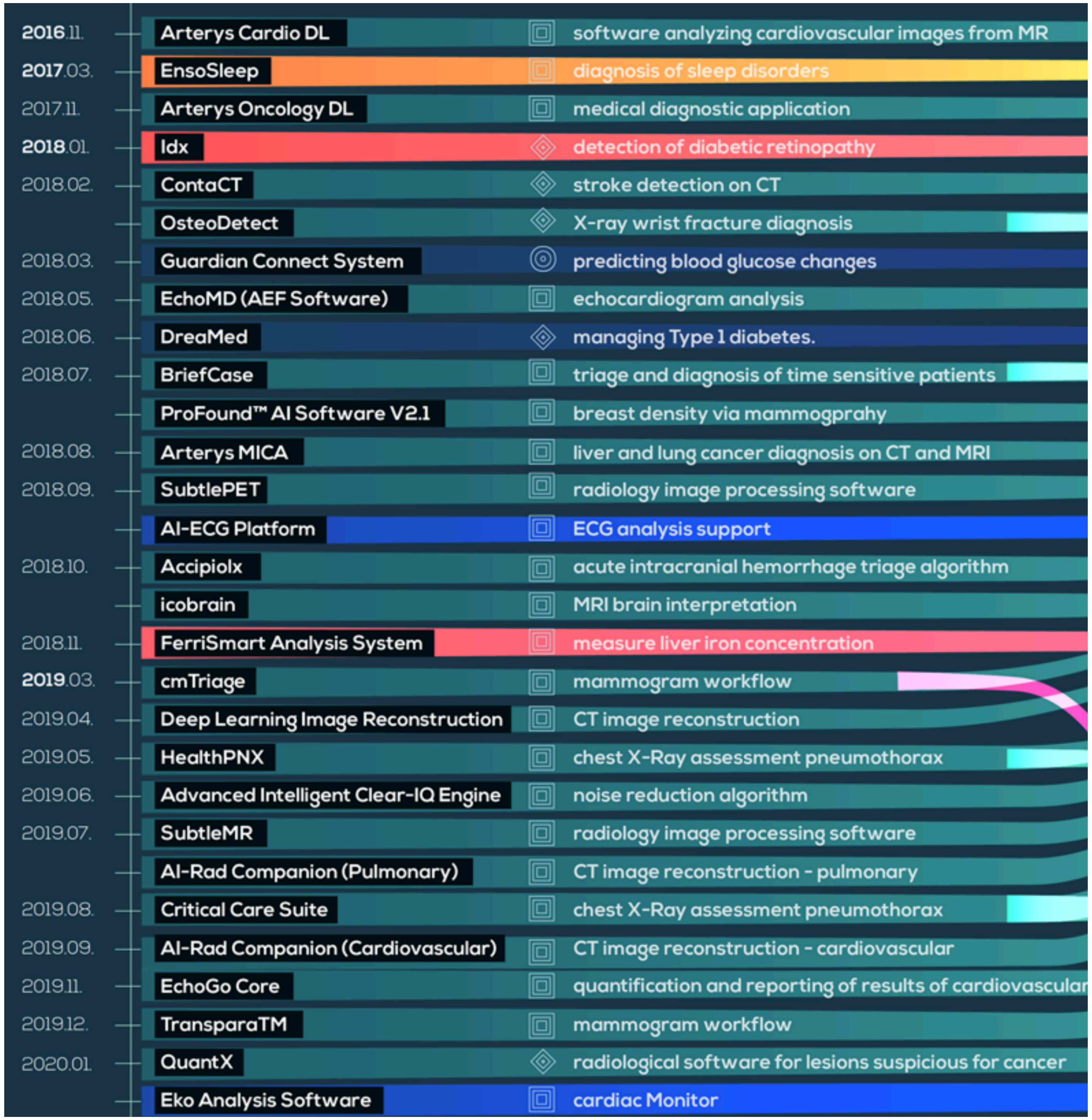
UCSF
University of California
San Francisco

Statistical Tools for Auditing Machine Learning Algorithms Across Subgroups and Time

Jean Feng
University of California, San Francisco

www.jeanfeng.com

FDA Approvals for Artificial Intelligence/ Machine Learning-based Software-as-a-Medical-Device (SaMD)



2016.11.	Arterys Cardio DL	software analyzing cardiovascular images from MR
2017.03.	EnsoSleep	diagnosis of sleep disorders
2017.11.	Arterys Oncology DL	medical diagnostic application
2018.01.	Idx	detection of diabetic retinopathy
2018.02.	ContaCT	stroke detection on CT
	OsteoDetect	X-ray wrist fracture diagnosis
2018.03.	Guardian Connect System	predicting blood glucose changes
2018.05.	EchoMD (AEF Software)	echocardiogram analysis
2018.06.	DreaMed	managing Type 1 diabetes.
2018.07.	BriefCase	triage and diagnosis of time sensitive patients
	ProFound™ AI Software V2.1	breast density via mammogprahy
2018.08.	Arterys MICA	liver and lung cancer diagnosis on CT and MRI
2018.09.	SubtlePET	radiology image processing software
	AI-ECG Platform	ECG analysis support
2018.10.	Accipiolx	acute intracranial hemorrhage triage algorithm
	icobrain	MRI brain interpretation
2018.11.	FerriSmart Analysis System	measure liver iron concentration
2019.03.	cmTriage	mammogram workflow
2019.04.	Deep Learning Image Reconstruction	CT image reconstruction
2019.05.	HealthPNX	chest X-Ray assessment pneumothorax
2019.06.	Advanced Intelligent Clear-IQ Engine	noise reduction algorithm
2019.07.	SubtleMR	radiology image processing software
	AI-Rad Companion (Pulmonary)	CT image reconstruction - pulmonary
2019.08.	Critical Care Suite	chest X-Ray assessment pneumothorax
2019.09.	AI-Rad Companion (Cardiovascular)	CT image reconstruction - cardiovascular
2019.11.	EchoGo Core	quantification and reporting of results of cardiovascular
2019.12.	TransparaTM	mammogram workflow
2020.01.	QuantX	radiological software for lesions suspicious for cancer
	Eko Analysis Software	cardiac Monitor

Benjamens et. al. 2020

Timeline of regulatory developments for AI/ML-based medical devices

- **2019**: Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback

- **2021**: Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Action Plan

- **2021:** Good Machine Learning Practice for Medical Device Development: Guiding Principles

- **2023**: Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions: Draft Guidance

How can we verify that an ML-based medical device is consistently safe and effective?

Timeline of regulatory developments for AI/ML-based medical devices

- **2019**: Proposed Regulatory Framework for Modifications to Artificial Intelligence/ Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback — 2
- **2021**: Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Action Plan — 1 2
- **2021:** Good Machine Learning Practice for Medical Device Development: Guiding Principles — 2
- **2023**: Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/ Machine Learning (AI/ML)-Enabled Device Software Functions: Draft Guidance — 1 2

1 **Performance across subgroups** (Algorithm Bias & Robustness)

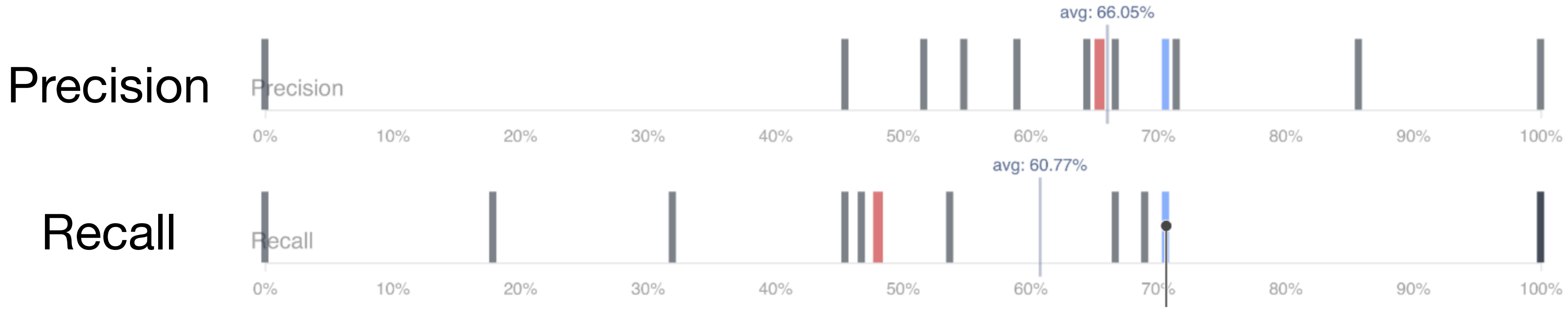
2 **Performance over time** (Real-World performance monitoring)

1 Performance across subgroups

FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning

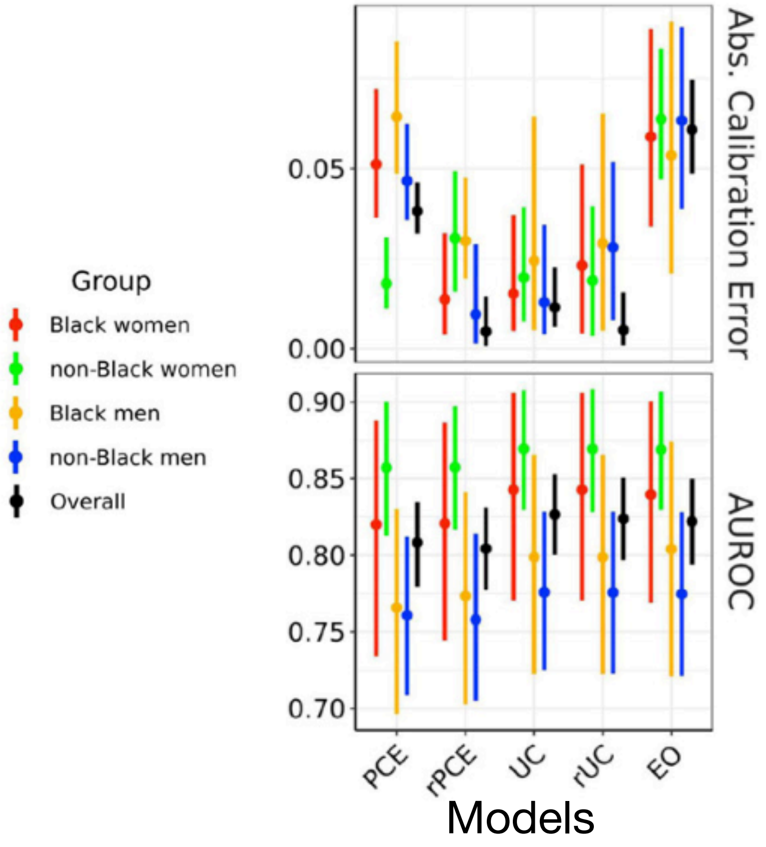
Ángel Alexander Cabrera Will Epperson Fred Hohman Minsuk Kahng
Jamie Morgenstern Duen Horng (Polo) Chau*

Georgia Institute of Technology



Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation

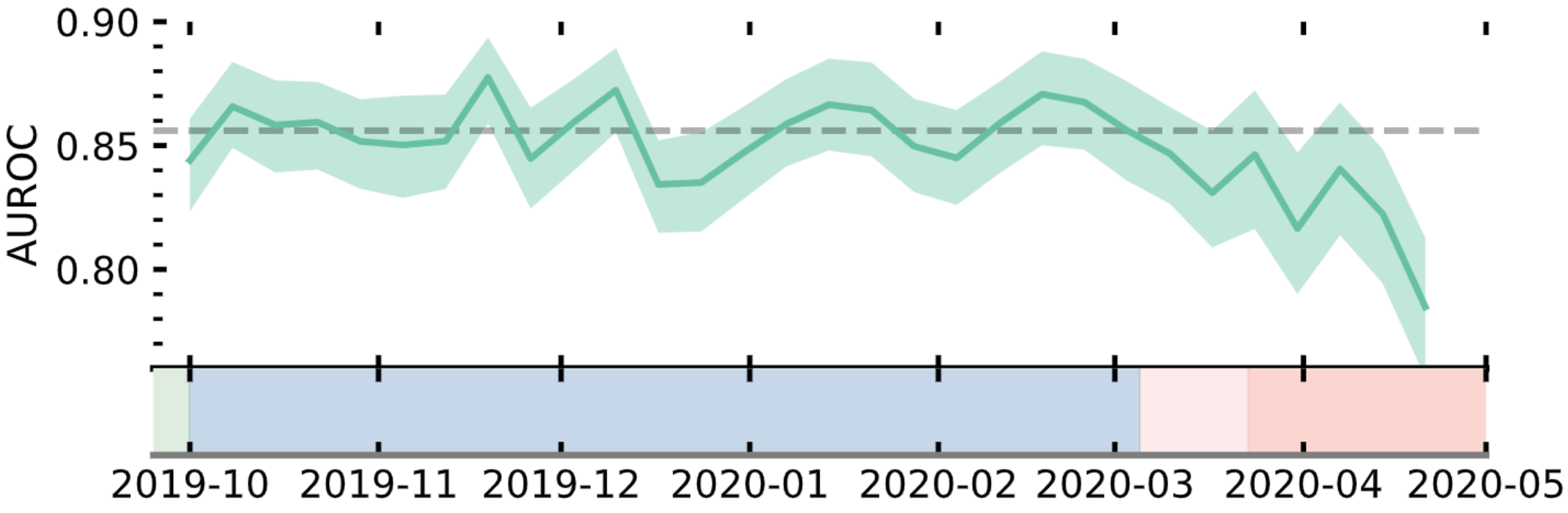
Agata Foryciarz [1,2] Stephen R Pfohl,[2] Birju Patel,[2] Nigam Shah [2]



2 Performance over time

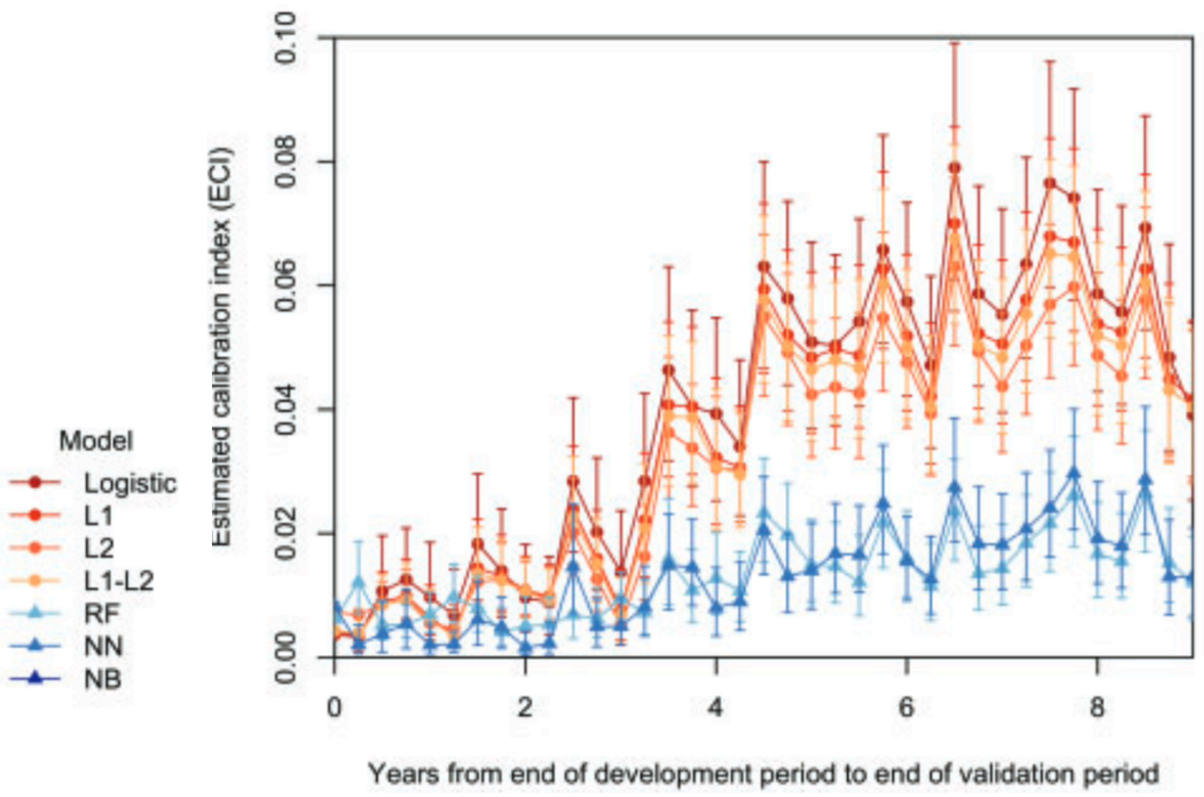
Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19

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]



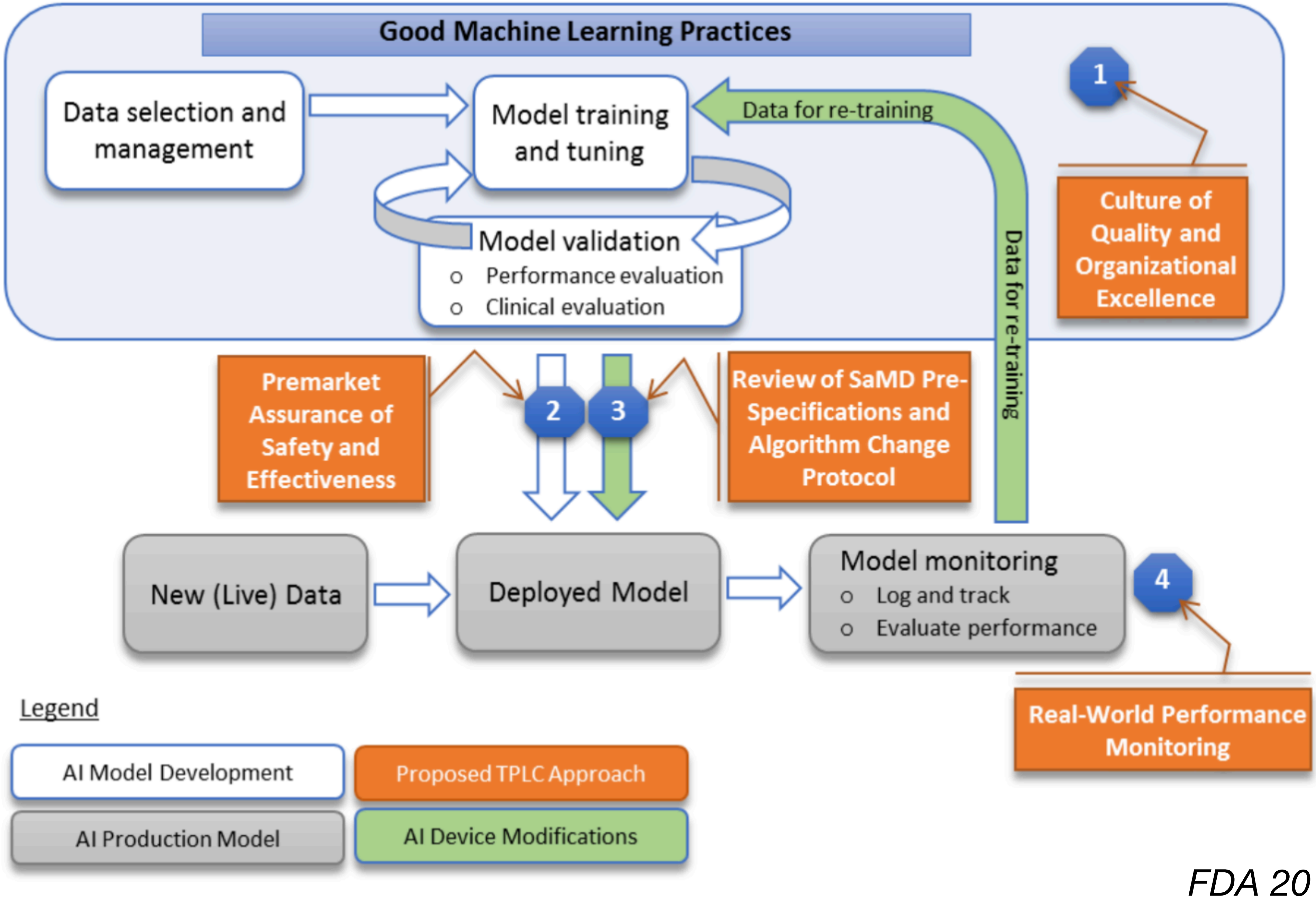
Calibration drift in regression and machine learning models for acute kidney injury

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



The role of model audits

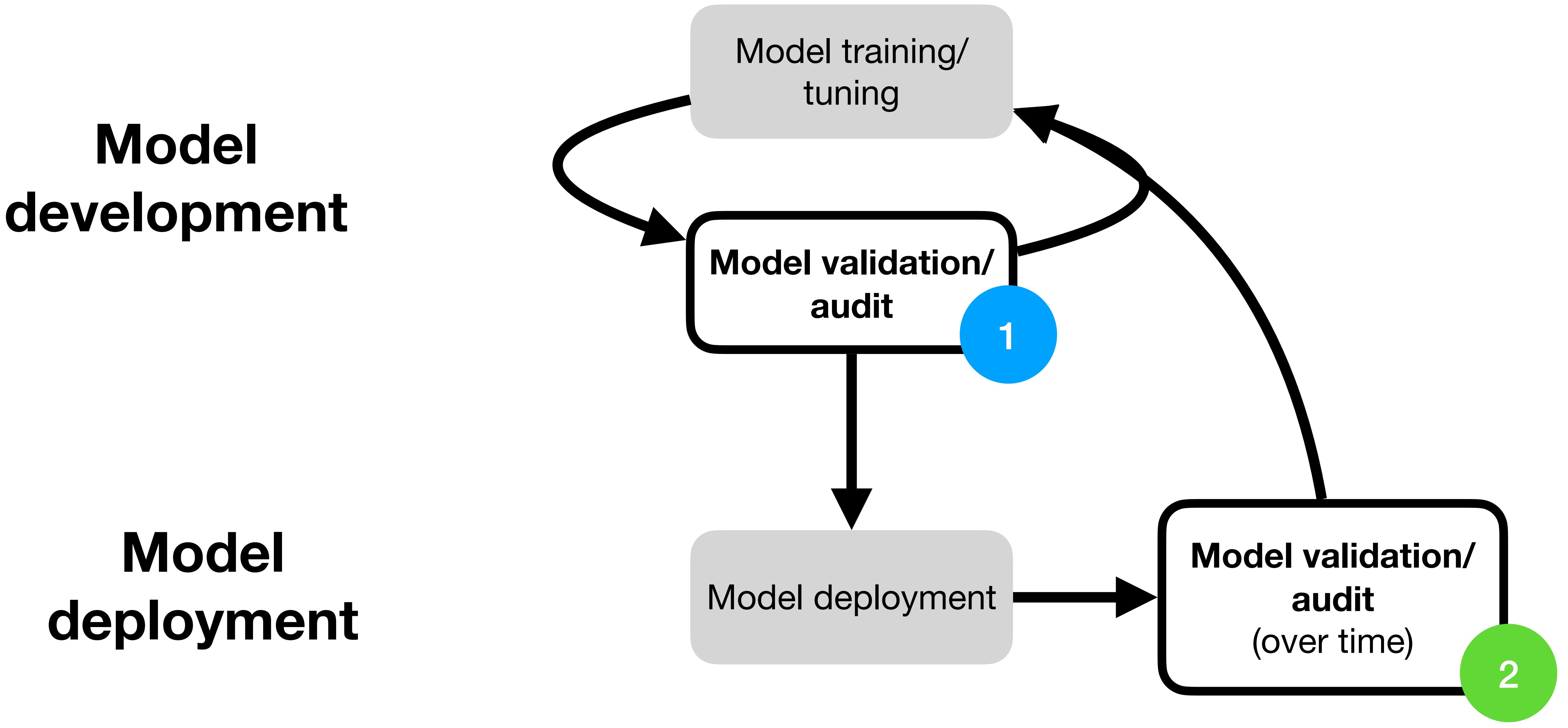
Model audits are the first step to ensuring the safety and effectiveness of ML-based medical devices.


Good Machine Learning Practices
Data selection and management
Model training and tuning
Data for re-training
Model validation
Performance evaluation
Clinical evaluation
Data for re-training
1
Culture of Quality and Organizational Excellence
Premarket Assurance of Safety and Effectiveness
2
3
Review of SaMD Pre-Specifications and Algorithm Change Protocol
New (Live) Data
Deployed Model
Model monitoring
Log and track
Evaluate performance
4
Real-World Performance Monitoring
Legend
AI Model Development
Proposed TPLC Approach
AI Production Model
AI Device Modifications

FDA 2019

The role of model audits

Model audits are the first step to ensuring the safety and effectiveness of ML-based medical devices.



1 **Performance across subgroups**
(Algorithm Bias & Robustness)

2 **Performance over time**
(Real-World performance monitoring)

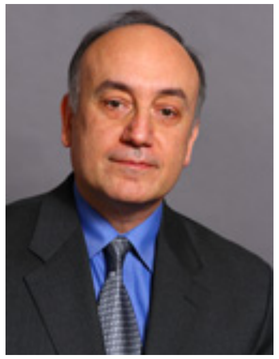
Outline

1. Auditing performance of ML algorithms across subgroups, *when the subgroups are unknown*

2. Auditing performance of ML algorithms over time, *in the presence of performativity*

Changepoint detection problems

Alexej Gossmann



Berkman Sahiner

Nicholas Petrick

Gene Pennello

Romain Pirracchio

Outline

1. Auditing performance of ML algorithms across subgroups, *when the subgroups are unknown*

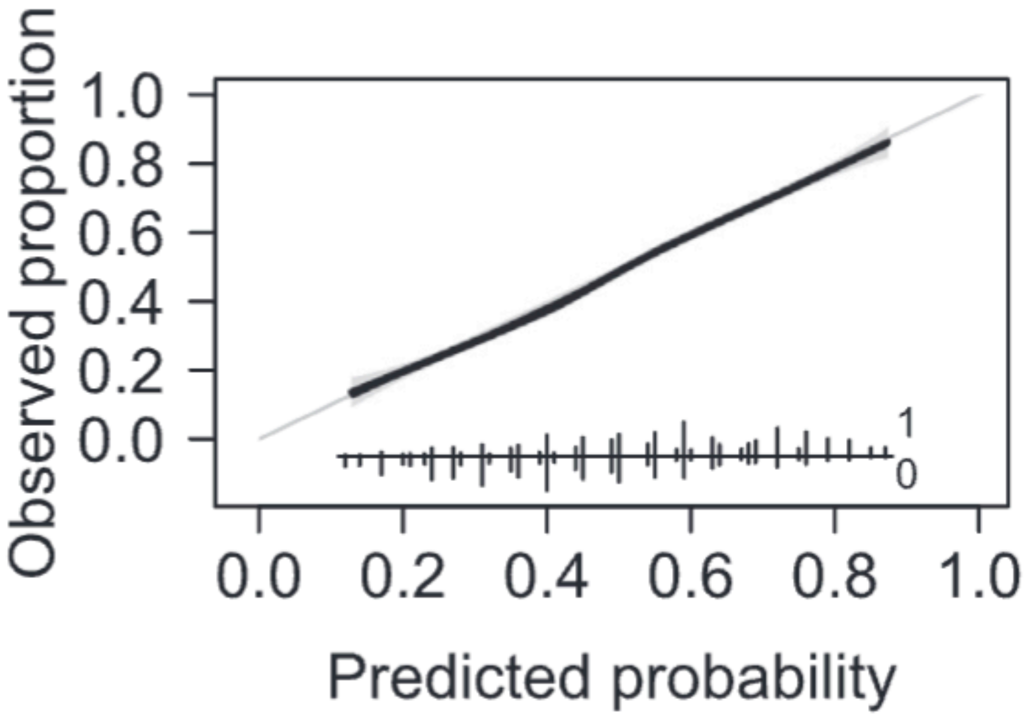
2. Auditing performance of ML algorithms over time, *in the presence of performativity*

Model calibration

When a risk prediction model $\hat{p}$ is used to inform medical decision making, a fundamental requirement is that the model is “reliable,” in that it is well-calibrated:

$$\Pr\left(Y = 1 \mid \hat{p}(X) = q\right) = q$$

$$\forall q \in [0,1]$$

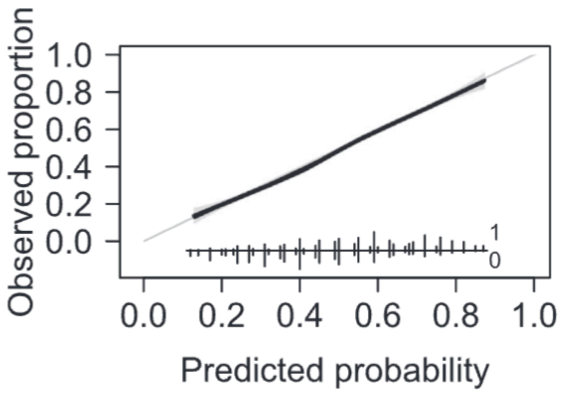


The calibration hierarchy

However, model calibration can vary across different subgroups.

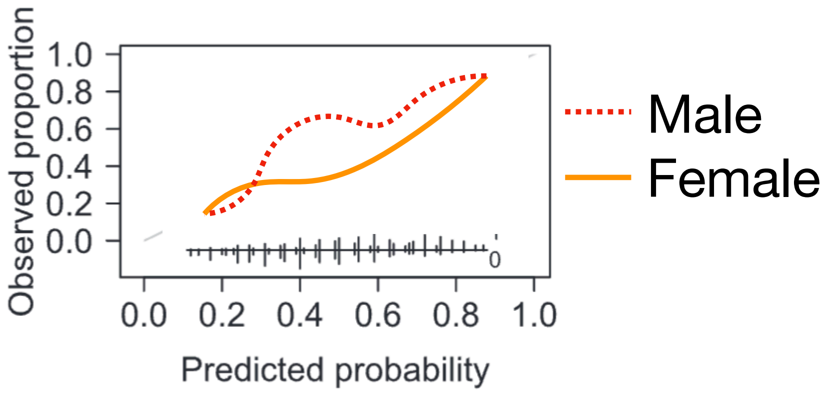
A model $\hat{p}$ that is well-calibrated across all subgroups is “strongly calibrated.”

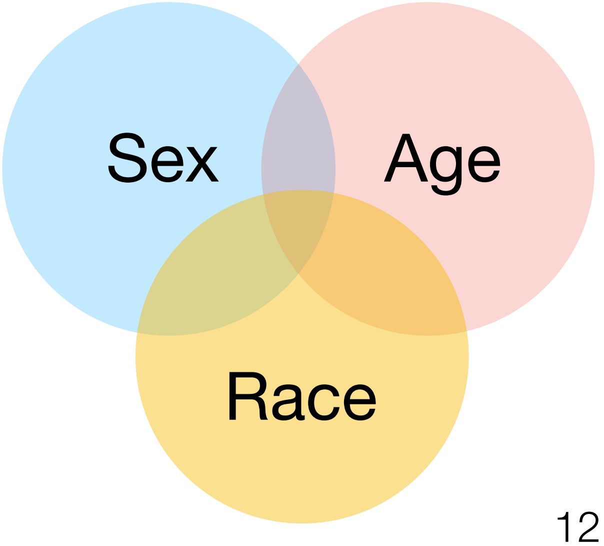
“Moderate” $\Pr\left(Y = 1 \mid \hat{p}(X) = q\right) = q$



“Strong” $\Pr\left(Y = 1 \mid \hat{p}(X) = q, X \in A\right) = q$

for all subgroups A



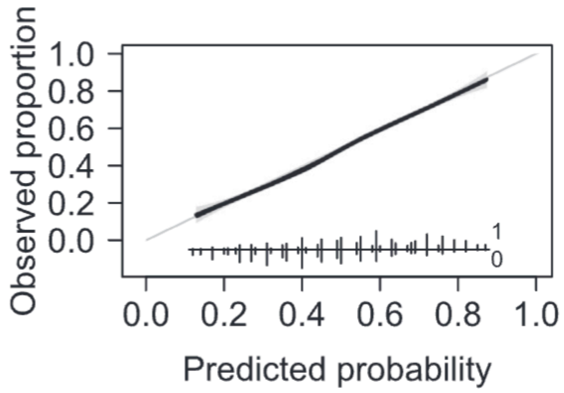


The calibration hierarchy

However, model calibration can vary across different subgroups.

A model $\hat{p}$ that is well-calibrated across all subgroups is “strongly calibrated.”

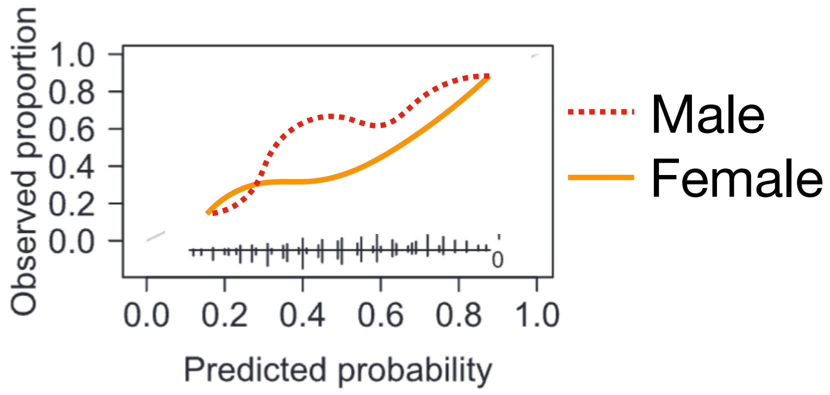
“Moderate” $\Pr\left(Y = 1 \mid \hat{p}(X) = q\right) = q$



“Strong” $\Pr\left(X \in A_\delta\right) \le \gamma$

where

$$\underbrace{A_\delta = \left\{X : \left|p_0(X) - \hat{p}(X)\right| > \delta\right\}}_{\textit{Poorly calibrated subgroup}}$$



Testing for strong calibration

- **Goal**: Construct an omnibus test that answers the question

  "Does a poorly-calibrated subgroup exist?"

$$H_0 : \Pr\left(X \in A_\delta\right) \leq \gamma \quad \text{where } A_\delta = \underbrace{\left\{X : \left|p_0(X) - \hat{p}(X)\right| > \delta\right\}}_{\textit{Poorly calibrated subgroup}}$$

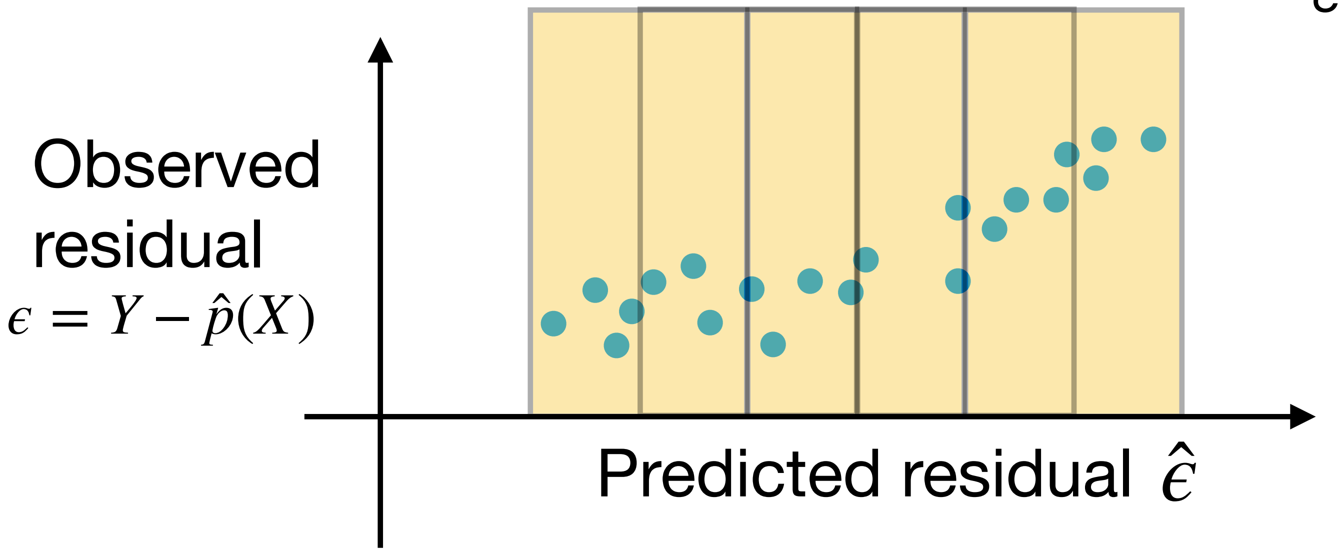
$$H_1 : \Pr\left(X \in A_\delta\right) > \gamma$$

- **Statistical challenges**: Power for identifying poorly-calibrated subgroups is often low because
  - Correction for multiple testing after searching over a large number of potential subgroups
  - Little remaining signal if a highly flexible model was fit (e.g. via machine learning)

Testing for strong calibration: Existing approach

- Suppose we trained a model $\hat{g}$ to predict the residual $\epsilon = Y - \hat{p}(X)$ at each X.

- Bin test observations by their predicted residuals and conduct a Chi-squared test (Goodness-of-fit Test)


Observed residual
$\epsilon = Y - \hat{p}(X)$
Predicted residual $\hat{\epsilon}$

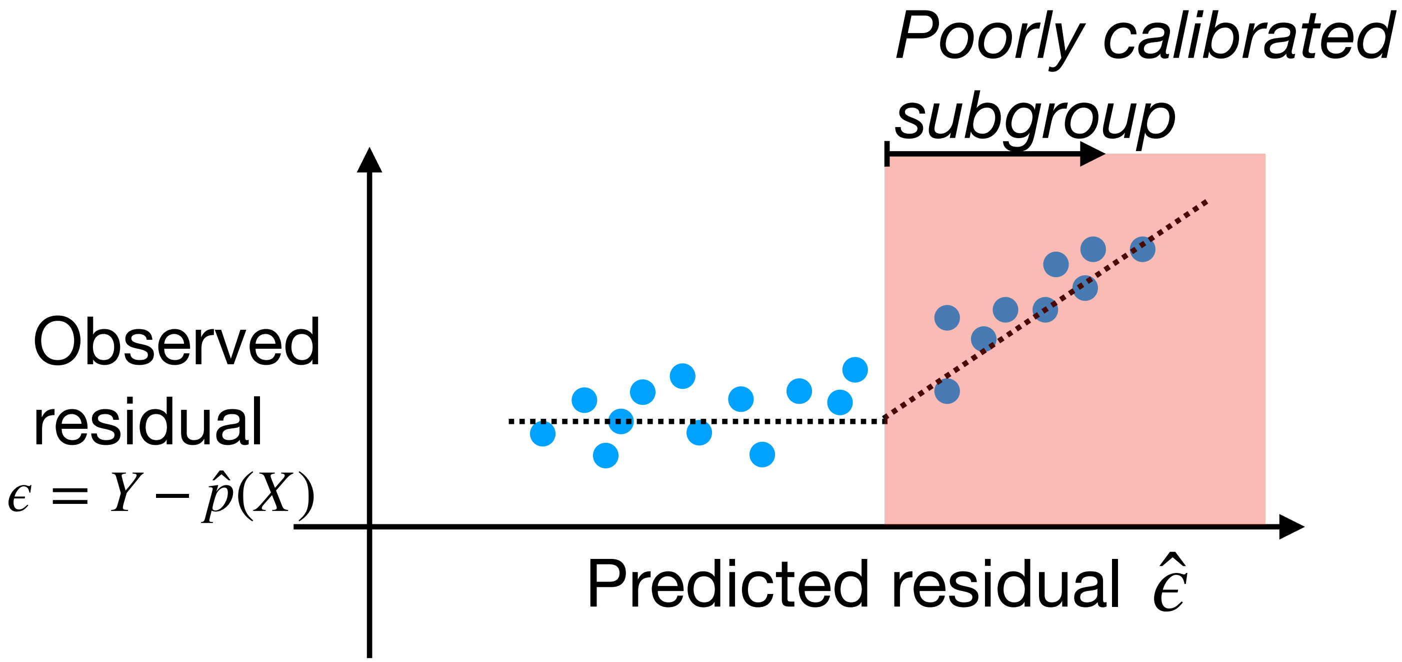
*Check if **observed** event rates matches **expected** event rates for each bin*

Zhang et. al. 2021

Testing for strong calibration = Testing for changepoints

- Suppose we trained a model $\hat{g}$ to predict the expected residual at each X.

- If we order test observations by their predicted residuals, we expect a drop in the association between the observed and predicted residuals…



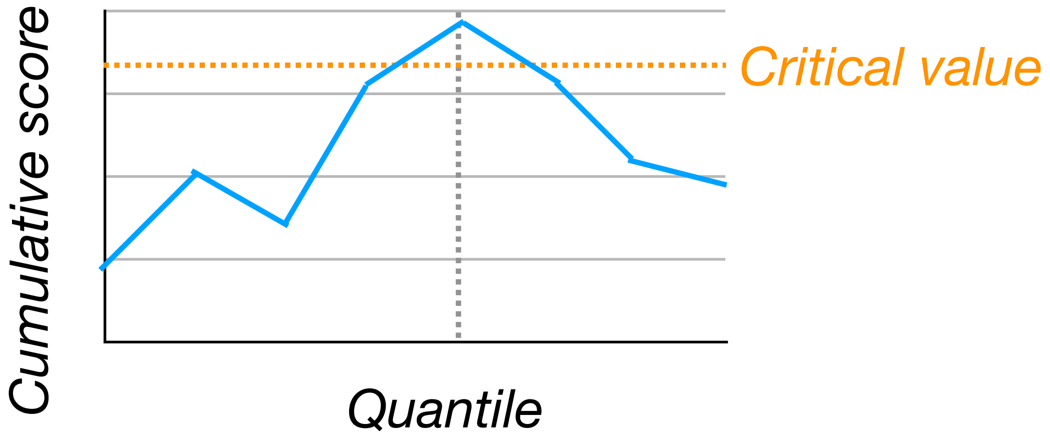
+ Avoids specifying subgroup size.

+ Detecting small subgroups $\Longleftrightarrow$ Detecting early changepoints

+ Respects structure learned by the residual model

Testing for strong calibration = Testing for changepoints

- Suppose we trained a model $\hat{g}$ to predict the expected residual at each X.



- If we order test observations by their predicted residuals, we expect a drop in the association between the observed and predicted residuals…

Test statistic: Score-based CUSUM

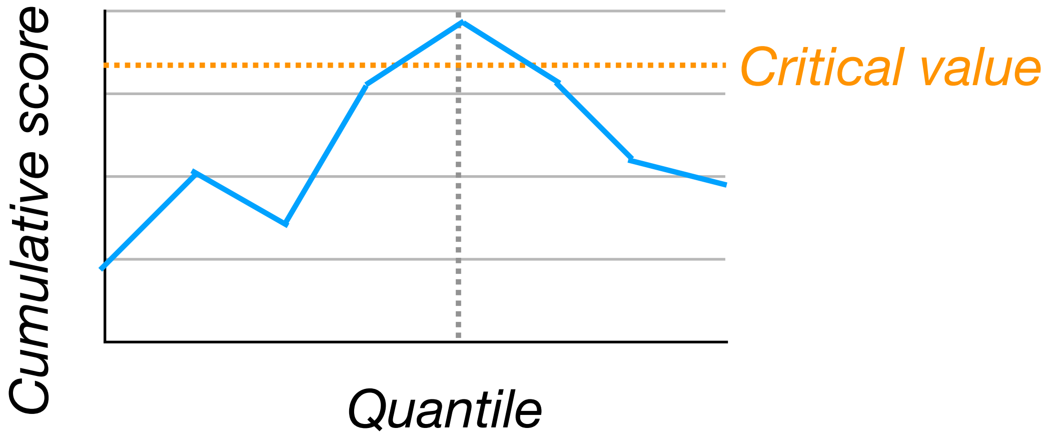
$$\max_{k=1,\cdots,K} \sup_{\gamma\geq 0} \frac{1}{n}\sum_{i=1}^{n} \underbrace{\big(Y_i - \hat{p}_\delta(Y_i|X_i)\big)\, \hat{g}_k(X_i)}_{\text{Score}}\, 1\{\hat{g}_k(X_i) \geq \gamma\}$$

Testing for strong calibration = Testing for changepoints

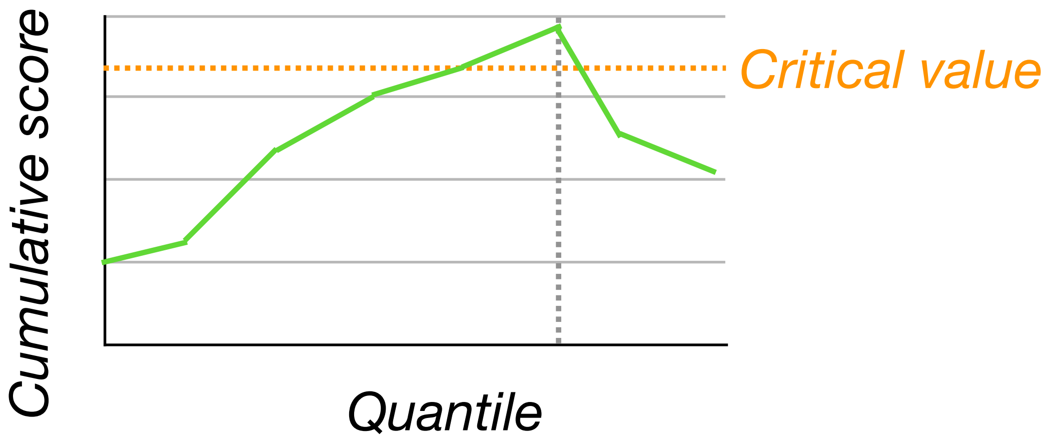
- Suppose we trained an ***ensemble of machine learning models*** $\{\hat{g}_k\}$ to predict the expected residual at each X.

- If we order test observations by their predicted residuals, we expect a drop in the association between the observed and predicted residuals…

Random Forest: $\hat{g}_{(1)}$

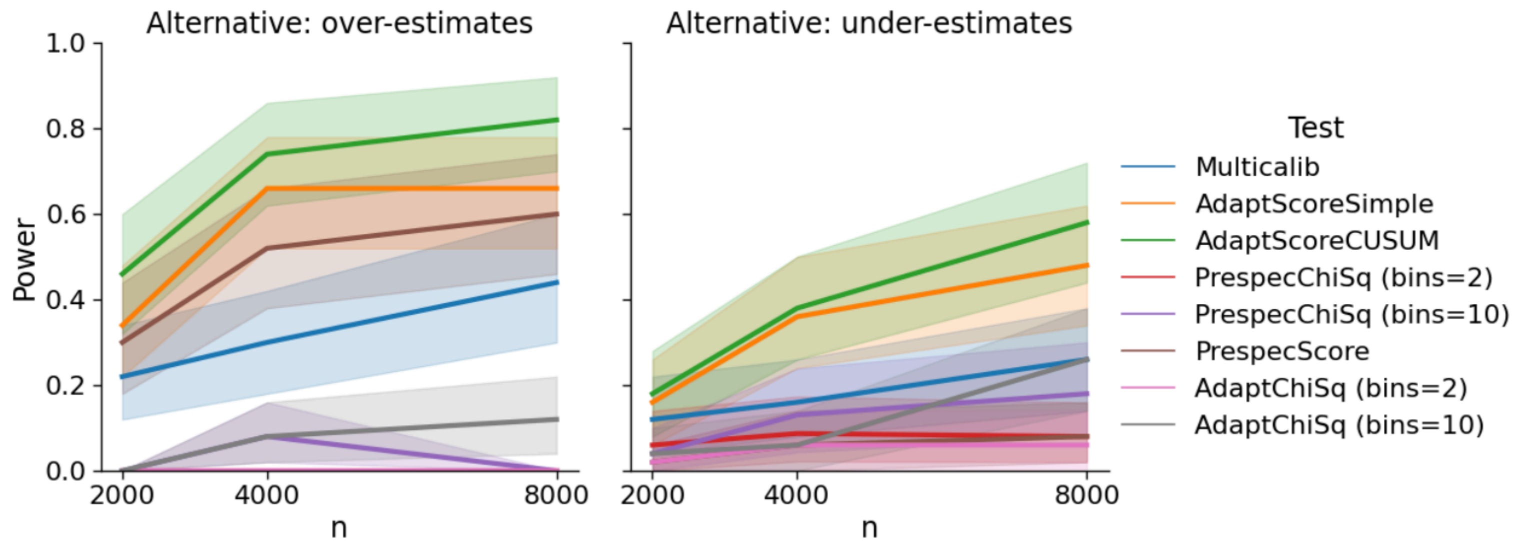


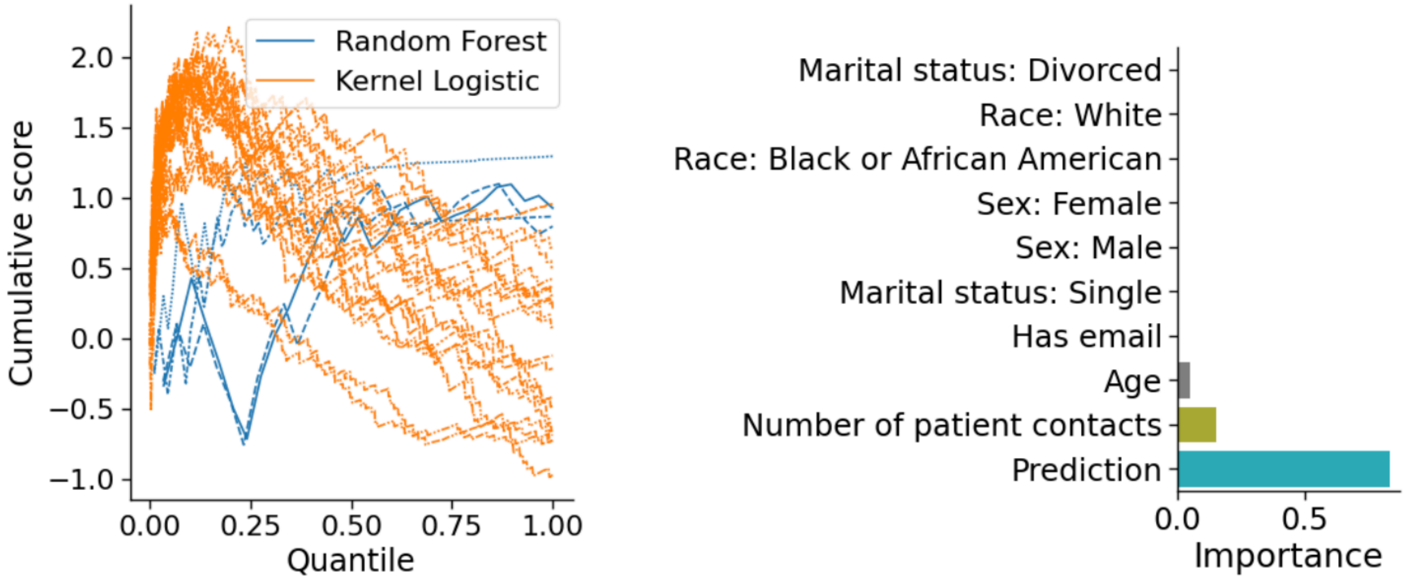
Kernel Logistic Regression: $\hat{g}_{(2)}$



Auditing a readmission model

- Trained a Random Forest (RF) that predicts risk of 30-day unplanned readmission using Electronic Health Records (EHR) from the Zuckerberg San Francisco General Hospital
- <u>Residual models</u>: Random Forests and Kernel Logistic Regression
- Audit the model for strong calibration with respect to the demographic variables ($\delta = 0.05$)



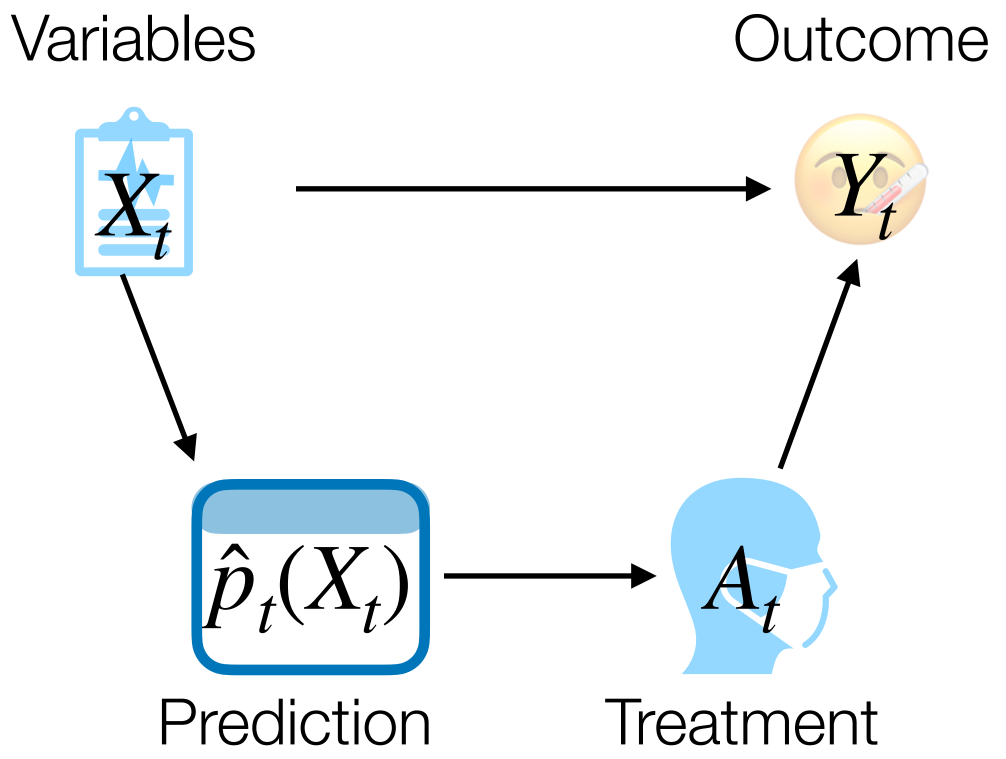


Outline

1. Auditing performance of ML algorithms across subgroups, *when the subgroups are unknown*
   - ➡ *We can reformulate this as a changepoint detection problem.*

2. Auditing performance of ML algorithms over time, *in the presence of performativity*

The problem of performativity



Suppose we have a model for predicting Post-operative Nausea and Vomiting (PONV)…

1. Alert! Patient is at high risk of PONV
2. Administer prophylactic treatment
3. Patient doesn't develop PONV

Notation

$\hat{p}_t : X_t \mapsto [0,1]$ ML-based risk prediction algorithm

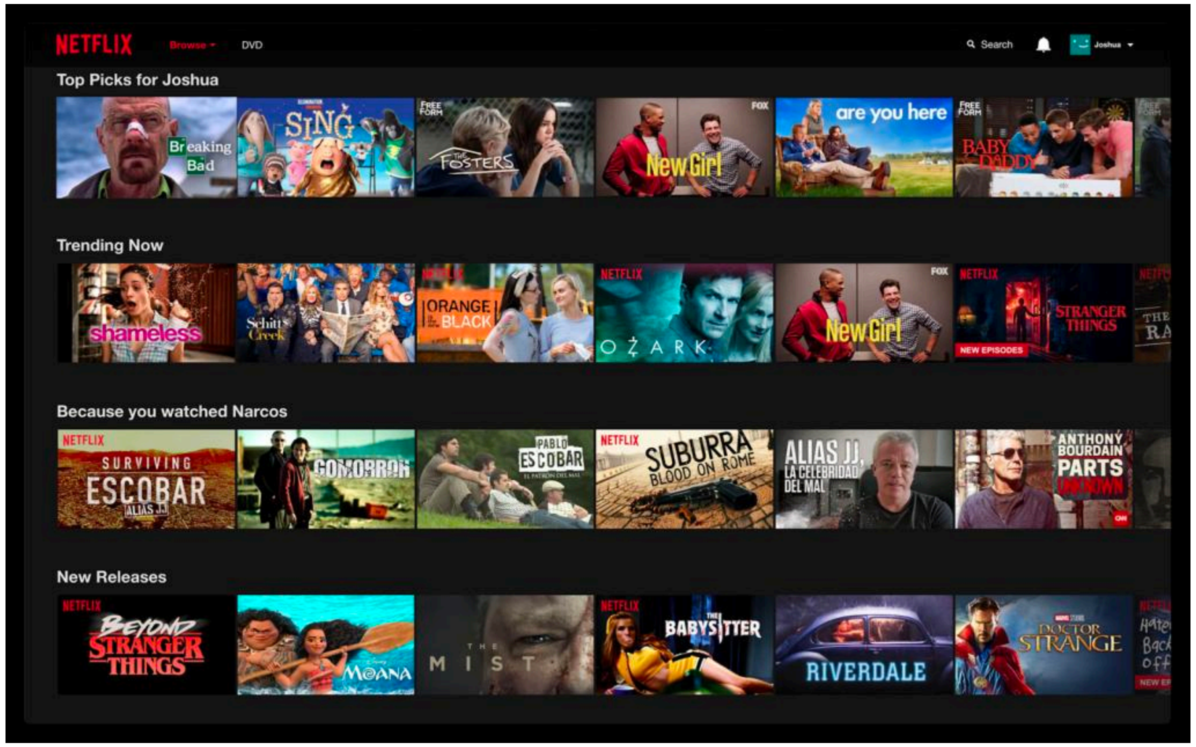
$$A_t = \begin{cases} 0 & \text{Standard-of-care (SOC)} \\ 1 & \text{Additional treatment} \end{cases}$$

$$Y_t = \begin{cases} 0 & \text{No PONV} \\ 1 & \text{PONV} \end{cases}$$

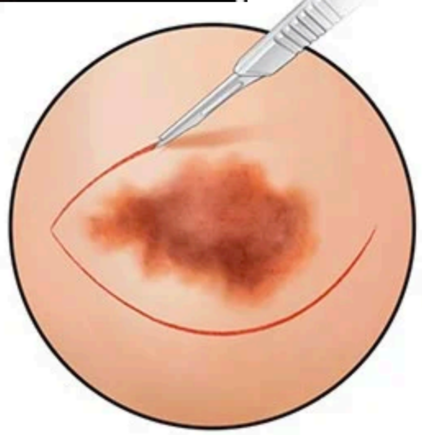
Was the model wrong or did the treatment make a difference?

The problem of performativity

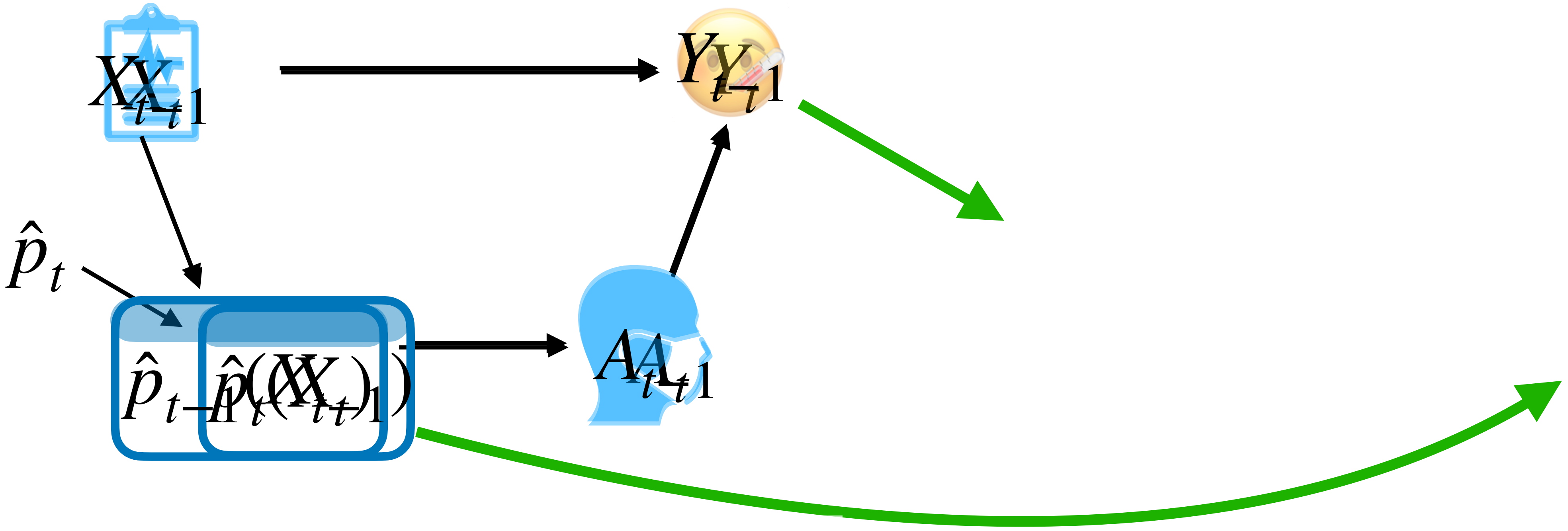
Recommendation engines



Diagnostic devices



Only monitor the data from patients receiving SOC?



Marginal performance

$$\mathbb{E}\left[\ell(Y_t(0), \hat{p}_t(X_t))\right]$$

- AUC
- Accuracy

- Requires highly accurate estimates of treatment propensities

tricky…

Conditional performance

$$Y_t(0) \mid \hat{p}_t(X_t)$$

- Model calibration
- PPV/NPV

- Conditions away components that are prone to distribution shifts

From monitoring in the “standard” setting to the performative setting

Hypothesis Test in the **standard** setting:

H_0 : There is no change in the conditional distribution, i.e.

$$\Pr\left(Y_t = 1 \mid Z_t = z\right) = g(z; \theta_0) \qquad \forall z \in \mathbb{R}, t = 1,2,\cdots$$

Hypothesis Test in the **performative** setting:

H_0 : There is no change in the conditional performance, i.e.

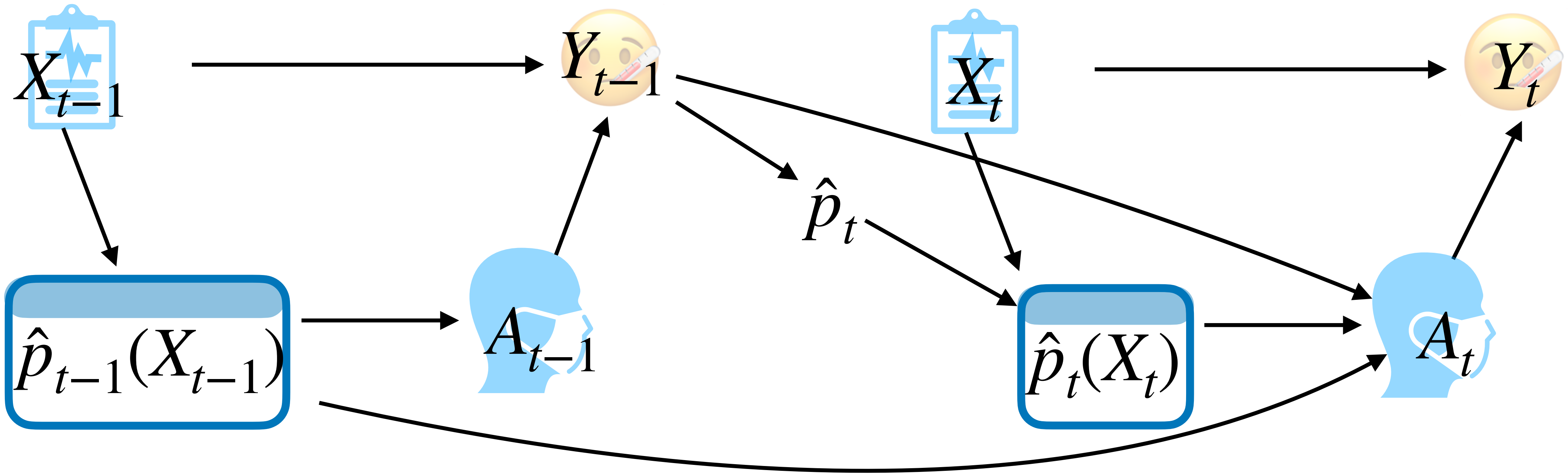
$$\Pr\left(Y_{\tau_i}(0) = 1 \mid \hat{p}_{\tau_i}(X_{\tau_i}) = q\right) = g(q; \theta_0) \quad \forall q \in \mathbb{R}, i = 1,2,\cdots$$

Ignoring performativity is valid if…

Conditional exchangeability:

A clinician's propensity to treat patient X_t only depends on the predicted risk and the clinician's past experiences interacting with the ML algorithm.

$$Y_t(0) \perp A_t \mid \hat{p}_t(X_t), \mathscr{F}_t$$



(We can extend this condition if treatment propensities depend on other variables as well.)

Monitoring solutions in the presence of performativity

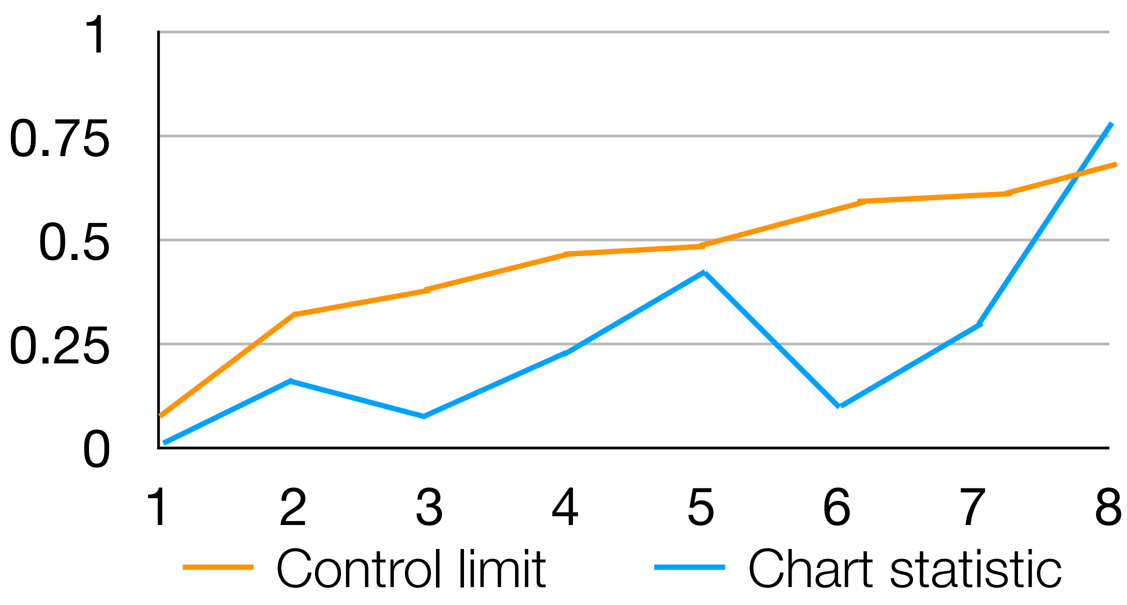
Frequentist

A score-based CUSUM procedure

Chart statistic at index i:

$$C(i) = \max_{s=1,\cdots,i} \left| \underbrace{\sum_{j=s}^{t} \nabla_\delta \log p\left(Y_{\tau_j} \mid \hat{p}_{\tau_j}(X_{\tau_j}); \hat{\theta}_{j-1}, \delta\right)\Big|_{\delta=0}}_{\text{Cumulative score from candidate changepoint } \tau_s} \right|$$

Control limit at index i: Dynamically calculated for a pre-specified alpha-spending function using a parametric Bootstrap.



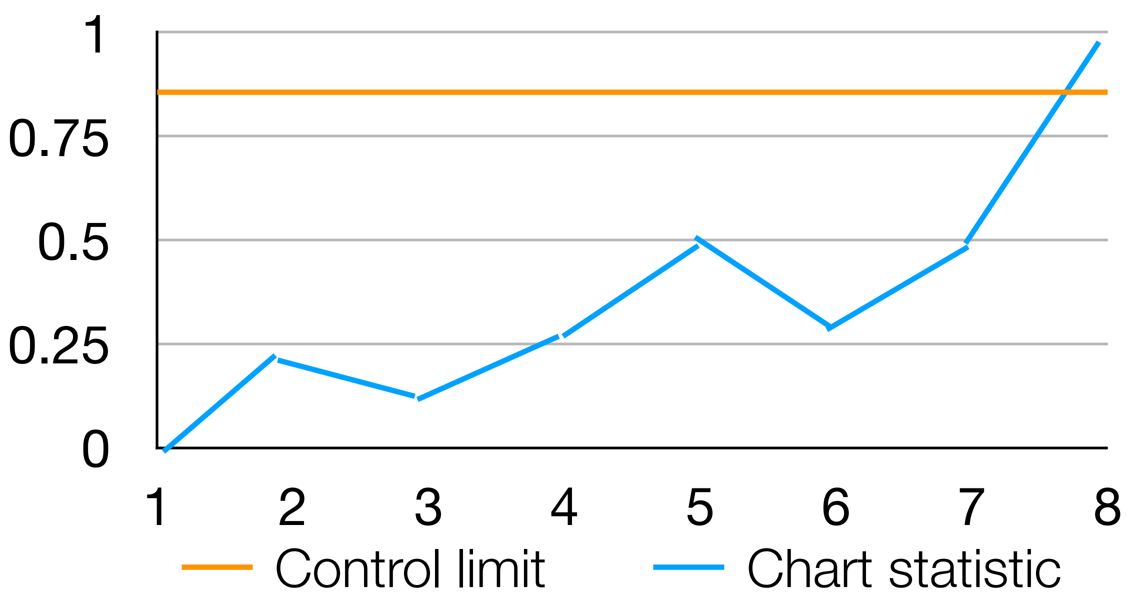
Bayesian

Full Bayesian inference

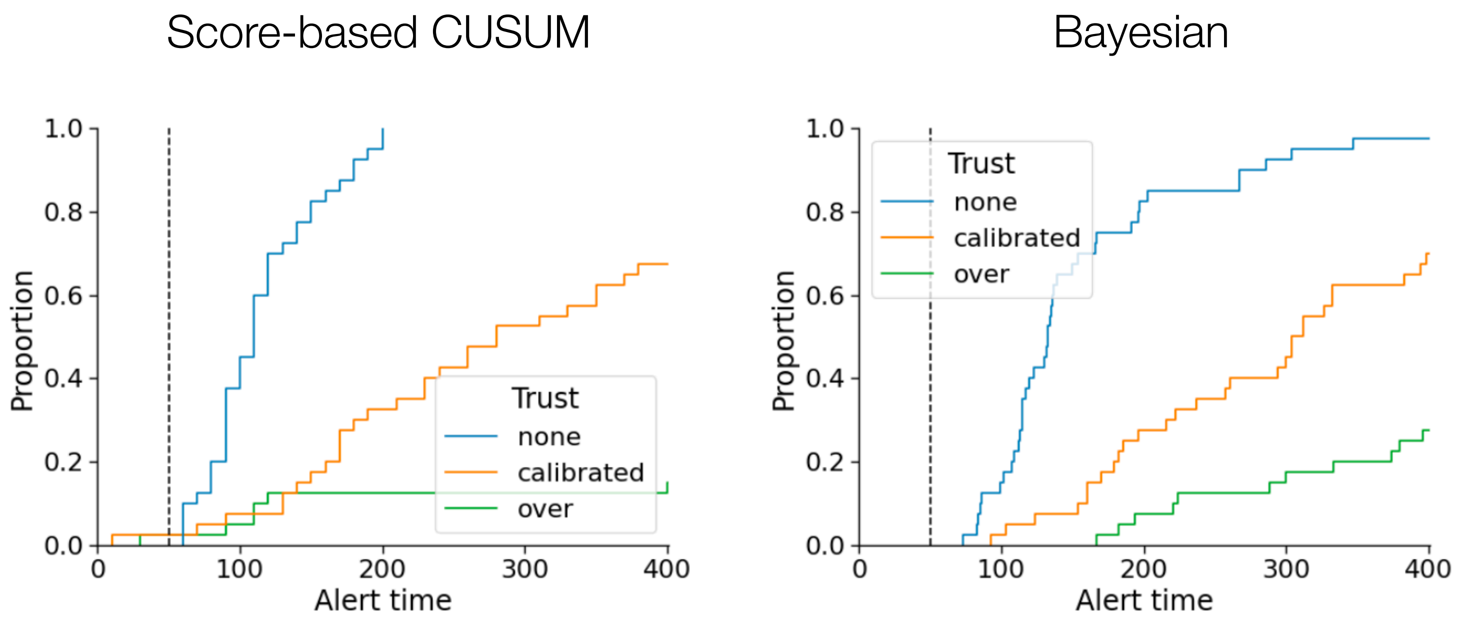
Chart statistic at index i:

$$C(i) = \underbrace{\Pr\left(\exists \kappa \leq \tau_i; \hat{p}_{\tau_1}(X_{\tau_1}), Y_{\tau_1}, \cdots, \hat{p}_{\tau_i}(X_{\tau_i}), Y_{\tau_i}\right)}_{\text{Posterior probability of there having been a changepoint}}$$

Control limit at index i: Fixed at $1 - \alpha$



Simulation: What is the impact of clinician trust?



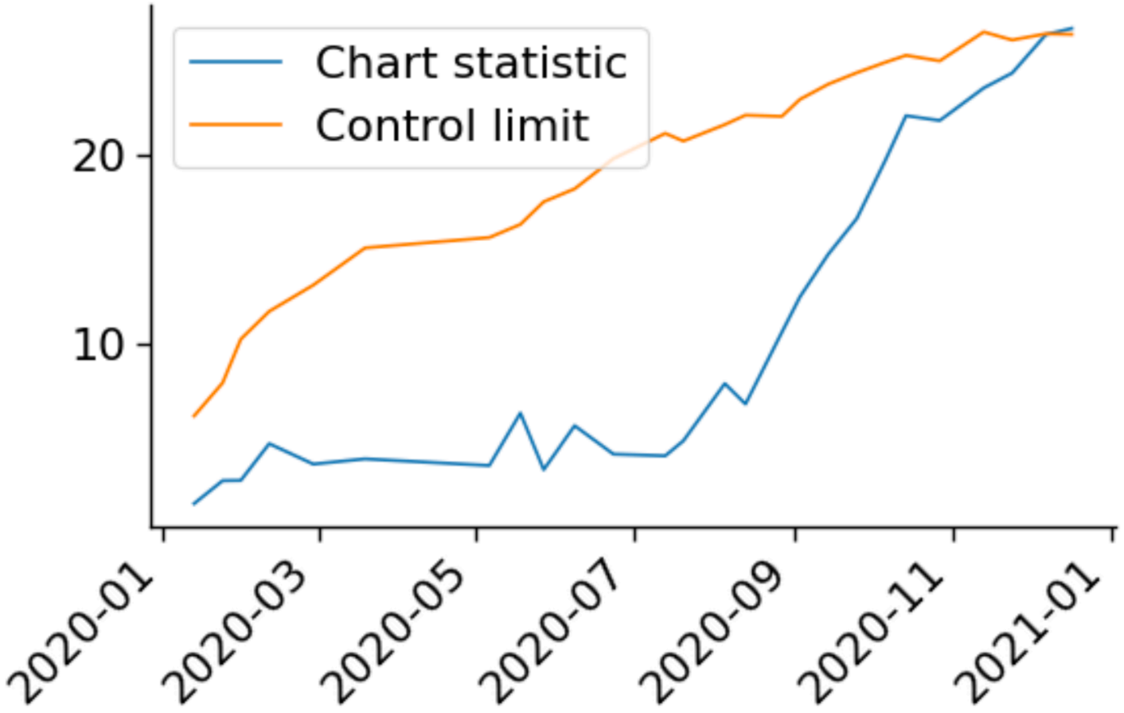
Calibration decay concentrated among patients unlikely to receive SOC

➡*When designing a ML monitoring system, determine if clinician trust is likely to interfere with our ability to detect performance decay. If so, consider designing a system that pulls in additional sources of data or actively increases the amount of information in the monitoring data.*

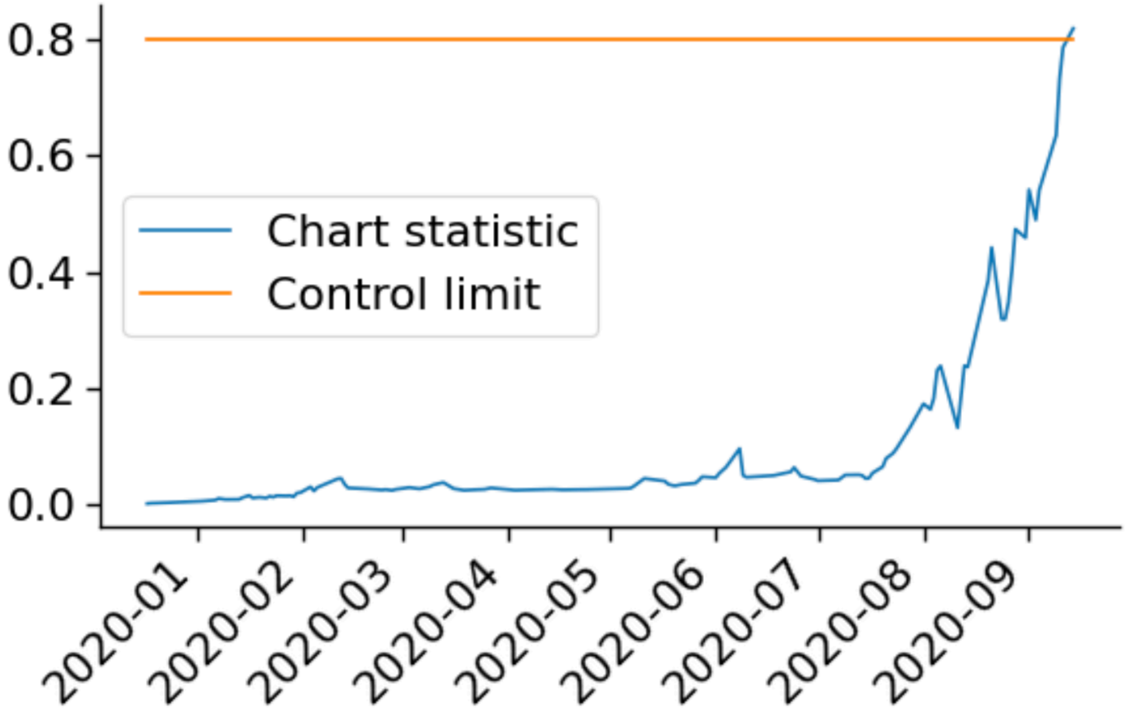
Case study: Post-operative Nausea and Vomiting (PONV)

- Data: UCSF Multicenter Perioperative Outcomes Group (MPOG)
- ML algorithm: A ***locked*** Random Forest using sex, smoking status, American Society of Anesthesiologists (ASA) classification, …

Score-based CUSUM



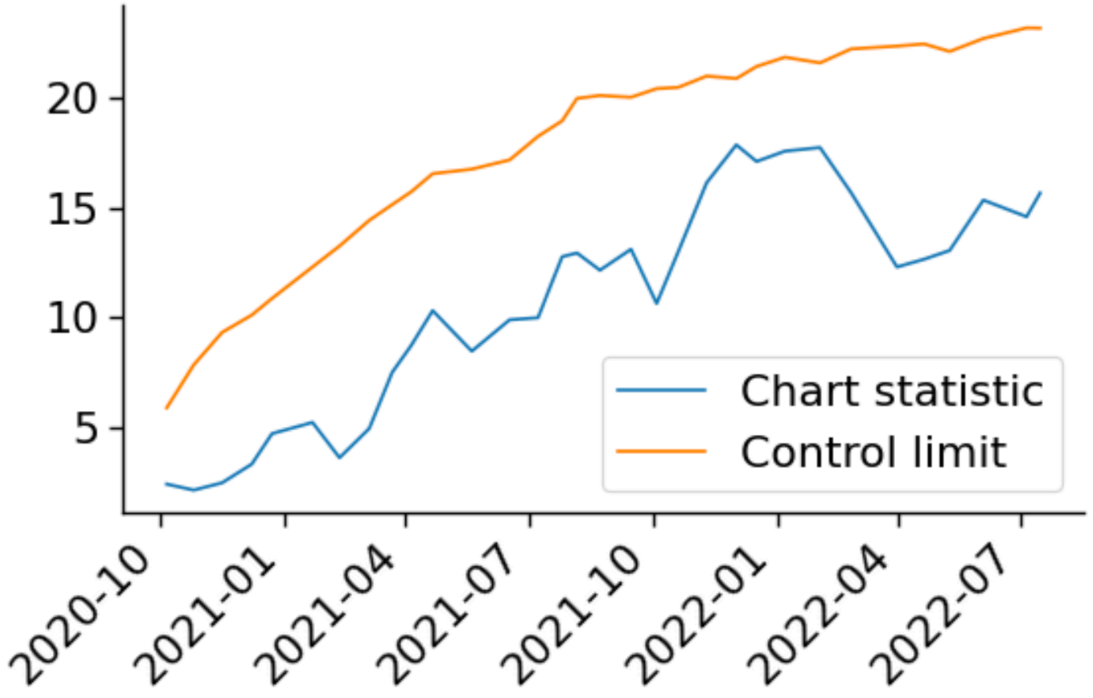
Bayesian monitoring

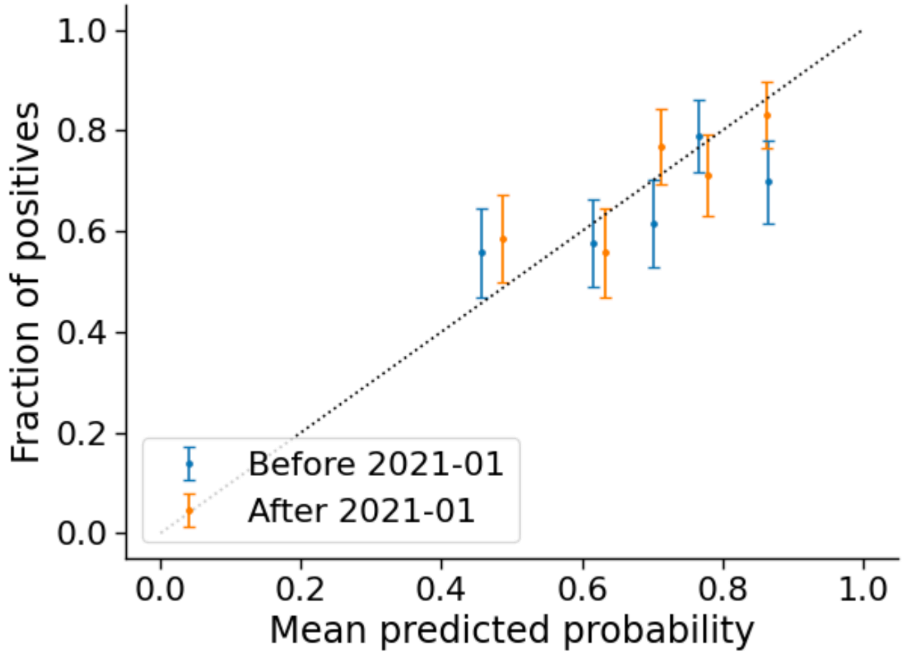


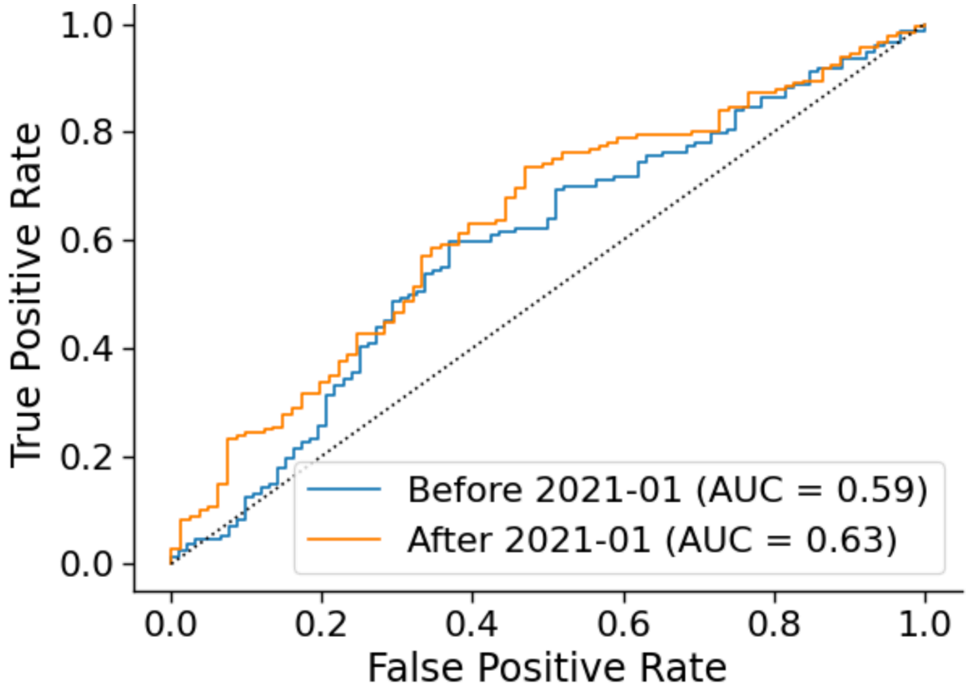
Case study: Post-operative Nausea and Vomiting (PONV)

- Data: UCSF Multicenter Perioperative Outcomes Group (MPOG)
- ML algorithm: A ***continually retrained*** Random Forest

Score-based CUSUM







Outline

1 Auditing performance of ML algorithms across subgroups, *when the subgroups are unknown*

➡ *We can reformulate this as a changepoint detection problem.*

➡ *http://arxiv.org/abs/2307.15247*

2 Auditing performance of ML algorithms over time, *in the presence of performativity*

➡ *By casting the online changepoint detection problem in the causal framework, we derive ignorability conditions and monitoring procedures.*

➡ *http://arxiv.org/abs/2211.09781*

Thank you!

Support from the UCSF-Stanford CERSI program

(Disclaimer: The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.)

www.jeanfeng.com