Intervention-Aware Early Warning

Dhivya Eswaran
Carnegie Mellon University
deswaran@cs.cmu.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Nina Mishra
Amazon
nmishra@amazon.com

Yonatan Naamad
Amazon
ynaamad@amazon.com

***Abstract*—How can we early warn against an impending student drop out or an adverse health condition in near real-time? More challengingly, how do we learn to early warn from data containing confounding interventions–e.g., tutoring or medicines–while remaining interpretable to the human decision maker?**

We consider the problem of learning to interpretably early warn from labeled data tainted by interventions. We first identify three principles that an early warning system should follow. We then propose SmokeAlarm which provably obeys these principles and produces early warning scores in an online manner. Notably, learned model is "bi-inspectable", i.e., it can be visualized both in the presence and in the absence of interventions. Experiments demonstrate the efficacy of SmokeAlarm over prior approaches.

I. Introduction

Early warnings have serious implications in a variety of domains. In medicine, warning of a forthcoming epileptic seizure or septic shock event can be life-altering. In data centers, warning of a pending server crash can alert cloud system administrators of downstream problems. Interventions can come to the rescue – particularly if they are delivered at an opportune moment. Pulling over to the side of the road before a seizure can prevent a car accident; antibiotics can avert organ failure; counseling a student may prevent their drop out.

Let us consider how algorithms trained on historical data can learn to automatically early warn. The classical solution is to train an algorithm with labels "fast-forwarded" in time. For instance, if the goal is to predict system failure Δ time steps in advance, then label a time step t with what happened at $t+\Delta$. However, as noted by [1], if interventions administered in the intervening duration averted (aided) an event, the observed labels underestimate (overestimate) the true early warning score. As a result, counter-intuitive results may ensue. We consider how past data, peppered with interventions, can be used to learn *intervention-aware* early warning scores.

To effectively aid human decision making in high-stakes domains like health care, interpretability of the models is a key concern [2]. As interpretability is not a monolithic concept [3], in this work, we focus on a model which "can be readily presented to the user with visual or textual artifacts" [4].

The approach we take in this paper is to first learn a function that predicts the probability of a future event under various future intervention regimens–accounting for the *stochastic* and *prolonged* effect of interventions–and then suitably time-decay this function to produce an early warning score. Our contributions are: **(i) Principles:** We identify three principles that an ideal early warning system should follow. **(ii) Algorithm:** We propose SmokeAlarm which provably obeys these principles, learns from past labeled data tainted by interventions offline and produces early warning scores online. **(iii) Interpretability:** A key novelty of SmokeAlarm is its "bi-inspectability", which is to say that the model can be visualized both in the presence and in the absence of an intervention (Fig. 1). Experiments on real-world data confirm that SmokeAlarm is better able to early warn than past approaches, outperforming by $16-38\%$ AUC with an average warning lead time of 6.1 hours before the onset of septic shock.

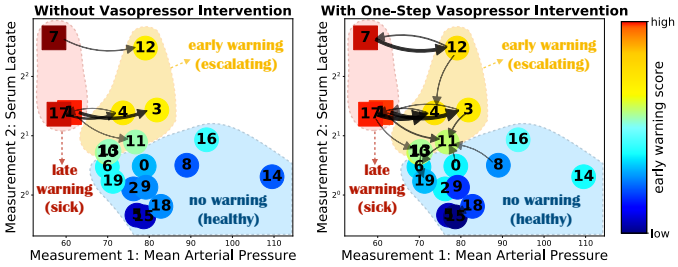


Fig. 1. **Interpretability:** SmokeAlarm model with states (numbered vertices) and early warning scores (vertex colors) shows that it alarms when the patient already has septic shock (red region), and does not warn when they are healthy (blue region). Importantly, it warns during the escalation to septic shock (yellow area), which is the opportune moment for intervention. SmokeAlarm also learns that *vasopressor* interventions tend to increase mean average pressure (MAP) as indicated by dark, thick rightward arrows (right) and thus decreases early warning scores of low-MAP states $\{1, 7, 17\}$.

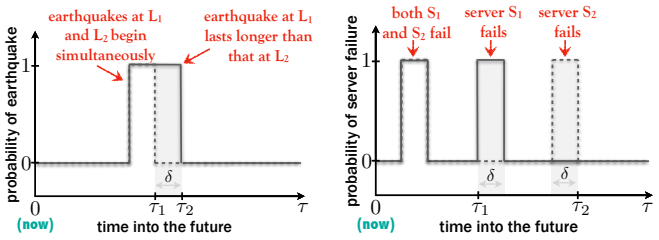
While primarily motivated by health care, our ideas can be more broadly applied, e.g., in warning of student drop outs (tutoring as interventions) or mechanical failures (repairs as interventions). In the interest of space, additional material including experiments on synthetic data, proofs and discussion are provided in the supplementary[1].

II. Related Work

Prior work has considered early warning against various adverse events such as sepsis [5], septic shock [6] and heart failure [7]. In general, these approaches do not account for confounding interventions which "can mask the ground truth labels needed to train and evaluate a prediction system" [1].

To cope with interventions, Caruana et. al. [2] advocates for *intelligible* models amenable to repairing by domain experts, e.g., by deleting incorrect rules such as asthma reduces the risk of pneumonia. [8] proposes a human-in-the-loop solution by seeking expert labels for pairwise comparisons of time points.

[1] www.cs.cmu.edu/~deswaran/papers/icdm19-smokealarm-sup.pdf



(a) Short vs. long earthquake (b) Early vs. late server failure

Fig. 2. The longer earthquake at L_1 and the server S_1 failing in quick successions require more urgent attention and should have a strictly higher early warning score than L_2 and S_2 respectively.

In contrast, we focus on solutions requiring minimal human labeling and post-processing efforts. More recently, [9] uses Counterfactual Gaussian Processes to forecast a single real-valued measurement in the presence of interventions. As such [9] does not address the early warning problem and scales poorly with input size due to the use of Gaussian Processes.

State-based methods can interpretably model the progression of trajectories. [10] infers a continuous-time Markov model for chronic obstructive pulmonary disease. [11] learns a probabilistic model to estimate the stages of chronic kidney disease. Both methods are unsupervised, ignore interventions and do not early warn. We also employ a state-based model, but explicitly account for interventions, illustrate bi-inspectability, and capitalize on labels to early warn.

Reinforcement learning [12] is a powerful tool to learn optimal intervention policies, e.g., for sepsis [13]. However, our focus is on a related but different problem of producing credible early warning scores to aid the human decision maker.

III. Preliminaries and Principles

Let $\mathbb{T}$ be a set of trajectories. Each trajectory $\mathcal{T} \in \mathbb{T}$ consists of real or categorical measurements $\mathbf{x}_t{=}(x_{t1}, \ldots, x_{tM})$, quantity of administered intervention $\mathbf{y}_t \in \{0, 1, \ldots, Y\}$ for a maximum dose Y and a binary label ℓ_t denoting event occurrence at each time step $t{=}1, 2, \ldots, T$. Some measurements and labels may be missing, but the interventions are fully observed. We focus on the case with a single intervention type. Let $\mathbf{x}_{:t}$ and $\mathbf{y}_{:t}$ be the observed values until time t and let $\mathbf{y}_{t+1:}$ denote the interventions after time t, i.e., $t{+}1$ onward.

Given a trajectory $\mathcal{T}{=}(\mathbf{x}_{:t}, \mathbf{y}_{:t})$, our goal is to produce an early warning score $w(\mathcal{T}) = w(\mathbf{x}_{:t}, \mathbf{y}_{:t})$ reflecting how soon and likely an event is, when not disturbed by future interventions. To formalize this intuition as verifiable principles, we need some definitions. First, a trajectory $\mathcal{T} = (\mathbf{x}_{:t}, \mathbf{y}_{:t})$ is said to have an **intervention-free future** iff no intervention is given after time t, i.e., $\mathbf{y}_{t+1:} = \mathbf{0}$. Next, a **future event probability function** $f{:}\mathbb{Z}_{\geq 0} \rightarrow [0, 1]$ is a function mapping a given $\tau{\geq}0$ to the probability of an event exactly τ steps into the future. Lastly, for functions $g_1, g_2 : \mathcal{D} \rightarrow \mathbb{R}$ where $\mathcal{D}$ is a countable set, g_1 is said to **dominate** g_2 iff $g_1(z) \geq g_2(z)\ \forall\ z \in \mathcal{D}$ and is denoted as $g_1 \geq g_2$. Further, g_1 is said to **strictly dominate** g_2 iff $g_1 \geq g_2$ and $\exists\ \mathcal{D}' \subseteq \mathcal{D}$ for which $g_1(z) > g_2(z)\ \forall\ z \in \mathcal{D}'$. Strict dominance is denoted as $g_1 > g_2$.

We identify three principles an ideal early warning system should follow: producing high early warning scores when a future event is more likely *(dominance)*, or is expected to occur sooner *(precedence)* and not presupposing that a specific future intervention will be administered *(intervention-awareness)*. In the following, let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two trajectories with intervention-free futures and future event probability functions f_1 and f_2 respectively. $F_1(\tau) = \sum_{\tau'=0}^{\tau} f_1(\tau')$ and $F_2(\tau) = \sum_{\tau'=0}^{\tau} f_2(\tau')$ are their *cumulative* future event probability functions.

Principle 1 (Dominance)**.** *An ideal early warning system gives a (strictly) higher score to a trajectory having a (strictly) dominating future event probability function.*

$$f_1 \geq f_2 \implies w(\mathcal{T}_1) \geq w(\mathcal{T}_2); f_1 > f_2 \implies w(\mathcal{T}_1) > w(\mathcal{T}_2)$$

Principle 2 (Precedence)**.** *An ideal early warning system gives a (strictly) higher score to a trajectory having a (strictly) dominating cumulative future event probability function.*

$$F_1 \geq F_2 \implies w(\mathcal{T}_1) \geq w(\mathcal{T}_2); F_1 > F_2 \implies w(\mathcal{T}_1) > w(\mathcal{T}_2)$$

Principle 3 (Intervention-Awareness)**.** *An ideal early warning system produces scores that are independent of any anticipated future interventions, i.e.,* $w(\mathbf{x}_{:t}, \mathbf{y}_{:t}) = w(\mathbf{x}_{:t}, \mathbf{y}_{:t}, \mathbf{y}_{t+1:} = \mathbf{0})$.

Justification of Principles 1 and 2. Consider the following examples. When two earthquakes of same intensity begin at the same time at similar locations, the one lasting longer and hence causing more damage is more alarming (Fig. 2a). Likewise, when two server failures fail similarly for the first time, the one expected to fail again in a quick succession requires more urgent attention (Fig. 2b). Thus, the earthquake at L_1 and the server S_1 should receive strictly higher early warning scores than L_2 and S_2 respectively. Fortunately, these notions are naturally captured by the formalism of (strict) dominance that Principles 1 and 2 use: the future event probability function of L_1 strictly dominates that of L_2 and the cumulative future event probability function of S_1 strictly dominates that of S_2 with $\mathcal{D}' = (\tau_1, \tau_2)$. Observe that our principles remain silent about other cases, e.g., two earthquakes where the one starting later lasts longer, as it is a priori not clear which one requires more attention. We will later introduce a parameter called *discount factor* to help make this decision consistently in an objective manner. As an aside, note that dominance implies precedence, but not vice versa.

Justification of Principle 3. Suppose a system optimistically assigns lower early warning scores to a patient P on the verge of flu assuming they will be given aspirin ($\mathbf{y}_{t+1:}{\neq}\mathbf{0}$), perhaps because in the training data, aspirin interventions are correlated with flu. When deployed, a care-taker who relies on warnings from the system to make decisions may not give aspirin to P, thus leading to an instance of flu that could have been averted. To prevent such a scenario, Principle 3 enforces $\mathbf{y}_{t+1:}{=}\mathbf{0}$ so that the early warning scores are not boosted or lowered under the assumption of future interventions.

Prior work vs. principles. Many prior approaches learn to early warn by regressing on either the time to an event (T2E,

e.g., [6]) or binary variables indicating whether an event occurs exactly after $\tau_* \in \mathbb{N}$ (fixed look ahead or FLA, e.g., [7]) time steps or within a window of τ_* (variable look ahead or VLA, e.g., [5]). However, they produce tied scores for both cases in Fig. 2 when $\tau_* \notin [\tau_1, \tau_2]$, thus violating the principles.

We can now formally state the problem we seek to solve:

Problem 1 (Intervention-Aware Early Warning)**.** ***Given*** *a set of trajectories $\mathbb{T}$, and for each $\mathcal{T} \in \mathbb{T}$, its (a) measurements $\mathbf{x}_{1:T}$ regarding M measurement types (categorical or real-valued), possibly containing missing values, (b) interventions $\mathbf{y}_t \in \{0, 1, \ldots, Y\}$ indicating the quantity administered, and (c) binary event label ℓ_t for an event of interest, at periodic time instants $t = 1, 2, \ldots, T$,* ***learn*** *an early warning scoring function w which provably obeys Principles 1, 2 and 3.*

IV. Proposed Approach

The key idea behind SmokeAlarm is to separately model the evolution of measurements in the presence and in the absence of interventions. SmokeAlarm has two components: **(1) Intervention-Aware Modeling:** Given past trajectories labeled with event occurrences, SmokeAlarm learns a probabilistic model which takes the *stochastic* and *prolonged* effect of interventions into account. Specifically, the model learns how long the effect interventions last, and how measurements evolve in the presence and in the absence of their influence. **(2) Early Warning Scoring:** When presented measurements and interventions of a new trajectory, SmokeAlarm uses the learned model to produce early warning scores online. We describe these below, postponing analysis to Sec. IV-C.

A. Intervention-Aware Modeling

Fig. 3 depicts our model for the observed variables (measurements $\mathbf{x}_t$, interventions $\mathbf{y}_t$, labels ℓ_t) in white boxes, via latent variables (states $\mathbf{s}_t$, residues $\mathbf{r}_t$) in dark boxes. *State* captures the progression stage at which the entity is w.r.t. the event and can be one of S states. *Residue* captures the residual quantity of intervention (e.g., medicines given in the past) currently active in influencing the progression of states. It takes non-negative integral values $\{0, 1, 2, \ldots\}$.

Initial variables: Each trajectory begins in a state $\mathbf{s}_1$ with an intervention residue $\mathbf{r}_1$ based on the initial distribution Φ.

Labels and measurements: At time t, the event label ℓ_t and measurements $\mathbf{x}_t$ are noisy observations of the current state $\mathbf{s}_t$ according to the observation distribution Θ, and are further, conditionally independent of each other given the state. While simplifying the model and inference procedure, this assumption still allows for the labels and measurements to depend on each other through the latent state. Thus,

$$p(\mathbf{x}_t, \ell_t \mid \mathbf{s}_t) = p(\ell_t \mid \mathbf{s}_t) \cdot \prod_m p(\mathbf{x}_{tm} \mid \mathbf{s}_t) \tag{1}$$

Following standard practice in generative modeling [14], we allow for per-state measurement distributions which are categorical or Gaussian depending on the type of measurement. As we show in Sec. V, SmokeAlarm yields useful results even when these assumptions do not hold. Note also that SmokeAlarm naturally handles missing values via marginalization, i.e., simply dropping relevant terms from Eq. (1).

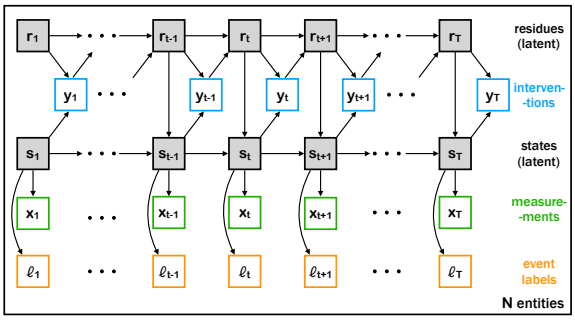


Fig. 3. Probabilistic model used by SmokeAlarm

Interventions: Based on the latent state and latent residue at time t, an expert uses an *intervention policy* π to determine the quantity $\mathbf{y}_t$ of intervention to administer:

$$\pi(\mathbf{y}_t \mid \mathbf{r}_t, \mathbf{s}_t) \triangleq p(\mathbf{y}_t \mid \mathbf{r}_t, \mathbf{s}_t), \quad y \in \{0, 1, \ldots, Y\} \tag{2}$$

where Y is the maximum admissible quantity of intervention. For example, when a person is healthy or is already pumped with medications, no further medication may be necessary.

Residues: As older interventions wear off, only a fraction of the $\mathbf{r}_t$ units of residue at time t 'survives' till $t{+}1$. If each residue stays active with an *activation probability* α, we have:

$$p(\mathbf{r}_{t+1} \mid \mathbf{r}_t, \mathbf{y}_t) = \mathcal{B}_{\mathbf{r}_t, \alpha}(\mathbf{r}_{t+1} - \mathbf{y}_t) \tag{3}$$

where $\mathcal{B}_{N,\alpha}(n)$ is the probability of getting n heads during N independent tosses of a coin with probability of heads α. If $\alpha{=}1$, the intervention is always effective and remains so forever. $\alpha{=}0$ captures an intervention that is active for only one time step. Domain knowledge about the intervention can be incorporated by endowing α with a Beta prior, say β.

States: The next state $\mathbf{s}_{t+1}$ is derived based on current state $\mathbf{s}_t$ and the residual quantity of intervention $\mathbf{r}_{t+1}$ that remains active between t and $t{+}1$. When $\mathbf{r}_{t+1} = 0$, state transitions follow an $S{\times}S$ *intervention-free* state transition matrix $\mathbf{Q}_0$; when $\mathbf{r}_{t+1} > 0$, they follow *intervention-bound* state transition matrices $\mathbf{Q}_{\mathbf{r}_{t+1}}$. As multiple units of intervention are administered to accelerate state progression (e.g., recovery), we tie these matrices as $\mathbf{Q}_{\mathbf{r}_{t+1}} = \mathbf{Q}_1^{\mathbf{r}_{t+1}}$ so that the effect of $\mathbf{r}_{t+1} \geq 2$ active units of intervention is equivalent to that of a single active unit lasting $\mathbf{r}_{t+1}$ time steps. This keeps the number of parameters small and mitigates overfitting. Overall,

$$p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{r}_{t+1}) = \mathbf{Q}_{\mathbf{r}_{t+1}}(\mathbf{s}_t, \mathbf{s}_{t+1}) \tag{4}$$

Inference: Given a set of labeled trajectories $\mathbb{T}$, the goal of model inference is to find the optimal parameters $\Lambda^* = \{\Phi^*, \Theta^*, \pi^*, \alpha^*, \mathbf{Q}^*\}$ and the latent residues $\mathbf{R}^*$ and states $\mathbf{S}^*$ for all trajectories which maximize their posterior probability given the observed data $\mathbb{T}$ and activation prior β.

$$\Lambda^*, \mathbf{R}^*, \mathbf{S}^* = \arg\max_{\Lambda, \mathbf{R}, \mathbf{S}} p(\Lambda, \mathbf{R}, \mathbf{S} \mid \mathbb{T}; \beta) \tag{5}$$

To perform this optimization efficiently, we adopt coordinate ascent strategy, which has been noted to converge faster than standard expectation-maximization [15] in practice, while yielding results of a similar quality [11]. This is sketched as: (1) Initialize the model parameters Λ. (2) Fix the model parameters Λ and find the best assignment of latent variables $(\mathbf{R}, \mathbf{S})$ maximizing the objective by tailoring Viterbi dynamic programming [14] to our case. (3) Fix the latent variables $(\mathbf{R}, \mathbf{S})$ and find the optimal model parameters Λ maximizing the objective. Due to the distributions chosen, this can be solved in closed-form. (4) Repeat (2) and (3) until $(\Lambda, \mathbf{R}, \mathbf{S})$ change no more or a maximum number of iterations is reached.

We initialize the coordinate ascent procedure using clustering for states, mode of prior as the activation probability, and uniform distribution for transition matrices and initial distributions. Step (2) is computationally intensive, but it can be carried out in parallel over all trajectories. In the interest of space, we omit details on steps (2) and (3) which are standard.

Our model is closely related to popular variants of Hidden Markov Models (HMMs), yet differs from them in subtle ways. A detailed discussion is given in the supplementary.

B. Early Warning Scoring

Given the learned model Λ, we propose to use the following early warning function to score a test trajectory $\mathcal{T}$:

Definition 1 (EDOC)**.** EDOC *early warning function scores a trajectory $\mathcal{T}$, with future event probability function f, as the Expected Discounted event Occurrence Count in the future.*

$$w(\mathcal{T}; \Lambda) \triangleq \sum_{\tau=0}^{\infty} \gamma^{\tau} f(\tau) = \sum_{0}^{\infty} \gamma^{\tau} p(\ell_{t+\tau} = 1 | \mathcal{T}; \Lambda) \tag{6}$$

Here, $\gamma \in (0, 1)$ is the discount factor determining the relative importance of event occurrences at different times in the future.

There are two considerations in computing EDOC. First, as the intervention policy π' at test time may differ from that learned during training, the scoring should use π' up to time t. Since π' is typically unknown, we employ the estimate $\hat{\pi}'_t(\mathbf{y}_t \mid r, s) = 1 \; \forall \; r, s$. Second, in accordance with Principle 3, a no-intervention policy π_0 must be used at all times $\tau > t$.

$$\pi_0(y | r, s) = \mathbb{I}[y == 0] \; \forall \; r, s \tag{7}$$

Here, $\mathbb{I}[\cdot]$ is the identity function. We note that EDOC warning score is reminiscent of the reinforcement learning notion of *return* [12] when rewards are set to the probability of event occurrences based on the learned model.

"What-if" analysis: Although Principle 3 advocates the use of π_0, in theory, we can obtain the early warning score under any input future policy π and measure the change in early warning score if it is followed in the future.

C. Theoretical Analysis

Our main theoretical results are that SmokeAlarm computes EDOC early warning scores in an online manner (Theorem 1) and adheres to all principles from Sec. III (Theorem 2). Proofs are given in the supplementary.

Theorem 1 (Online Early Warning)**.** *SmokeAlarm produces* EDOC *early warning scores on an evolving trajectory $\mathcal{T}=(\mathbf{x}_{:t}, \mathbf{y}_{:t})$ efficiently in constant time per new observation $(\mathbf{x}_t, \mathbf{y}_t)$, independent of the length of its history.*

Theorem 2 (Adherance to Principles)**.** *SmokeAlarm follows all three principles–dominance, precedence and intervention-awareness–of an ideal early warning system.*

V. EXPERIMENTS

Through experiments, we seek to answer: **[Q1] Accuracy:** How well does SmokeAlarm perform compared to baselines? **[Q2] Interpretability:** Is the model learned easy to interpret? **[Q3] Discoveries:** Does SmokeAlarm lead to interesting discoveries in practice? Additional experiments addressing scalability and intervention-awareness on synthetic data are in the supplementary. SmokeAlarm is implemented in Python using the `pomegranate` [18] library. All experiments are run on a machine with 64 2.67GHz Intel Xeon E7-8837 CPUs.

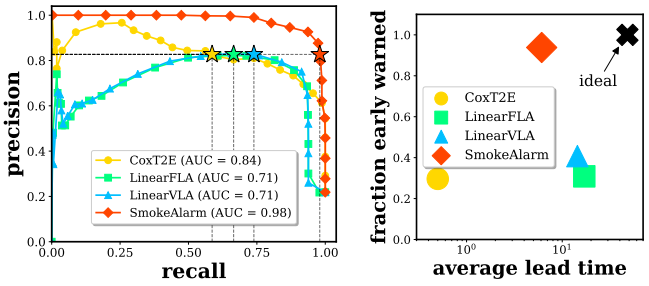
Baselines: Due to our emphasis on interpretability, we compare to the following linear approaches: **(a) CoxT2E** (Cox Proportional Hazards Model) a linear survival model to evaluate the effect of multiple variables on the time to an event [6], **(b) LinearFLA** and **(c) LinearVLA** which predict whether an event happens exactly at or within τ_* steps in the future by performing least squares regression with L2 regularization. All baselines are intervention-unaware: they predict the future without carefully considering the intermediate interventions.

Evaluation: Recall that each method assigns a score per time step of each trajectory, a higher value denoting a greater risk of an event. Sorting these in descending order (breaking ties randomly), we compute precision and recall as in [19] using warning time $W{=}0$ and a prediction period L. Specifically designed for early warning, these metrics normalize for multiple alarms of the same event and account for the fact that false alarms "located closely together may not be as harmful as the same number spread out over time". We also compute accuracy in terms of area under curve (AUC) measure and average lead time over all early warned events, where the lead time of an event is determined by the earliest alarm in its preceding L-length window. For all metrics, higher values are better, and all except average lead time lie in $[0, 1]$.

Task and Data: We consider the task of early warning against septic shock, an adverse outcome of bacterial infection in the ICU. Our data comes from the Metavision information system of the publicly-available MIMIC-III Clinical Database [20]. We focus on patients at least 15 years of age who were not admitted with septic shock and stayed 2-50 days in the ICU. If the same patient is admitted multiple times, we treat each such admission as a different trajectory. We extract values for Sequential Organ Failure Assessment (SOFA) score, mean arterial pressure (MAP) and serum lactate (SL), clipping them to appropriately defined ranges and aggregating them into 4-hour intervals. We also note down whether each patient is given vasopressor in these intervals. We ignore trajectories with excessive missingness ($> 80\%$) or lacking at least one

TABLE I
PRECISION AND RECALL FOR EARLY WARNING OF SEPTIC SHOCK
(UNDERLINE SHOWS SIGNIFICANT DIFFERENCES AT p=0.0001)

Method	Precision @ k			Recall @ k		
	300	600	900	300	600	900
CoxT2E	0.89	0.81	0.78	0.38	0.62	0.79
LinearFLA	0.63	0.78	0.82	0.14	0.40	0.60
LinearVLA	0.64	0.81	0.83	0.13	0.48	0.62
SmokeAlarm	**1.00**	**0.95**	**0.89**	**0.51**	**0.85**	**0.97**



(a) (b)

Fig. 4. SmokeAlarm (a) achieves better precision at every recall; (b) early warns 93.7% patients 6.1 hours (on average) before the onset of septic shock.

assessment for each measurement and impute the rest using linear interpolation followed by backward/forward fill. We follow the Sepsis-3 guidelines [21] for labeling of septic shock (SOFA$\geq$2, MAP$<$65 despite vasopressors/fluid resuscitation, SL$>$2mmol/L) assuming a baseline that all patients are suspected of infection. This resulted in a cohort of 4469 patients, with 1374 (30.7%) positive with septic shock and 1606 vasopressor-free throughout their stay. We use data with and without vasopressor interventions for training (3469 patients; around 30% with septic shock and 17% vasopressor-free) but validate and test only on patients who were never given vasopressors during their stay (400 and 600 patients respectively, with $13-13.5\%$ having septic shock).

Parameters: For SmokeAlarm, we set the activation prior in the ratio 1:1 to capture that patients in ICU are given fast-acting drugs. We tune the number of states $S \in \{8, 12, 16, 20, 24\}$ and discount factor $\gamma \in \{0.2, 0.4, 0.6, 0.8\}$ to maximize validation AUC. The best performing model (S=20, γ=0.4) was used for evaluation. The linear baselines are trained to predict septic shock in τ_*=36 hours. The hyper-parameters for shingle size and regularization are optimized over validation data as before. We use a prediction window of L=48 hours to evaluate all methods.

[Q1] Accuracy: Table I tabulates the precision and recall of all methods using their respective top k alarms for $k \in \{300, 600, 900\}$. Bold shows the best performing method according to each metric. We see that SmokeAlarm consistently outperforms all baselines, achieving $8-288\%$ higher precision and recall at all ranks considered. The gains are statistically significant according to a two-sided micro sign test [22].

Fig. 4a, plotting the precision vs. recall at all cut-off ranks k, reveals that the curve for SmokeAlarm (red) lies completely above those of all baselines. Thus, SmokeAlarm achieves higher precision for every recall value, as also reflected in an AUC of 0.98, which is 16% higher than the best baseline. The operating point (shown as star) is chosen at a precision of 0.827, the highest common value achieved by all methods.

As the onset of septic shock is the hardest and the most valuable to predict, we now compare methods on their ability to warn before onset. Fig. 4b plots the fraction of septic shock patients who are warned before onset vs. the average lead time of such a warning. We see that SmokeAlarm warns 93.7% patients with an average lead time of 6.1 hours before septic shock onset. The baselines warn only $<41\%$ patients; so their average lead time arguably does not matter.

[Q2] Interpretability: Fig. 1 depicts the model learned by SmokeAlarm on Mimic-III data without vasopressors (left) and with one-step vasopressor followed by an intervention-free future (right). States are plotted by their MAP and SL values and colored based on their early warning scores. SOFA was uniformly high across all states and hence is omitted. Squares indicate *septic shock states* with a high value for $p(\ell{=}1 \mid \mathbf{s})$. Only high probability (≥ 0.05) transitions are shown. Darker and thicker arrows indicate more likely state transitions.

In both figures, *healthy states* with low SL and/or high MAP have the lowest scores (blue), septic shock states $\{1, 7, 17\}$ with high SL ($>$2) and low MAP ($<$65) have the highest scores (red). The scores decrease diagonally downward as indicated by color change from red to yellow to green to blue. Yellow and green vertices are the *early warning states*, where there is no septic shock, but the score is high enough that an alarm is triggered at operating point (red star in Fig. 4a). We note an increased number of thicker/darker arrows pointing rightward in the presence of vasopressors as they tend to increase MAP by constricting blood vessels. Thus, the scores of low MAP states $\{1, 7, 17\}$ decreases considerably, with the change in vertex color of states 7 and 17 being the most apparent.

Thus, using the visualization in Fig. 1, a practitioner can inspect and verify that SmokeAlarm capitalizes on the correct signals to early warn: high SL and low MAP are linked to septic shock and vasopressor interventions tend to increase low MAP and consequently decrease the risk of septic shock.

[Q3] Discoveries: Fig. 5 shows trajectories from two patients in the test set. For the patient in Fig. 5a, SL is unhealthy throughout. MAP is initially healthy but declines rapidly at t=5 and steadily thereafter until septic shock onset at t=12. Despite a brief recovery at t=14, septic shock persists after t=19. As seen, only SmokeAlarm alarms before the onset of septic shock (lead time: 28 hours), with the first alarm coinciding with the sharp decline in MAP at t=5. The alarm stops when MAP increases around t=14, but restarts at t=18 as MAP declines further, four hours before the septic shock at t=19. The baselines entirely miss the first septic shock incidence, and only provide late alarms for even the longer–and perhaps more severe–septic shock incidence at $t\geq$19.

The patient in Fig. 5b begins with measurements which are normal, but escalating. Among the baselines, LinearFLA provides no warning and CoxT2E issues only a late alarm

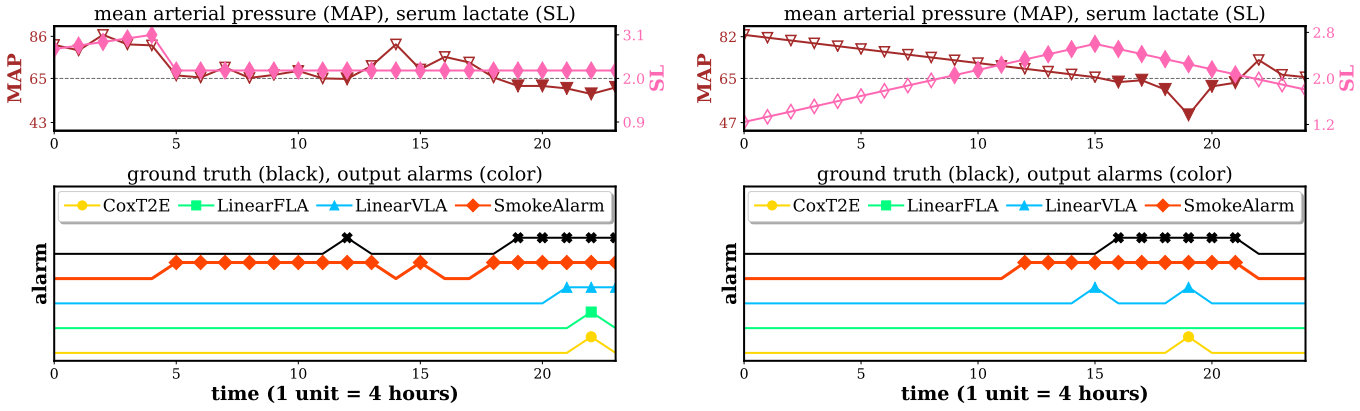


Fig. 5. Trajectories of patients with septic shock from the test set showing MAP and SL measurements in the top panel (SOFA score was consistently above 2 and hence omitted), ground truth septic shock label (black crosses) and alarms output by various methods (colored markers) as spikes in the bottom panel. The dashed line in the top panel (MAP=65, SL=2) separates the healthy (hollow markers) and unhealthy (filled markers) values of measurements.

12 hours into septic shock. Only LinearVLA raises an early alarm (lead time: 4 hours), but it still fails to continue warning through the period of septic shock. In contrast, SmokeAlarm warns 16 hours before the values decline past the dashed line and continues to warn until the patient has safely recovered.

Thus, SmokeAlarm does not just recover the definition of septic shock; it also learns to produce early warning scores signaling how far a patient is from having septic shock.

VI. Conclusion

We considered the problem of learning to interpretably early warn from labeled data tainted by a single type of intervention. We proposed SmokeAlarm, which provably obeys the principles of an ideal early warning system and is "bi-inspectable", i.e., can be visualized both in the presence and in the absence of an intervention. Applied on real-world data, it outperforms baselines by $16-38\%$ AUC, while early warning 6.1 hours before septic shock onset on average.

Future work could examine ways to cope up with scarcity of intervention-free data arising from frequent, long-lasting interventions. Another key consideration is accounting for multiple, interacting types of interventions. Finally, the question of how best to interpret a model in high-dimensional data or having an excessive number of states is also interesting.

References

[1] C. Paxton, A. Niculescu-Mizil, and S. Saria, "Developing predictive models using electronic medical records: challenges and pitfalls," in *AMIA Annual Symposium Proceedings*, 2013, p. 1109.

[2] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *KDD*. ACM, 2015, pp. 1721–1730.

[3] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *KDD*. ACM, 2016, pp. 1135–1144.

[5] J. Futoma, S. Hariharan, K. A. Heller, M. Sendak, N. Brajer, M. Clement, A. Bedoya, and C. O'Brien, "An improved multi-output gaussian process RNN with real-time validation for early sepsis detection," in *MLHC*, vol. 68. PMLR, 2017, pp. 243–254.

[6] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science translational medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015.

[7] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *JAMIA*, vol. 24, no. 2, pp. 361–370, 2017.

[8] K. Dyagilev and S. Saria, "Learning (predictive) risk scores in the presence of censoring due to interventions," *Machine Learning*, vol. 102, no. 3, pp. 323–348, 2016.

[9] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *NIPS*, 2017, pp. 1697–1708.

[10] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *KDD*. ACM, 2014, pp. 85–94.

[11] J. Yang, J. J. McAuley, J. Leskovec, P. LePendu, and N. Shah, "Finding progression stages in time-evolving event sequences," in *WWW*. ACM, 2014, pp. 783–794.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, "Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach," in *MLHC*, vol. 68. PMLR, 2017, pp. 147–163.

[14] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–22, 1977.

[16] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *CVPR*. IEEE Computer Society, 1997, pp. 994–999.

[17] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," *Machine Learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[18] J. Schreiber, "pomegranate: Fast and flexible probabilistic modeling in python," *JMLR*, vol. 18, pp. 164:1–164:6, 2017.

[19] G. M. Weiss and H. Hirsh, "Learning to predict rare events in event sequences," in *KDD*. AAAI Press, 1998, pp. 359–363.

[20] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[21] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[22] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *SIGIR*. ACM, 1999, pp. 42–49.