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*

Editors: Francesco Locatello and Vanessa Didelez

Abstract

After a machine learning (ML)-based system is deployed, monitoring its performance is important to ensure the safety and effectiveness of the algorithm over time. When an ML algorithm interacts with its environment, the algorithm can affect the data-generating mechanism and be a major source of bias when evaluating its standalone performance, an issue known as performativity. Although prior work has shown how to *validate* models in the presence of performativity using causal inference techniques, there has been little work on how to *monitor* models in the presence of performativity. Unlike the setting of model validation, there is much less agreement on which performance metrics to monitor. Different monitoring criteria impact how interpretable the resulting test statistic is, what assumptions are needed for identifiability, and the speed of detection. When this choice is further coupled with the decision to use observational versus interventional data, ML deployment teams are faced with a multitude of monitoring options. The aim of this work is to highlight the relatively under-appreciated complexity of designing a monitoring strategy and how causal reasoning can provide a systematic framework for choosing between these options. As a motivating example, we consider an ML-based risk prediction algorithm for predicting unplanned readmissions. Bringing together tools from causal inference and statistical process control, we consider six monitoring procedures (three candidate monitoring criteria and two data sources) and investigate their operating characteristics in simulation studies. Results from this case study emphasize the seemingly simple (and obvious) fact that *not all monitoring systems are created equal*, which has real-world impacts on the design and documentation of ML monitoring systems.

Keywords: Statistical process control, model monitoring, performativity

1. Introduction

After a machine learning (ML)-based system is deployed, performance monitoring of the algorithm is necessary to minimize the release of misleading or outdated predictions (Breck et al., 2017; Finlayson et al., 2021; Feng et al., 2022). Nevertheless, compared to the topic of model development and pre-deployment evaluation, the topic of post-deployment performance monitoring remains relatively understudied (Sculley et al., 2015; Paleyes et al., 2022; Zhang et al., 2022). Although prior works have suggested monitoring using sequential testing methods from statistical process control (SPC), these works largely assume an "ideal data setting" where the ML algorithm does not interact with its environment (Montgomery, 2013; Qiu, 2013). However, ML algorithms are often induce

* jean.feng@ucsf.edu

changes in their environment, which can be a major source of bias when evaluating their performance. This phenomenon, also known as performativity, algorithmic confounding, and feedback loops, has been discovered across various contexts including medicine, recommendation engines, and more (Bottou et al., 2013; Paxton et al., 2013; Chaney et al., 2018; Perdomo et al., 2020).

As a motivating example, consider a clinical risk prediction model for predicting 30-day unplanned readmission (see e.g. Horwitz et al. (2019)). Many models have been developed to help hospitals reduce their readmission rates (Burke et al., 2017; Barbieri et al., 2020; Mahmoudi et al., 2020), spurred by the Hospital Readmissions Reduction Program. Prior works have documented various interventions hospitals may consider to reduce their readmission rates, such as post-discharge follow-ups. Here we suppose that after the algorithm's predicted readmission risk for a patient is shown to the clinician, the clinician may choose between one of two treatments: scheduling a 5-day follow-up appointment or discharging with no plans for additional follow-up. Performativity occurs when clinicians trust the algorithm, so that patients with high predicted risks tend to have high rates of follow-up. As a consequence of this self-fulfilling prophecy, estimating performance using standard measures like AUC and accuracy without adjusting for (post-deployment) treatment propensities will be biased. Prior work has shown how to obtain unbiased estimates of model performance, adjusted for performativity, using techniques from causal inference (Sperrin et al., 2019). Nevertheless, there has been little discussion on how one should design an online monitoring procedure in the post-deployment setting.

One view of model monitoring is that it simply conducts model validation repeatedly over time, where the goal is to detect if the performance of a model has strayed from its performance at the time of initial model validation (Nishida and Yamauchi, 2007; Klaise et al., 2020; Schröder and Schulz, 2022; Corbin et al., 2023). For instance, many models are evaluated in terms of their average performance, such as accuracy and AUC. One may consider detecting changes in such quantities by combining causal inference with SPC and reweighting the monitoring data to match the target population (Steiner and MacKay, 2001a; Sun et al., 2014; Cook et al., 2015). However, there are many other ways to characterize model performance, including performance within particular subgroups (Mitchell et al., 2021) and conditional measures like a model's calibration curve (Van Calster et al., 2016). Thus, monitoring mean performance is neither the only nor necessarily the best option.

A contrasting view is that the goal of model monitoring is not just to estimate changes in model performance but to minimize exposure to a faulty model. This is particularly true in safety-critical settings like healthcare where elevated model error rates have severe consequences. In such settings, a much stronger emphasis is placed on detection speed (Montgomery, 2013). Alerts produced by a monitoring procedure are handled by a quality assurance team, which investigates potential causes of performance decay and decides the most appropriate corrective action to take (e.g., updates to the model or data processing pipeline) (Breck et al., 2017; Raji et al., 2020; Feng et al., 2022).

Monitoring procedures vary in their *monitoring criterion* (i.e., what they aim to detect). Taking a frequentist view in this paper, the monitoring criterion corresponds to a hypothesis test about a model's performance over time. Monitoring criteria vary in how sensitive they are, their sample size requirements, and how interpretable their test statistics are. In the presence of performativity (and other sources of bias), there are additional considerations, as the monitoring target must be cast within a causal framework to define the downstream effects of the ML algorithm: now our interest is in a *causal* monitoring criterion, which impacts the types of data that can be used (observational versus interventional) and assumptions necessary for identifiability. For instance, monitoring changes in AUC may be highly interpretable, but it is slow, as AUC is a relatively "coarse" metric

(Pencina et al., 2008), and will require assumptions such as positivity in the observed data. On the other hand, methods for monitoring algorithmic fairness and conditional performance measures can require weaker (or even no) assumptions regarding positivity, but may be less interpretable (Feng et al., 2024a).

Although the importance of model monitoring has been recognized by model developers, users, and regulators alike (Eaneff et al., 2020; U.S. Food and Drug Administration and Health Canada, 2021; Schröder and Schulz, 2022), previous discussions of model monitoring have lacked precision in terms of what the target estimand is, how it should be selected, and how it should be monitored. The goal of this work is to demonstrate the value of causal reasoning when monitoring models in the presence of performativity, the underappreciated complexity of designing a monitoring strategy, and how different strategies can have vastly different consequences. We do this by conducting an in-depth case study of the unplanned readmission model described earlier, though the conclusions in this work apply more broadly. We walk through various procedures by considering different candidate monitoring criteria (Section 4.1) and candidate data sources (Section 4.2), and enumerate potential ways the readmission model can interact with its environment. We then focus on the specific problem of "interfering (medical) interventions" (IMI) and compare the different monitoring strategies in an extensive simulation study (Section 4.4). Through this case study, we illustrate that *not all monitoring systems are created equal*. Designing an effective monitoring system requires a systematic causally-informed approach, and the (expected) operating characteristics of the deployed system should be transparent to users and other stakeholders. Finally, we conclude with a discussion of open research problems at the intersection of causal inference and model monitoring. Code is available at `https://github.com/jjfeng/monitoring_causally`.

2. Related Work

Although the term "performativity" was recently coined in (Perdomo et al., 2020), this phenomenon has been described across various problem domains. One of the earliest types of performativity was perhaps described in (Begg and Greenes, 1983; Zhou, 1998; Alonzo and Kittelson, 2006): when assessing a diagnostic test, a well-known difficulty is that the test result itself can affect whether the patient receives a definitive assessment for disease status (e.g., a biopsy). This problem, known as *verification bias* or *work-up bias*, is typically viewed within a missing data framework. More modern works focus on ML algorithms that aim to predict future outcomes but where their predictions can affect this very outcome (Bottou et al., 2013; Paxton et al., 2013; Chaney et al., 2018; Liley et al., 2021). This type of performativity is typically formalized within a causal inference framework (Mendler-Dünner et al., 2022; Wu et al., 2022). However, there are many similarities in the approaches taken, due to connections between missing data and causal inference (Mohan and Pearl, 2021).

Although there exist methods for one-time model evaluation in the presence of performativity (Wu et al., 2022), the only work we are aware of that addresses monitoring in the presence of performativity is Feng et al. (2024a). However, Feng et al. (2024a) only describes one specific monitoring procedure. *The goal of this work is to provide a more global view of the monitoring landscape to describe the multitude of monitoring options one may consider in practice.* More generally, there are few other works in the sequential process control literature that discuss the role of causal inference, as the field is primarily concerned with monitoring observational data quantities. The most relevant methods to this work are those for monitoring differentially censored outcomes

(Steiner and MacKay, 2001a,b; Sun et al., 2014), again due to connections between missing data and causal inference.

3. Background: sequential monitoring

Given a monitoring criterion that defines how a data-generating process is expected to behave (also known as *in-control* in the SPC literature), the goal of a sequential monitoring procedure is to detect when this process no longer behaves as expected (i.e. *out-of-control*), given an incoming sequence of observations. Formally, one can define the sequential monitoring problem as a hypothesis test, where the null hypothesis is that the process is in-control, and the sequential monitoring procedure as a sequential hypothesis testing procedure characterized by two functions: chart statistic $\hat{C} : \{1, 2, \cdots\} \to \mathbb{R}$ and control limit $h : \{1, 2, \cdots\} \to \mathbb{R}$ (Zhang and Woodall, 2015). When the chart statistic first exceeds the control limit, i.e., when $\hat{C}(t) > h(t)$ for some time t, an alarm is fired and the null hypothesis is rejected. This is typically visualized via a control chart, which plots the two curves against each other (Figure 2).

In the following sections, we explore different configurations for the three key ingredients of any monitoring system: (i) the monitoring criterion, (ii) the monitoring data, and (iii) the sequential hypothesis testing procedure. We discuss the different options and their impacts on the operating characteristics of the overall monitoring procedure. Here our interest is in procedures that maintain Type I error—the probability that it fires an alarm under the null for a given time span—and maximize power—the probability of firing an alarm after the process is out-of-control (Chu et al., 1996; Zeileis, 2005). One may also consider other operating characteristics depending on the problem setting (Qiu, 2013).

4. The case study: ML-based risk prediction algorithms

Our central case study is an algorithm for predicting patients' risk of unplanned readmission within 30 days. Suppose the initial algorithm was evaluated based on its average positive predictive value (PPV) and negative predictive value (NPV). The ML algorithm is allowed to evolve based on prior observations. At each time point, we observe a new patient with variables X_t. Upon querying the ML algorithm, the clinician is faced with two "treatment options": they may either schedule a 5-day follow-up appointment for that patient (intervention) or discharge them with no additional follow-up (standard of card, SOC).

Notation. Let X_t represent variables for a patient drawn from the population at time t, uniformly at random. A_t is a binary "treatment" variable indicating whether the patient does or does not have a follow-up appointment ($a_t = 1$ versus $a_t = 0$, respectively). Using potential outcomes notation, $Y_t(a_t)$ is a binary outcome which indicates if a 30-day unplanned readmission occurs under treatment option a_t, where $Y_t(a_t) = 1$ means the patient is readmitted. For convenience, let $Y_t := Y_t(A_t)$ denote the observed outcome under the consistency assumption (Hernán and Robins, 2010). $\hat{f}_t(x, a)$ is the predicted risk from the ML algorithm at time t, given patient variables x and treatment a. In addition, let $\hat{y}_t$ be the binarized version of $\hat{f}_t$, e.g. $\hat{y}_t(x, a) = \mathbb{1}\{\hat{f}_t(x, a) > b\}$ for some threshold b. Let Z_t denote non-patient variables that may affect treatment decisions at time t, such as past performance of the ML algorithm. To formally describe how information accumulates over time, we define the process $\{(Z_t, X_t, A_t, Y_t(0), Y_t(1), \hat{f}_t) : t = 1, 2, \cdots\}$ as being adapted

to the filtration $\{\mathcal{F}_t : t = 1, 2, \cdots\}$, where $\mathcal{F}_t$ is the sigma field generated by the collection of observations up to time t, i.e. $(\hat{f}_t, Y_{t-1}, A_{t-1}, X_{t-1}, Z_{t-1}, \cdots, Y_1, A_1, X_1, Z_1)$.

4.1. Candidate monitoring criteria

When an ML system interacts with its environment to affect downstream outcomes, a causal framework is necessary to isolate the *standalone* performance of the algorithm. For instance, in this case study, the standalone PPV/NPV of the algorithm is $\Pr\left(Y_t(a) = \upsilon | \hat{y}_t(X_t, a) = \upsilon; \hat{f}_t\right)$ for $a, \upsilon \in \{0, 1\}$. Estimation of these measures directly from the observational data by restricting to the subpopulations who received the same treatment and binary classification, i.e. $\Pr\left(Y_t = \upsilon | \hat{y}_t(X_t, a) = \upsilon, A_t = a; \hat{f}_t\right)$ for $a, \upsilon \in \{0, 1\}$, is biased. For instance, if clinicians tend to treat patients predicted to be at high risk of readmission, the observational PPV will be biased downwards while the observational NPV will be biased upwards. As shown in later sections, this can lead to inflated Type I error and/or decreased power. So rather than presenting typical monitoring criteria for observational quantities, we present *causal* monitoring criteria.

In the following section, we describe three candidate criteria for monitoring model performance. All candidate criteria are based on PPV/NPV, as the original algorithm was approved based on its overall PPV/NPV. The criteria are listed from the most interpretable and least strict to the least interpretable and most strict, where stringency refers to strength in terms of model calibration (Van Calster et al., 2016) and algorithmic fairness requirements.

Criterion 1: The average PPV/NPVs should be maintained above some thresholds. For some set of predefined thresholds $\{c_{a\upsilon} : a, \upsilon \in \{0, 1\}\}$, we consider the causal monitoring criterion

$$H_{0,C1}^{\text{causal}} : \Pr\left(Y_t(a) = \upsilon | \hat{y}_t(X_t, a) = \upsilon, \mathcal{F}_t\right) \geq c_{a\upsilon} \quad \forall t, a, \upsilon. \tag{1}$$

The key benefit of this criterion is its interpretability, as it is a standard metric used in this case to approve the initial model. This criterion is less strict in that the model's subgroup-specific performance can vary over time without any variation in its overall PPV/NPV. (Conditioning on $\mathcal{F}_t$ in (1) lets us condition on the ML algorithm $\hat{f}_t$, which avoids the situation where $\hat{f}_t$ is a random variable that needs to be marginalized over.)

Criterion 2: Subgroup-specific PPV/NPVs should be maintained above their respective thresholds. Motivated by concerns regarding algorithmic fairness, another criterion is that the PPV/NPV is maintained across predefined subgroups $\mathcal{S}_1, \cdots, \mathcal{S}_K$, such as those defined by protected attributes. That is, for predefined subgroup-specific thresholds $\{c_{a\upsilon k}\}$, this criterion corresponds to checking the null hypothesis

$$H_{0,C2}^{\text{causal}} : \Pr\left(Y_t(a) = \upsilon | \hat{y}_t(X_t, a) = \upsilon, X_t \in \mathcal{S}_k, \mathcal{F}_t\right) \geq c_{a\upsilon k} \quad \forall t, a, \upsilon, k. \tag{2}$$

This criterion is stricter than (1) and more sensitive to distribution shifts. However, to decide the thresholds $c_{a\upsilon k}$, we may need to collect more data to estimate the PPV/NPV per subgroup, which may delay the time to model deployment. Another concern is that any procedure that tests (2) must account for multiple testing, so it is unclear how powerful a monitoring procedure for (2) would ultimately be.

Criterion 3: The predicted probabilities should not be over-confident for any subgroup. Finally, an even stricter criterion than checking subgroup-specific PPV/NPVs is to check that the algorithm's risk predictions are not overly extreme, in that the predicted risk should not be more

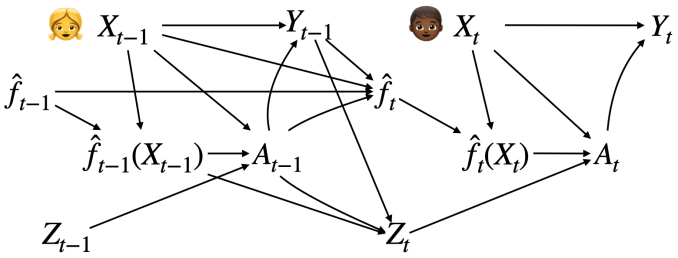

Figure 1: Causal model describing interfering medical interventions induced by an evolving ML-based risk prediction algorithm $\hat{f}_t$. X_t denotes variables for the patient being queried for at time t. Z_t denotes non-patient variables that may affect treatment decisions, such as past performance of the ML algorithm. A_t denotes treatment assignment and Y_t denotes the patient's outcome. Note that $\hat{f}_t(X_t) = (\hat{f}_t(X_t, 0), \hat{f}_t(X_t, 1))$.

extreme than the true adverse event rate for *any* subgroup for some tolerance $\delta \geq 0$. Formally, we check the null hypothesis:

$$H_{0,C3}^{\text{causal}} : \hat{s}_t(x, a) \left(\hat{f}_t(x, a) - \Pr(Y_t(a) = 1|x) \right) \leq \delta \quad \forall t, a, x, \tag{3}$$

where $\hat{s}_t(x, a) = \text{sign} \left(\hat{f}_t(x, a) - 0.5 \right)$ is the predicted class assuming a threshold of 0.5. This test can also be viewed as checking the strong calibration requirement, where the predicted probabilities are required to be within some distance of the true risk for each individual; models that satisfy the strong calibration requirement also satisfy strong notions of algorithmic fairness (Van Calster et al., 2016; Hebert-Johnson et al., 2018; Feng et al., 2024b). It is the strictest criterion among the three and, thus, the most sensitive to dataset shifts. Although this can accelerate detection, the null hypothesis can be quite sensitive, particularly for small values of δ, and may be violated only for a very small subgroup. Thus, we must be careful when interpreting a rejection of (3), as it could indicate a decay in model performance or that the model was not well-calibrated to begin with.

4.2. Data sources and causal models

To check monitoring criteria from Section 4.1, we must consider the types of data available and the biases they exhibit. For this case study, suppose we can collect data from either an observational setting where healthcare providers are free to decide how to respond to the ML algorithm or an interventional setting where patients are randomized to the two treatment options. As shown in Table 1, the number of potential biases can be large, and some may even interact. Causal reasoning provides a framework for conceptualizing these biases and determining how best to adjust for them. Stakeholders can also help identify which biases are less likely to occur or can be prevented through proper implementation of the ML system.

For simplicity, suppose we are mainly concerned with the problem of "interfering (medical) interventions" (IMI) (Paxton et al., 2013; Dyagilev and Saria, 2016; Lenert et al., 2019; Liley et al., 2021).[1] IMI is the problem where the ML algorithm modifies treatment patterns, resulting in dependent "censoring" of counterfactual outcomes. For instance, if we estimated the NPV of $\hat{y}_t(\cdot, a)$ by directly calculating the NPV among patients who received treatment a, this estimate would be

1. This problem has been previously referred to as "confounding medical interventions." However, strictly speaking, predictions from the ML algorithm do not confound the relationship between A and Y. Hence, we have modified the name of this phenomenon in this paper.

Potential sources of bias
Interfering medical interventions: Patients are scheduled for follow-up appointments with differing rates, driven in part by recommendations from the ML algorithm.
Selection bias: The ML algorithm is only queried for a non-random subset of patients (Ladapo et al., 2013; US Food and Drug Administration, 2007), such as only the more difficult cases or subpopulations the algorithm is believed to perform well in.
Off-label use: ML algorithm may be queried in settings that are not recommended. For instance, the algorithm may be queried too early during an inpatient stay, or it might be queried during an outpatient visit even though it is only suggested for analyzing inpatient stays.
Patient trust: The usage of the ML algorithm or its performance over time may motivate patients to leave or enter a hospital system (Hashimoto et al., 2018).
Insurance coverage and billing: The dollar amount covered by insurance companies for a given treatment may vary based on a patient's predicted risk, thereby modifying downstream medical decision making.
Circular definitions: Hospitals may be more likely to readmit emergency department patients who were previously predicted to have high-risk of readmission.

Table 1: Potential sources of bias when monitoring an ML algorithm for predicting risk of 30-day unplanned readmission. Many of these biases may vary over time, due to changes in how the clinician interacts with the algorithm and/or updates to the algorithm.

biased upwards because patients who are assigned treatment a will tend to have lower risk if the algorithm is performing as expected. As illustrated in the simulation study in Section 4.4, this can lead to unnecessarily long detection delays.

To define the data-generating mechanism for the observational setting more formally, we assume data follows the directed acyclic graph (DAG) shown in Figure 1. IMI is caused by the arrows entering the treatment decision A_t. More specifically, we assume that A_t depends on patient variables X_t, output from the ML algorithm $\hat{f}_t(X_t, \cdot)$, and non-patient factors Z_t that influence treatment decisions (e.g. recent performance of the ML algorithm).

4.3. Candidate monitoring strategies

We now describe various sequential procedures for detecting violations of the monitoring criteria listed in Section 4.1. This section presents variants of the popular CUSUM control chart (Page, 1954) for each candidate criteria and data source. The methods vary in what identifiability assumptions they require and how the chart statistics are defined. The Appendix describes a unified procedure for constructing control limits that control the Type I error and provides detailed identification assumptions.

4.3.1. MONITORING THE AVERAGE PPV/NPV (CRITERION 1)

To motivate the chart statistic for monitoring (1), let us first suppose the counterfactual outcomes are observed. Over the time window from t_1 to t_2, the difference between the threshold and the true

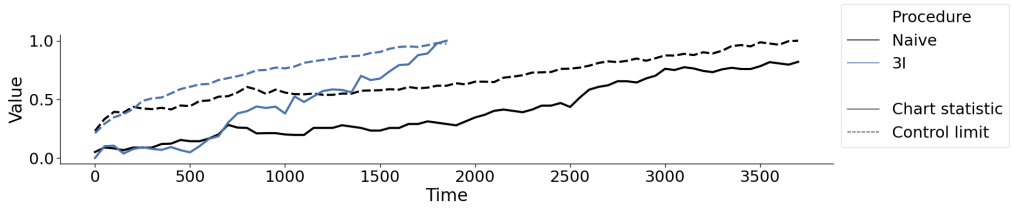

Figure 2: Example control charts, which plot the chart statistic (solid line) and control limit (dashed line) over time. When the chart statistics exceeds the control limit an alarm is fired.

PPV/NPV can be estimated by

$$c_{av} - \frac{\sum_{i=t_1}^{t_2} \mathbb{1}\{Y_i(a) = v\}\mathbb{1}\{\hat{y}_i(X_i, a) = v\}}{\sum_{i=t_1}^{t_2} \mathbb{1}\{\hat{y}_i(X_i, a) = v\}} \quad \text{for } a, v = 0, 1. \tag{4}$$

Equivalently, we can rescale by the denominator and monitor $\Pr(Y_i(a) = v, \hat{y}_i(X_i, a) = v|\mathcal{F}_i)$ per

$$\frac{1}{t_2 - t_1 + 1} \sum_{i=t_1}^{t_2} (c_{av} - \mathbb{1}\{Y_i(a) = v\})\, \mathbb{1}\{\hat{y}_i(X_i, a) = v\}. \tag{5}$$

As the actual change time is unknown, the CUSUM procedure (Page, 1954) defines a scan statistic that calculates the maximum cumulative sum of residuals over all possible changepoints:

$$\hat{C}^{\text{causal}}(t) = \max_{a,\tau,v} \sum_{i=\tau}^{t} (c_{av} - \mathbb{1}\{Y_i(a) = v\})\mathbb{1}\{\hat{y}_i(X_i, a) = v\}. \tag{6}$$

Under the null, the conditional mean of each summand in (6) is zero, which implies that the CUSUM chart statistic converges to a Brownian motion; this facilitates the construction of control limits that control the Type I error (Chu et al., 1996; Zeileis, 2005). Under the alternative, the conditional mean of each summand is positive, so (6) grows at an $O_p(t)$ rate and has an asymptotic power of one.

Option 1N: A naïve monitoring procedure If one was to ignore the impact of IMI, one may choose to directly monitor the average PPV/NPV in the observed data using the chart statistic

$$\hat{C}^{\text{1N}}(t) = \max_{a,\tau,v} \sum_{i=\tau}^{t} (c_{av} - \mathbb{1}\{Y_i = v\})\mathbb{1}\{\hat{y}_i(X_i, A_i) = v, A_i = a\}. \tag{7}$$

However, it is clear from the causal model that this quantity can be biased upwards or downwards depending on the treatment propensities. Thus, the alarm rate for such a procedure may be too high or too low.

Option 1I (Interventional) To avoid biases introduced by IMI, one option is to randomize treatment with known weights $p_i(a|x, z, \hat{f}) \in (0, 1)$. By design, we have through inverse propensity weighting (IPW)

$$\mathbb{E}\left[\left(c_{av} - \frac{\mathbb{1}\{Y_i = v, A_i = a\}}{p_i(A_i|X_i, Z_i, \hat{f}_i)}\right) \mathbb{1}\left\{\hat{y}_i(X_i, a) = v\right\}\right] = \mathbb{E}\left[(c_{av} - \mathbb{1}\{Y_i(a) = v\})\, \mathbb{1}\left\{\hat{y}_i(X_i, a) = v\right\}\right], \tag{8}$$

without needing to make any additional identifiability assumptions. Consequently, we can monitor the CUSUM chart statistic with IPW

$$\hat{C}^{1\mathrm{I}}(t) = \max_{\tau,a,v} \sum_{i=\tau}^{t} \left(c_{av} - \frac{\mathbb{1}\{Y_i = v, A_i = a\}}{p_i(A_i|X_i, Z_i, \hat{f}_i)} \right) \mathbb{1}\{\hat{y}_i(X_i, a) = v\}. \tag{9}$$

Option 1O (Observational) Let us continue with the above example, but we do not randomize treatment this time. Although (8) no longer holds in general, it does hold if we assume positivity—i.e. $p_t(A_t = 1|X_t, Z_t, \hat{f}_t) \in (0, 1)$ almost everywhere for the oracle model p_t—and (sequential) conditional exchangeability—i.e. $Y_t(a) \perp A_t|X_t, Z_t, \mathcal{F}_t$—for all times t. Under these identifiability assumptions, we may extend (9) to the observational setting via a two-part solution similar to that in Steiner and MacKay (2001a), where we (i) monitor PPV/NPVs assuming the propensity model is constant and (ii) monitor for changes in the propensity model. As the treatment propensities are unknown, we approximate (9) using estimates of the propensities instead. That is, letting $\hat{p}_t$ denote our estimate of the propensity model at time t, the chart statistic is defined as

$$\hat{C}_a^{1\mathrm{O}}(t) = \max_{\tau,a,v} \sum_{i=\tau}^{t} \left(c_{av} - \frac{\mathbb{1}\{Y_i = v, A_i = a\}}{\hat{p}_i(A_i|X_i, Z_i, \hat{f}_i)} \right) \mathbb{1}\{\hat{y}_i(X_i, a) = v\}. \tag{10}$$

Because (10) requires an estimate of the propensity model at $t = 0$, this monitoring procedure requires conducting a "pre-monitoring study" after the deployment of the algorithm, where the sole purpose is to learn treatment assignment patterns (Feng et al., 2024a). During this phase, healthcare providers are asked to make treatment decisions based on predictions from the ML algorithm, but we only begin monitoring upon completion of this phase. (This pre-monitoring study is akin to Phase I in SPC (Qiu, 2013).) Another drawback of this procedure is that we cannot verify the identifiability assumptions (Pearl et al., 2016). In fact, violations of the positivity condition are more likely to occur the more accurate the ML algorithm is (Lenert et al., 2019).

4.3.2. MONITORING SUBGROUP-SPECIFIC PPV/NPVS (CRITERION 2)

We now extend the procedures from above to monitor subgroup-specific PPV/NPVs instead. For the interventional setting (**Option 2I**), we define the chart statistic as the maximum of (9) calculated for each subgroup, i.e.

$$\hat{C}^{2\mathrm{I}}(t) = \max_{\tau,a,v,k} w_{kv} \sum_{i=\tau}^{t} \left(c_{avk} - \frac{\mathbb{1}\{Y_i = v, A_i = a\}}{p_i(A_i|X_i, Z_i, \hat{f}_i)} \right) \mathbb{1}\{\hat{y}_i(X_i, a) = v, X_i \in \mathcal{S}_k\}. \tag{11}$$

where w_{kv} is the weight associated with subgroup $\mathcal{S}_k$ and label v. For the observational setting (**Option 2O**), the chart statistic $\hat{C}^{2\mathrm{O}}$ is exactly the same as $\hat{C}^{2\mathrm{I}}$ except that the oracle propensity model is replaced by its estimate.

Compared to tests for Criterion 1, these tests can be more powerful when performance decay is confined to only one of the specified subgroups. Also, in the observational setting, the major benefit of Option 2O over Option 1O is that the subgroup weights w_{kv} can be selected to downweight subgroups with (near-)violations of the positivity condition; we can even remove such subgroups altogether. For instance, the empirical analyses in Section 4.4 set w_{kv} to be the inverse of the

estimated standard deviation of the summand in (11) for subgroup $\mathcal{S}_k$, with respect to the pre-change distribution. This way, the variance of the CUSUM statistic in each subgroup has roughly the same variance.

Potential concerns of these procedures are that: (i) their power depends on the overlap between $\{\mathcal{S}_k\}$ and the actual subgroup that experiences performance decay, (ii) more data needs to be collected upfront to estimate the initial PPV/NPV per subgroup and select the subgroup-specific thresholds c_{ak}, and (iii) we must perform multiplicity correction both across subgroups and over time to control the Type I error.

4.3.3. CHECKING FOR OVER-CONFIDENT RISK PREDICTIONS (CRITERION 3)

To test (3), we note that the expected residual

$$\mathbb{E}\left[\hat{s}_t(X_t, a)\left(\hat{f}_t(X_t, a) - Y_t(a)\right) - \delta \mid X_t = x, \mathcal{F}_t\right] \tag{12}$$

is no larger than zero for almost every x under the null. On the other hand, (12) will take on positive values at some values of x under the alternative. So to check for violations of (3), one approach is to check if (12) when averaged over *any* marginal distribution of $X_t \mid \mathcal{F}_t$ is large. In particular, **Option 3I** monitors the following chart statistic in the interventional setting:

$$\hat{C}^{3\mathrm{I}}(t) = \max_{\tau,k} w_{kv} \sum_{i=\tau}^{t} \left[\hat{s}_t(X_t, A_i)\left(\hat{f}_i(X_i, A_i) - Y_i\right) - \delta\right] \mathbb{1}\{X_i \in \mathcal{S}_k\}. \tag{13}$$

Compared to Options 1I and 2I, a major benefit of this approach is that there are no inverse weights, so it does not require the positivity assumption. However, the drawback is that (13) does not monitor a standard performance metric and is thus not very interpretable. Nevertheless, it may still be effective as a monitoring metric.

Another drawback of this approach is its sensitivity to miscalibration in the model, even in small subgroups. In addition, this procedure places more weight on checking predictions for treatments with higher propensities. So compared to the procedures with inverse weights, it will have lower power when distribution shifts concentrate in low-propensity regions.

In the observational setting, **Option 3O** assumes sequential conditional exchangeability so that (12) is equal to the observational quantity $\mathbb{E}[\hat{s}_t(X_t, A_t)(Y_t - \hat{f}_t(X_t, A_t)) - \delta \mid A_t = a, X_t = x, \mathcal{F}_t]$ for almost every (x, a). As such, the chart statistic $\hat{C}^{3\mathrm{O}}$ is defined exactly the same as $\hat{C}^{3\mathrm{I}}$. Compared to Options 1O and 2O, the benefits of this approach are that: (i) there is no need to model the propensity and the propensity model may even vary over time, (ii) we do not need to conduct a pre-monitoring study (as opposed to Option 2O), and (iii) the positivity assumption is not needed to control the Type I error.

4.4. Comparing candidate strategies: a simulation study

To design the most suitable monitoring solution, one needs to compare the multitude of monitoring options with respect to various dimensions, including their interpretability, fairness, data requirements, identification assumptions, hyperparameters, and operating characteristics. Many of these properties can be summarized using a table, as shown in Table 2. Given that we do not know how exactly the data will evolve over time, the monitoring procedures should be evaluated across a variety of simulated data settings; active stakeholder engagement is necessary to ensure this set

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

Table 2: Properties of different monitoring procedures

of simulations is sufficiently comprehensive. In this section, we illustrate how such simulations can illuminate differences in the operating characteristics between the aforementioned monitoring options.

For this case study, suppose we believe healthcare providers will closely follow the ML algorithm's recommendations in the observational setting. We compare against an interventional setting where randomization weights are defined using a logistic regression model that favors the recommended treatment but with less extreme propensities. An alternative view of this comparison is that it illustrates how different levels of clinician trust can impact detection delay. We simulate a sudden shift in the conditional distribution at time $t = 500$ and vary the following factors:

- Whether the conditional distribution of $Y(X, a)$ shifts for treatment $a = 0$ versus $a = 1$. Treatment $a = 0$ is assigned at a higher rate in the population in the observational setting. (`Treatment A=0` vs `A=1 Shift`)

- The subpopulation experiencing calibration decay. We simulate a shift in the conditional distribution of $Y(a)$ for all X (`Subgroup all`), a known subgroup with prevalence of 40% (`Subgroup known`), and a misspecified subgroup with prevalence of 35% (`Subgroup misspec`).

- How much the risk $\Pr(Y(X, a) = 1|X)$ increased (`Magnitude 10%` versus `20%`).

In the Appendix, we also present results when the simulated distribution shift is gradual.

For all the monitoring criteria, the null hypotheses allow for only a 2% drop in the PPV/NPV or 2% difference in the true risk. As such, none of the null hypotheses are true. The control limit for each procedure controls the Type I error rate over 4000 time points at 10%. We implement procedures for Criteria 2 and 3 to monitor shifts with respect to all patients, the known subgroup, and the complement of the known subgroup. Full simulation details are provided in the Appendix.

As shown in Figure 5, there are substantial differences in power across the twelve simulation settings. The performance of the naïve procedure was highly erratic because it fails to adjust for

IMI. It attains the fastest time to detection when there is a shift in the entire population, because the PPV drops substantially once the ML algorithm is deployed and even more after the shift. However, the naïve procedure can be one of the slowest detectors in other scenarios. Among the methods that correctly adjust for IMI, monitoring criterion 3 was consistently the most powerful approach. The gap in performance between criterion 3 versus 1 and 2 tended to be larger when the magnitude and/or prevalence of performance decay was smaller. The ranking between methods for monitoring criterion 1 and 2 alternated based on how widespread calibration decay was and whether it was contained within a known subgroup.

Interestingly, collecting interventional data was beneficial only in certain circumstances. In particular, it led to faster detection when distribution shifted with respect to treatment $A = 1$ but slower detection when the shift was with respect to treatment $A = 0$. This is because the rate of assigning treatment $A = 0$ was higher in the observational setting. Randomization tended to be more helpful for criteria 1 and 2 than 3, because the inverse propensity weights became less extreme. In contrast, detection speed and power using Options 3O versus 3I were generally quite similar.

By conducting a thorough analysis of various monitoring options, we can weigh the utility of various procedures. In practice, one would likely simulate numerous other data settings in addition to those shown. For instance, one may investigate the impact of changes in clinician trust over time, model updating over time, violations of the assumptions made by the monitoring procedures, and more. For this case study, Table 2 and the simulation results suggest that Option 3O is a reasonable monitoring strategy. This would significantly simplify the deployment of the ML algorithm, though it may require recalibration of the initial model to ensure that it is strongly calibrated.

5. Discussion

Although the problem of performativity complicates monitoring the performance of ML algorithms, we show in this work that causal inference provides a way to conceptualize biases induced by ML algorithms. However, the question when designing a monitoring strategy is not only how to adjust for performativity but also *what* we should be monitoring. Not all monitoring systems are created equal. There are a multitude of monitoring strategies to choose from, that vary in their data source, identifiability assumptions, interpretability of their test statistics, the number of hyperparameters, and more. For instance, we find in this case study that checking for fairness violations can be a powerful approach to detecting model decay early, revealing an interesting connection between algorithmic fairness and performance monitoring. More generally, ML quality teams should conduct systematic evaluations and seek input from diverse stakeholders to choose between various options, as the choice is often not clear upfront. After a monitoring system has been put in place, documentation should also be available to users to understand the operating characteristics of the monitoring procedures.

There are still many open areas for research, many of which we believe can be answered with the help of causal inference. First, the interplay between model monitoring and other types of performativity warrants further investigation, such as verification bias due to diagnostic algorithms as well as automation bias when using Large Language Models (LLMs) (Lee et al., 2023). Moreover, with the proliferation of ML systems, it becomes increasingly important to study interactions between multiple ML algorithms. Second, different types of experimental designs should be considered. In this work, we only consider collecting data from fully observational or fully interventional set-

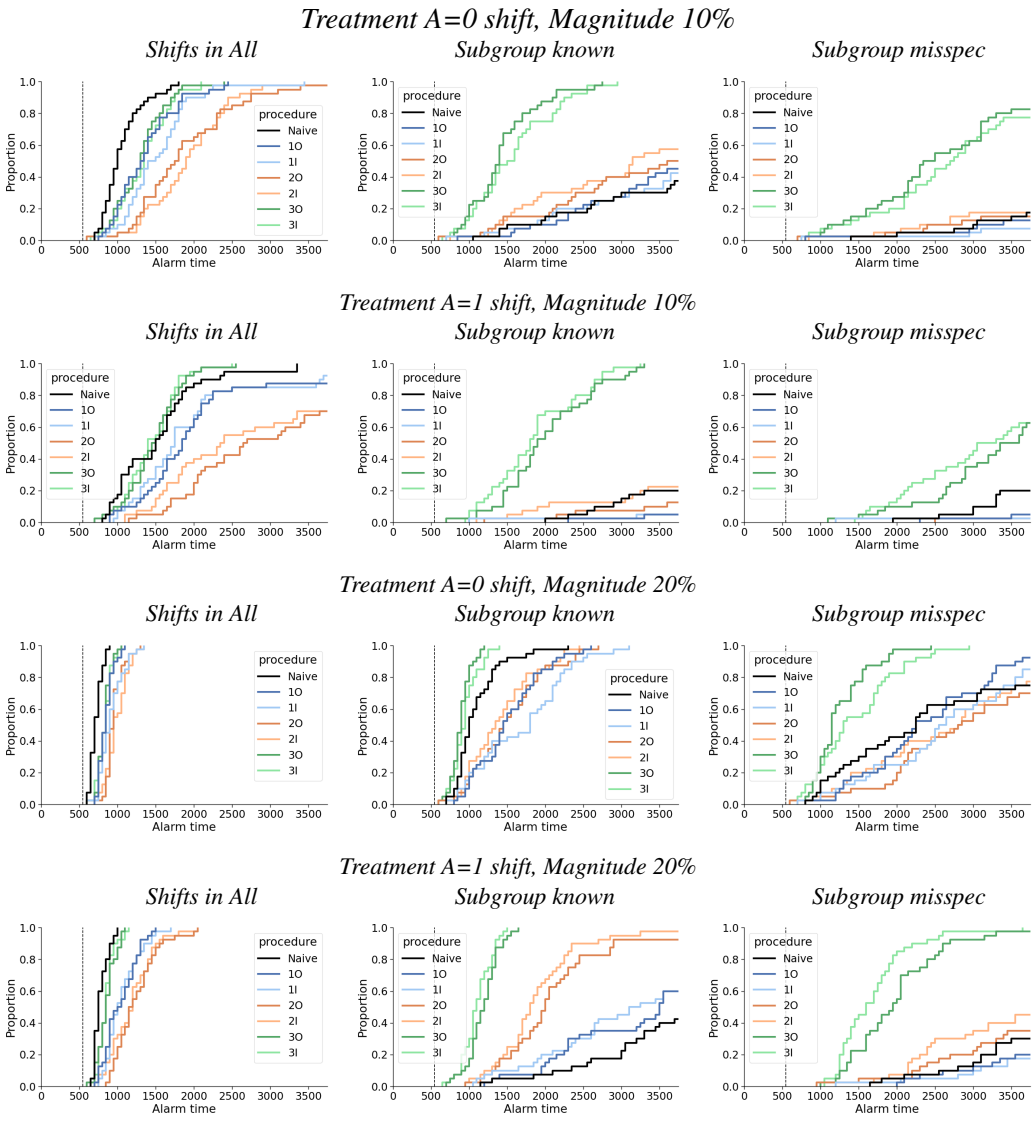

Figure 3: Statistical power of different procedures, as characterized by the proportion of alarms fired at each time point. Dashed vertical line is the time of the distribution shift.

tings. One may also consider combining both data sources (Bareinboim and Pearl, 2016), adaptive randomization (Pallmann et al., 2018), or even quasi-experimental designs such as encouragement designs (Hirano et al., 2000) and instrumental variable analyses (Baiocchi et al., 2014). More generally, one can even consider other types of estimands. We have focused on the predictive value of an ML-based risk prediction algorithm, but one may also be interested in the overall effect of the device on patient outcomes, which may require different randomization schemes (Bossuyt et al., 2000; U.S. Food and Drug Administration, 2022). Finally, this work studies the performance of variants of the CUSUM, but future work should consider other types of monitoring procedures including those based on anytime inference (Grünwald et al., 2019; Shekhar and Ramdas, 2023) and Bayesian inference (West and Harrison, 1997).

Acknowledgments

This work was funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2022C1-25619). The views presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of the PCORI®, its Board of Governors or Methodology Committee, and the Food and Drug Administration.

References

Todd A Alonzo and John M Kittelson. A novel design for estimating relative accuracy of screening tests when complete disease verification is not feasible. *Biometrics*, 62(2):605–612, June 2006.

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Stat. Med.*, 33(13):2297–2340, June 2014.

Sebastiano Barbieri, James Kemp, Oscar Perez-Concha, Sradha Kotwal, Martin Gallagher, Angus Ritchie, and Louisa Jorm. Benchmarking deep learning architectures for predicting readmission to the ICU and describing Patients-at-Risk. *Sci. Rep.*, 10(1):1111, January 2020.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. U. S. A.*, 113(27):7345–7352, July 2016.

Colin B Begg and Robert A Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39(1):207–215, 1983. URL http://www.jstor.org/stable/2530820.

P M Bossuyt, J G Lijmer, and B W Mol. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet*, 356(9244):1844–1847, November 2000.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *J. Mach. Learn. Res.*, 14(101):3207–3260, 2013.

Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D Sculley. The ML test score: A rubric for ML production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1123–1132. ieeexplore.ieee.org, December 2017. URL http://dx.doi.org/10.1109/BigData.2017.8258038.

Robert E Burke, Jeffrey L Schnipper, Mark V Williams, Edmondo J Robinson, Eduard E Vasilevskis, Sunil Kripalani, Joshua P Metlay, Grant S Fletcher, Andrew D Auerbach, and Jacques D Donzé. The HOSPITAL score predicts potentially preventable 30-day readmissions in conditions targeted by the hospital readmissions reduction program. *Med. Care*, 55(3):285–290, March 2017.

Allison J B Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 224–232, New York, NY, USA, September 2018. Association for Computing Machinery. URL https://doi.org/10.1145/3240323.3240370.

Chia-Shang James Chu, Maxwell Stinchcombe, and Halbert White. Monitoring structural change. *Econometrica*, 64(5):1045–1065, 1996. URL http://www.jstor.org/stable/2171955.

Andrea J Cook, Robert D Wellman, Jennifer C Nelson, Lisa A Jackson, and Ram C Tiwari. Group sequential method for observational data by using generalized estimating equations: application to vaccine safety datalink. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 64(2):319–338, 2015. URL http://www.jstor.org/stable/24771896.

Conor K Corbin, Rob Maclay, Aakash Acharya, Sreedevi Mony, Soumya Punnathanam, Rahul Thapa, Nikesh Kotecha, Nigam H Shah, and Jonathan H Chen. DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record. *J. Am. Med. Inform. Assoc.*, 30(9):1532–1542, August 2023. URL http://dx.doi.org/10.1093/jamia/ocad114.

Holger Dette and Josua Gösmann. A likelihood ratio approach to sequential change point detection for a general class of parameters. *J. Am. Stat. Assoc.*, 115(531):1361–1377, July 2020. URL https://doi.org/10.1080/01621459.2019.1630562.

Kirill Dyagilev and Suchi Saria. Learning (predictive) risk scores in the presence of censoring due to interventions. *Mach. Learn.*, 102(3):323–348, March 2016. URL https://doi.org/10.1007/s10994-015-5527-7.

Stephanie Eaneff, Ziad Obermeyer, and Atul J Butte. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA*, September 2020.

Jean Feng, Rachael V Phillips, Ivana Malenica, Andrew Bishara, Alan E Hubbard, Leo A Celi, and Romain Pirracchio. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *npj Digital Medicine*, 5(1):1–9, May 2022. URL https://www.nature.com/articles/s41746-022-00611-y.

Jean Feng, Alexej Gossmann, Gene Pennello, Nicholas Petrick, Berkman Sahiner, and Romain Pirracchio. Monitoring machine learning-based risk prediction algorithms in the presence of performativity. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, May 2024a. URL http://arxiv.org/abs/2211.09781.

Jean Feng, Alexej Gossmann, Romain Pirracchio, Nicholas Petrick, Gene Pennello, and Berkman Sahiner. Is this model reliable for everyone? testing for strong calibration. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, May 2024b. URL http://arxiv.org/abs/2211.09781.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.*, 385(3):283–286, July 2021. URL http://dx.doi.org/10.1056/NEJMc2104626.

Gombay. Parametric sequential tests in the presence of nuisance parameters. *Theory Stoch. Process.*, 2002. URL https://www.researchgate.net/profile/Edit-Gombay/publication/228881435_Parametric_sequential_tests_in_the_

presence_of_nuisance_parameters/links/0912f51193d6682ce5000000/Parametric-sequential-tests-in-the-presence-of-nuisance-parameters.pdf.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. June 2019.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) masses. *International Conference on Machine Learning*, 80:1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Miguel A Hernán and James M Robins. Causal inference, 2010.

K Hirano, G W Imbens, D B Rubin, and X H Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, March 2000.

Leora I Horwitz, Masha Kuznetsova, and Simon A Jones. Creating a learning health system through Rapid-Cycle, randomized testing. *N. Engl. J. Med.*, 381(12):1175–1179, September 2019. URL http://dx.doi.org/10.1056/NEJMsb1900856.

Janis Klaise, Arnaud Van Looveren, Clive Cox, Giovanni Vacanti, and Alexandru Coca. Monitoring and explainability of models in production. In *Workshop on Challenges in Deploying and Monitoring Machine Learning Systems*, July 2020. URL http://arxiv.org/abs/2007.06299.

Joseph A Ladapo, Saul Blecker, Michael R Elashoff, Jerome J Federspiel, Dorice L Vieira, Gaurav Sharma, Mark Monane, Steven Rosenberg, Charles E Phelps, and Pamela S Douglas. Clinical implications of referral bias in the diagnostic performance of exercise testing for coronary artery disease. *J. Am. Heart Assoc.*, 2(6):e000505, December 2013. URL http://dx.doi.org/10.1161/JAHA.113.000505.

Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.*, 388(13):1233–1239, March 2023.

Matthew C Lenert, Michael E Matheny, and Colin G Walsh. Prognostic models will be victims of their own success, unless.... *J. Am. Med. Inform. Assoc.*, 26(12):1645–1650, December 2019. URL http://dx.doi.org/10.1093/jamia/ocz145.

James Liley, Samuel Emerson, Bilal Mateen, Catalina Vallejos, Louis Aslett, and Sebastian Vollmer. Model updating after interventions paradoxically introduces bias. *International Conference on Artificial Intelligence and Statistics*, 130:3916–3924, 2021. URL http://proceedings.mlr.press/v130/liley21a.html.

Elham Mahmoudi, Neil Kamdar, Noa Kim, Gabriella Gonzales, Karandeep Singh, and Akbar K Waljee. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ*, 369:m958, April 2020.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating performativity by predicting from predictions. *Conference on Neural information processing systems*, August 2022.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annu. Rev. Stat. Appl.*, 8(1):141–163, March 2021.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *J. Am. Stat. Assoc.*, 116(534):1023–1037, April 2021. URL https://doi.org/10.1080/01621459.2021.1874961.

Douglas C Montgomery. *Statistical quality control*. John Wiley & Sons, Nashville, TN, 7 edition, 2013.

Kyosuke Nishida and Koichiro Yamauchi. Detecting concept drift using statistical testing. In *Discovery Science*, pages 264–269. Springer Berlin Heidelberg, 2007. URL http://dx.doi.org/10.1007/978-3-540-75488-6_27.

E S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. URL http://www.jstor.org/stable/2333009.

Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.*, 55(6):1–29, December 2022. URL https://doi.org/10.1145/3533378.

Philip Pallmann, Alun W Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V Hampson, Jane Holmes, Adrian P Mander, Lang'o Odondi, Matthew R Sydes, Sofía S Villar, James M S Wason, Christopher J Weir, Graham M Wheeler, Christina Yap, and Thomas Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med.*, 16(1):29, February 2018.

Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Annu. Symp. Proc.*, 2013:1109–1115, November 2013. URL https://www.ncbi.nlm.nih.gov/pubmed/24551396.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, January 2016. URL https://play.google.com/store/books/details?id=I0V2CwAAQBAJ.

Michael J Pencina, Ralph B D'Agostino, Sr, Ralph B D'Agostino, Jr, and Ramachandran S Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.*, 27(2):157–72; discussion 207–12, January 2008. URL http://dx.doi.org/10.1002/sim.2929.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé Iii and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Peihua Qiu. *Introduction to Statistical Process Control*. Chapman and Hall/CRC, 1st edition edition, October 2013. URL https://www.taylorfrancis.com/books/mono/10.1201/b15016/introduction-statistical-process-control-peihua-qiu.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 33–44, New York, NY, USA, January 2020. Association for Computing Machinery.

Tim Schröder and Michael Schulz. Monitoring machine learning models: a categorization of challenges and methods. *Data Science and Management*, 5(3):105–116, September 2022. URL https://www.sciencedirect.com/science/article/pii/S2666764922000303.

D Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In C Cortes, N Lawrence, D Lee, M Sugiyama, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential changepoint detection via backward confidence sequences. *International Conference on Machine Learning*, June 2023. URL https://openreview.net/pdf?id=QT5Cphscf2.

Matthew Sperrin, David Jenkins, Glen P Martin, and Niels Peek. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Inform. Assoc.*, 26 (12):1675–1676, December 2019.

Stefan H Steiner and R Jock MacKay. Monitoring processes with data censored owing to competing risks by using exponentially weighted moving average control charts. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 50(3):293–302, September 2001a. URL https://academic.oup.com/jrsssc/article-pdf/50/3/293/48750653/jrsssc_50_3_293.pdf.

Stefan H Steiner and R Jock MacKay. Detecting changes in the mean from censored lifetime data. *Frontiers in Statistical Quality Control*, 2001b.

Rena Jie Sun, John D Kalbfleisch, and Douglas E Schaubel. A weighted cumulative sum (WCUSUM) to monitor medical outcomes with dependent censoring. *Stat. Med.*, 33(18):3114–3129, August 2014. URL http://dx.doi.org/10.1002/sim.6139.

US Food and Drug Administration. Statistical guidance on reporting results from studies evaluating diagnostic tests. 2007. URL https://www.fda.gov/files/medical%20devices/published/Guidance-for-Industry-and-FDA-Staff---Statistical-Guidance-on-Reporting-Result 28PDF-Version%29.pdf.

U.S. Food and Drug Administration. Clinical performance assessment: Considerations for Computer-Assisted detection devices applied to radiology images and radiology device data in premarket notification (510(k)) submissions. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-dev September 2022. Accessed: 2023-9-20.

U.S. Food and Drug Administration and Health Canada. Good machine learning practice for medical device development, October 2021. URL https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.*, 74:167–176, June 2016. URL http://dx.doi.org/10.1016/j.jclinepi.2015.12.005.

Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, New York, NY, 1997.

Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. *IJCAI*, January 2022.

Achim Zeileis. A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Economet. Rev.*, 24(4):445–466, October 2005. URL https://doi.org/10.1080/07474930500406053.

Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Trans. Software Eng.*, 48(1):1–36, January 2022. URL http://dx.doi.org/10.1109/TSE.2019.2962027.

Xiang Zhang and William H Woodall. Dynamic probability control limits for risk-adjusted bernoulli CUSUM charts. *Stat. Med.*, 34(25):3336–3348, November 2015. URL http://dx.doi.org/10.1002/sim.6547.

X H Zhou. Comparing correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias. *Biometrics*, 54(2):453–470, June 1998.

Appendix A. Calculating control limits

For the procedures monitoring Criterion 3, we use the Monte Carlo procedure described in (Zhang and Woodall, 2015; Feng et al., 2024b) to construct dynamic control limits (DCL). At each time t, we bootstrap Y_t^* given X_t under the worst-case null hypothesis satisfying (3), which corresponds to the case where $\Pr(Y_t^* = 1|X_t, \hat{f}_t) = \hat{f}_t(X_t, A_t) + \delta$. For each of the bootstrap sequences, we calculate the corresponding chart statistics. The DCL is then chosen to satisfy an alpha spending function. Here, we use an alpha spending function that uniformly spends the α over time. This procedure provably controls the Type I error in finite samples (Feng et al., 2024b).

In general, Criteria 1 and 2 are weaker than 3. To conduct a more fair comparison in the simulation studies, we construct DCLs for their corresponding test statistics to control the Type I error rate under Criteria 3. Thus the sensitivity of the methods for monitoring Criteria 1 or 2 are actually higher than they would be otherwise. Also, for illustrative purposes, we make the simplifying assumption in the simulation studies that the estimation error for the propensity model is negligible; this is true, for instance, if we had a sufficiently long pre-monitoring phase. To formally adjust for estimation error, one can use ideas such as that in (Gombay, 2002; Dette and Gösmann, 2020; Feng et al., 2024a).

Appendix B. Identification assumptions

The identification assumptions needed for procedures 1I and 2I are as follows:

Condition 1 (Positivity) *For some $\epsilon > 0$, weights $p_t(a_t|x_t, z_t, \hat{f}_t)$ satisfy $p_t(a_t|x_t, z_t, \hat{f}_t) \in (\epsilon, 1-\epsilon)$ for all time points t.*

Condition 2 (Conditional Exchangeability) *The potential outcome $Y_t(a)$ is conditionally independent of treatment assignment A_t, i.e. $Y_t(a) \perp A_t|X_t, Z_t, \mathcal{F}_t$.*

Appendix C. Simulation settings

There are two steps to conducting this simulation study. First, we must simulate a hypothetical ML algorithm for predicting a patient's risk of readmission. Second, we simulate data to investigate the operating characteristics of various monitoring procedures. We discuss each step in turn.

Simulating the algorithm. We consider a setup with only two patient variables $X \in \mathbb{R}^{10}$, generated independently from a normal distribution with mean zero and variance 4. In the pre-deployment setting, we suppose the treatment was assigned uniformly at random. The data is generated from the logistic regression model

$$\text{logit}(y = 1|x, a) = -0.5x_1 - x_2 + 0.5a + x_1a + 2x_2a.$$

A random forest classifier $\hat{f}$ is trained using 5000 observations and locked thereafter. The model outputs the predicted risk (probability) of a readmission under a particular treatment.

Treatment propensities. We suppose that treatment decisions are assigned according to a logistic regression model of the form

$$\text{logit}(a = 1|x, \hat{f}_t) = \beta \left[\hat{f}_t(x, a = 1) - \hat{f}_t(x, a = 0))\right]$$

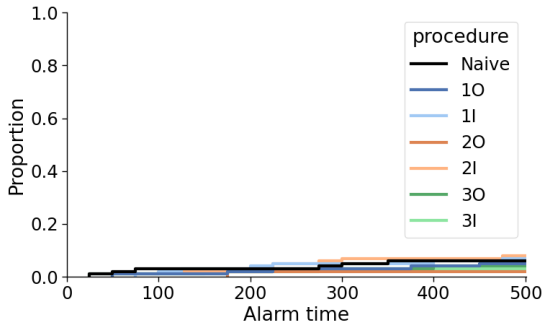

Figure 4: Assessing Type I error for monitoring procedures

for some $\beta \in \mathbb{R}$. We suppose $\beta = -6$ in the observational setting (i.e. clinicians closely follow recommendations from the ML algorithm) and $\beta = -2$ in the interventional setting (i.e. randomization weights favor following the ML algorithm).

Shifts in the outcome distribution. We simulate shifts in the conditional distribution of $Y(a)$ under treatment a for $a = 0$ versus $a = 1$ in the following subgroups:

- `Subgroup all`: all x
- `Subgroup known`: $x_1 \in (-1, 2)$ and $|x_2| < 2.5$
- `Subgroup misspec`: $|x_1| < 1.5$ and $|x_2| < 1.5$

The simulations vary in the magnitude of the increase ($c = 10\%$ versus $c = 20\%$). Specifically, for subgroup $\mathcal{S}$ and treatment a, the conditional distribution shifts to

$$\begin{aligned} p_1(Y = 1|X, A) :=& p_0(Y = 1|X, A) - \mathbb{1}\{p_0(Y = 1|X, A) > 0.5, X \in \mathcal{S}, A = a\} * c \\ &+ \mathbb{1}\{p_0(Y = 1|X, A) < 0.5, X \in \mathcal{S}, A = a\} * c \end{aligned}$$

where p_0 is the pre-change probability and p_1 is the post-change probability.

Pre-specified subgroups in the monitoring procedures. The subgroups considered by monitoring procedures 2I, 2O, 3I, and 3O are: $\mathcal{S}_1 = \{x : x \in \mathbb{R}\}$, $\mathcal{S}_2 = \{x : x_1 \in (-1, 2), |x_2| < 2.5\}$, and $\mathcal{S}_3 = \mathcal{S}_1 \setminus \mathcal{S}_2$. For computational speed, observations were monitored in batches of 50.

Appendix D. Additional simulation results

We also ran a simulation to verify Type I error control of the monitoring procedures. The nominal rate was set to $\alpha = 0.1$ and the data was simulated to be IID over time. Results shown in Figure 4 are for $\delta = 0.02$.

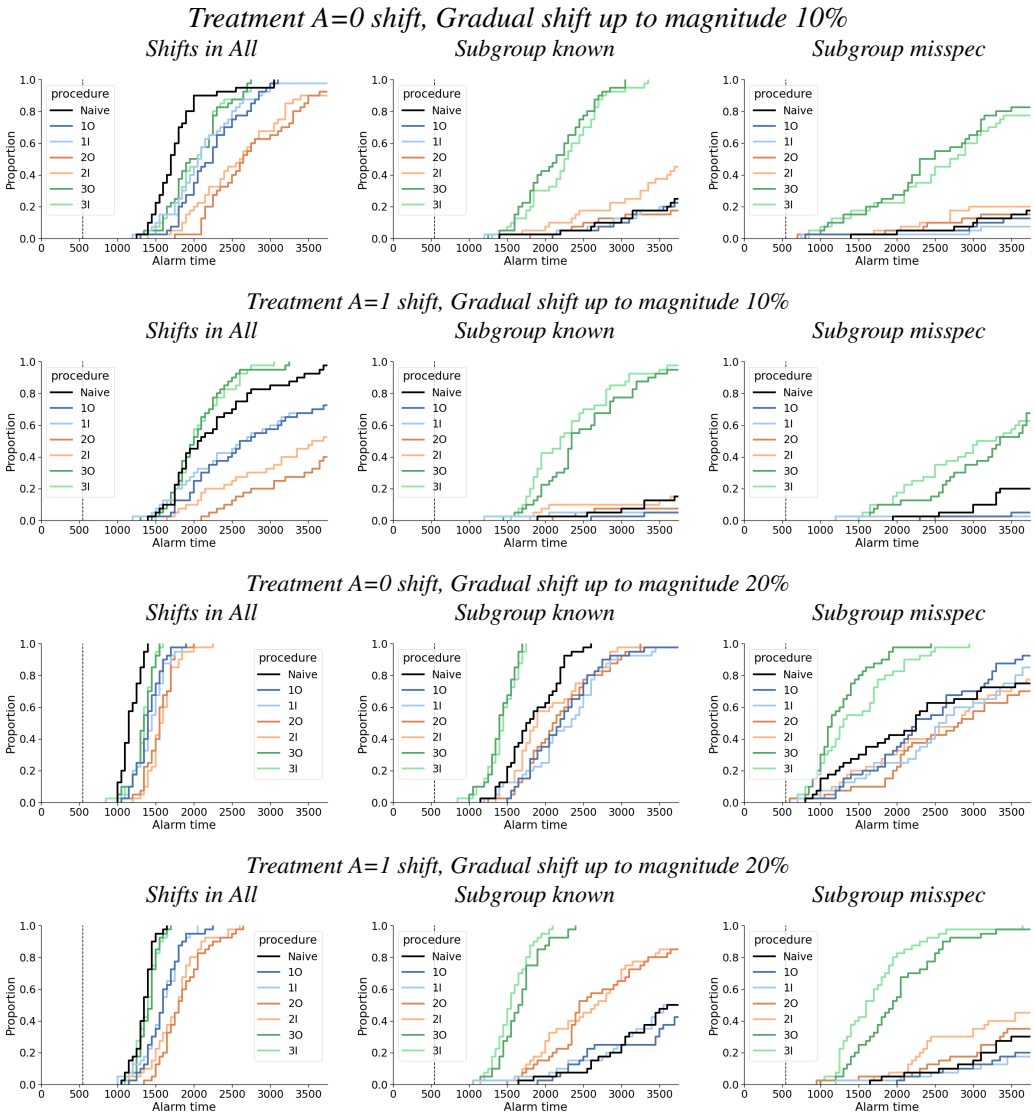

Figure 5: Statistical power of different procedures, as characterized by the proportion of alarms fired at each time point. The simulated shift in the conditional distribution of the outcome is gradual over time, with the start time of the shift indicated by the dashed vertical line.