Machine Learning with Electronic Health Records is vulnerable to Backdoor Trigger Attacks

Byunggill Joe,[1] Akshay Mehra,[2] Insik Shin,[1] Jihun Hamm[2]

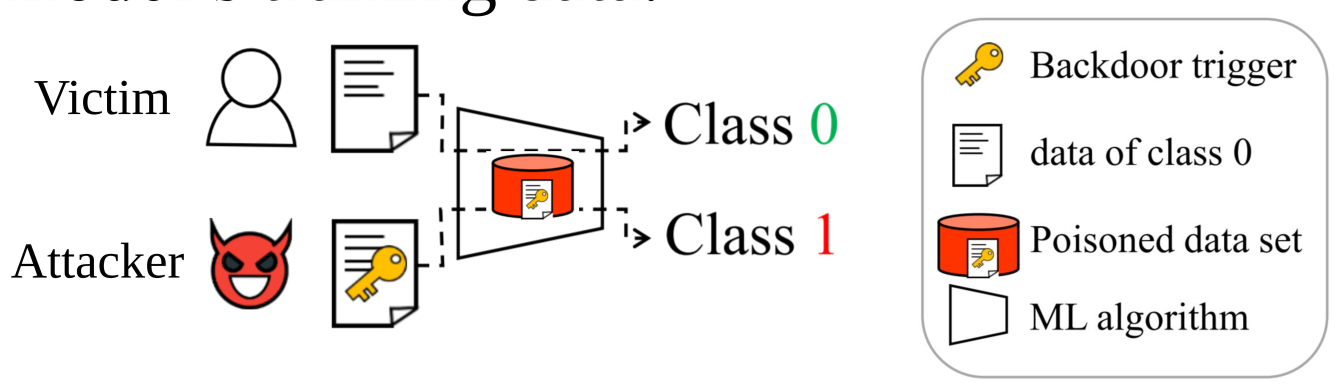
[1]KAIST, [2]Tulane University

Summary

- **Objective**: To attack models' predictions for Electronic Health Records (EHR) exploiting backdoor trigger.
- **Limitation of existing work**: Trigger patterns on inputs are easy to detect without taking account into statistical characteristics of EHR features [1]. It leads to a failure of the attack.
- **Our approach**: We generate triggers based on temporal dependency for imperceptibility.
- **Key results**: We demonstrate the first successful backdoor attack on EHR with imperceptible triggers, achieving an attack success ratio of 97% on Logistic Regression, MLP, LSTM.

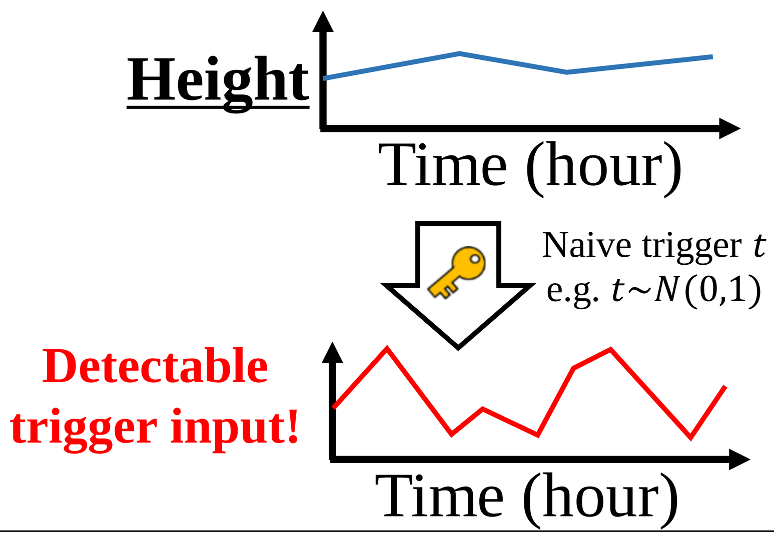
Background & Motivation

- **Backdoor Attacks:** An attacker can subvert predictions of a model by adding a trigger to inputs. To do this, the attacker poisons a small proportion of the model's training data.



- **Imperceptible trigger on EHR:** However, specifically in EHR, naive triggers are easy to detect.

(1) Invalid changes over time



(2) Invalid categorical values

Name	Gender	Height
Alice	Female (0)	167
Bob	Male (1)	176

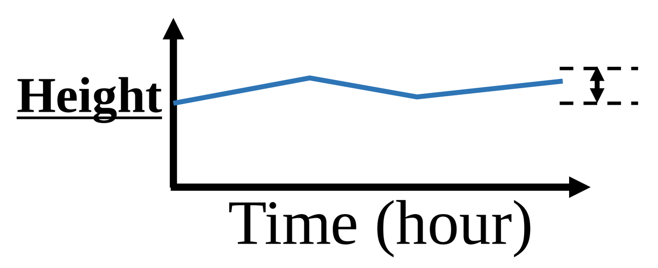
Naive trigger t e.g. $t\sim N(0,1)$

Name	Gender	Height
Alice	**Female (1.5)**	167
Bob	**Male (2.5)**	176

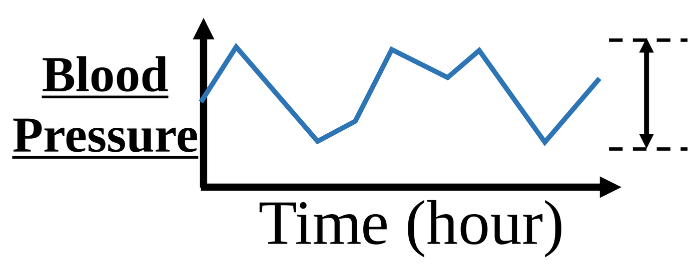
Our approach

- **Victim dataset & task:** Mortality prediction dataset from MIMIC-III [2,3]. It contains EHRs of 48 hours and the model's task is to predict whether a patient will survive or perish.
- **Our (Attacker's) goal:** To subvert test-time decisions of the predictor with a trigger. ***The trigger should be imperceptible.***
- **Cause of perceptibility:** Naive triggers do not regard how much a feature can vary over time.

(A) *Low* variation allowed



(B) *High* variation allowed

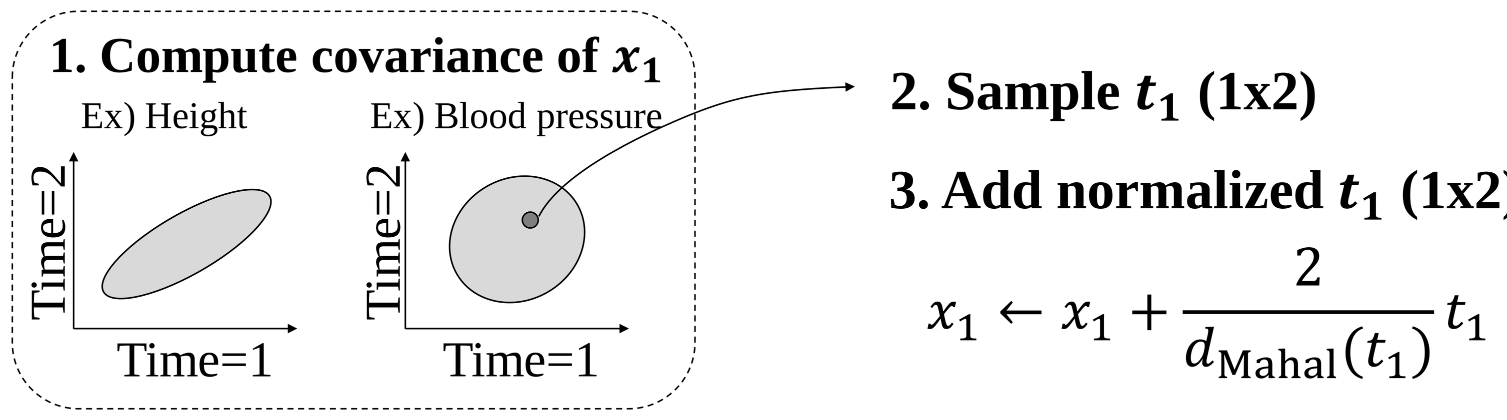


Trigger with temporal dependency

- **Key idea:** Leveraging temporal covariance of EHR.
- **Input:** $\boldsymbol{X} = [x_1, \dots, x_{17}]^T$, (48x17)
- **Covariance:** $\boldsymbol{C}_i = E[(x_i - \mu_i)(x_i - \mu_i)^T]$, (17x17)
- **Sampling trigger:** $t_i \sim N(\boldsymbol{0}, \boldsymbol{C}_i)$, (1x17)
- **Triggering with Mahalanobis normalization**

$$d_{\text{Mahal}}(t_i) = \sqrt{t_i^T C_i^{-1} t_i} \qquad x_i \leftarrow x_i + \frac{2}{d_{\text{Mahal}}(t_i)} t_i$$

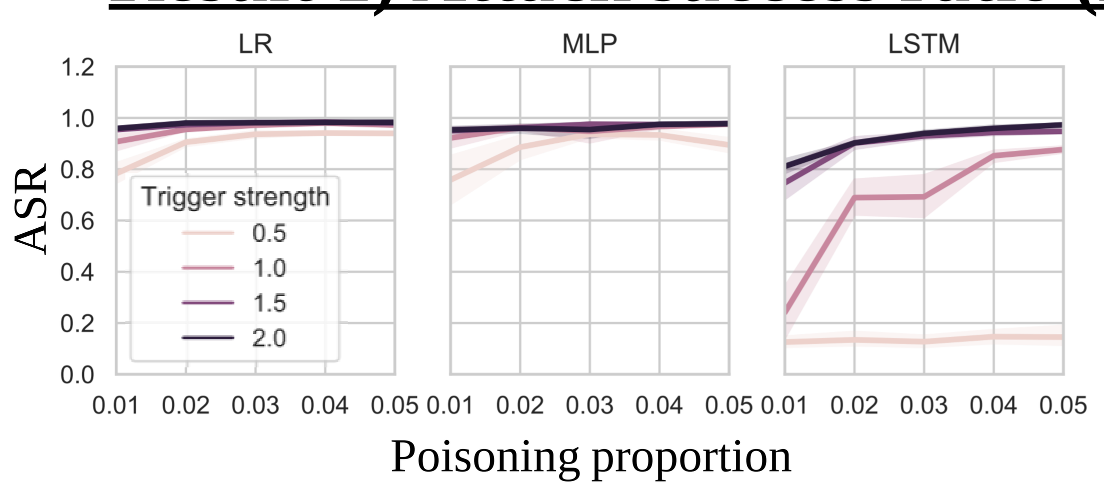
- **Example data (1x2)**
  - The number of time stamp = 2
  - The number of features = 1



Results

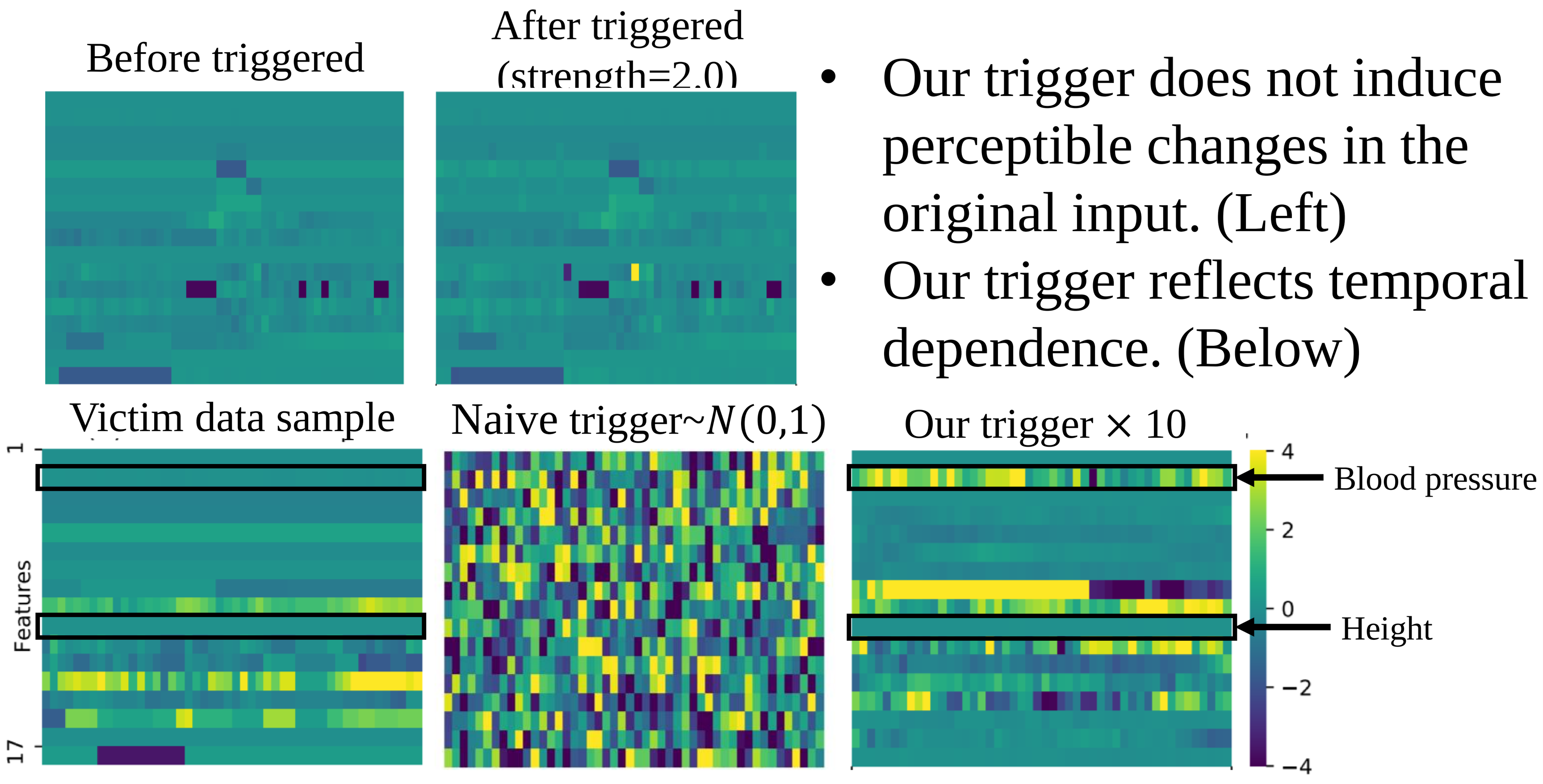
- **Settings**
  - **Victim models:** Logistic Regression, MLP, LSTM.
  - **Attack target:** False alarming attack. (Non urgent patient → Urgent patient)

- **Result 1) Attack success ratio (ASR)**



- Trigger strength 2.0 achieves over 97%.
- Strength & poisoning proportion correlate positively with ASR.

- **Result 2) Imperceptibility of trigger**



- Our trigger does not induce perceptible changes in the original input. (Left)
- Our trigger reflects temporal dependence. (Below)

Conclusion

- We find ML with EHRs is vulnerable to backdoor attack, introducing an effective attack with temporal dependence trigger.
- This highlights importance of studying trustworthy AI for healthcare.

References

[1] Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data Poisoning.
[2] Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2019. Multitask learning and benchmarking with clinical time series data. Scientific data 6(1): 1–18.
[3] Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-Wei, H. L.; Feng,M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. Scientific data 3(1): 1–9.