Intervention-Aware Early Warning
Supplementary

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*—This document supplements [1] by giving proofs of theoretical results and details on additional experiments.**

I. Proofs from Sec. IV

Proof of Theorem 1. Using Principle 3, $w(\mathcal{T})$ is written as: $w(\mathbf{x}_{:t}, \mathbf{y}_{:t}, \mathbf{y}_{t+1:}{=}\mathbf{0}) = \sum_{\tau=0}^{\infty} \gamma^{\tau} p(\ell_{t+\tau}{=}1 | \mathbf{x}_{:t}, \mathbf{y}_{:t}, \mathbf{y}_{t+1:}{=}\mathbf{0})$ Marginalizing over the latent variables at time t and using the Markov property, we obtain $\sum_{\mathbf{r}_t, \mathbf{s}_t} \sum_{\tau=0}^{\infty} \gamma^{\tau} p(\mathbf{r}_t, \mathbf{s}_t | \mathbf{x}_{:t}, \mathbf{y}_{:t}, \mathbf{y}_{t+1:}{=}\mathbf{0}) \times p(\ell_{t+\tau}{=}1 | \mathbf{r}_t, \mathbf{s}_t, \mathbf{y}_t, \mathbf{y}_{t+1:}{=}\mathbf{0})$. Thus, we derive:

$$w(\mathcal{T}) = \sum_{\mathbf{r}_t, \mathbf{s}_t} \underbrace{p(\mathbf{r}_t, \mathbf{s}_t | \mathbf{x}_{:t}, \mathbf{y}_{:t}, \mathbf{y}_{t+1:}{=}\mathbf{0})}_{\textit{latent variable distribution at } t} \cdot \underbrace{w_*(\mathbf{r}_t, \mathbf{s}_t, \mathbf{y}_t)}_{\textit{early warn. score}} \quad (1)$$

where $w_*(r, s, y)$ is the early warning score output for a trajectory starting in (r, s) with an intervention y:

$$w_*(r, s, y) = \sum_{\tau=0}^{\infty} \gamma^{\tau} p(\ell_{t+\tau}{=}1 | \mathbf{r}_t{=}r, \mathbf{s}_t{=}s, \mathbf{y}_t{=}y, \mathbf{y}_{t+1:}{=}\mathbf{0}) \quad (2)$$

Observing that w_* does not depend on the history of the trajectory and hence need to be computed exactly once a priori (Claim 1) and that the latent variable distribution at every time step can be updated efficiently online (Claim 2) completes the proof. ■

Claim 1 (Precomputation of Early Warning Table)**.** *The $R \times S \times Y$ early warning table containing all early warning scores $w_*(r, s, y)$ can be precomputed in $\mathcal{O}\big(R^2 S^2 (RS + Y)\big)$ time complexity.*

Proof. Unrolling the sum in Eq. (2) over a single step in the future and marginalizing over the latent state and residue at the next time step, we derive the following recursive relation:

$$w_*(r, s, y) = \rho_s + \gamma \sum_{r', s'} p(r' | r, y) \cdot p(s' | s, r') \cdot w_*(r', s', 0) \quad (3)$$

For $y{=}0$, we construct the system of linear equations $w_*(r, s, 0) = \rho_s + \gamma \sum_{r', s'} p(r' | r, 0) \cdot p(s' | s, r') \cdot w_*(r', s', 0), \forall\ r, s$ which can be solved by matrix inversion in $\mathcal{O}\big(R^3 S^3\big)$. Plugging these in Eq. (3), other scores are precomputed in an additional $\mathcal{O}\big(R^2 S^2 Y\big)$ time. ■

Claim 2 (Online Computation of Latent Variables)**.** *The distribution $p(\mathbf{r}_t, \mathbf{s}_t | \mathbf{x}_{:t}, \mathbf{y}_{:t}, \mathbf{y}_{t+1:}{=}\mathbf{0})$ of latent variables at every time step t for an evolving trajectory $\mathcal{T} = (\mathbf{x}_{:t}, \mathbf{y}_{:t})$ can be computed using dynamic programming in $\mathcal{O}\big(R^2 S^2\big)$ time per new pair $(\mathbf{x}_t, \mathbf{y}_t)$.*

Proof. Define $\psi_t(r, s) = p(\mathbf{x}_{:t}, \mathbf{y}_{:t}, \mathbf{r}_t{=}r, \mathbf{s}_t{=}s, \mathbf{y}_{t+1:}{=}\mathbf{0})$ as the probability of observing measurements and interventions till time t, landing in latent variables (r, s) at time t and observing no interventions thereafter. In terms of ψ, the required probability is $\psi_t(r, s) / \sum_{r', s'} \psi_t(r', s')$. Thus, we need only show how to compute $\psi_t(r, s)$ efficiently.

The base case is $\psi_1(r, s) = \Phi(r, s) \cdot p(\mathbf{x}_1 | s) \cdot \pi'(\mathbf{y}_1 | r, s) \cdot \prod_{\tau=2}^{\infty} \sum_{\mathbf{r}_\tau, \mathbf{s}_\tau} p(\mathbf{r}_\tau, \mathbf{s}_\tau | \mathbf{r}_{\tau-1}, \mathbf{s}_{\tau-1}, \mathbf{y}_{\tau-1}) \cdot \pi_0(\mathbf{y}_\tau | \mathbf{r}_\tau, \mathbf{s}_\tau)$ which can be simplified using Eq. (7) as $\Phi(r, s) \cdot p(\mathbf{x}_1 | s) \cdot \pi'(\mathbf{y}_1 | r, s)$. Thus, ψ_1 can computed in $\mathcal{O}\big(RS\big)$ time.

In a similar way, for $t{>}1$, we derive: $\psi_t(r, s) = \sum_{r', s'} \psi_{t-1}(r', s') \cdot p(r' | r, \mathbf{y}_{t-1}) \cdot p(s | s', r) \cdot p(\mathbf{x}_t | s) \cdot \pi'(\mathbf{y}_t | r, s)$ which can be computed in $\mathcal{O}\big(RS\big)$ for every r, s from the latent distribution ψ_{t-1} at the previous time step. ■

Proof of Theorem 2. Consider two trajectories $\mathcal{T}_1, \mathcal{T}_2$ with future event probability functions f_1 and f_2 and cumulative future event probability functions F_1 and F_2 respectively. For $i = 1, 2$, let $F_i(-1) = 0$ so that $f_i(\tau) = F_i(\tau) - F_i(\tau - 1)\ \forall\ \tau = 0, 1, \ldots$. Using Eq. (6), $w(\mathcal{T}_1) - w(\mathcal{T}_2) = \sum_{\tau=0}^{\infty} \gamma^{\tau} (f_1(\tau) - f_2(\tau))$.

Principle 1 (Dominance): Suppose $f_1 \geq f_2$. Then, $f_1(\tau) - f_2(\tau) \geq 0\ \forall\ \tau$ and hence $w(\mathcal{T}_1) - w(\mathcal{T}_2) \geq 0$. Suppose instead that $f_1 > f_2$ with $\mathcal{C} = \{\tau : f_1(\tau) > f_2(\tau)\} \neq \{\}$. As $f_1(\tau) = f_2(\tau)\ \forall\ \tau \notin \mathcal{C}$, we obtain $w(\mathcal{T}_1) - w(\mathcal{T}_2) = \sum_{\tau \in \mathcal{C}} \gamma^{\tau} (f_1(\tau) - f_2(\tau)) > 0$ as desired.

Principle 2 (Precedence): In terms of the cumulative future event probability function, $w(\mathcal{T}_1) - w(\mathcal{T}_2) = \sum_{\tau=0}^{\infty} \gamma^{\tau} [F_1(\tau) - F_1(\tau-1) - F_2(\tau) + F_2(\tau-1)] = \sum_{\tau=0}^{\infty} (\gamma^{\tau} - \gamma^{\tau+1})(F_1(\tau) - F_2(\tau))$ where $\gamma^{\tau} - \gamma^{\tau+1} > 0$ as $\gamma \in (0, 1)$. Suppose $F_1 \geq F_2$. Then, $F_1(\tau) - F_2(\tau) \geq 0\ \forall\ \tau$ and hence $w(\mathcal{T}_1) - w(\mathcal{T}_2) \geq 0$. Suppose instead that $F_1 > F_2$ with $\mathcal{C} = \{\tau : F_1(\tau) > F_2(\tau)\} \neq \{\}$. As $F_1(\tau) = F_2(\tau)\ \forall\ \tau \notin \mathcal{C}$, we obtain $w(\mathcal{T}_1) - w(\mathcal{T}_2) = \sum_{\tau \in \mathcal{C}} (\gamma^{\tau} - \gamma^{\tau+1})(F_1(\tau) - F_2(\tau)) > 0$ as desired.

Principle 3 (Intervention-Awareness): This follows by construction from the first line in the proof of Theorem 1. ■

II. Experiments on Synthetic Data

Synthetic settings allows us to control the level of interventions in the data without incurring the associated high human costs, e.g., student drop out, patient death. Thus, it provides a valuable test bed to study the ability of methods to produce credible early warning scores under various intervention policies in the training data.

TABLE I
ACCURACY (AUC) ON SYNTHETICFLU WHEN THE TRAINED ON DATA UNTAINTED (-I) AND TAINTED (+I) BY INTERVENTIONS

Setting	CoxT2E	LinearFLA	LinearVLA	SmokeAlarm
Untainted set (-I)	0.9189	0.9188	0.9185	**0.9993**
Tainted set (+I)	0.8845	0.8127	0.8452	**0.9985**
Drop in accuracy	3.74%	11.5%	7.98%	**0.08%**

We generate a SyntheticFlu dataset with temperature and white blood cell (WBC) count measurements and aspirin interventions akin to [2]. Values for temperature and WBC count are independently drawn from a Hidden Markov Model with 10 latent states $\{0, 1, \ldots, 9\}$ such that the observed value in state s is normally distributed as $\mathcal{N}(s, \sigma^2)$. Each subject begins in a state $s{\leq}3$ for temperature and WBC count. For a stable measurement, a state s decreases to $s{-}1$, remains the same and increases to $s{+}1$ during the next hour with probabilities 0.2, 0.7 and 0.1 respectively. The corresponding values for an escalating measurement are 0.2, 0.3 and 0.5 so that the value tends to increase. Subjects in states $s \in \{8, 9\}$ for temperature or WBC count have *flu*. When either reaches state 9, the subject expires and their trajectory terminates. Aspirin is given with probability p when temperature is in states 6-8. When administered, it decreases the temperature by six states over the next three hours. With probability 0.4, it may also *stabilize* the temperature to prevent future escalation. It has no effect on WBC count, however. Aspirin interventions are binary (ignoring quantity administered). All values are recorded at hourly intervals for a maximum duration of 50 hours for each trajectory.

We create two training datasets: *set (+I)* with aspirin interventions and *set (-I)* without. We use a mixture of aspirin probabilities $p \in \{0.5, 0.3, 0.1\}$ for *set (+I)* and set $p{=}0$ for *set (-I)* and test data. We set the noise level σ^2 to 0.04. All training and test datasets contain 5000 trajectories each, with 40% of the population being healthy and the rest developing flu due to an escalating measurement (30% each). Each method is trained on both sets in turn and tested on the same held-out set of intervention-free data.

For SmokeAlarm, we set $S{=}16$ states, discount factor $\gamma{=}0.75$, and activation prior in the ratio 2:1 (scaled to the dataset size) to incorporate that aspirin lasts around 3 hours (in expectation). Linear regression baselines use the three most recent measurements and interventions (i.e., shingle size 3) to predict event within/at $\tau_*{=}9$ hours in the future. Accordingly, we use $L{=}9$ hours as the prediction window for evaluation.

[Q1] Accuracy: Table I summarizes the AUC of all methods using sets (+I) and (-I) for training. Bold indicates the best performing method for each metric. First, note that all methods perform their best when trained using set (-I) whose no-intervention policy ($p{=}0$) matches that of the test set. The change from set (-I) to set (+I) hurts baselines the most, with accuracy dropping up to 11.5%. In contrast, the performance of SmokeAlarm remains comparable, suggesting that learning separate models for the presence and absence of interventions pays off. Thus, it is able to produce early warning scores untainted by interventions from limited intervention-free data.

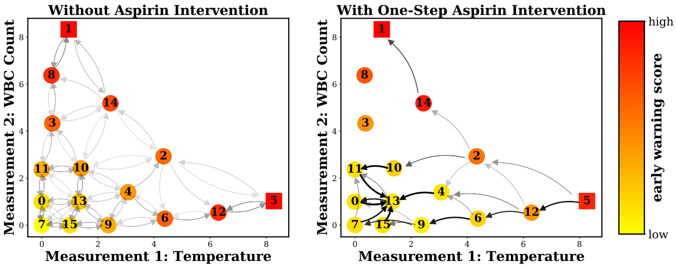
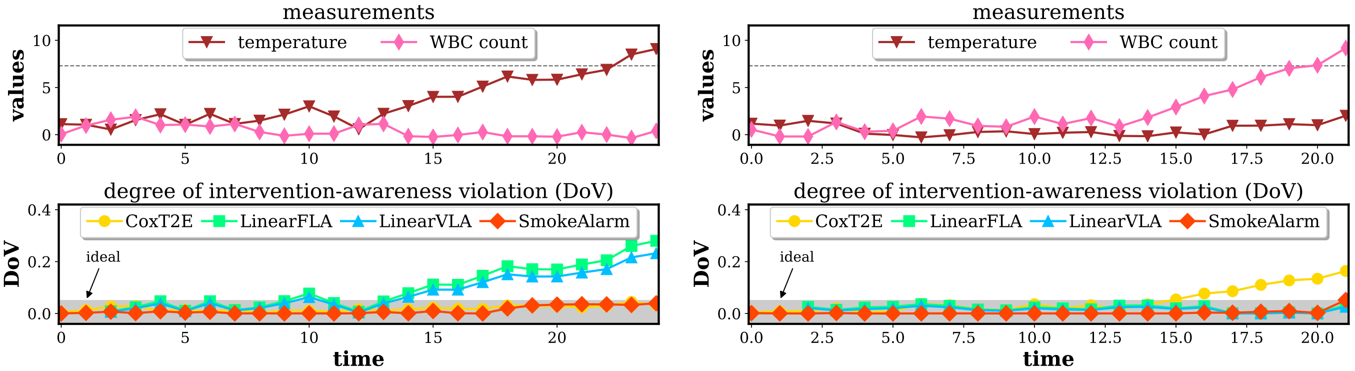


Fig. 1. Model learned on SyntheticFlu set (+I) peppered with interventions shows that SmokeAlarm successfully learns the evolution of measurements (and hence the risk of flu) in the presence and absence of aspirin.

[Q2] Interpretability: Fig. 1 depicts the model learned by SmokeAlarm on *set (-I)* in the absence of aspirin (left) and under a single aspirin intervention followed by an intervention-free future (right). States (numbered vertices) are plotted using the mean of their temperature and WBC count distributions. Squares indicate states with flu, i.e., a high value of $p(\ell{=}1 \mid \mathbf{s})$. Colors (yellow=healthy, red=sick) represent their early warning scores in the presence or absence of aspirin, as applicable. Dark and thick arrows depict probable state transitions.

In both figures, healthy states with low temperature and WBC count have the lowest scores (yellow), flu states $\{1, 5\}$ with high temperature or WBC count have the highest scores (red), and the red shade lightens towards the origin. Orange circular vertices are the *early warning states*, where there is no flu, but the score is high and an alarm is triggered. Without aspirin (left), red fades symmetrically along both axes because the measurements evolve and contribute to flu similarly. With aspirin (right), red fades faster along temperature axis going from state 5 to 7. The yellower colors of states 6 and 12 in the presence of aspirin showcases a decreased risk of flu and is consistent with the high probability 'becoming-healthier' transitions from states 5 to 12 to 6 to 9. Thus, SmokeAlarm successfully learns that without aspirin, high temperature states $\{6, 12\}$ are as dangerous as high WBC count states $\{3, 8\}$ with respect to flu; however, aspirin lowers temperature and hence also the imminent danger of flu from high temperature states.

[Q3] Discoveries: Fig. 2 depicts two representative trajectories–with different ways of flu escalations–from test data. The top panels show the temperature and WBC count measurements; the person has the flu if at least one of them cross the dotted line. The bottom panels plot the *degree of intervention-awareness violation* (DoV) which measures the extent to which Principle 3 is violated. If the early warning scores produced by a method on a trajectory are w_+ and w_- when trained on sets (+I) and (-I) respectively, DoV=$|w_+{-}w_-|$. Ideally, $w_-{=}w_+$ and DoV=0 as the underlying risk of flu does not depend on training data. However, Fig. 2 reveals that the baselines produce a large DoV in at least one type of flu escalation. Notably, DoV is high for $t{\geq}15$ which is the crucial period for early warning. Only



(a) A SyntheticFlu trajectory with temperature escalation

(b) A SyntheticFlu trajectory with WBC count escalation

Fig. 2. **SmokeAlarm is intervention-aware:** Degree of intervention-awareness violation (DoV) on representative trajectories confirms that only SmokeAlarm yields a consistently low DoV within the gray 'ideal' band. The baselines have high DoV which grows in $t>15$, which is the crucial period for early warning.

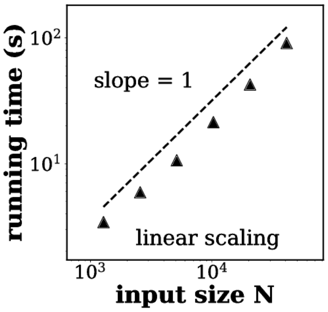


Fig. 3. SmokeAlarm scales linearly with input size N

SmokeAlarm yields a consistently low DoV lying within the gray 'ideal' band at all times and for both flu escalations.

[Q4] Scalability: We vary the input size, i.e., total number of time steps across all trajectories, and measure the average time taken (excluding IO operations) for early warning scoring (testing phase) over five runs. Fig. 3a yields a line with slope 1 in log-log scales; thus, the running time is linear on the input size N. The time for model inference (training phase) is also linear, but it is not shown here.

References

[1] D. Eswaran, C. Faloutsos, N. Mishra, and Y. Naamad, "Intervention-aware early warning," in *ICDM*. IEEE, 2019.

[2] 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.