# Monitoring machine learning-based risk prediction algorithms in the presence of performativity

**Jean Feng**[1] **Alexej Gossmann**[2] **Gene Pennello**[2]
**Nicholas Petrick**[2] **Berkman Sahiner**[2] **Romain Pirracchio**[1]
[1]University of California, San Francisco
[2]U.S. Food and Drug Administration, Center for Devices and Radiological Health

## Abstract

Performance monitoring of machine learning (ML)-based risk prediction models in healthcare is complicated by the issue of performativity: when an algorithm predicts a patient to be at high risk for an adverse event, clinicians are more likely to administer prophylactic treatment and alter the very target that the algorithm aims to predict. A simple approach is to ignore performativity and monitor only the untreated patients, whose outcomes remain unaltered. In general, ignoring performativity may inflate Type I error because (i) untreated patients disproportionally represent those with low predicted risk, and (ii) changes in the clinician's trust in the ML algorithm and the algorithm itself can induce complex dependencies that violate standard assumptions. Nevertheless, we show that valid inference is still possible when monitoring *conditional* rather than marginal performance measures under either the assumption of conditional exchangeability or time-constant selection bias. Finally, performativity can vary over time and induce nonstationarity in the data, which presents challenges for monitoring. To this end, we introduce a new score-based cumulative sum (CUSUM) monitoring procedure with dynamic control limits. Through extensive simulation studies, we study applications of the score-based CUSUM and how it is affected by various factors, including the efficiency of model updating procedures and the level of clinician trust. Finally, we apply the procedure to detect calibration decay of a risk model during the COVID-19 pandemic.

## 1 INTRODUCTION

After a machine learning (ML)-based system is deployed in clinical practice, real-world monitoring of the algorithm for potential performance degradation is necessary for mitigating risk and is an important aspect of good machine learning practice (GMLP) (Breck et al., 2017; U.S. Food and Drug Administration and Health Canada, 2021). Locked ML algorithms may gradually become outdated and output misleading risk predictions; continual learning procedures are exposed to the additional risk of incorporating deleterious model updates (Klaise et al., 2020; Feng et al., 2020). Various methods for performance monitoring are available, which can generally be categorized into those based on statistical process control (Kahn et al., 1996; Gama et al., 2014; Feng et al., 2022) and sliding window comparisons (Bifet and Gavaldà, 2007; Nishida and Yamauchi, 2007). All these procedures assume an ideal data setting in which the prediction target is observed. However, the data available for monitoring an ML-based risk prediction algorithm are often subject to performativity, where predictions from the algorithm can alter the very outcome that it aims to predict (Paxton et al., 2013; Lenert et al., 2019; Perdomo et al., 2020; Liley et al., 2021).

As an example, consider the Targeted Real-time Early Warning System (TREWS) sepsis risk prediction algorithm (Adams et al., 2022), which estimates the probability of a patient developing septic shock if they only receive standard of care (SOC) and no additional interventions. This algorithm was recently shown to reduce in-hospital mortality by increasing the likelihood that high-risk patients receive antibiotics (Henry et al., 2022). Clinicians' likelihood to interact with the TREWS system depended on their previous interactions, and the authors hypothesize that clinician trust,

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

and thus the impact of performativity, will evolve with increased exposure to ML-based systems. Monitoring TREWS is especially important because it depends on electronic health record (EHR) data, which is prone to distribution shifts. A simple monitoring solution is to compare the predicted risk to the actual outcome *only among patients who received SOC*, as counterfactual outcomes for patients receiving antibiotics are unknown. However, in the likely scenario where patients predicted to be at high risk are preferentially selected to receive an intervention, procedures that ignore performativity can be biased because the ML algorithm itself is a major source of confounding. As demonstrated in the Appendix, naïvely monitoring marginal performance measures such as misclassification rate without adjusting for treatment propensities can significantly delay detection.

In the offline setting, one can address the mismatch between the "SOC-only" and general target population by weighting the data by the inverse of their treatment propensities. Prior works have suggested combining inverse weights with sequential monitoring procedures to adjust for the similar problem of censoring (Steiner and MacKay, 2001; Sun et al., 2014). However, proper error rate control is contingent on knowing the exact weights, which is unlikely to hold in our setting. Moreover, it is difficult to anticipate how clinician trust in the ML algorithm varies over time, so it may not even be possible to estimate propensity weights accurately when performativity varies over time. Finally, treatment propensities are more extreme (close to one or zero) the better an ML algorithm is, which reduces power as shown in our empirical analyses.

Given the difficulties of monitoring *marginal* performance measures, we propose monitoring *conditional* performance measures. The intuition is that we can ignore biases induced by performativity by conditioning on variables likely to experience shifts. For binary-valued classifiers, we monitor conditional measures such as positive and negative predictive values. For risk prediction models, we monitor (stratified) calibration curves, as calibration is a popular measure of model reliability and one of the most common measures to decay in real-world settings (Hickey et al., 2013; Davis et al., 2017). Formally, we establish two conditions for ignorability: the first is a variant of the conditional exchangeability assumption (Rubin, 1976), and the second is time-constant selection bias. Under either condition, we show one can directly apply procedures intended for settings where there is no performativity to settings where there is. Thus one can analyze SOC-only data and avoid estimating treatment propensities altogether. This provides monitoring teams with a simple first step toward addressing performativity.

In addition, we address two general challenges that arise when monitoring ML-based risk prediction models. First, performativity can change over time as the ML algorithm and clinician trust in the algorithm evolve. So the population receiving SOC changes over time, and the predictor sequence is nonstationary. Second, the exact performance characteristics of the ML algorithm may not be known upfront and must be estimated. However, many monitoring algorithms assume the exact pre-change data distribution is known (Tartakovsky et al., 2014). Although there exist procedures that partially address these challenges (Dette and Gösmann, 2020; Zeileis and Hornik, 2007; Gombay, 2017), we are unaware of a frequentist monitoring procedure that adequately addresses both. To this end, we introduce a new nonanticipating score-based CUSUM chart statistic (Page, 1954) and a computationally efficient procedure for generating dynamic control limits (DCLs) (Zhang and Woodall, 2015; Driscoll et al., 2021). (Although Bayesian approaches naturally address these challenges (Fearnhead and Liu, 2007; Adams and MacKay, 2007), exact posterior inference is computationally challenging/expensive for the setting with binary outcomes and the more complex forms of performance decay considered in this work (Shiryaev, 1963; West, 1986; Bhattacharya, 1994).)

The main contributions of this work are: we (i) formalize the problem of monitoring conditional performance in the presence of performativity as a hypothesis test within a causal framework, (ii) prove conditions under which performativity is ignorable, (iii) propose a new score-based CUSUM procedure, and (iv) investigate operating characteristics of the proposed procedure in extensive simulation studies and a real-world dataset spanning the COVID-19 pandemic. We show that by wrapping evolving black-box ML algorithms within a monitoring framework, they may continue learning and improving in model discrimination (e.g., AUC), as long as they remain well-calibrated. Finally, while performativity is the primary motivation for this work, all of the results apply to the more general problem of performance monitoring in the presence of differential treatment propensities, which can occur even when no ML algorithm is deployed. Code for reproducing the paper is available at `https://github.com/jjfeng/monitoring_ML_performativity`.

## 2 RELATED WORKS

**Causal inference in sequential settings**: Prior works have highlighted how performativity can be defined formally using causal inference (Bottou et al., 2013; Chaney et al., 2018; Liley et al., 2021; Mendler-

Dünner et al., 2022). There is now a growing literature on sequential inference for causal quantities using observational data (Li et al., 2011; Cook et al., 2015; Waudby-Smith et al., 2021). Nearly all existing works target a time-constant estimand and require the positivity assumption (propensities bounded away from zero), which is likely to be practically violated in our setting. The only work we are aware of that specifically considers the connection between causal inference and performance monitoring is the review in Feng et al. (2023) of the tradeoffs between different design choices for a monitoring system. Beyond this, the closest methods are those that address the related problem of censoring (Steiner and MacKay, 2001; Sun et al., 2014).

**Online changepoint/concept drift detection:** We refer the reader to prior reviews of changepoint detection methods (Tartakovsky et al., 2014; Gama et al., 2014). Methods vary in which aspects of a data distribution they monitor, including marginal versus conditional components. The two ignorability conditions established in this work are widely applicable: Under either condition, we can directly apply likelihood-based methods for detecting conditional shifts in non-performative settings to performative settings. Nevertheless, monitoring ML-based risk prediction algorithms also requires one to account for unknown pre-change parameters and nonstationary predictors. These issues have only been addressed individually in prior works, including the score-based CUSUM in (Gombay, 2017). We extend the score-based CUSUM as its structure lends itself nicely to theoretical analyses and efficient computation. Score-based detection methods have been used in other contexts (Gombay, 2003; Liu et al., 2021; Wu et al., 2023).

# 3 TWO MONITORING PROBLEMS

To formalize the monitoring problem in the presence of performativity, we begin with defining the monitoring problem in the "standard" setting, in which outcomes are unaffected by treatment decisions. If there is a monitoring procedure for addressing the "standard" hypothesis test, can we directly apply it to the setting with performativity to analyze SOC-only data, ignoring treatment propensities entirely? It's not immediately clear what hypothesis we are trying to test through such an approach. This section shows how defining the hypothesis test in the presence of performativity requires being mathematically precise about the causal quantities being monitored.

We focus on testing a single changepoint, because the current clinical ML workflow is to deploy model monitoring, analyze the root cause when an alarm is raised, take corrective actions (e.g., update the model and/or data pre-processing), and reset the monitoring procedure to begin the cycle again. Table 1 in the Appendix summarizes notation used in this paper.

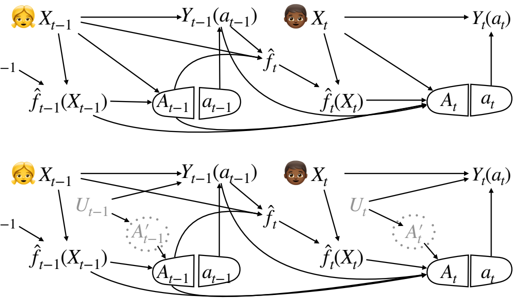


Figure 1: Example Single World Intervention Graphs (SWIGs) (Richardson and Robins, 2013) that imply ignorability of the propensity model. For time $t$, $\hat{f}_t$ is the ML algorithm, $X_t$ is subject covariates, $A_t$ is the assigned treatment, and $Y_t(a)$ is the potential outcome under treatment $a$. The top SWIG satisfies conditional exchangeability, in that $Y_t(0)$ and $A_t$ are conditionally independent given the risk prediction and observed data before time $t$. The bottom SWIG satisfies time-constant selection bias per the assumptions listed in Example 1 of the Appendix, where gray variables indicate unobserved variables. $U_t$ is an unobserved confounder, and $A_t'$ is a tentative treatment decision based solely on $U_t$ which affects $A_t$ through a determinative causation structure, as indicated by the dotted circle.

## 3.1 The standard monitoring problem

Here we describe the standard hypothesis test for a locked ML algorithm $\hat{f}$ mapping patient variables from domain $\mathcal{X} \subseteq \mathbb{R}^p$ to $\mathcal{Q} = [0,1]$. Suppose a new patient is observed at each time $t$ with covariates $X_t$ (index $t$ corresponds to different patients, not repeated observations). The clinician observes the patient's predicted risk $\hat{f}(X_t)$ and their true outcome $Y_t$ thereafter. Patients are distinct in this setup, so we suppose patient outcomes are conditionally independent, i.e., $Y_t \perp Y_{t'} | X_t$ for all $t' < t$.

Following the framework in Chu et al. (1996), we monitor for changes in the conditional distribution $Y_t|\hat{f}(X_t)$ or, equivalently, changes in the calibration curve $\mathbb{E}[Y_t|\hat{f}(X_t) = q]$ over $q \in [0,1]$. We suppose that performance drift is unlikely to occur initially, so data from time $t = 1$ to $m$ is considered "noncontaminated." We monitor from time $t = m+1$ to $mK$ for some fixed integer $K > 1$. The conditional distribution $Y_t|\hat{f}(X_t)$ is assumed to follow some model $g$ with parameters $(\theta, \delta\mathbb{1}\{t > \kappa\})$, where parameter $\theta \in \mathbb{R}^p$ describes the pre-change distribution and $\delta \in \mathbb{R}^d$ describes the shift after changepoint $\kappa = \lfloor m\kappa^{\text{rel}} \rfloor$ for some $\kappa^{\text{rel}} \in [1, K]$. So $\kappa^{\text{rel}} = K$ and/or $\delta = 0$ means there is no shift. Thus, the hypothesis test in the standard setting can be formalized as

$$\begin{aligned} &H_0 : Y_i|\hat{f}(X_i) = q \sim g_{\theta,0}(q) \quad \forall i = 1, \cdots, mK \\ &H_1 : \exists \delta, \kappa \text{ s.t. } Y_i|\hat{f}(X_i) = q \sim g_{\theta,\delta \mathbb{1}\{i>\kappa\}}(q) \\ &\quad \forall i = 1, \cdots, mK. \end{aligned} \tag{1}$$

We discuss example choices for $g$ in the Appendix.

A sequential monitoring procedure is defined by its chart statistic $C_m(t)$ and dynamic control limit (DCL) $h_m(t)$ at times $t$. A procedure fires an alarm when the chart statistic first exceeds the control limit, i.e., $\hat{T}_m = \inf\{t : C_m(t) > h_m(t)\}$. In this work, we characterize procedures by their Type I error rate—defined as the probability of firing an alarm before the changepoint, $\Pr\left(\hat{T}_m < \kappa\right)$—and asymptotic consistency—whether $\lim_{m\to\infty} \Pr\left(\hat{T}_m \leq mK\right)$ equals one under $H_1$.

### 3.2 The problem of performativity

We now formalize the hypothesis test in the setting with performative predictions, with the additional complication that the ML algorithm may evolve over time. We focus on algorithms that only predict the outcome under SOC, such as TREWS. Such algorithms are common in practice because it is simpler to model outcomes under SOC than that under every possible treatment option. Nevertheless, results in this work can be extended to monitor algorithms that predict treatment-specific outcomes; we provide a discussion of this in Section B.5 of the Appendix.

Data is now generated as follows. After observing the risk prediction $\hat{f}_t(X_t)$ of a new patient at time $t$, the clinician takes into account various factors—such as the prediction, patient covariates, and prior experience with the algorithm—to decide treatment $A_t$, where $A_t = 0$ indicates SOC and $A_t = 1$ indicates additional intervention. Following the potential outcomes framework, let $Y_t(a)$ indicate the patient outcome if treatment $a$ is administered. Under usual consistency assumptions, we denote the observed patient outcome as $Y_t = Y_t(A_t)$. Let filtration $\mathcal{F}_t$ denote the sigma-algebra generated by all data prior to time $t$, i.e., $(\hat{f}_1, A_1, X_1, Y_1, \cdots, \hat{f}_{t-1}, A_{t-1}, X_{t-1}, Y_{t-1}, \hat{f}_t)$. So $(Y_t(a), X_t, A_t)$ is adapted to filtration $\mathcal{F}_t$. Let $\tau_i$ be the timepoint of the $i$th patient receiving SOC, which can be viewed as a random stopping time. Thus SOC-only data is defined as $\{\tau_i : i = 1, \cdots, mK\}$.

When we monitor the performance of an ML algorithm using SOC-only data, we must be careful in delineating the causal hypothesis. It can be tempting to replace index $i$ in (1) with the random stopping time $\tau_i$ and the observed outcome with the potential outcome, so that the monitoring target is $Y_{\tau_i}(0)|\hat{f}_{\tau_i}(X_{\tau_i})$. However, this is no different from monitoring the *observed* distribution because $Y_{\tau_i}(0) = Y_{\tau_i}$.

To monitor the *causal* quantity of interest, let $(\tilde{X}_t, \tilde{Y}_t(a))$ represent independent observations from the target population at time $t$. These observations are solely a mathematical construct (a counterfactual in some sense) and are never observed. In the presence of performativity, the causal analog of (1) is instead

$$\begin{aligned} &H_0 : \tilde{Y}_{\tau_i}(0)|\hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \mathcal{F}_{\tau_i} \sim g_{\theta,0}(q) \\ &H_1 : \exists \delta, \kappa \text{ s.t. } \tilde{Y}_{\tau_i}(0)|\hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \mathcal{F}_{\tau_i} \sim g_{\theta,\delta \mathbb{1}\{i>\kappa\}}(q) \\ &\quad \forall i = 1, \cdots, mK. \end{aligned} \tag{2}$$

A number of mathematical subtleties in (2) are worth discussing. First, after introducing the independent observations, our monitoring target $\tilde{Y}_{\tau_i}(0)|\hat{f}_t(\tilde{X}_{\tau_i}), \mathcal{F}_{\tau_i}$ now correctly corresponds to the causal quantity of interest. We also condition on the filtration $\mathcal{F}_{\tau_i}$ so that the monitoring target is the calibration curve of the *realized ML algorithm* with respect to the *realized population.* Without the filtration, we would be monitoring a *random ML algorithm* for a *random population*, which is not of practical interest and mathematically more difficult to analyze. Next, when we apply a standard monitoring procedure to the SOC-only data, the absolute time of the changepoint is technically random. This may seem odd mathematically but still plausible in practice. Finally, the decision to define the distribution shift using $i > \kappa$ rather than $i \geq \kappa$ is critical in the causal setting, as the latter contradicts assumptions such as conditional exchangeability (Section 4.1). Thus, distribution shifts can only occur *before* treatment is assigned, not immediately *after* treatment assignment (but before the outcome).

Finally, an important extension is to monitor for changes in the more detailed conditional distribution by conditioning on some variable subset $X_{t,\mathtt{S}}$ for $\mathtt{S} \subseteq \{1, \cdots, p\}$:

$$\begin{aligned} &H_0 : \tilde{Y}_{\tau_i}(0)|\hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \tilde{X}_{\tau_i,\mathtt{S}} = x_{\mathtt{S}}, \mathcal{F}_{\tau_i} \sim g_{\theta,0}(q, x_{\mathtt{S}}) \\ &H_1 : \exists \delta, \kappa \text{ s.t.} \\ &\quad \tilde{Y}_{\tau_i}(0)|\hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \tilde{X}_{\tau_i,\mathtt{S}} = x_{\mathtt{S}}, \mathcal{F}_{\tau_i} \sim g_{\theta,\delta \mathbb{1}\{i>\kappa\}}(q, x_{\mathtt{S}}) \\ &\quad \forall i = 1, \cdots, mK. \end{aligned} \tag{3}$$

For risk prediction models, this corresponds to monitoring *stratified* calibration curves—a stricter notion of calibration (Van Calster et al., 2016)—as well as stronger notions of algorithmic fairness (Hebert-Johnson et al., 2018). Testing this hypothesis is also more feasible in certain settings, as discussed in the following section.

# 4 IGNORABILITY ASSUMPTIONS

Under what conditions can we ignore performativity and directly apply a standard monitoring procedure to analyze SOC-only data? To establish operating characteristics of a monitoring procedure in the non-performative setting, the standard approach is to factorize $\Pr\left(\hat{f}(X_1), Y_1, \cdots, \hat{f}(X_t), Y_t\right)$ into

$$\prod_{i=1}^{t} \Pr\left(Y_i \mid \hat{f}(X_i); \theta, \delta \mathbb{1}\{i > \kappa\}\right) \prod_{i=1}^{t} \Pr\left(\hat{f}(X_i) \mid \mathcal{F}_i; \eta\right) \tag{4}$$

for some nuisance parameter $\eta$ and analyze the large sample behavior of the (log) first product using the martingale central limit theorem (Zeileis, 2005). So if we prove that the observed data distribution in the presence of performativity admits the same factorization but with the *causal* quantities of interest, operating characteristics transfer from the standard setting to that with performativity. Although (4) follows directly from conditional independencies that hold in the standard setting, it is less clear when a similar factorization would hold in the presence of performativity.

We present two options under which performativity is ignorable: either conditional exchangeability or time-constant selection bias. Throughout, we make the assumption that distinct patients $(X_t, A_t, Y_t(a))$ and $(\tilde{X}_t, \tilde{A}_t, \tilde{Y}_t(a))$ from the same target population are IID, conditional on $\mathcal{F}_t$ (IID+SUTVA). Proofs for all theoretical results are in the Appendix.

## 4.1 Conditional exchangeability

The conditional exchangeability assumption states that treatment assignment is conditionally independent of the potential outcome at each time $t$. The simplest version only conditions on the risk prediction:

$$Y_t(0) \perp A_t \mid \hat{f}_t(X_t), \mathcal{F}_t \quad \forall t = 1, 2, \cdots \tag{5}$$

More complex conditions are discussed Section 4.3. Condition (5) holds, for instance, if treatment decisions only depend on the risk prediction. A more complex example is shown in Figure 1, where both treatment propensities and ML algorithm evolve. Establishing ignorability requires careful reasoning about random stopping times (e.g., when can we replace $t$ with $\tau_i$?). In particular, we prove the following.

**Theorem 1.** *Assuming IID+SUTVA, consistency, and* (5)*, we have for all* $t$ *that*

$$\begin{aligned} &\Pr\left(Y_{\tau_1}, \hat{f}_{\tau_t}(X_{\tau_1}), \cdots, Y_{\tau_t}, \hat{f}_{\tau_t}(X_{\tau_t})\right) \\ \propto &\prod_{i=1}^{t} \Pr\left(\tilde{Y}_{\tau_i}(0) | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}), \mathcal{F}_{\tau_i}; \theta, \delta \mathbb{1}\{i > \kappa\}\right). \end{aligned}$$

## 4.2 Time-constant selection bias

When unmeasured confounders $U_t$ exist, conditional exchangeability no longer holds and conditioning on $A_t = 0$ results in selection bias. Nevertheless, performativity is ignorable if selection bias remains constant over time. That is, suppose there exists some function $h : \mathcal{Q} \mapsto \mathbb{R}$ such that for all $t$ and $q$, we have

$$\begin{aligned} &\mathbb{E}\left[Y_t(0) | \hat{f}_t(X_t) = q, \mathcal{F}_t\right] - \mathbb{E}\left[Y_t | \hat{f}_t(X_t) = q, A_t = 0, \mathcal{F}_t\right] \\ &= h(q). \end{aligned} \tag{6}$$

The idea is that we can recover shifts in the conditional risk in the target population using SOC-only data because this bias cancels out, i.e., $\mathbb{E}\big[Y_t | \hat{f}_t(X_t){=}q, A_t{=}0, \mathcal{F}_t\big] - \mathbb{E}\big[Y_1 | \hat{f}_1(X_1){=}q, A_1{=}0, \mathcal{F}_1\big] = \mathbb{E}\big[Y_t(0) | \hat{f}_t(X_t){=}q, \mathcal{F}_t\big] - \mathbb{E}\big[Y_1(0) | \hat{f}_1(X_1){=}q, \mathcal{F}_1\big]$. Note that time-constant selection bias can hold *even if treatment propensities vary over time*. Example 1 (Figure 1 bottom) provides one such set of conditions, in which treatment decisions follow a determinative causation structure (Hernán et al., 2004; VanderWeele and Robins, 2007, 2009). Under this assumption and again through careful reasoning of random stopping times, we can establish ignorability.

**Theorem 2.** *Suppose model class* $g$ *is defined such that for any two parameter sets* $(\theta, \delta)$ *and* $(\theta', \delta')$*, we have that* $\delta = \delta'$ *as long as the absolute shift in the conditional risk is the same for all* $q \in \mathcal{Q}$*. Assuming IID+SUTVA and* (6)*, we have for all* $t$ *that*

$$\begin{aligned} &\Pr\left(Y_{\tau_1}, \hat{f}_{\tau_1}(X_{\tau_1}), \cdots, Y_{\tau_t}, \hat{f}_{\tau_t}(X_{\tau_t})\right) \\ &\propto \prod_{i=1}^{t} \Pr\left(\tilde{Y}_{\tau_i} | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}), \mathcal{F}_{\tau_i}; \theta', \delta \mathbb{1}\{i > \kappa\}\right). \end{aligned} \tag{7}$$

Unlike Theorem 1, the factorization in (7) is now with respect to $\theta'$, which may not coincide with $\theta$. Nevertheless, $\theta$ is a nuisance parameter and the actual monitoring targets—the causal shift parameter $\delta$ and changepoint $\kappa$—remain unbiased.

## 4.3 Conditioning on additional confounders

In practice, treatment decisions may not only rely on the risk prediction but also other patient factors $X_{t,\mathtt{S}}$. In this case, we can extend the conditional exchangeability assumption by further conditioning on $X_{t,\mathtt{S}}$, i.e.,

$$Y_t(0) \perp A_t \mid \hat{f}_t(X_t), X_{t,\mathtt{S}}, \mathcal{F}_t \quad \forall t = 1, 2, \cdots. \tag{8}$$

Following the same arguments as those for Theorem 1, we have

$$\Pr\left(X_{\tau_1,\mathtt{S}}, Y_{\tau_1}, \hat{f}_{\tau_1}(X_{\tau_1}), \cdots, X_{\tau_t,\mathtt{S}}, Y_{\tau_t}, \hat{f}_{\tau_t}(X_{\tau_t})\right)$$
$$\propto \prod_{i=1}^{t} \Pr\left(\tilde{Y}_{\tau_i}(0) | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}), \tilde{X}_{\tau_i,\mathtt{S}}; \theta, \delta\mathbb{1}\{i > \kappa\}\right).$$

and ignorability holds when testing the extended hypothesis (3). Similarly, we can extend the time-constant selection bias assumption in (6) as well as the results in Theorem 2 by additionally conditioning on $X_{t,\mathtt{S}}$, which let us directly test (3) using a standard monitoring procedure.

## 5 THE SCORE-BASED CUSUM

Having established ignorability, there are two other issues we must account for when monitoring ML-based risk prediction algorithms: (i) predictor sequences can be nonstationary due to changes in performativity over time, and (ii) the algorithm's performance prior to the changepoint is typically estimated, which is an additional source of variability. To address these two issues, we introduce a new score-based CUSUM procedure with DCLs. To motivate the approach, we begin with the setting where the pre-change parameter $\theta$ is known and then extend it to the setting where $\theta$ is unknown. The former setting may occur when the model is deployed on the very population it was trained on. However, the pre-change parameter will likely be unknown when a model is deployed at a different site from what it was trained on.

To simplify notation, we will describe the CUSUM procedures for the standard setting, using $Z_t$ to represent variables to the right of the conditioning bar, e.g., $Z_t = \hat{f}_t(X_t)$ or $Z_t = (\hat{f}_t(X_t), X_{t,\mathtt{S}})$. We can directly apply these to SOC-only data in the presence of performativity as long as either of the aforementioned ignorability conditions holds.

### 5.1 Pre- and post-change model class

Before presenting the monitoring methods, we discuss model classes $g$ that can be used to test for shifts in the conditional distribution of the outcome over time. For the conditional exchangeability assumption, any model class can be used, as long as it is differentiable with respect to $\theta$ and $\delta$. For the time-constant selection bias assumption, the functional form of $g$ is restricted to those that satisfy the constancy assumption.

For concreteness, we present the following two models used in empirical analyses of this paper. The first describes the pre-change distribution and the structural change on the log odds scale using logistic regression, i.e.,

$$\begin{aligned}&\Pr\left(Y_t = 1 \mid Z_t; \theta, \delta\mathbb{1}\{t > \kappa\}\right)\\ &= \frac{1}{1 + \exp\left(-(\theta + \delta\mathbb{1}\{t > \kappa\})^\top Z_t\right)}.\end{aligned} \tag{m.1}$$

The second describes the pre-change distribution on the log odds scale but the structural change on the risk scale using

$$\begin{aligned}&\Pr\left(Y_t = 1 \mid Z_t; \theta, \delta\mathbb{1}\{t > \kappa\}\right)\\ &= \left[\frac{1}{1 + \exp\left(-\theta^\top Z_t\right)} + (\delta\mathbb{1}\{t > \kappa\})^\top Z_t\right]_{[0,1]},\end{aligned} \tag{m.2}$$

where $[x]_{[0,1]} = \min(1, \max(0, x))$. We use either (m.1) or (m.2) in examples where conditional exchangeability holds, and (m.2) in examples where time-constant selection bias holds.

### 5.2 Known pre-change parameter

Suppose the true value of $\theta$, denoted $\theta_0$, is known. For observation $(Z_t, Y_t)$, the score (i.e., gradient of the log likelihood) with respect to $\delta$ at $\delta_0 = 0$ is $\nabla_\delta \log p\left(Y_t \mid Z_t; \theta_0, \delta_0\right)$. Because the conditional mean of the score is zero prior to the changepoint and nonzero after, we monitor for shifts in the average score using a score-based CUSUM. In particular, for candidate changepoint $t'$, the cumulative score up to time $t$ is $\psi_m^{(\text{known})}(t', t) = \sum_{i=t'}^{t} \nabla_\delta \log p\left(Y_i \mid Z_i; \theta_0, \delta_0\right)$. Since the true changepoint time is unknown, the score-based CUSUM with respect to norm $\|\cdot\|$ is defined as

$$C_m^{(\text{known})}(t) = \max_{t'=m+1,\cdots,t} \left\|\psi_m^{(\text{known})}(t', t)\right\|. \tag{9}$$

In our empirical analyses, we use $\|\cdot\|_1$, though one can consider other norms. For instance, $\|\cdot\|_2$ is similar to using Rao's score statistic.

A major benefit of the score-based CUSUM is the computational ease for constructing DCLs. Here we define DCLs $h_m(t)$ recursively using an alpha spending approach. Let $\alpha^{\text{rel}} : [1, K] \mapsto [0, 1]$ be the alpha-spending function, where $\alpha^{\text{rel}}$ is continuous and monotonically non-decreasing. Then $h_m(t)$ is the minimal threshold at which the conditional false alarm rate up to time $t$ matches the prespecified alpha-spending rate under the null, i.e., $\Pr\left(\exists t' \in \{m+1, \cdots, t\} \text{ s.t. } C_m^{(\text{known})}(t') > h_m(t') | Z_1, \cdots, Z_t\right) \leq \alpha^{\text{rel}}(t/m)$. Because the pre-change parameter is known, we can calculate the distribution of the chart statistic under the null by resampling outcomes $Y_t^*$ given $Z_t$. By constructing sufficiently many sequences $\{(Z_t, Y_t^{*(b)}) : t = m+1, \cdots, mK\}$ for $b = 1, \cdots, B$ and computing their corresponding chart statistics, we can construct DCLs with exact Type I error control. Of note, Type I error control holds *without* assuming positivity. Under the alternative, this procedure is consistent as long as the average score after the changepoint is bounded away from zero (see Appendix).

### 5.3 Unknown pre-change parameter

When the pre-change parameter $\theta$ is unknown, we need to estimate its value and adjust the DCLs to reflect this additional source of uncertainty. The key idea is to use a *nonanticipative* chart statistic, in that the score for the $i$-th observation is calculated with respect to an estimate of $\theta$ using only historical data (Lorden and Pollak, 2005). This preserves the martingale structure of the chart statistic even when $\theta$ for the pre-change distribution is continually re-estimated.

To this end, let our estimate $\hat{\theta}_{m,t}$ of $\theta$ at time $t$ be the solution to the estimating equation $\sum_{i=1}^{t} \nabla_\theta \log p\left(Y_i \mid Z_i; \theta, \delta_0\right) = 0$. We define the score-based CUSUM chart statistic

$$C_m^{(\text{plugin})}(t) = \max_{t'=m+1,\cdots,t} \left\| \psi_m^{(\text{plugin})}(t', t) \right\| \qquad (10)$$

where

$$\psi_m^{(\text{plugin})}(t', t) = \sum_{i=t'}^{t} \nabla_\delta \log p\left(Y_i \mid Z_i; \hat{\theta}_{m,i-1}, \delta_0\right). \qquad (11)$$

The score for observation $(Z_i, Y_i)$ uses $\hat{\theta}_{m,i-1}$ rather than $\hat{\theta}_{m,i}$, so (11) is nonanticipative.

To determine the operating characteristics of $C_m^{(\text{plugin})}$ under the null, we require the following assumptions, in addition to Assumptions 5 and 6 in the Appendix.

**Assumption 1.** *Under the null, there is a zero-mean $(p+d)$-dimensional non-degenerate gaussian process $U = (U_\theta, U_\delta)$ such that*

$$\max_{m+1 \le i \le mK} \left\| \frac{1}{\sqrt{m}} \sum_{j=1}^{i} \nabla_{\theta,\delta} \log p(Y_j | Z_j; \theta_0, \delta_0) - U(i/m) \right\| = o_p(1).$$

**Assumption 2.** *Under the null, $\hat{\theta}_{m,i}$ is asymptotically linear with a remainder term that converges uniformly to zero, i.e.,*

$$\max_{m < i \le mK} \sqrt{m} \left\| \hat{\theta}_{m,i} - \theta_0 - \Lambda_m^{-1}(i) \sum_{j=1}^{i} \nabla_\theta \log p(Y_j | Z_j; \theta_0, \delta_0) \right\| = o_p(1)$$

*where* $\Lambda_m(i) = \mathbb{E}\left[ -\sum_{j=1}^{i} \nabla_\theta^2 \log p\left(Y_j \mid Z_j; \theta_0, \delta_0\right) \right]$.

These assumptions hold, for instance, under piecewise local stationarity (Wu and Zhou, 2018; Horváth et al., 2021). We can then prove that under the null, $\psi_m^{(\text{plugin})}(t_1, t_2)$ is well-approximated by $\phi_m(t_1, t_2) = \sum_{i=t_1}^{t_2} \nabla_\delta \log p\left(Y_i \mid Z_i; \theta_0, \delta_0\right) + \sum_{i=t_1}^{t_2} \bar{V}_0\left(\frac{i}{m}\right) \Lambda_0^{-1}\left(\frac{i-1}{m}\right) \sum_{j=1}^{i-1} \nabla_\theta \log p\left(Y_j \mid Z_j; \theta_0, \delta_0\right)$ for $\bar{V}_0$ and $\Lambda_0$ defined in the Appendix, and converges to a Gaussian process, giving us the following result.

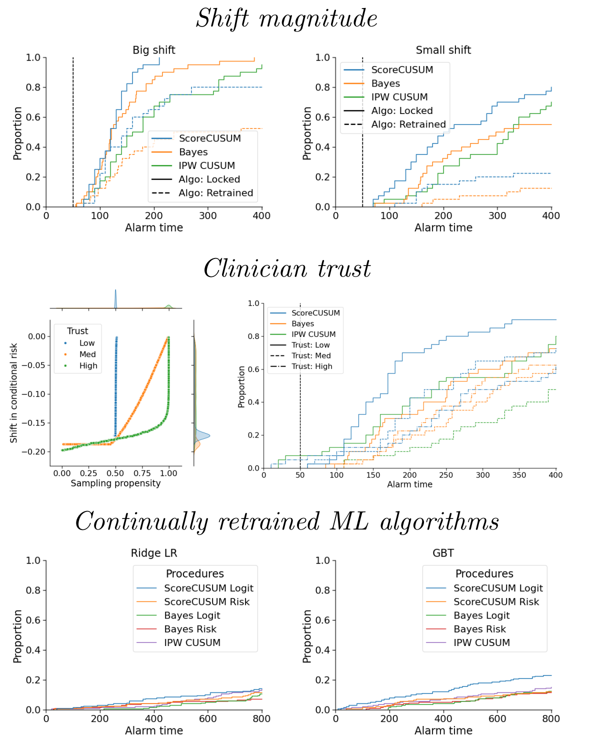


Figure 2: Simulation results, plotting alarm rate at each time point. Dashed vertical line indicates time of calibration decay. Left plot in "Clinician trust" shows how different trust levels sample subgroups with varying amounts of calibration decay at different rates.

**Theorem 3.** *Suppose the null hypothesis is true and that Assumptions 3, 4, 5 and 6. Then*

$$\max_{m < t_1, t_2 \le mK} \frac{1}{\sqrt{m}} \left\| \psi_m^{(plugin)}(t_1, t_2) - \phi_m(t_1, t_2) \right\| = o_p(1).$$

*and* $\left\{ (\nu_1, \nu_2) \mapsto \frac{1}{\sqrt{m}} \psi_m^{(plugin)}(\lfloor m\nu_1 \rfloor, \lfloor m\nu_2 \rfloor) \right\}_{(\nu_1,\nu_2) \in \Delta} \Rightarrow$ $\left\{ (\nu_1,\nu_2) \mapsto U_\delta(\nu_2) - U_\delta(\nu_1) + \int_{\nu_1}^{\nu_2} \bar{V}_0(v) \Lambda_0^{-1}(v) U_\theta(v) dv \right\}_{(\nu_1,\nu_2) \in \Delta}$ *where* $\Delta = \{(\nu_1, \nu_2) : \nu_1 < \nu_2, \nu_1 \in [1, K], \nu_2 \in [1, K]\}$.

In addition, Theorem 4 in the Appendix proves that the procedure is consistent if analogous assumptions hold under the alternative.

Given Theorem 3, we can approximate the distribution of the chart statistic under the null by calculating $C_m^{(\text{approx})}(t) = \max_{t'=m+1,\cdots,t} \|\phi_m(t', t)\|$ for resampled sequences $\{(Z_t, Y_t^{*(b)})\}$. One caveat is that this technically requires sampling from the true $\theta$, but we only have an estimate for its value. As such, we perform the parametric bootstrap and resample outcomes using the most recent estimate. See the Appendix for pseudocode and additional implementation details.

# 6 SIMULATION STUDIES

We now explore various applications of the score-based CUSUM and the impact of various factors on its operating characteristics through a series of simulation studies. The following sections investigate how the magnitude of performance decay, clinician trust, and continual model retraining affect alarm rates. All data in the main manuscript is simulated under the conditional exchangeability assumption. The Appendix presents several other experiments, including simulations where the data is generated under the time-constant selection bias assumption, verification of Type I error control under the null, gradual as opposed to sudden decay in model calibration, sensitivity analysis to assumptions, and a comparison to procedures that inappropriately ignore performativity. The Appendix also describes simulation settings in more detail, including how the data and clinician trust were simulated. The nominal false alarm rate for the score-based CUSUM is set to $\alpha = 0.1$ in all simulations.

To compare against methods that monitor marginal performance measures, we implement the CUSUM based on (Steiner and MacKay, 2001; Sun et al., 2014) to monitor the Brier score, a *marginal* measure of model calibration (`IPW CUSUM`). Nevertheless, this procedure requires knowing the true propensities, which we plugin; thus, one should consider this an oracle procedure. Because monitoring marginal performance is unsuitable in certain settings, we exclude it from particular simulations. As previously mentioned, there are few methods for monitoring conditional performance measures that adequately control false alarm rates in our setting. The closest comparator to this work is Bayesian monitoring (`Bayesian`) (Adams and MacKay, 2007), which we implement using `Stan` (Carpenter et al., 2017). To make the procedure comparable, we have tuned its control limits to match the frequentist false alarm rate. We note that Bayesian monitoring has some practical limitations: posterior inference is much more computationally expensive than the CUSUM, posterior inference is challenging for complex forms of distribution shift, and the procedure is sensitive to the choice of the prior.

## 6.1 Shift magnitude

We first assess how the magnitude of calibration decay affects detection delay. We generated the outcome using a logistic regression (LR) model and simulated big versus little shifts in its conditional distribution by shifting the coefficients of the oracle LR model. The fitted model was a LR model as well. In both cases, the score-based CUSUM had the highest power. Unsurprisingly, statistical power for detecting calibration decay is higher for larger shifts (Fig 2 top).

Next, we continually retrained the ML model using an exponentially weighted average forecaster (EWAF) that adapts to adversarial data shifts (Cesa-Bianchi and Lugosi, 2006; Feng, 2021). Model retraining generally increases the time to an alarm, though it depends on the reaction time of the retraining procedure. After a small shift, the EWAF recalibrated the model quickly and significantly extended the total lifetime of the ML algorithm. After a big shift, the EWAF was slow to recalibrate and did not significantly delay detection by the monitoring procedure. In summary, the interaction between a model monitoring and updating procedure can be viewed as a competition.

## 6.2 Clinician trust

We now investigate how clinician trust can interfere with our ability to detect performance decay. We simulate three levels of clinician trust (low, medium, and high) by defining treatment propensities using LR, with the predicted logit as its sole input variable and coefficients 0.01, 1, and 6, respectively. The simulated shift in the true conditional risk is largest among patients with the highest predicted risk.

The score-based CUSUM had substantially higher power than the other methods in all but the high trust setting, in which it had similar performance as Bayesian monitoring (Fig 2 middle). The IPW-based CUSUM did poorly, with zero power in the high trust setting due to near-violations of the positivity assumption. In general, increasing clinician trust increases detection delay, which supports recent suggestions to educate healthcare providers on the appropriate use of ML and warn against over-reliance (Finlayson et al., 2021; Harris et al., 2022). Nevertheless, clinician trust does not always interfere with model monitoring, as illustrated in the Appendix.

## 6.3 Retraining in stationary settings

Ideally, continual retraining of an ML algorithm will steadily improve model discrimination and maintain calibration. Although we generally expect this behavior in IID data settings, there are no such guarantees for black-box algorithms. To this end, we test if one can continually retrain ridge-penalized logistic regression (LR) and gradient-boosted tree (GBT) models while keeping their alarm rates close to the nominal rate. Because a basic GBT model can be poorly calibrated, we recalibrated model updates using Platt scaling (Platt, 1999). We also considered monitoring on the logit versus risk scale by defining the data model using (m.1) versus (m.2), respectively.

For both ML algorithms, alarm rates of the monitoring

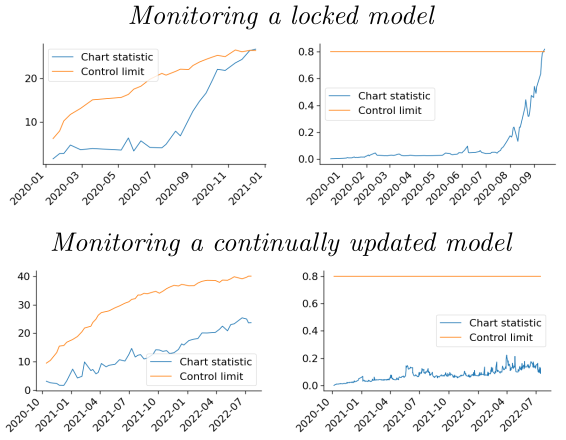


Figure 3: Control charts for monitoring a locked (top) and continually updated (bottom) ML-based risk prediction model for PONV using the score-based CUSUM (left) and Bayesian inference (right).

procedures did not exceed 20% (Fig 2 bottom). The score-based CUSUM achieved the nominal rate of 10% when monitoring on the risk scale, and was more sensitive when monitoring on the logit scale. The Appendix shows how this behavior can be explained by plotting calibration curves on the logit versus risk scale. As such, we suggest using the risk scale in practice, which is typically more relevant to end-users of the model.

## 7 MONITORING A PONV RISK CALCULATOR

Postoperative nausea and vomiting (PONV) is a common side effect of anesthesia, which has prompted the development of risk calculators to guide the use of antiemetic medications (Apfel et al., 2012). Here, we simulate monitoring an ML-based PONV risk model using data from the UCSF MPOG registry ($n = 2434$).

Using data from January 2018 to May 2019, we trained a random forest (RF) to predict the risk of PONV based on preoperative variables. We locked the model, used the first 200 patients as noncontamination data, and started monitoring in mid-December 2019. We supposed conditional exchangeability (5) holds. Control limits were set so $\alpha = 0.2$. We did not run the IPW-based CUSUM as treatment propensities were unknown. The score-based CUSUM and Bayesian monitoring fired alarms in late 2020 (Fig 3). The detection of calibration decay is not unexpected and may be explained by many causes: (i) the anesthesia department had implemented changes in antiemetic medication administration; (ii) there was a shift in the type of patients who received surgery during the COVID-19 pandemic; and (iii) exposure to the SARS-Cov2 virus affected the overall health of many patients.

When we instead continually retrained the RF and started monitoring in October 2020, no calibration decay was detected by either procedure and the AUC of the continually retrained RF increased from 0.54 to 0.59 (See Fig 11 in the Appendix). This real-world example illustrates how wrapping online learning methods within a monitoring framework can ensure model reliability while allowing steady improvement in model discrimination.

## 8 DISCUSSION

Although performativity can complicate monitoring of ML algorithms, we have shown that performativity is ignorable when monitoring conditional performance measures under either the assumption of conditional exchangeability or time-constant selection bias. Because performativity may evolve over time, we introduced a new score-based CUSUM procedure. Future work includes integrating ideas in this work with other types of monitoring procedures including those for open-end settings (e.g., conformal inference (Volkhonskiy et al., 2017) and anytime inference (Shekhar and Ramdas, 2023)); increasing robustness to model misspecification and violations to assumptions; and incorporating other data sources to reduce the impact of clinician trust on our ability to monitor.

## ACKNOWLEDGMENTS

This work was greatly improved by helpful suggestions and feedback from Alan Hubbard, Fan Xia, and Si Wen. We are grateful to Daniel Lazzareschi for sharing the UCSF MPOG data.

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 financial assistance award Center of Excellence in Regulatory Science and Innovation grant to University of California, San Francisco (UCSF) and Stanford University, U01FD005978. 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.

### References

Roy Adams, Katharine E Henry, Anirudh Sridharan, and Hossein et al. Soleimani. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat. Med.*, 28(7):1455–1460, July 2022. URL `http://dx.doi.org/10.1038/s41591-022-01894-0`.

Ryan Prescott Adams and David J C MacKay.

Bayesian online changepoint detection. October 2007. URL http://arxiv.org/abs/0710.3742.

C C Apfel, F M Heidrich, S Jukar-Rao, L Jalota, C Hornuss, R P Whelan, K Zhang, and O S Cakmakkaya. Evidence-based analysis of risk factors for postoperative nausea and vomiting. *Br. J. Anaesth.*, 109(5):742–753, November 2012. URL http://dx.doi.org/10.1093/bja/aes276.

P K Bhattacharya. Some aspects of change-point analysis. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, Lecture notes-monograph series, pages 28–56. Institute of Mathematical Statistics, Hayward, CA, 1994. URL https://projecteuclid.org/ebook/download?urlId=10.1214/lnms/1215463112&isFullBook=false.

Albert Bifet and Ricard Gavaldà. Learning from Time-Changing data with adaptive windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 443–448. Society for Industrial and Applied Mathematics, April 2007.

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.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *J. Stat. Softw.*, 76(1), 2017. URL https://www.osti.gov/biblio/1430202.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, March 2006. URL https://play.google.com/store/books/details?id=zDnRBlazhfYC.

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.

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.

Sharon E Davis, Thomas A Lasko, Guanhua Chen, Edward D Siew, and Michael E Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Inform. Assoc.*, 24(6):1052–1061, November 2017. URL http://dx.doi.org/10.1093/jamia/ocx030.

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.

Anne R Driscoll, William H Woodall, and Changliang Zou. Use of conditional false alarm metric in statistical process monitoring. In *Frontiers in Statistical Quality Control 13*, pages 3–12. Springer International Publishing, 2021. URL http://dx.doi.org/10.1007/978-3-030-67856-2_1.

Paul Fearnhead and Zhen Liu. On-line inference for multiple changepoint problems. *J. R. Stat. Soc. Series B Stat. Methodol.*, 69(4):589–605, September 2007.

Jean Feng. Learning to safely approve updates to machine learning algorithms. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, pages 164–173, New York, NY, USA, April 2021. Association for Computing Machinery. URL https://doi.org/10.1145/3450439.3451864.

Jean Feng, Scott Emerson, and Noah Simon. Approval policies for modifications to machine learning-based software as a medical device: a study of bio-creep. *Biometrics*, 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, Adarsh Subbaswamy, Alexej Gossmann, Harvineet Singh, Berkman Sahiner, Mi-Ok Kim, Gene Pennello, Nicholas Petrick, Romain Pirracchio, and Fan Xia. Towards a Post-Market monitoring framework for machine learning-based medical devices: A case study. *Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems*, November 2023.

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.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):1–37, March 2014.

Edit Gombay. Sequential Change-Point detection and estimation. *Seq. Anal.*, 22(3):203–222, January 2003.

Edit Gombay. Editor's special invited paper: On the efficient score vector in sequential monitoring. *Sequential Analysis*, 36(4):435–466, October 2017. URL https://doi.org/10.1080/07474946.2017.1394728.

Steve Harris, Tim Bonnici, Thomas Keen, Watjana Lilaonitkul, Mark J White, and Nel Swanepoel. Clinical deployment environments: Five pillars of translational machine learning for health. *Frontiers in Digital Health*, 4, 2022. URL https://www.frontiersin.org/articles/10.3389/fdgth.2022.939292.

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.

Katharine E Henry, Roy Adams, and Cassandra et al. Parent. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nat. Med.*, 28(7):1447–1454, July 2022. URL http://dx.doi.org/10.1038/s41591-022-01895-z.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, September 2004. URL http://dx.doi.org/10.1097/01.ede.0000135174.63482.43.

Graeme L Hickey, Stuart W Grant, Gavin J Murphy, Moninder Bhabra, Domenico Pagano, Katherine McAllister, Iain Buchan, and Ben Bridgewater. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardiothorac. Surg.*, 43(6): 1146–1152, June 2013. URL http://dx.doi.org/10.1093/ejcts/ezs584.

Lajos Horváth, Curtis Miller, and Gregory Rice. Detecting early or late changes in linear models with heteroscedastic errors. *Scand. Stat. Theory Appl.*, 48(2):577–609, June 2021. URL https://onlinelibrary.wiley.com/doi/10.1111/sjos.12507.

M G Kahn, T C Bailey, S A Steib, V J Fraser, and W C Dunagan. Statistical process control methods for expert system performance monitoring. *J. Am. Med. Inform. Assoc.*, 3(4):258–269, July 1996. URL http://dx.doi.org/10.1136/jamia.1996.96413133.

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.

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.

Lingling Li, Martin Kulldorff, Jennifer C Nelson, and Andrea J Cook. A propensity Score-Enhanced sequential analytic method for comparative drug safety surveillance. *Statistics in Biosciences*, 2011. URL https://link.springer.com/content/pdf/10.1007/s12561-011-9034-5.pdf.

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.

Lang Liu, Joseph Salmon, and Zaid Harchaoui. Score-Based change detection for Gradient-Based learning machines. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4990–4994, June 2021. URL http://dx.doi.org/10.1109/ICASSP39728.2021.9414085.

Gary Lorden and Moshe Pollak. Nonanticipating estimation applied to sequential analysis and changepoint detection. *Annals of Statistics*, 33(3):1422–1454, June 2005.

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

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.

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.

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.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10 (3):61–74, 1999.

T S Richardson and J M Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences*, 2013. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.294.7647&rep=rep1&type=pdf.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, December 1976. URL https://academic-oup-com.ucsf.idm.oclc.org/biomet/article-pdf/63/3/581/756166/63-3-581.pdf.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential changepoint detection via backward confidence sequences. *International Conference on Machine Learning*, June 2023.

A N Shiryaev. On optimum methods in quickest detection problems. *Theory Probab. Appl.*, 8(1):22–46, January 1963. URL https://doi.org/10.1137/1108002.

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 2001.

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.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press, August 2014. URL https://play.google.com/store/books/details?id=zhsbBAAAQBAJ.

U.S. Food and Drug Administration and Health Canada. Good machine learning practice for medical device development, October 2021.

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.

Tyler J VanderWeele and James M Robins. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am. J. Epidemiol.*, 166(9):1096–1104, November 2007. URL http://dx.doi.org/10.1093/aje/kwm179.

Tyler J VanderWeele and James M Robins. Minimal sufficient causation and directed acyclic graphs. *Annals of Statistics*, 37(3):1437–1465, June 2009.

Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alexander Gammerman, and Vladimir Vovk. Inductive conformal martingales for Change-Point detection. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos, editors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 132–153. PMLR, 2017.

Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H Kennedy, and Aaditya Ramdas. Doubly robust confidence sequences for sequential causal inference. March 2021. URL http://arxiv.org/abs/2103.06476.

Mike West. Bayesian model monitoring. *J. R. Stat. Soc.*, 48(1):70–78, September 1986. URL https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1986.tb01391.x.

Suya Wu, Enmao Diao, Taposh Banerjee, Jie Ding, and Vahid Tarokh. Score-based quickest change detection for unnormalized models. *International Conference on Artificial Intelligence and Statistics*, 2023.

Weichi Wu and Zhou Zhou. Gradient-based structural change detection for nonstationary time series M-estimation. *Ann. Stat.*, 46(3):1197–1224, 2018. URL https://www.jstor.org/stable/26542822.

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.

Achim Zeileis and Kurt Hornik. Generalized m-fluctuation tests for parameter instability. *Stat. Neerl.*, 61(4):488–508, November 2007. URL `https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2007.00371.x`.

Xiang Zhang and William H Woodall. Dynamic probability control limits for risk-adjusted bernoulli CUSUM charts. *Stat. Med.*, 34(25):3336–3348, November 2015.

## CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No – the methods were run using the CPU of an ordinary laptop]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not applicable – no new assets]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# A Ignorability conditions

## A.1 Conditional exchangeability

*Proof of Theorem 1.* By chain rule, we have

$$\begin{aligned}&\Pr\left(Y_{\tau_1} = y_1, \hat{f}_{\tau_t}(X_{\tau_1}) = q_1, \cdots, Y_{\tau_t} = y_t, \hat{f}_{\tau_t}(X_{\tau_t}) = q_t\right)\\ =&\prod_{i=1}^{t} \Pr\left(Y_{\tau_i} = y_i | \hat{f}_{\tau_i}(X_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) \Pr\left(\hat{f}_{\tau_i}(X_{\tau_i}) = q_i | \mathcal{F}_{\tau_i}, \eta\right).\end{aligned} \tag{12}$$

So it is sufficient to prove that

$$\Pr\left(Y_{\tau_i} = y_i | \hat{f}_{\tau_i}(X_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) = \Pr\left(\tilde{Y}_{\tau_i}(0) = y_i | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}; \theta, \delta\mathbb{1}\{i > \kappa\}\right), \tag{13}$$

in that the conditional distribution on the right hand side is distributed per model $g$ with parameters $(\theta, \delta\mathbb{1}\{i > \kappa\})$. To see this, we have

$$\Pr\left(Y_{\tau_i} = y_i | \hat{f}_{\tau_i}(X_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) \tag{14}$$

$$= \sum_{t'=\tau_{i-1}}^{\infty} \Pr\left(Y_{\tau_i} = y_i | \hat{f}_{\tau_i}(X_{\tau_i}) = q_i, \tau_i = t', \mathcal{F}_{\tau_i}\right) \Pr\left(\tau_i = t' | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) \tag{15}$$

$$= \sum_{t'=\tau_{i-1}}^{\infty} \Pr\left(Y_{\tau_i}(0) = y_i | \hat{f}_{\tau_i}(X_{\tau_i}) = q_i, \tau_i = t', \mathcal{F}_{\tau_i}\right) \Pr\left(\tau_i = t' | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) \tag{16}$$

$$= \sum_{t'=\tau_{i-1}}^{\infty} \Pr\left(\tilde{Y}_{\tau_i}(0) = y_i | \tilde{A}_{\tau_i} = 0, \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \tau_i = t', \mathcal{F}_{\tau_i}\right) \Pr\left(\tau_i = t' | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) \tag{17}$$

$$= \sum_{t'=\tau_{i-1}}^{\infty} \Pr\left(\tilde{Y}_{\tau_i}(0) = y_i | \tau_i = t', \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) \Pr\left(\tau_i = t' | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right), \tag{18}$$

where (16) follows by consistency, (17) follows by the IID+SUTVA assumption, and (18) follows by (5). Finally, $\Pr(\tilde{Y}_{\tau_i}(0) = y_i | \tau_i = t', \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i})$ is distributed per model $g$ with parameters $(\theta, \delta\mathbb{1}\{i > \kappa\})$ for all $t'$, giving us our desired result. □

## A.2 Time-constant selection bias

*Proof of Theorem 2.* Under SUTVA and assumption (6), the pre-change distribution of the observational SOC-only data is described by

$$\mathbb{E}\left[\tilde{Y}_{\tau_i} \mid \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \tilde{A}_{\tau_i} = 0, \mathcal{F}_{\tau_i}\right] = \mathbb{E}\left[\tilde{Y}_{\tau_i}(0) \mid \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \mathcal{F}_{\tau_i}\right] + h(q). \tag{19}$$

Thus the pre-change distribution of the observational SOC-only data is also constant prior to the changepoint. Assuming a sufficiently large model class $g$, there exists some $\theta'$ such that

$$\tilde{Y}_{\tau_i} \mid \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \tilde{A}_{\tau_i} = 0, \mathcal{F}_{\tau_i} \sim g_{\theta',0}(q) \qquad \forall i < \kappa. \tag{20}$$

Moreover, under assumption (6), the shift in the conditional risk for the observed data distribution

$$\mathbb{E}\left[\tilde{Y}_t \mid \hat{f}_t(\tilde{X}_t) = q, \tilde{A}_t = 0, \mathcal{F}_t\right] - \mathbb{E}\left[\tilde{Y}_1 \mid \hat{f}_1(\tilde{X}_1) = q, \tilde{A}_1 = 0, \mathcal{F}_1\right] \tag{21}$$

is equal to the shift in the conditional risk for the causal data distribution

$$\mathbb{E}\left[\tilde{Y}_t(0) \mid \hat{f}_t(\tilde{X}_t) = q, \mathcal{F}_t\right] - \mathbb{E}\left[\tilde{Y}_1(0) \mid \hat{f}_1(\tilde{X}_1) = q, \mathcal{F}_1\right] \tag{22}$$

for all $q \in \mathcal{Q}$. So we have that

$$\tilde{Y}_{\tau_i} \mid \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \tilde{A}_{\tau_i} = 0, \mathcal{F}_{\tau_i} \sim g_{\theta', \delta \mathbb{1}\{i > \kappa\}}(q) \qquad \forall i = 1, \cdots, mK, \tag{23}$$

where $\delta$ and $\kappa$ are the same parameters as those for the causal distribution $\tilde{Y}_{\tau_i}(0) \mid \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q, \mathcal{F}_{\tau_i}$. Thus

$$\Pr\left(Y_{\tau_1} = y_1, \hat{f}_{\tau_t}(X_{\tau_1}) = q_1, \cdots, Y_{\tau_t} = y_t, \hat{f}_{\tau_t}(X_{\tau_t}) = q_t\right) \tag{24}$$

$$\propto \prod_{i=1}^{t} \Pr\left(Y_{\tau_i} = y_i | \hat{f}_{\tau_i}(X_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}\right) \tag{25}$$

$$\propto \prod_{i=1}^{t} \Pr\left(\tilde{Y}_{\tau_i} = y_i | \hat{f}_{\tau_i}(\tilde{X}_{\tau_i}) = q_i, \mathcal{F}_{\tau_i}; \theta', \delta \mathbb{1}\{i > \kappa\}\right) \tag{26}$$

where (26) follows the IID+SUTVA assumption and (23). □

### A.3 Example satisfying the time-constant selection bias assumption

**Example 1.** *Consider the bottom single world intervention graph (SWIG) in Figure 1, where $U_t$ is an unmeasured confounder. Suppose a determinative cause structure, where $A'_t = 1$ implies that $A_t = 1$. Suppose the distribution $(U_t, A'_t)$ is constant over time; as such, we will drop the time indices when denoting their marginal and conditional distributions. For the conditional risk model, assume there is no-additive interaction with respect to time and $U_t$, i.e.*

$$\mathbb{E}\left[Y_t(0) \mid X_t = x, U_t = u, \mathcal{F}_t\right] = g_0(x, u; \theta_0) + g_1(x; \delta)\mathbb{1}\{t > \kappa\}.$$

*First, we note that by d-separation, the assumption that $(X_t, A_t, Y_t(a))$ and $(\tilde{X}_t, \tilde{A}_t, \tilde{Y}_t(a))$ are IID conditional on $\mathcal{F}_t$ holds. Under the determinative cause assumption, we have that $A_t = 0$ implies $A_{t'} = 0$, so $U_t \perp X_t \mid f_t(X_t) = q, A_t = 0$. Then for all times $t$ and $q \in \mathcal{Q}$, we have that*

$$\mathbb{E}\left[Y_t(0) \mid \hat{f}_t(X_t) = q, \mathcal{F}_t\right] - \mathbb{E}\left[Y_t \mid \hat{f}_t(X_t) = q, A_t = 0, \mathcal{F}_t\right] \tag{27}$$

$$= \int \left(g_0(x_t, u; \theta_0) + g_1(x_t; \delta_1)\mathbb{1}\{t > \kappa\}\right) p(x_t \mid \hat{f}_t(x_t) = q, \mathcal{F}_t)\left[p(u) - p(u \mid a' = 0)\right] dx_t du \tag{28}$$

$$= \int g_0(x_t, u; \theta_0) p(x_t \mid \hat{f}_t(x_t) = q, \mathcal{F}_t)\left[p(u) - p(u \mid a' = 0)\right] dx_t du. \tag{29}$$

*There are various conditions under which (29) is time-constant. One option is that $g_0(x, u; \theta_0)$ is additive, in that $g_0(x, u; \theta_0) = g_{0,0}(x; \theta_0) + g_{0,1}(u; \theta_0)$. Alternatively, time-constancy holds if the ML algorithm is locked ($\hat{f}_t = \hat{f}$ for all $t$) and the distribution of $X_t$ does not vary over time. Crucially, note that the propensity model is still allowed to vary over time.*

## B The Score-based CUSUM

### B.1 Example models for the conditional distribution

For the theoretical analyses, we suppose the parametric model for $Y_t | Z_t$ is correctly specified. Here we provide some example models, which we use in the empirical analyses. The first model describes both the pre-change distribution and the structural change on the log odds scale using logistic regression, i.e.

$$\Pr\left(Y_t = 1 \mid Z_t; \theta, \delta \mathbb{1}\{t > \kappa\}\right) = \frac{1}{1 + \exp\left(-(\theta + \delta \mathbb{1}\{t > \kappa\})^\top Z_t\right)}. \tag{m.1}$$

The second model describes the pre-change distribution on the log odds scale but the structural change on the risk scale using

$$\Pr\left(Y_t = 1 \mid Z_t; \theta, \delta \mathbb{1}\{t > \kappa\}\right) = \left[\frac{1}{1 + \exp\left(-\theta^\top Z_t\right)} + (\delta \mathbb{1}\{t > \kappa\})^\top Z_t\right]_{[0,1]}, \tag{m.2}$$

where $[x]_{[0,1]} = \min(1, \max(0, x))$. When conditional exchangeability holds, we can monitor for structural change on any scale and use either (m.1) or (m.2). If only time-constant selection bias holds, we use (m.2) to model shifts on the risk scale.

### B.2 Known pre-change parameter

Consistency of the monitoring procedure holds when there is some $c > 0$ and $K' \in (\kappa^{\text{rel}}, K]$ such that

$$\lim_{m\to\infty} \left\| \frac{1}{m} \sum_{t=\lfloor m\kappa^{\text{rel}}\rfloor}^{\lfloor mK'\rfloor} \mathbb{E}\left[\nabla_\delta \log p\left(Y_t|Z_t;\theta_0,\delta\right)|_{\delta=0}\right] \right\| \geq c, \tag{30}$$

and the martingale

$$\sum_{t=\lfloor m\kappa^{\text{rel}}\rfloor}^{\lfloor mK'\rfloor} \nabla_\delta \log p\left(Y_t|Z_t;\theta_0,\delta\right)|_{\delta=0} - \mathbb{E}\left[\nabla_\delta \log p\left(Y_t^*|Z_t;\theta_0,\delta\right)|_{\delta=0} \mid Z_t\right] \tag{31}$$

satisfies the martingale central limit theorem.

### B.3 Unknown pre-change parameter: Assumptions and proofs

Let $\mathcal{Z}$ and $\mathcal{Y}$ be the domains for the predictors and outcomes. Let $(\theta_0, \delta_0)$ and $(\theta_0, \delta_1)$ parameterize the pre-change and post-change distribution, where $\delta_0 = 0$ and $\delta_1 \neq 0$. We assume that $p(y|z;\theta,\delta)$ is 3-times continuously differentiable with respect to $(\theta, \delta)$. For convenience, denote

$$\Lambda_m(i) = \mathbb{E}\left[-\sum_{j=1}^{i} \nabla_\theta^2 \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right]$$
$$V_m(i) = \mathbb{E}\left[\nabla_\theta\nabla_\delta \log p\left(Y_i \mid Z_i;\theta_0,\delta_0\right)|Z_i\right].$$

We use the symbol $\Rightarrow$ to mean weak convergence in the space under consideration. Throughout, we will use $c$ (sometimes with subscripts) to denote constants, which may vary across contexts. When we write $Z_m \leq_p c$, this means that asymptotically as $m \to \infty$, the random variable $Z_m$ is bounded by some constant $c$ with probability 1.

#### B.3.1 Asymptotics under the null

Here we prove asymptotic convergence of the chart statistic under the null. We restate the assumptions from the main manuscript for clarity.

**Assumption 3.** *Under the null, there is a zero-mean $(p+d)$-dimensional non-degenerate gaussian process $U$ such that*

$$\max_{m+1\leq i\leq mK} \left\| \left[ \frac{1}{\sqrt{m}} \sum_{j=1}^{i} \begin{pmatrix} \nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right) \\ \nabla_\delta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right) \end{pmatrix} \right] - \begin{pmatrix} U_\theta(i/m) \\ U_\delta(i/m) \end{pmatrix} \right\| = o_p\left(1\right).$$

*where $U_\theta$ and $U_\delta$ are p- and d-dimensional, respectively.*

**Assumption 4.** *Under the null, $\hat{\theta}_{m,i}$ is asymptotically linear with a remainder term that converges uniformly to zero, i.e.*

$$\max_{m<i\leq mK} \sqrt{m} \left\| \left(\hat{\theta}_{m,i} - \theta_0\right) - \mathbb{E}\left[-\sum_{j=1}^{i} \nabla_\theta^2 \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right]^{-1} \sum_{j=1}^{i} \nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right) \right\| = o_p\left(1\right).$$

In addition, we require the following assumptions.

**Assumption 5.** *Under the null, there exist functions $\Lambda_0 : [1,K] \mapsto \mathbb{R}^{p\times p}$ and $\bar{V}_0 : [1,K] \mapsto \mathbb{R}^{d\times p}$ such that*

$$\max_{m<i\leq mK} \left\| \Lambda_0^{-1}\left(\frac{i}{m}\right) - m\mathbb{E}\left[-\sum_{j=1}^{i} \nabla_\theta^2 \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right]^{-1} \right\| = o_p(1)$$
$$\bar{V}_0(t) = \mathbb{E}\left[\nabla_\theta\nabla_\delta \log p\left(Y_{\lfloor mt\rfloor} \mid Z_{\lfloor mt\rfloor};\theta_0,\delta_0\right)\right] \quad \forall t \in [1,K].$$

**Assumption 6.** *There exist constants $c_1, c_2 > 0$ and some neighborhood $B(\theta_0, r)$ centered at $\theta_0$ with radius $r > 0$ such that*

$$\sup_{\tilde{\theta} \in B(\theta_0, r)} \sup_{y \in \mathcal{Y}, z \in \mathcal{Z}} \left\| \nabla_\theta^2 \nabla_\delta \log p \left( y \mid z; \tilde{\theta}, \delta_0 \right) \right\|_\infty \leq c_1 \tag{32}$$

$$\sup_{y \in \mathcal{Y}, z \in \mathcal{Z}} \left\| \nabla_\theta \nabla_\delta \log p \left( y \mid z; \theta_0, \delta_0 \right) \right\|_\infty \leq_p c_2. \tag{33}$$

To prove Theorem 3, we first prove the following lemma.

**Lemma 1.** *Suppose Assumptions 3, 4, 5, and 6 hold. Define*

$$\tilde{\phi}_m(t_1, t_2) = \sum_{i=t_1}^{t_2} \nabla_\delta \log p \left( Y_i \mid Z_i; \theta_0, \delta_0 \right) + \sum_{i=t_1}^{t_2} V_m(i) \Lambda_m^{-1}(i-1) \sum_{j=1}^{i-1} \nabla_\theta \log p \left( Y_j \mid Z_j; \theta_0, \delta_0 \right).$$

*Under the null, we have*

$$\max_{m < t_1, t_2 \leq mK} \frac{1}{\sqrt{m}} \left\| \psi_m^{(plugin)}(t_1, t_2) - \tilde{\phi}_m(t_1, t_2) \right\|_2 = o_p(1).$$

*Proof.* Consider the decomposition

$$\frac{1}{\sqrt{m}} \left( \psi_m^{(\text{plugin})}(t_1, t_2) - \tilde{\phi}_m(t_1, t_2) \right) = R_m^{(1)}(t_1, t_2) + R_m^{(2)}(t_1, t_2) + R_m^{(3)}(t_1, t_2)$$

where

$$\begin{aligned} R_m^{(1)}(t_1, t_2) = & \frac{1}{\sqrt{m}} \sum_{i=t_1}^{t_2} \Bigg[ \nabla_\delta \log p \left( Y_i \mid Z_i; \hat{\theta}_{m,i-1}, \delta_0 \right) - \nabla_\delta \log p \left( Y_i \mid Z_i; \theta_0, \delta_0 \right) \\ & - \nabla_\theta \nabla_\delta \log p \left( Y_i \mid Z_i; \theta_0, \delta_0 \right) \left( \hat{\theta}_{m,i-1} - \theta_0 \right) \Bigg] \\ R_m^{(2)}(t_1, t_2) = & \frac{1}{\sqrt{m}} \sum_{i=t_1}^{t_2} V_m(i) \left[ \hat{\theta}_{m,i-1} - \theta_0 - \Lambda_m^{-1}(i-1) \sum_{j=1}^{i-1} \nabla_\theta \log p \left( Y_j \mid Z_j; \theta_0, \delta_0 \right) \right] \\ R_m^{(3)}(t_1, t_2) = & \frac{1}{\sqrt{m}} \sum_{i=t_1}^{t_2} \left[ \nabla_\theta \nabla_\delta \log p \left( Y_i \mid Z_i; \theta_0, \delta_0 \right) - V_m(i) \right] \left( \hat{\theta}_{m,i-1} - \theta_0 \right). \end{aligned}$$

We will prove that each term in this decomposition is negligible, i.e.

$$\max_{m < t_1 < t_2 \leq mK} \left\| R_m^{(j)}(t_1, t_2) \right\|_2 = o_p(1) \quad \forall j = 1, 2, 3. \tag{34}$$

**First remainder term.** For any $\epsilon > 0$, we have that

$$\begin{aligned} \Pr \left( \max_{m < t_1 < t_2 \leq mK} \left\| R_m^{(1)}(t_1, t_2) \right\|_2 > \epsilon \right) \leq & \Pr \left( \max_{m < i \leq mK} \left\| \hat{\theta}_{m,i} - \theta_0 \right\|_2 > c \frac{\log m}{\sqrt{m}} \right) \\ & + \Pr \left( \max_{m < t_1 < t_2 \leq mK} \left\| R_m^{(1)}(t_1, t_2) \right\|_2 > \epsilon, \max_{m < i \leq mK} \left\| \hat{\theta}_{m,i} - \theta_0 \right\|_2 \leq c \frac{\log m}{\sqrt{m}} \right). \end{aligned} \tag{35}$$

The first summand on the RHS of (35) goes to zero because Assumptions 3, 4, and 5 imply

$$\max_{m < i \leq mK} \left\| \hat{\theta}_{m,i} - \theta_0 \right\|_2 = o_p \left( \frac{\log m}{\sqrt{m}} \right). \tag{36}$$

To bound the second summand, we have by Taylor's theorem and Assumption 6 that for sufficiently large $m$

$$\begin{aligned}\left\|R_m^{(1)}(t_1,t_2)\right\|_2 &\leq \frac{1}{2\sqrt{m}}\sum_{i=t_1}^{t_2}\left(\max_{\tilde{\theta}\in B(\theta_0,r)}\max_{z\in\mathcal{Z},y\in\mathcal{Y}}\left\|\nabla_\theta^2\nabla_\delta \log p\left(y \mid z;\tilde{\theta},\delta_0\right)\right\|_\infty\right)\left\|\hat{\theta}_{m,i-1}-\theta_0\right\|_2^2\\
&\leq \frac{c_1}{2\sqrt{m}}\sum_{i=t_1}^{t_2}\left\|\hat{\theta}_{m,i-1}-\theta_0\right\|_2^2\\
&\leq \frac{c_1}{2}\sqrt{m}\,(K-1)\max_{m<i\leq mK}\left\|\hat{\theta}_{m,i}-\theta_0\right\|_2^2\\
&= o_p\left((K-1)\,(\log m)^2/\sqrt{m}\right)\end{aligned}$$

for all $(t_1,t_2)$ where $m<t_1\leq t_2\leq mK$. So (34) holds for $j=1$.

**Second remainder term.** By Assumption 6 and the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}\left\|R_m^{(2)}(t_1,t_2)\right\|_2 &\leq \frac{c_2}{\sqrt{m}}\sum_{i=t_1}^{t_2}\max_{m<i\leq mK}\left\|\hat{\theta}_{m,i-1}-\theta_0-\Lambda_m^{-1}(i-1)\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right\|_2\\
&\leq c\,(K-1)\sqrt{m}\max_{m<i\leq mK}\left\|\hat{\theta}_{m,i}-\theta_0-\Lambda_m^{-1}(i)\sum_{j=1}^{i}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right\|_2.\end{aligned}$$

Then by Assumption 4, this term is $o_p(1)$. So (34) holds for $j=2$.

**Third remainder term.** For any $\epsilon>0$, we have that

$$\begin{aligned}&\Pr\left(\max_{t_1,t_2}\left\|\sum_{i=t_1}^{t_2}\left(\nabla_\theta\nabla_\delta \log p\left(Y_i \mid Z_i;\theta_0,\delta_0\right)-V_m(i)\right)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\right\|_2\geq\epsilon\sqrt{m}\right)\\
\leq&\Pr\left(\max_{m<i\leq mK}\left\|\hat{\theta}_{m,i}-\theta_0\right\|_2>c\frac{\log m}{\sqrt{m}}\right)\\
&+\Pr\left(\max_{t_1,t_2}\left\|\sum_{i=t_1}^{t_2}\left(\nabla_\theta\nabla_\delta \log p\left(Y_i \mid Z_i;\theta_0,\delta_0\right)-V_m(i)\right)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\right\|_2\geq\epsilon\sqrt{m},\max_{m<i\leq mK}\left\|\hat{\theta}_{m,i}-\theta_0\right\|_2\leq c\frac{\log m}{\sqrt{m}}\right).\end{aligned}\tag{37}$$

Per (36), the first summand on the RHS of (37) goes to zero. The second summand is bounded by

$$\Pr\left(\max_{t_1,t_2}\left\|\sum_{i=t_1}^{t_2}\left(\nabla_\theta\nabla_\delta \log p\left(Y_i \mid Z_i;\theta_0,\delta_0\right)-V_m(i)\right)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\mathbb{1}\left\{\|\hat{\theta}_{m,i-1}-\theta_0\|_2\leq c\frac{\log m}{\sqrt{m}}\right\}\right\|_2\geq\epsilon\sqrt{m}\right).\tag{38}$$

Because the outcome $Y_i$ is conditionally independent of past data given $Z_i$, the elements in this summation form a martingale difference sequence, i.e.

$$\mathbb{E}\left[G_m(i)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\mathbb{1}\left\{\left\|\hat{\theta}_{m,i-1}-\theta_0\right\|_2\leq c\frac{\log m}{\sqrt{m}}\right\}\middle|\mathcal{F}_i\right]=0$$

where we use the notational shorthand

$$G_m(i)=\nabla_\theta\nabla_\delta \log p\left(Y_i \mid Z_i;\theta_0,\delta_0\right)-V_m(i)$$

and

$$\mathcal{F}_i=\left(Z_1,Y_1,\cdots,Z_{i-1},Y_{i-1},Z_i\right).$$

Moreover, by Assumption 6, $G_m(i)$ is sub-Gaussian. That is, there is some $\sigma^2 > 0$ such that for all $\lambda > 0$, we have for all unit vectors $u, v$ that

$$\mathbb{E}\left[\exp\left(\lambda v^\top G_m(i) u\right)\middle|\mathcal{F}_{i-1}\right] \le \mathbb{E}\left[\exp\left(\lambda^2\sigma^2\lambda_{\max}\left(G_m(i)\right)\right)\middle|\mathcal{F}_{i-1}\right].$$

By the law of total expectations, we then have for any unit vector $v$ that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda v^\top \sum_{i=t_1}^{t_2} G_m(i)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\mathbb{1}\left\{\left\|\hat{\theta}_{m,i-1}-\theta_0\right\|_2 \le c\frac{\log m}{\sqrt{m}}\right\}\right)\right]\\
=&\mathbb{E}\Bigg[\mathbb{E}\left[\exp\left(\lambda v^\top G_m(t_2)\left(\hat{\theta}_{m,t_2-1}-\theta_0\right)\mathbb{1}\left\{\left\|\hat{\theta}_{m,t_2-1}-\theta_0\right\|_2 \le c\frac{\log m}{\sqrt{m}}\right\}\right)\middle|\mathcal{F}_{t_2-1}\right]\\
&\quad\times\exp\left(\lambda v^\top \sum_{i=t_1}^{t_2-1} G_m(i)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\mathbb{1}\left\{\left\|\hat{\theta}_{m,i-1}-\theta_0\right\|_2 \le c\frac{\log m}{\sqrt{m}}\right\}\right)\Bigg]\\
\le&\exp\left(\lambda^2c^2\sigma^2\frac{(\log m)^2}{m}\right)\mathbb{E}\left[\exp\left(\lambda v^\top \sum_{i=t_1}^{t_2-1} G_m(i)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\mathbb{1}\left\{\left\|\hat{\theta}_{m,i-1}-\theta_0\right\|_2 \le c\frac{\log m}{\sqrt{m}}\right\}\right)\right]\\
\le&\exp\left(\lambda^2c^2\sigma^2\left(K-1\right)\left(\log m\right)^2\right).
\end{aligned}$$

Using the Chernoff bound, we have that (38) is bounded by

$$\begin{aligned}
&\sum_{m<t_1\le t_2\le mK}\Pr\left(\left\|\sum_{i=t_1}^{t_2}\left(\nabla_\theta\nabla_\delta \log p\left(Y_i \mid Z_i;\theta_0,\delta_0\right)-V_m(i)\right)\left(\hat{\theta}_{m,i-1}-\theta_0\right)\mathbb{1}\left\{\left\|\hat{\theta}_{m,i-1}-\theta_0\right\|_2 \le c\frac{\log m}{\sqrt{m}}\right\}\right\|_2 \ge \epsilon\sqrt{m}\right)\\
\le& m^2\left(K-1\right)^2\exp\left(\lambda^2c^2\sigma^2\left(K-1\right)\left(\log m\right)^2-\epsilon^2 m\right).
\end{aligned}$$

This converges to zero as $m\to\infty$, so (34) holds for $j=3$. □

Using Lemma 1 above, we are now ready to prove Theorem 3.

*Proof of Theorem 3.* Consider the decomposition

$$\tilde{\phi}_m(t_1,t_2) = \phi_m(t_1,t_2) + R_m^{(1)}(t_1,t_2) + R_m^{(2)}(t_1,t_2) \tag{39}$$

with remainder terms defined as

$$\begin{aligned}
R_m^{(1)}(t_1,t_2) =& \frac{1}{\sqrt{m}}\sum_{i=t_1}^{t_2}\left(V_m(i)-\bar{V}_0\left(i/m\right)\right)\Lambda_m^{-1}(i-1)\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\\
R_m^{(2)}(t_1,t_2) =& \frac{1}{\sqrt{m}}\sum_{i=t_1}^{t_2}\bar{V}_0\left(i/m\right)\left(\Lambda_m^{-1}(i-1)-\frac{1}{m}\Lambda_0^{-1}\left(\frac{i-1}{m}\right)\right)\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right).
\end{aligned}$$

We first show the two remainder terms are negligible, i.e.

$$\max_{m<t_1<t_2\le mK}\|R_m^{(j)}(t_1,t_2)\| = o_p(1) \quad \forall j=1,2. \tag{40}$$

**First remainder term.** To bound the first remainder, note that the partial sums form a martingale due to Assumption 5, i.e.

$$\mathbb{E}\left[\left(V_m(i)-\bar{V}_0\left(i/m\right)\right)\Lambda_m^{-1}(i-1)\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\middle|\mathcal{F}_i\right] = 0.$$

Moreover, $\left(V_m(i) - \bar{V}_0(i/m)\right)\Lambda_m^{-1}(i)$ is sub-Gaussian and

$$\max_{m<i\leq mK}\left\|\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right\|_2 = o_p\left(\frac{\log m}{\sqrt{m}}\right)$$

per Assumptions 3 and 5. As such, we can use a similar martingale argument as the previous lemma to prove that (40) is satisfied for $j=1$.

**Second remainder term.** By the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
&\frac{1}{\sqrt{m}}\sum_{i=t_1}^{t_2}\bar{V}_0\left(i/m\right)\left(\Lambda_m^{-1}(i-1)-\frac{1}{m}\Lambda_0^{-1}\left(\frac{i-1}{m}\right)\right)\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\\
&\leq c\sum_{i=t_1}^{t_2}\left\|\Lambda_m^{-1}(i)-\frac{1}{m}\Lambda_0^{-1}(i/m)\right\|_2\left\|\frac{1}{\sqrt{m}}\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right\|_2\\
&\leq c\left(K-1\right)\left(\max_{i=m+1,\cdots,mK}\left\|m\Lambda_m^{-1}(i)-\Lambda_0^{-1}(i/m)\right\|_2\right)\left(\max_{m<i\leq mK}\left\|\frac{1}{\sqrt{m}}\sum_{j=1}^{i-1}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\right\|_2\right)
\end{aligned}$$

By Assumptions 3 and 5, it follows that (40) for $j=2$.

In addition, by Assumption 3, we have that

$$\left\{\nu \mapsto \frac{1}{\sqrt{m}}\sum_{j=1}^{\lfloor m\nu\rfloor}\begin{pmatrix}\nabla_\theta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\\ \bar{V}_0(\nu)\Lambda_0^{-1}(\nu)\nabla_\delta \log p\left(Y_j \mid Z_j;\theta_0,\delta_0\right)\end{pmatrix}\right\} \Rightarrow \left\{\nu \mapsto \begin{pmatrix}U_\theta(i/m)\\ \bar{V}_0(\nu)\Lambda_0^{-1}(\nu)U_\delta(i/m)\end{pmatrix}\right\}.$$

So by Slutsky's theorem and the continuous mapping theorem, we have weak convergence of the process $\phi_m$ with respect to the space of bounded functions $f:\Delta \mapsto \mathbb{R}^d$ as follows

$$\left\{(\nu_1,\nu_2)\mapsto\phi_m(\nu_1,\nu_2)\right\}_{(\nu_1,\nu_2)\in\Delta} \Rightarrow \left\{(\nu_1,\nu_2)\mapsto U_\delta(\nu_2)-U_\delta(\nu_1)+\int_{\nu_1}^{\nu_2}\bar{V}_0(v)\Lambda_0^{-1}(v)U_\theta(v)dv\right\}_{(\nu_1,\nu_2)\in\Delta}. \quad (41)$$

Combining this result with Lemma 1 and (40), the process $\psi_m^{(\text{plugin})}$ converges weakly to the same limit as $\phi_m$. □

### B.3.2 Asymptotics under the alternative

Suppose there is some $K' \in (\kappa, K]$ that satisfies the following assumptions. Assumptions 7, 8, and 9 can be viewed as analogous but simplified versions of the assumptions in Section B.3.1. Assumption 10 assumes the cumulative score process is characterized by some non-zero drift under the alternative for some time period, even when we continually update the plugin estimator $\hat{\theta}_{m,i}$. For example, this is likely to hold for values of $K'$ that are slightly larger than $\kappa$, since the plugin estimators up to time $\lfloor mK'\rfloor$ will not have strayed too far from its actual value of $\theta_0$ prior to the changepoint. For $v\in[1,K]$, define the limit of the MLEs as $\bar{\hat{\theta}}(v) = \lim_{m\to\infty}\hat{\theta}_{m,\lfloor mv\rfloor}(v)$.

**Assumption 7.** *Under the alternative, suppose that*

$$\frac{1}{\sqrt{m}}\sum_{j=\lfloor m\kappa\rfloor}^{\lfloor mK'\rfloor}\nabla_\delta \log p\left(Y_j \mid Z_j;\bar{\hat{\theta}}\left(\frac{j-1}{m}\right),\delta_0\right) - \mathbb{E}_{\theta_0,\delta_1\mathbb{1}\{i\geq m\kappa\}}\left[\nabla_\delta \log p\left(Y_j \mid Z_j;\bar{\hat{\theta}}\left(\frac{j-1}{m}\right),\delta_0\right)\middle| Z_j\right] = O_P(1).$$

**Assumption 8.** *Under the alternative, suppose that*

$$\max_{i=\lfloor m\kappa\rfloor,\cdots,\lfloor mK'\rfloor}\left\|\sqrt{m}\left(\hat{\theta}_{m,i}-\bar{\hat{\theta}}(i/m)\right)\right\|_2 = O_p(1).$$

**Assumption 9.** *There is some $c > 0$ and neighborhood $B$ that includes the set $\left\{\bar{\hat{\theta}}\left(v\right) : v \in [\tau, K']\right\}$ with nonzero radius such that*

$$\sup_{\theta \in B} \sup_{y \in \mathcal{Y}, z \in \mathcal{Z}} \left\|\nabla_\theta \nabla_\delta \log p\left(y \mid z; \theta, \delta_0\right)\right\|_\infty \leq_p c.$$

**Assumption 10.** *There is some $c > 0$ such that*

$$\lim_{m \to \infty} \left\| \frac{1}{m} \sum_{j=\lfloor m\kappa \rfloor}^{\lfloor mK' \rfloor} \mathbb{E}_{\theta_0, \delta_1}\left[\nabla_\delta \log p\left(Y_j \mid Z_j; \bar{\hat{\theta}}(t), \delta_0\right)\right] \right\|_2 \geq c.$$

**Theorem 4.** *Suppose Assumptions 7 to 10 hold. Then under the alternative hypothesis, we have*

$$\lim_{m \to \infty} \Pr\left(\exists t \in \{m+1, \cdots, mK\} \textit{ such that } C_m^{(plugin)}(t) > h_m(t)\right) = 1.$$

*Proof.* By the definition of $C_m^{(\text{plugin})}(t)$, suffices to prove that

$$\frac{1}{\sqrt{m}} \psi_m^{(\text{plugin})}(\lfloor m\kappa \rfloor, \lfloor mK' \rfloor) = \frac{1}{\sqrt{m}} \sum_{i=\lfloor m\kappa \rfloor}^{\lfloor mK' \rfloor} \nabla_\delta \log p\left(Y_i \mid Z_i; \hat{\theta}_{m,i-1}, \delta_0\right)$$

goes to infinity. Consider the following decomposition

$$\begin{aligned}
&\frac{1}{\sqrt{m}} \psi_m^{(\text{plugin})}(\lfloor m\kappa \rfloor, \lfloor mK' \rfloor) \\
=& \frac{1}{\sqrt{m}} \sum_{i=\lfloor m\kappa \rfloor}^{\lfloor mK' \rfloor} \mathbb{E}_{\theta_0, \delta_1}\left[\nabla_\delta \log p\left(Y_i \mid Z_i; \bar{\hat{\theta}}\left(\frac{i-1}{m}\right), \delta_0\right)\right] \\
&+ \frac{1}{\sqrt{m}} \sum_{i=\lfloor m\kappa \rfloor}^{\lfloor mK \rfloor} \left\{\nabla_\delta \log p\left(Y_i \mid Z_i; \bar{\hat{\theta}}\left(\frac{i-1}{m}\right), \delta_0\right) - \mathbb{E}_{\theta_0, \delta_1}\left[\nabla_\delta \log p\left(Y_i \mid Z_i; \bar{\hat{\theta}}\left(\frac{i-1}{m}\right), \delta_0\right)\right]\right\} \\
&+ R_m
\end{aligned} \tag{42}$$

The first term on the right hand side must diverge to infinity by Assumption 10. Per Assumption 7, the second term is $O_p(1)$. Per Assumption 8 and Taylor's theorem, the remainder $R_m$ satisfies

$$\lim_{m \to \infty} \|R_m\|_2 \leq \lim_{m \to \infty} \frac{c_1}{\sqrt{m}} \sup_{\{\tilde{\theta}_{m,i-1} : i = \lfloor m\kappa \rfloor, \cdots \lfloor mK' \rfloor\} \in B} \left\| \sum_{i=\lfloor m\kappa \rfloor}^{\lfloor mK' \rfloor} \nabla_\theta \nabla_\delta \log p\left(Y_i \mid Z_i; \tilde{\theta}_{m,i-1}, \delta_0\right) \left(\hat{\theta}_{m,i-1} - \bar{\hat{\theta}}\left(\frac{i-1}{m}\right)\right) \right\|_2. \tag{43}$$

Combined with Assumption 9, we have that

$$\lim_{m \to \infty} \|R_m\|_2 \leq \lim_{m \to \infty} c_2 \sqrt{m} \max_{i = \lfloor m\kappa \rfloor, \cdots, \lfloor mK' \rfloor} \left\| \hat{\theta}_{m,i-1} - \bar{\hat{\theta}}\left(\frac{i-1}{m}\right) \right\|_2 = O_p(1). \tag{44}$$

As such, the right hand side of (42) diverges to infinity, which means its left hand side must also diverge to infinity. Thus we have our desired result. □

### B.4 Implementation details

There are a number of implementation decisions to make. First, the number of sequences $B$ should be set to a value large enough such that the estimated DCLs converge. In our simulations, we chose $B$ so that the chart statistic for five or more bootstrapped sequences exceeded the DCL at each time step. Next, one can maximize statistical power by tuning the shape of the alpha-spending function. Here we simply use a linear alpha-spending function, but future work may explore nonlinear functions instead. Finally, our theoretical results allow for monitoring in a fully sequential manner or in batches of observations. We found that batching had a negligible impact on detection delay and improved both computational efficiency and convergence to the asymptotic distribution. As such, we recommend setting the batch size so a significant change is unlikely to occur within a batch. In our experiments, the batch size is set to 10.

### B.5 Extension to monitor treatment-specific outcomes

It is straightforward to extend results in this work to test for shifts in the conditional distribution of treatment-specific outcomes, i.e.,

$$\begin{aligned}
&H_0 : \tilde{Y}_i(a)|\hat{f}_i(\tilde{X}_i) = q, \mathcal{F}_i \sim g_{\theta,0}(q,a) \qquad \forall i \in \{1, \cdots, mK\}, q \in [0,1], a \in \{0,1\}\\
&H_1 : \exists \delta, \kappa \text{ s.t. } \tilde{Y}_i(a)|\hat{f}_i(\tilde{X}_i) = q, \mathcal{F}_i \sim g_{\theta,\delta \mathbb{1}\{i>\kappa\}}(q,a) \qquad \forall i \in \{1, \cdots, mK\}, q \in [0,1], a \in \{0,1\}.
\end{aligned} \tag{45}$$

One can again establish ignorability under either an extension of the assumption of conditional exchangeability or time-constant selection bias to both treatment options, e.g.,

$$(Y_t(0), Y_t(1)) \perp A_t|\hat{f}_t(X_t), \mathcal{F}_t \forall t = 1, 2, \cdots \tag{46}$$

Under ignorability, one can directly apply the score-based CUSUM described in Section 5 to all monitoring data (not just SOC-only data).

## C Simulation details and additional results

For each simulation, we generate a $p$-random vector $X_t$ and random variable $U_t$. All these variables are drawn independently from the uniform distribution from -1 to 1. The outcome $Y_t(0)$ was generated using either (m.1) or (m.2) with $Z_t = (X_t, U_t, 1)$. Treatments were assigned using one of two models. The first is a logistic regression model with $(\hat{f}_t(X_t), X_t, U_t)$ with coefficients and intercept denoted by $\gamma_{\text{LR}}$. The second sets treatment to $A_t = \max(A_t^{(1)}, A_t^{(2)})$, where $A_t^{(1)}$ and $A_t^{(2)}$ are generated using logistic regression models with inputs $(\hat{f}_t(X_t), X_t, U_t)$, with coefficients and intercept denoted by $\gamma^{(1)}$ and $\gamma^{(2)}$, respectively. The model parameters used to generate outcomes and treatment assignments are given in Table 2.

### C.1 Comparator: Bayesian changepoint monitoring

We implemented the Bayesian monitoring procedure as follows. The chart statistic in Bayesian monitoring is the posterior probability of there having been a change, i.e. $\hat{C}^{\text{bayes}}(t) = \Pr(\kappa \le t \mid Y_1, \cdots, Y_t, Z_1, \cdots, Z_t)$. We used a static control limit of $1 - \alpha$. Given the minimax optimality of the Shiryaev-Roberts procedure (Shiryaev, 1963), we define a modified geometric prior over $\kappa$, in which

$$\pi(\kappa = t) \propto p(1-p)^{t-1} \quad \forall t = m+1, \cdots, mK \tag{47}$$

with $p = 1/mK$ and the probability of there being no changepoint is set to 0.5. We assume a normal prior for $\theta$ with the mean and covariance matrix set to the results from maximum likelihood estimation on the non-contaminated data. We also place a normal prior for $\delta$ with mean zero and a diagonal covariance matrix. In the simulations, we set the diagonal matrix so that the mean norm of $\delta$ in the prior is close to that of the actual shift. In practice, such information is not known and one must rely on prior knowledge.

### C.2 Additional simulation: Naïve monitoring of the misclassification rate

Here we present an additional simulation that compares the false alarm rate of score-based monitoring with a naïve CUSUM procedure that monitors the overall misclassification rate to our proposed score-based CUSUM procedure that monitors the conditional distribution of $Y_t(0)|\hat{f}_t(X_t)$. We assume conditional exchangeability with respect to $\hat{f}_t(X_t)$. The outcome is generated per (m.1) with $Z_t = X_t$ and $\delta = (2, 1, 1, 1, \vec{0}_4, 0)$. Treatment is assigned using a logistic regression model with input as $\hat{f}_t(X_t)$ with parameter $\gamma_{\text{LR}} = (1, -0.5)$ up to time $t = 200$ and $\gamma_{\text{LR}} = (5, -2.5)$ thereafter. Rather than fitting a risk prediction model, we fit a locked binary classifier for simplicity and classify any observation to be positive if the predicted risk exceeds 0.7.

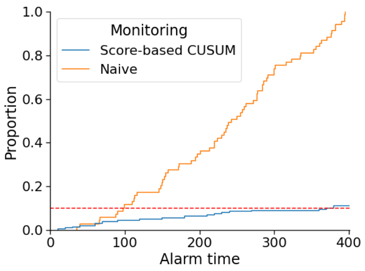


Figure 4: Comparison of naïve CUSUM procedure that monitors the unadjusted overall misclassification rate versus score-based CUSUM monitoring of the conditional distribution $Y_t(0)|\hat{f}_t(X_t)$

As shown in Figure 4, the naïve CUSUM has a substantially inflated false alarm rate because the overall misclassification rate, without further adjustment, is sensitive to shifts in clinician trust. On the other hand, the proposed score-based CUSUM procedure controls the false alarm rate at the desired level.

### C.3 Additional simulation: Verifying false alarm rate control

We evaluate false alarm rate control of score-based CUSUM monitoring in finite samples. The data are simulated under the null and satisfy either the assumption of conditional exchangeability or time-constant selection bias. A shift in the treatment propensities is introduced halfway through the monitoring period.

Here we consider a locked ML algorithm. In the `Conditional Exchangeability` simulation, the treatment propensities are generated according to a logistic regression model with only $\hat{f}(X_t)$ and $\tilde{X}_t$ as inputs; the outcome is generated according to (m.1). In the `Time-constant selection bias` simulation, the treatment propensities and outcome are generated per Example 1 of the Appendix. We consider two versions of both assumptions: one that only conditions on $\hat{f}(X_t)$ and another that conditions on $(\hat{f}(X_t), \tilde{X}_t)$. The former leads us to test (2) and the latter leads us to test (3).

We vary the size of the noncontamination dataset size $m$ and monitor for $mK$ time points, with $K$ set to 4. As shown in Figure 5, Type I error is inflated for small values of $m$, but converges to the nominal rate once $m$ is sufficiently large.

### C.4 Additional details: Shift magnitude (Section 6.1)

For the evolving ML algorithm, we trained an Exponentially Weighted Averaging Forecaster (EWAF) that was an ensemble of three ML algorithms: one continually retrained on all prior data, one continually retrained on the most recent 200 observations, and one continually retrained on the most recent 400 observations. The models were retrained after every 10 observations. Calibration curves of the locked versus evolving models are shown in Figure 6.

### C.5 Additional simulation: Clinician trust (Section 6.2)

Not all types of clinician trust will substantially delay detection of performance decay. In particular, consider a setting where the calibration decay is highest among patients who have the highest *and* lowest predicted risks (we refer to this simulation as "symmetric" calibration decay). Patients with the lowest predicted risks are very likely to receive SOC. As such, even in settings with high clinician trust (strong performativity), patients experiencing calibration decay are frequently sampled by the monitoring procedure and time to detection does not significantly vary across different levels of clinician trust (Figure 7).

### C.6 Addition details: Retraining in stationary settings (Section 6.3)

For this experiment, we simulated higher-dimensional data with $X \in \mathbb{R}^{50}$. For computational speed, batch size for continual retraining was 10.

Performance characteristics of the model updates are shown in Figure 8. Note that the calibration curves plotted

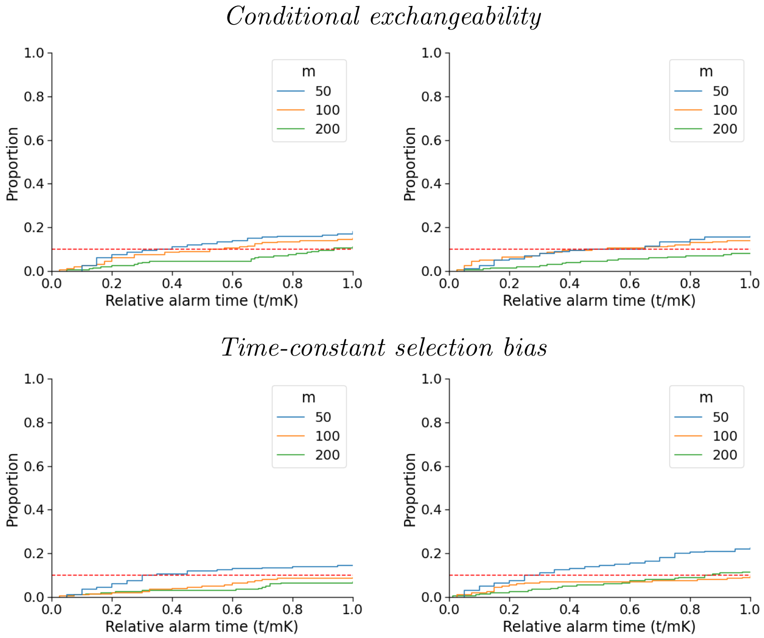


Figure 5: Cumulative distribution of alarm times for score-based CUSUM monitoring of a locked model $\hat{f}$ under the null. The assumption of conditional exchangeability and time-constant selection bias are satisfied in the top and bottom rows, respectively. In the left column, the assumptions hold when conditioning on $\hat{f}(X_t)$; in the right column, the assumptions hold when conditioning on $(\hat{f}(X_t), \tilde{X}_t)$. The target false alarm rate is 0.1, which is achieved as the size of the non-contaminated dataset $m$ increases.

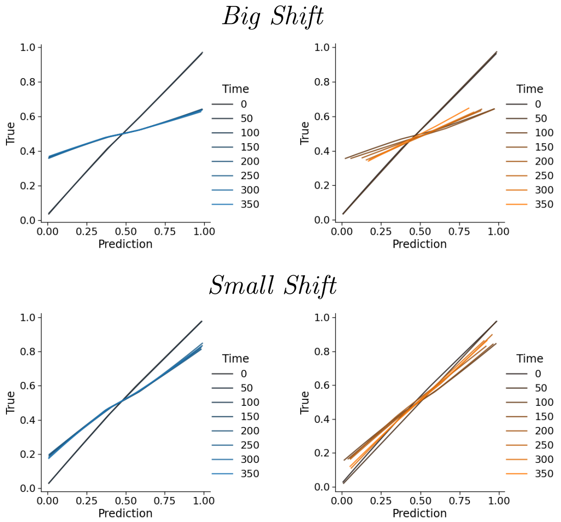


Figure 6: Calibration curves of the locked and continually retrained models (first and second columns, respectively) in the presence of big and small distribution shifts (top and bottom rows, respectively). Calibration curves are plotted over time, including before and after the distribution shift.

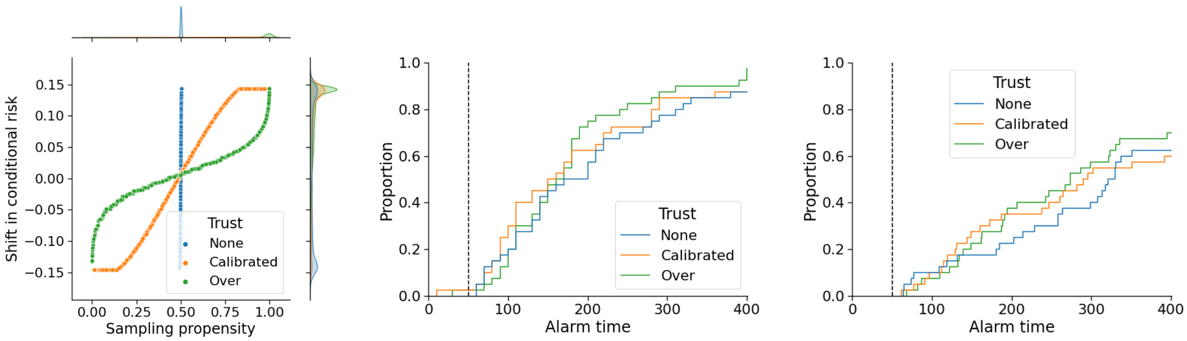
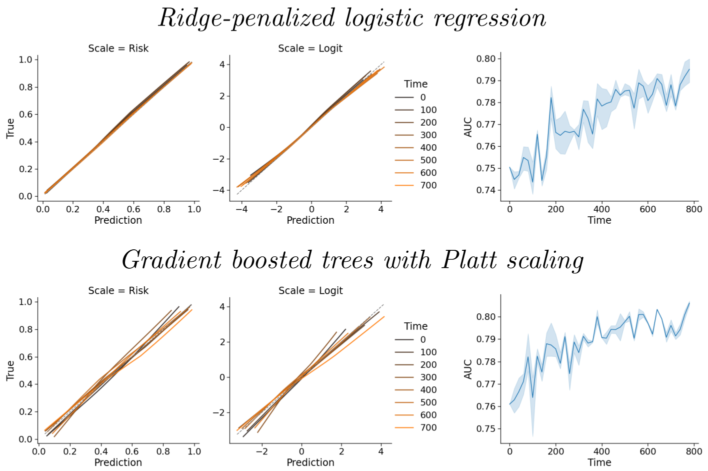


Figure 7: The left column shows how shifts in the conditional risk vary with respect to the probability of a patient being assigned SOC, and thus their probability of being sampled for monitoring. The middle and right columns show the cumulative distribution of alarm times using score-based CUSUM and Bayesian monitoring, respectively.


*Ridge-penalized logistic regression*

*Gradient boosted trees with Platt scaling*

Figure 8: Calibration curves of continually retrained models, where the data stream is stationary. Calibration curves are plotted on the risk and logit scales in the first two columns. Average AUCs of the model updates are shown in the right column.

on the logit scale are farther away from the ideal diagonal line, compared to calibration curves plotted on the risk scale.

### C.7 Additional simulation: Gradual decay of model calibration

The simulation setup is the same as Section 6.1 except that the structural change is gradual. Results are shown in Figure 9. Compared to the results in Section 6.1 for a sudden shift, the median time to detection is longer for all the methods. Nevertheless, the relative ranking of the monitoring procedures in terms of their median time to detection is the same.

### C.8 Additional simulation: Sensitivity analysis of the time-constant selection bias assumption

We explore how violations of the time-constant selection bias assumption can impact detection delay of a structural change. We first simulate data that satisfies this time-constancy assumption based on Example 1. Then we introduce violations of this assumption by adding an edge from the unmeasured confounder $U_t$ to the final treatment decision $A_t$ in the DAG and setting a non-zero edge weight at times $t = 100$ or 300. Such shifts in the propensity model could occur if, say, a clinician suspects performance of the ML algorithm has decayed and decides to place more weight in the unmeasured risk factor $U_t$.

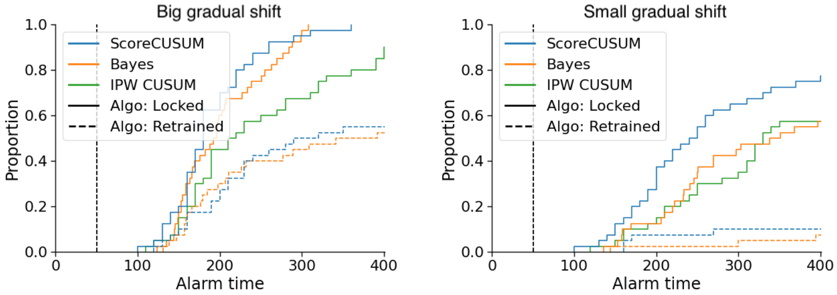


Figure 9: Monitoring structural change starting with the same setup as Section 6.1 but with *gradual* structural change. Compared to the results for a sudden shift, there is a longer time to detection. Nevertheless, the relative ranking of the methods in terms of their median time to alarm is the same.

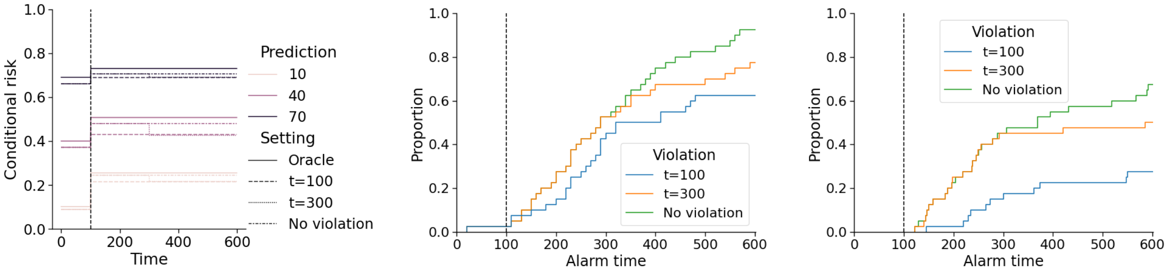


Figure 10: Monitoring a structural change at time $t = 100$ (dashed vertical line) when the assumption of time-constant selection bias is violated. We introduce a shift in the treatment propensities that violates the time-constant selection bias assumption at times $t = 100$ and $t = 300$. We also simulate no violation of the time-constant selection bias by never introducing this shift in the treatment propensities. For different risk prediction values, we plot the conditional risks among the SOC-only population over time, which are biased for those among the general population. We also plot the oracle conditional risk for comparison. The middle and right columns show the cumulative distribution of alarm times for score-based CUSUM monitoring and Bayesian monitoring, respectively.

Because the simulated violation is designed to dampen the shift observed in the data, we find that detection delay increases as the violation occurs earlier in time (Figure 10). In the worst case scenario, the structural change *and* the shift in the treatment propensities occurs at the same time ($t = 100$) and power drops by 30%. Nevertheless, such a scenario is unlikely to happen in practice since it assumes clinicians know exactly when performance decays.

In this simulation, we find that the power of Bayesian monitoring is much lower than that for the score-based CUSUM. This is likely due to the sensitivity of Bayesian inference to model misspecification: the monitoring model assumes a single changepoint, whereas the observed data distribution shifts at two time points. Consequently, its power drops by over 40%. In fact, even without violations of the time-constancy assumption, the power of the Bayesian procedure is much lower than that of the score-based CUSUM. This may be due to difficulties in performing posterior inference for (m.2), which is only partially differentiable.

## D Monitoring a PONV risk calculator

Data was obtained from UCSF MPOG data repository, under the MPOG data sharing agreement. Input to features to the ML algorithm included preoperative variables, including biological sex, smoking status, age, ASA score, and blood test results. A patient is defined as receiving additional care if they received at least two antiemetics. The list of antiemetics that counted towards intervention were: Propofol infusion, Metoclopramide, Aprepitant, Scopolamine patch, and Haloperidol.

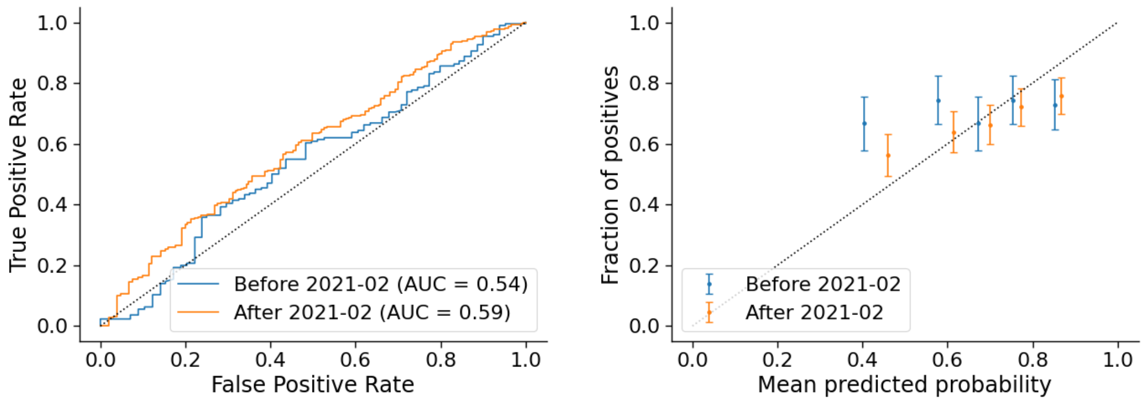


Figure 11: ROC and calibration curves of the evolving ML-based risk prediction model for PONV.

| Symbol | Meaning |
|---|---|
| $X_t$ | Patient covariates that go into the ML algorithm |
| $Y_t$ | Patient outcome |
| $A_t$ | Treatment assignment |
| $\tau_1, \tau_2, \cdots$ | Indices for the subsequence of times at which the patient was assigned standard-of-care (SOC) |
| $Z_t$ | Predictors in the standard monitoring setting |
| $\hat{f}_t$ | The ML algorithm at time $t$ |
| $\kappa^{\text{rel}} \in (1, K)$ | Position of changepoint in relative time |
| $\kappa$ | Changepoint in absolute time, equal to $\lfloor m\kappa^{\text{rel}} \rfloor$ |
| $\theta$ | Parameter indexing the pre-change distribution |
| $\delta$ | Parameter indexing the structural change |
| $m$ | Size of dataset needed for initialization of monitoring procedures (also known as non-contaminated data) |
| $m+1, \cdots, mK$ | The time period for monitoring structural change |
| $C_m(t)$ | The chart statistic of a monitoring procedure at time $t$ |
| $h_m(t)$ | The control limit of a monitoring procedure at time $t$ |
| $\hat{T}_m$ | Alarm time of a monitoring procedure, i.e. when the chart statistic first exceeds the control limit |

Table 1: Mathematical symbols

**Locked model** The random forest (RF) was trained using data from January 2018 to May 2019. The first 200 patients were used to initialize monitoring procedures and monitoring began mid-December 2019. Conditional exchangeability was assumed to hold with respect to $\hat{f}_t(X_t)$ and the data was modeled using (m.1).

**Retrained model** The RF was retrained every 10 observations. The monitoring procedures were initialized using observations from 200 patients starting July 2019 and began monitoring October 2020. Performance characteristics of the continually retrained model are shown in Figure 11. AUC of the retrained model is calculated by weighting the monitoring data by the estimated inverse propensity weights.

**Algorithm 1** Pseudocode for score-based CUSUM procedure with dynamic control limits

```
Select time factor $K$, alpha spending function $\alpha^{\text{rel}}$, and number of bootstrap sequences $B$.
Let $\mathcal{B}_m = \{1, ..., B\}$ represent the bootstrap sequences that have not been rejected at time $m$.
Observe non-contaminated data $\{(Z_t, Y_t) : t = 1, \cdots, m\}$.
Calculate MLE $\hat{\theta}_{m,m}$
for $b = 1, ..., B$ do
    Resample outcome $Y_t^{*(b)}$ given $Z_t$ for $t = 1, \cdots, m$, with $\theta = \hat{\theta}_{m,m}$ and $\delta = 0$.
end for
for $t = m + 1, ..., mK$ do
    Observe $(Z_t, Y_t)$.
    Calculate chart statistic $C_m^{(plugin)}(t)$ and MLE $\hat{\theta}_{m,t}$.
    for $b \in \mathcal{B}_{t-1}$ do
        Resample outcome $Y_t^{*(b)}$ given $Z_t$ with $\theta = \hat{\theta}_{m,t-1}$ and $\delta = 0$.
        Compute $\phi_m(t', t)$ for $t' = m + 1, \cdots, t - 1$ for the $b$-th bootstrap sequence.
        Calculate $C_m^{*,(b)}(t) = \max_{t' \in \{m+1, \cdots, t\}} \phi_m(t', t)$.
    end for
    Set $h_m(t)$ such that the proportion of bootstrap chart statistics exceeding the DCL is
        $$\left| \left\{ b : b \in \mathcal{B}_{t-1}, C_m^{*,(b)}(t) > h_m(t) \right\} \right| / B = \alpha^{\text{rel}} \left( t/m \right) - \alpha^{\text{rel}} \left( (t-1)/m \right).$$
    Define $\mathcal{B}_t = \{b : b \in \mathcal{B}_{t-1}, C_m^{*,(b)}(t) \leq h_m(t)\}$.
    if $C_m^{(\text{plugin})}(t) > h_m(t)$ then
        Fire an alarm. Break.
    end if
end for
```

| Section | Experiment(s) | Outcome model | Treatment model |
|---|---|---|---|
| C.3 | CE with respect to $\hat{f}_t(X_t)$ | (m.1) with $\theta = (2,1,1,1,\vec{0}_4,0,0,0)$ and $\delta = \vec{0}$ | $\gamma_{\text{LR}} = (0.3, \vec{0}_8, 0, 0, 0)$ up to $t = mK/2$.<br>$\gamma_{\text{LR}} = (0.6, \vec{0}_8, 0, 0, 0)$ after $t = mK/2$. |
| C.3 | CE with respect to $(\hat{f}_t(X_t), X_t)$ | (m.1) with $\theta = (2,1,1,1,\vec{0}_4,1,0,0)$ and $\delta = \vec{0}$ | $\gamma_{\text{LR}} = (0.3, \vec{0}_8, 0.1, 0, 0)$ prior to $t = mK/2$.<br>$\gamma_{\text{LR}} = (0.6, \vec{0}_8, 0.2, 0, 0)$ after $t = mK/2$. |
| C.3 | TC with respect to $\hat{f}_t(X_t)$ | (m.2) with $\theta = (2,1,1,1,\vec{0}_4,0,1,0)$ and $\delta = \vec{0}$ | $\gamma^{(1)} = (0, \vec{0}_8, 0, 1, -2)$ at all time points.<br>$\gamma^{(2)} = (0.2, \vec{0}_8, 0, 0, 0)$ up to $t = mK/2$.<br>$\gamma^{(2)} = (0.4, \vec{0}_8, 0, 0, 0)$ after $t = mK/2$. |
| C.3 | TC with respect to $(\hat{f}_t(X_t), X_t)$ | (m.2) with $\theta = (2,1,1,1,\vec{0}_4,1,1,0)$ and $\delta = \vec{0}$ | $\gamma^{(1)} = (0, \vec{0}_8, 0, 1, -2)$ at all time points.<br>$\gamma^{(2)} = (0.2, \vec{0}_8, 0.3, 0, 0)$ up to $t = mK/2$.<br>$\gamma^{(2)} = (0.4, \vec{0}_8, 0.6, 0, 0)$ after $t = mK/2$. |
| 6.1 | Big shift and small shift. | (m.1) with $\theta = (2,1,1,1,\vec{0}_4,0,0,0)$ and $\delta = (-1.6,-0.8,-0.8,-0.8,\vec{0}_4,0,0,0)$ or $\delta = (-1,-0.5,-0.5,-0.5,\vec{0}_4,0,0,0)$ for the big or small shifts, respectively. | $\gamma_{\text{LR}} = (2, \vec{0}_8, 0, 0, 0)$ |
| C.7 | Big gradual shift and small gradual shift. | (m.1) with $\theta = (2,1,1,1,\vec{0}_4,0,0,0)$ and $\delta_t = \min(1, 0.01 * (t-50)) * (-1.6,-0.8,-0.8,-0.8,\vec{0}_4,0,0,0)$ or $\delta_t = \min(1, 0.01 * (t-50)) * (-1,-0.5,-0.5,-0.5,\vec{0}_4,0,0,0)$ for the big or small gradual shifts, respectively. | $\gamma_{\text{LR}} = (2, \vec{0}_8, 0, 0, 0)$ |
| 6.2 | Risks either shifted symmetrically or only among those with high-risk. | (m.1) with $\theta = (2,1,1,1,\vec{0}_4,0,0,0)$ and $\delta = (-1,-0.5,-0.5,-0.5,\vec{0}_4,0,0,0)$ or $\delta = (-0.2,-0.15,-0.15,-0.15,\vec{0}_4,0,0,-0.75)$ for symmetric- or high-risk shifts, respectively. | Low: $\gamma_{\text{LR}} = (0.01, \vec{0}_8, 0, 0, 0)$<br>Med: $\gamma_{\text{LR}} = (1, \vec{0}_8, 0, 0, 0)$<br>High: $\gamma_{\text{LR}} = (6, \vec{0}_8, 0, 0, 0)$ |
| 6.3 | Model retrained using ridge-penalized logistic regression or gradient boosted trees | (m.1) with $\theta = (2,1,1,\vec{0}_{47},0,0,0)$ and $\delta = \vec{0}$ | $\gamma_{\text{LR}} = (0.5, \vec{0}_{50}, 0, 0, 0)$ |
| C.8 | Violations of the time-constant selection bias assumption are introduced at time $t'$ | (m.2) with $\theta = (2,1,1,1,\vec{0}_4,0,0,0)$ and $\delta = (-0.1,-0.02,\vec{0}_6,0,0,0)$ | $\gamma^{(1)} = (0, \vec{0}_8, 0, 1, -1)$ up to $t'$.<br>$\gamma^{(1)} = (-0.5, \vec{0}_8, 0, 1, -1)$ after $t'$.<br>$\gamma^{(2)} = (0.8, \vec{0}_8, 0, 0, 0)$ at all time points. |

Table 2: Model parameters used to generate outcomes and treatment assignments in the simulations.