Towards a Post-Market Monitoring Framework for Machine Learning (ML)-based Medical Devices: A case study

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; Paper at https://arxiv.org/abs/2311.11463

Post-market monitoring: not as simple as it looks!

- **Performance monitoring is a regulatory requirement**. Per CFR Title 21 Part 820, all medical devices should institute quality systems, which includes “monitor[ing] production processes to ensure that a device conforms to its specifications.”
- **What is the optimal performance metric to monitor?** Although it may be natural to monitor the same metrics used for initial model approval (e.g. AUC), such metrics may be suboptimal when the goal is to detect shifts in performance *as quickly as possible*.
- **What data should we use?** Observational data is most convenient, but exhibits well-known biases. Interventional data explicitly eliminates such biases.
- **What assumptions are needed?** To overcome biases in observational data, assumptions are needed to identify quantities of interest.
- **Given the range of monitoring strategies, a framework for assessing and comparing different monitoring strategies is needed.** Merging ideas from causal inference with statistical process control, we propose four basic steps:

1. Define potential monitoring criteria
2. Enumerate biases. Define the causal model
3. Describe candidate monitoring strategies
4. Compare pros/cons of candidate strategies

The case study

- Postoperative Nausea and Vomiting (PONV) is a common side effect of anesthesia.
- Consider a ML algorithm that predicts a patient’s risk of developing PONV if they are or are not given anti-nausea medication. $\hat{f}_t$ is the algorithm at time t that outputs a risk. $\hat{y}_t$ is the binarized version.
- Suppose the algorithm was approved initially based on its positive and negative predictive values (PPV and NPV, respectively).

▷Step 1. 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.

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

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

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

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

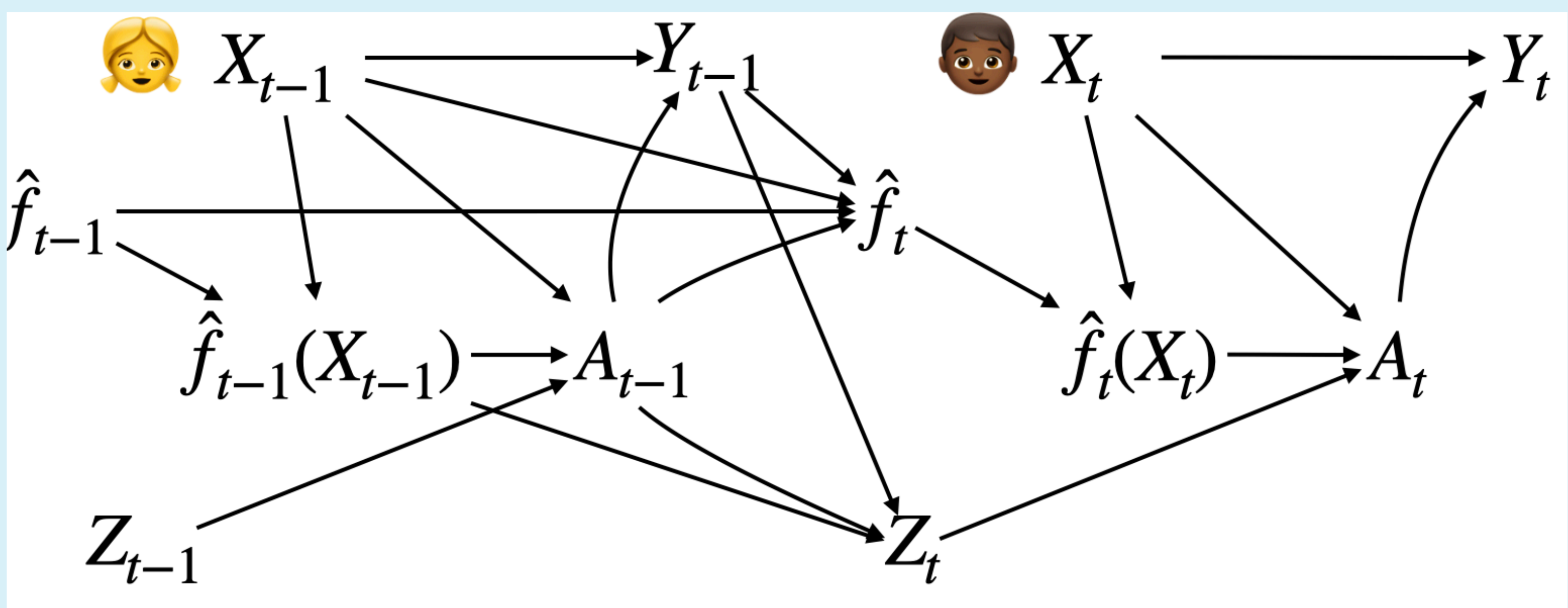
$$H_0^{(3)} : \left| \Pr(Y_t(a) = 1 \,|\, x) - \hat{f}_t(X_t, a) \right| \leq \delta \forall t, a, x$$

▷Step 2. Biases and the causal model

Potential sources of bias that exist across different data sources can be enumerated using the target trial emulation framework. This list can then be filtered and ranked based on expert opinion.

Study Population	**Spectrum/referral bias**: ML algorithm is only queried for a subset of patients, e.g. subpopulations the algorithm is believed to perform well in.
Conditions of use	**Off-label use:** ML algorithm queried in settings that are not recommended, e.g. too early or late during a surgical case.
Benchmark/ Outcomes	**Interfering medical interventions (IMI):** Patients are treated with differing rates, driven by recommendations from the ML algorithm.
	Circular definitions: Outcome label is biased by the algorithm’s predictions, e.g. PONV is more likely to be documented for cases predicted to be at high risk.

This case study assumes the main source of bias is from interfering medical interventions (IMI), described by the following Causal Directed Acyclic Graph (DAG):



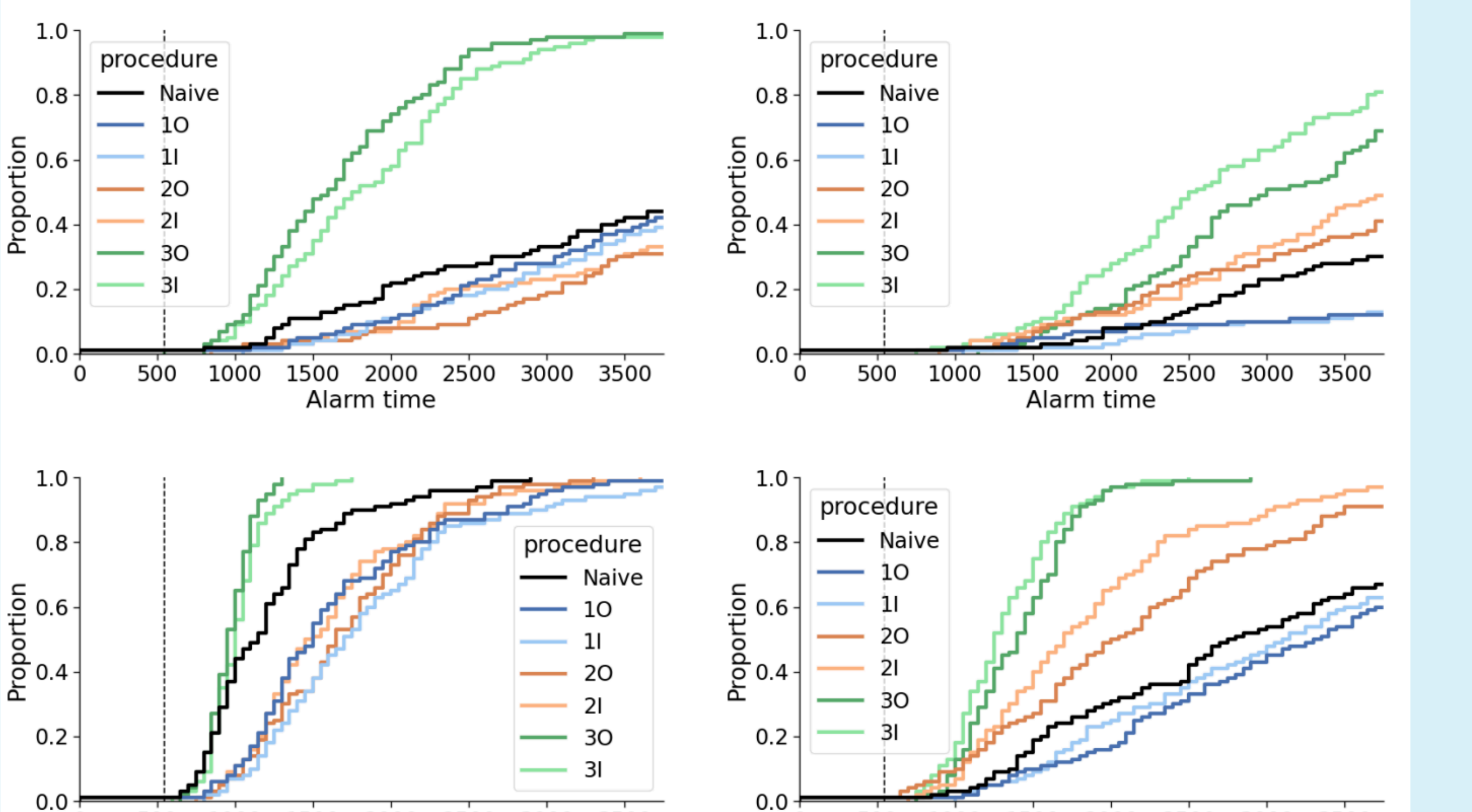
▷Step 3. Monitoring strategies

Each of the three aforementioned criteria can be monitored using interventional (I) or observational (O) data under suitable identifiability assumptions and certain data requirements, leading to 3x2 candidate monitoring strategies.

<u>*Example*</u>: Procedure 1I monitors C1 given interventional data using chart statistic $C_{1I}(t) = \max_{\tau,a,v} \sum_{i=\tau}^{t} \left(c_{av} - \frac{1\{Y_i = v, A_i = a\}}{p_i(A_i = a \,|\, X_i, Z_i, \hat{f}_i)} \right) 1\{\hat{y}_i(X_i, a) = v\}$ where the propensities are known a priori. Procedure 1O monitors C1 given observational data using the same statistic, but plugs in *estimated* propensities.

▷Step 4. Comprehensive comparison

Comparison of time to detection (statistical power)



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

In this case study, procedures 3I and 3O are the most powerful procedures and would be reasonable choices. 3I is better at controlling the worst-case detection delay, whereas 3O is much more convenient.

Funding: This work was funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2022C1-25619). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by PCORI, FDA/HHS, or the U.S. Government.