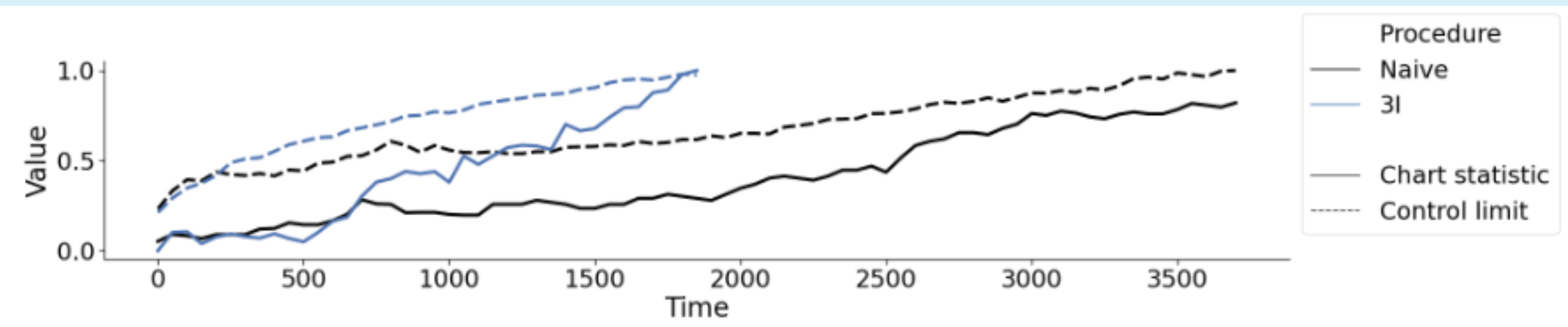
Designing monitoring strategies for deployed ML algorithms: navigating performativity through a causal lens

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*

Monitoring: not as easy as you think!

- Although there is widespread agreement on the need to monitor ML algorithms for performance decay, the immense complexity of designing a monitoring strategy has been relatively under-appreciated.
- Prior works have lacked precision in terms of what the target estimand is, how it should be selected, and how it should be monitored.
- **Contribution of this work:**
  - Highlights the wide range of monitoring strategies, even in a relatively simple case study.
  - Demonstrates the importance of a systematic causally-informed approach to enumerate candidate monitoring strategies.
  - Merges ideas from causal inference with statistical process control to account for ***performativity,*** the phenomena where an ML algorithm interacts with its environment to affect downstream data-generating mechanisms.



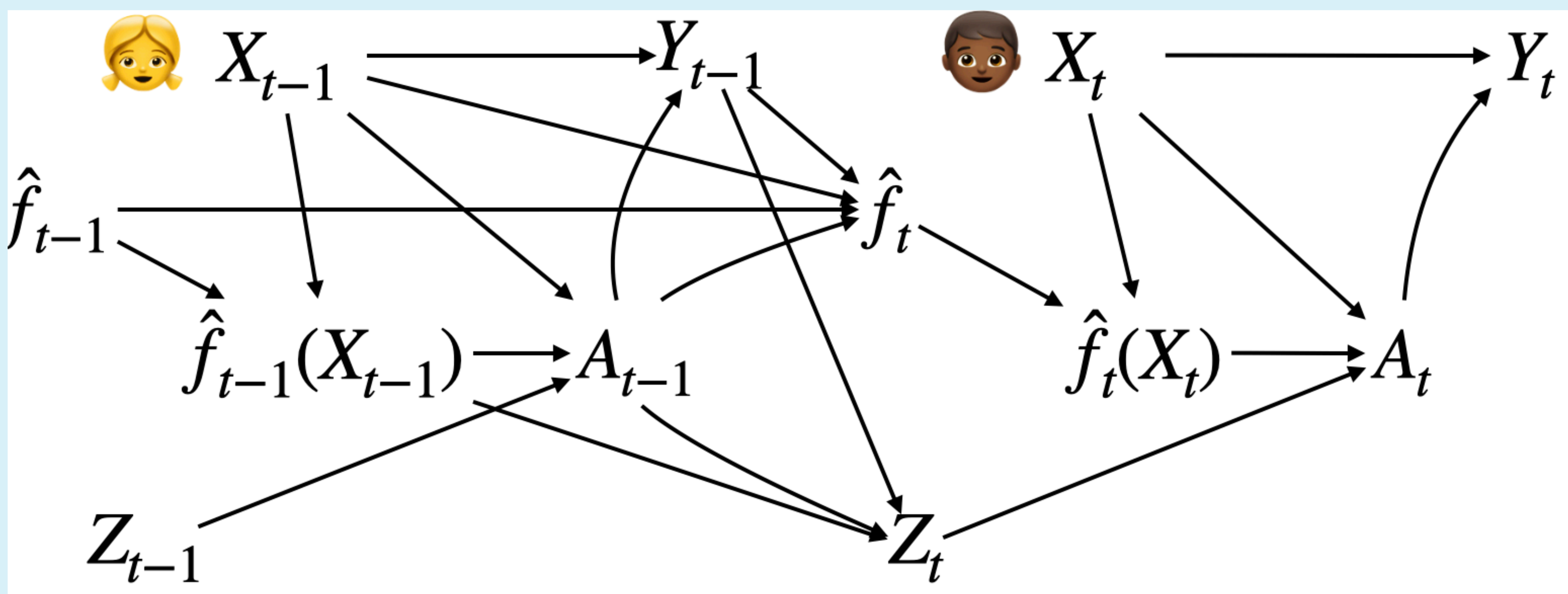
Example monitoring charts. An alarm is fired when the chart statistic exceeds the control limit.

A case study

- Consider a ML algorithm that predicts a patient's risk of unplanned readmission if a follow-up appointment is or is not scheduled. $\hat{f}_t$ is the algorithm at time t. $\hat{y}_t$ is the binarized prediction.
- The potential biases induced by this ML algorithm are numerous and varied, including:

Study Population	**Spectrum/referral bias**: ML algorithm is only queried for a subpopulation of patients.
Conditions of use	**Off-label use**: ML algorithm is queried in settings that are not recommended.
Benchmark/ Outcomes	**Interfering medical interventions (IMI):** Patients are treated with differing rates, driven by recommendations from the ML algorithm.

- Suppose the main source of bias is from *interfering medical interventions (IMI)*…



Funding: This work was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a Center of Excellence in Regulatory Science and Innovation grant to University of California, San Francisco (UCSF) and Stanford University. 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.

3 Candidate monitoring criteria

Each monitoring criterion can be formulated as a hypothesis test involving causal estimands. Examples:

- **C1**: The average PPV/NPVs should be maintained above specified thresholds.

$$H_0^{(1)}: \Pr(Y_t(a) = v \mid \hat{y}_t(X_t, a) = v, F_t) \geq c_{a,v} \forall t, a, v$$

- **C2**: The PPV/NPV for subgroups $S_1, \cdots, S_k$ should be maintained above their respective thresholds.

$$H_0^{(2)}: \Pr(Y_t(a) = v \mid \hat{y}_t(X_t, a) = v, X_t \in S_k, F_t) \geq c_{a,v} \forall t, a, v, k$$

- **C3**: The predicted probabilities should be well-calibrated with respect to *any* subgroup (strong calibration), for tolerance $\delta \geq 0$.

$$H_0^{(3)}: \left| \Pr(Y_t(a) = 1 \mid x) - \hat{f}_t(X_t, a) \right| \leq \delta \forall t, a, x$$

3x2 Candidate monitoring strategies

Each of the three aforementioned criteria can be monitored using interventional (I) or observational (O) data under suitable identifiability assumptions and certain data requirements.

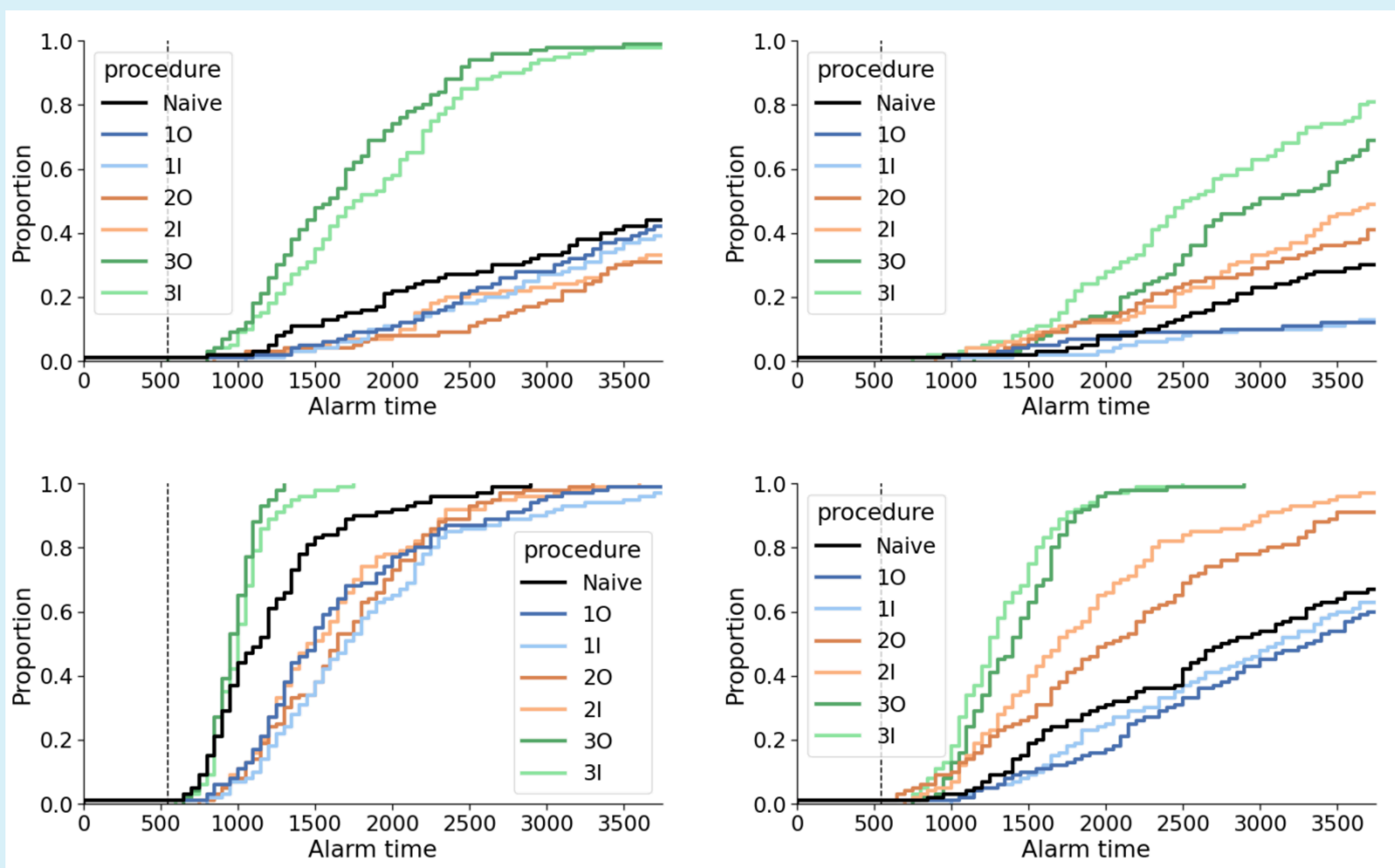
Example: Procedure 1I monitors C1 given interventional data using chart statistic

$$C_{1I}(t) = \max_{\tau,a,v} \sum_{i=\tau}^{t} \left(c_{av} - \frac{1\{Y_i = v, A_i = a\}}{p_i(A_i = a \mid X_i, Z_i, \hat{f}_i)} \right) 1\{\hat{y}_i(X_i, a) = v\}$$

where the propensities are known a priori. Procedure 1O monitors C1 given observational data using the same statistic, but plugs in *estimated* propensities.

Comparison of candidate strategies

Comparison of time to detection



Comparison of properties/requirements

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