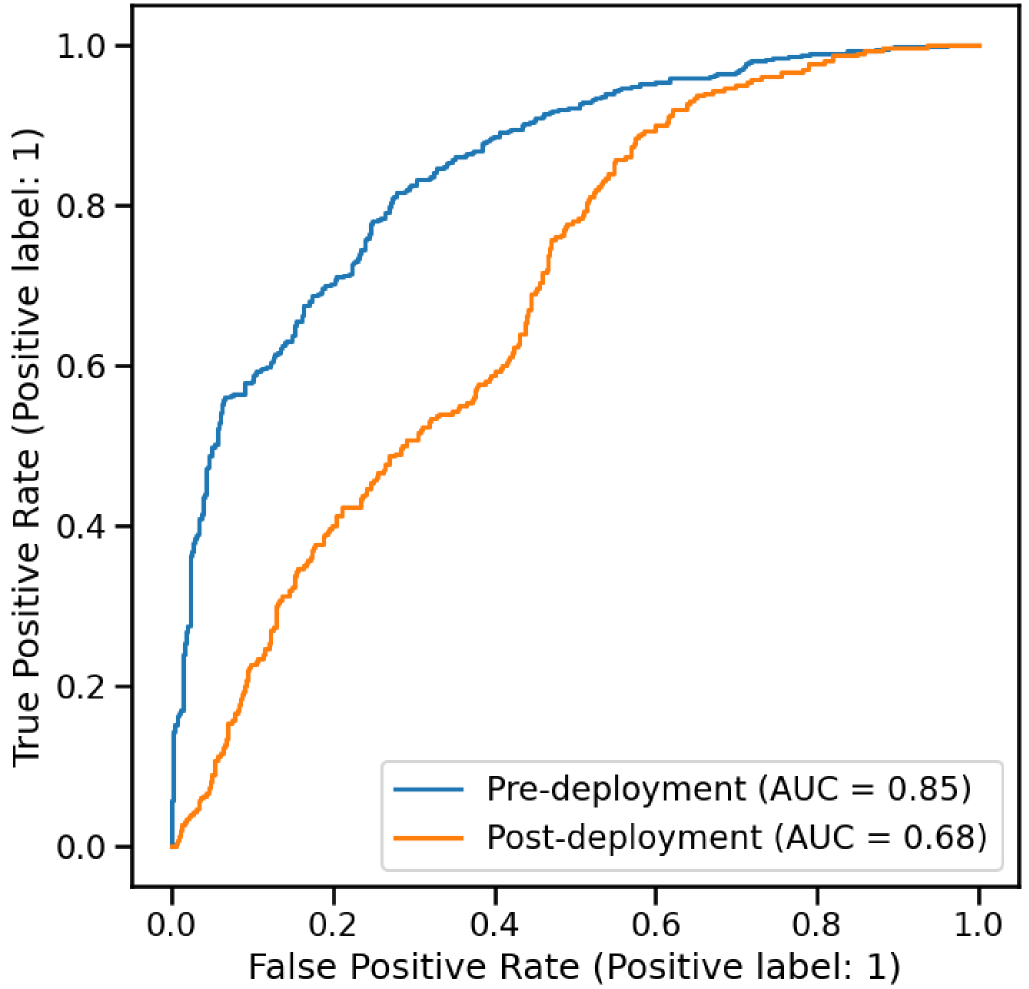
Open challenges

- How do we adjust our monitoring strategies when the ML algorithm impacts its environment?
- Not all monitoring procedures are created equal
- Is there a way to *continuously* update models while guaranteeing model safety and effectiveness?
- If performance deterioration is observed, can we identify the cause?
- How does one monitor/update algorithms when the true labels are unobserved or observed only after a substantial delay?
- How do we monitor generative AI algorithms?
- …

Challenge 1: When ML algorithms impact their environment



Lenert et al 2019



Challenge 1: When ML algorithms impact their environment

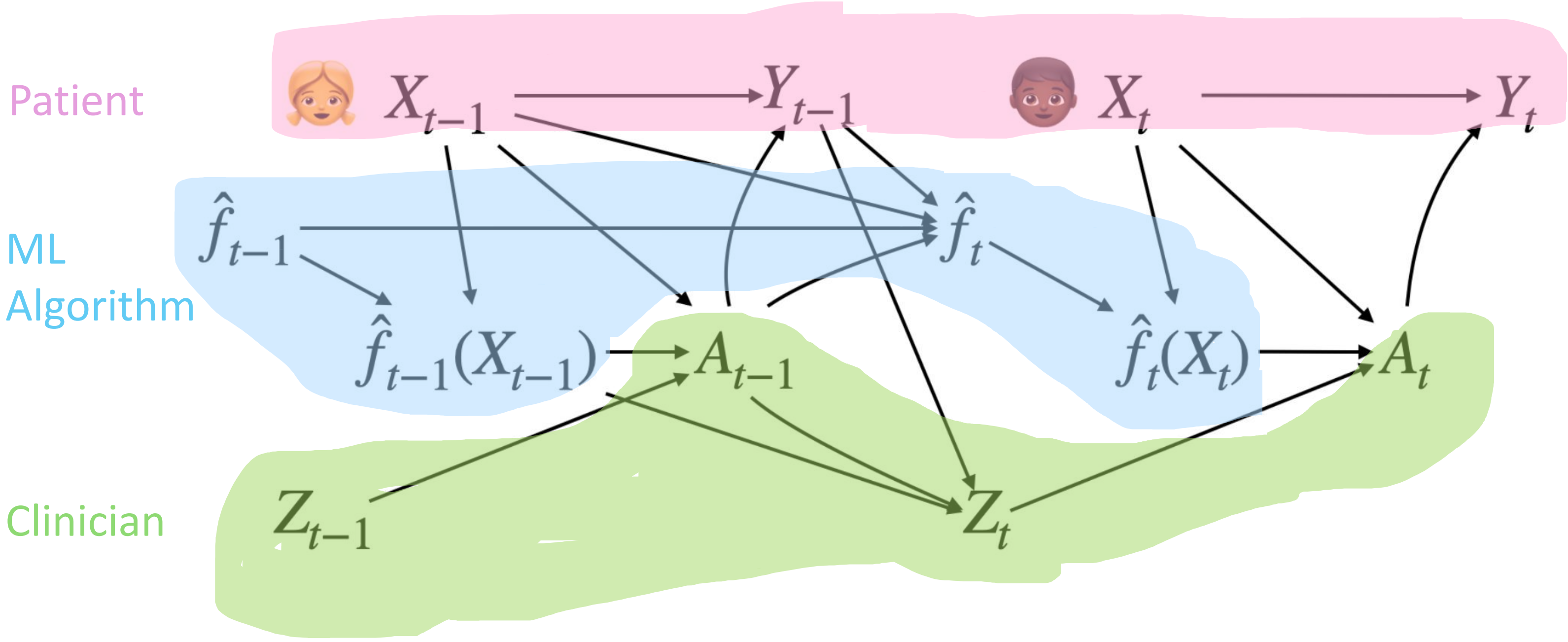
- Causal inference methods help us answer the crucial question “what is the performance of the ML algorithm if it did not modify clinician behavior?”
- **Algorithms for model monitoring and updating need to integrate causal reasoning**

Designing monitoring strategies for deployed machine learning algorithms: navigating performativity through a causal lens

Jean Feng[1*], Adarsh Subbaswamy[2], Alexej Gossmann[2], Harvineet Singh[1], Berkman Sahiner[2], Mi-Ok Kim[1], Gene Pennello[2], Nicholas Petrick[2], Romain Pirracchio[1], Fan Xia[1]

[1] *University of California, San Francisco*

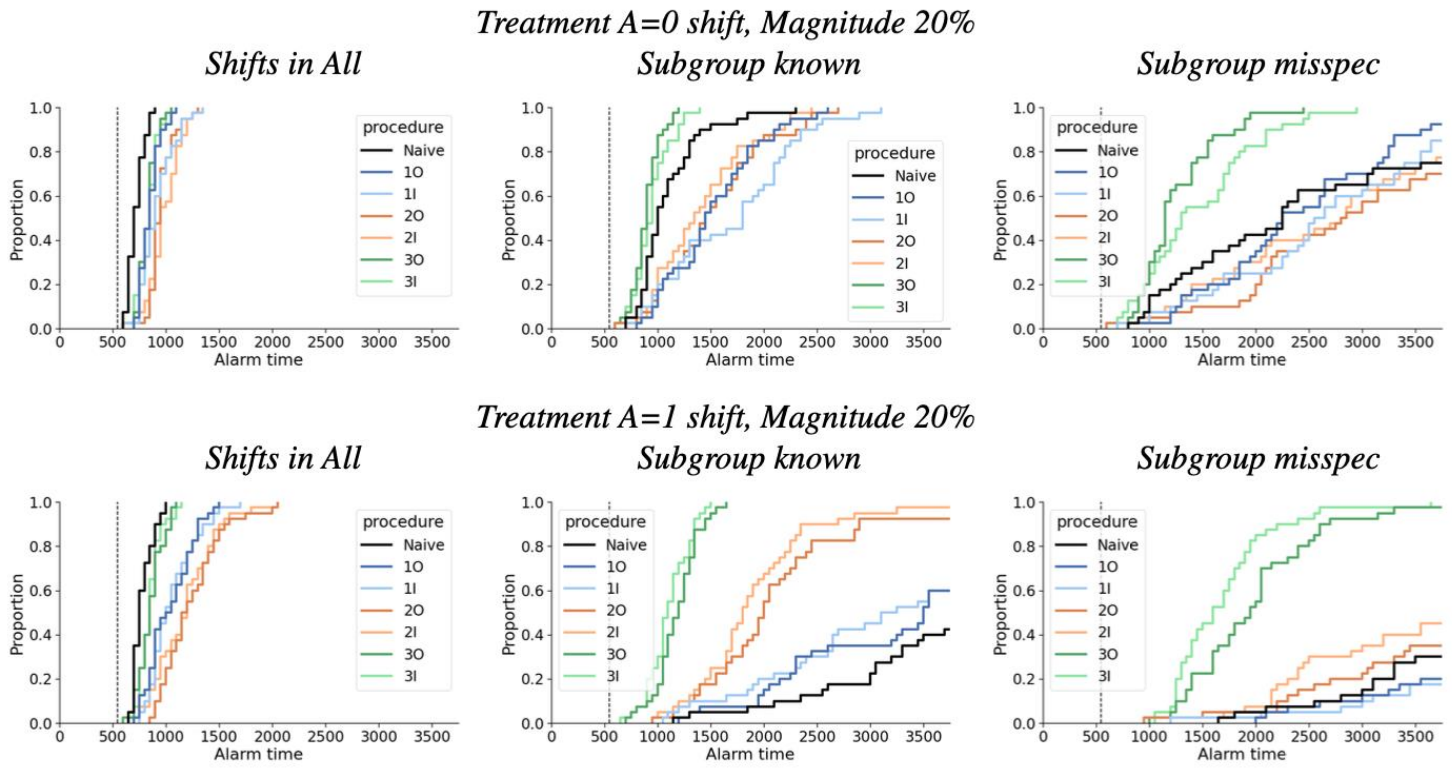
[2] *U.S. Food and Drug Administration, Center for Devices and Radiological Health*



Challenge 2: Not all monitoring algorithms are created equal

- There are a **multitude** of ways to monitor the same algorithm, including:
  - Which performance metrics are monitored
  - What data is collected
  - What assumptions are needed
- Procedures can vary widely in their operating characteristics

Procedure	Interpretability	Fairness	Data requirements	Assumptions	Hyperparameters
1I	High	None	Interventional	Positivity	None
1O	High	None	Observational, Must conduct pre-monitoring phase	Positivity, Conditional Exchangeability	None
2I	High	Moderate	Interventional	Positivity	Subgroups, subgroup PPV/NPV
2O	High	Moderate	Observational, Must conduct pre-monitoring phase	Positivity, Conditional Exchangeability	Subgroups, subgroup PPV/NPV
3I	Medium	Strong	Interventional	None	Subgroups, tolerance level
3O	Medium	Strong	Observational, No pre-monitoring phase	Conditional Exchangeability	Subgroups, tolerance level



Challenge 2: Not all monitoring algorithms are created equal

- The existence of a monitoring strategy does not automatically imply that an ML system is safe and effective.
- Encourage proper design of monitoring solutions through:
  - **Guidance on comprehensive evaluation of ML monitoring strategies**
  - Transparency

Roadmap towards comprehensive evaluation of ML monitoring systems

1. **Define potential monitoring criteria**
2. **Enumerate data sources and define the causal models**
3. **Describe candidate monitoring strategies**
4. **Compare the pros and cons of candidate strategies**

Select final strategy after discussion with team members and stakeholders

Open challenges

- How do we adjust our monitoring strategies when the ML algorithm impacts its environment?
- Not all monitoring procedures are created equal
- Is there a way to *continuously* update models while guaranteeing model safety and effectiveness?
- If performance deterioration is observed, can we identify the cause?
- How does one monitor/update algorithms when the true labels are unobserved or observed only after a substantial delay?
- How do we monitor generative AI algorithms?
- …

Thanks!


Harvineet Singh


Romain Pirracchio


Andrew Bishara


Fan Xia


Mi-Ok Kim


Leo Celi


Alan Hubbard


Nicholas Petrick


Berkman Sahiner


Gene Pennello


Adarsh Subbaswamy


Alexej Gossmann