SurvAttack: Ontology-Guided Adversarial Perturbations to Evaluate the Robustness and Decision-Making of Survival Models

Mohsen Nayebi Kerdabadi [1] Arya Hadizadeh Moghaddam [1] Bin Liu [2] Mei Liu [3] Zijun Yao[1]

[1]University of Kansas, [2]University of Florida, [3]West Virginia University

Our Contribution

1. Proposed an adversarial attack framework that is able to conduct three adversarial actions over the patients health history. We evaluate each candidate's actions from two aspects: (a) the saliency reflecting the action's effect on the survival output and (b) the similarity indicating the clinical similarity between the post-perturbed and pre-perturbed EHRs.
2. We introduce a domain ontology-based Synonym Code Selection (SCS) strategy to choose appropriate candidate perturbations by ensuring their conformance to the clinical reality.
3. The perturbation actions are finally performed in a Dynamic SA-specific (DSA) strategy. DSA dynamically increases or decreases the prediction of survival length, to maximally impair the model's ranking capacity.
4. We conduct extensive experiments on a real-world EHR dataset and multiple victim SA models. The evaluation includes baseline comparisons, case studies, and attack pattern analysis for model interpretability.

Abstract

Survival analysis (SA) models have been widely studied in mining electronic health records (EHRs), particularly in forecasting the risk of critical conditions for prioritizing high-risk patients. However, their vulnerability to adversarial attacks is much less explored in the literature. Developing black-box perturbation algorithms and evaluating their impact on survival models can not only evaluate the robustness of these models in the pre-deployment stage but also provide counterfactual insights into the clinical interpretation of model predictions. In this work, we introduce SurvAttack, a black-box adversarial attack framework leveraging clinically compatible and semantically consistent subtle perturbations on longitudinal EHRs to degrade survival models' predictive performance. We specifically develop a method to manipulate medical codes with various adversarial actions throughout a patient's medical history. These adversarial actions are prioritized using a composite scoring strategy based on multi-aspect perturbation quality. The proposed perturbation algorithm is then used in an efficient SA-specific strategy to attack a survival model when estimating the temporal ranking of survival urgency for patients.

Preliminary of Survival Analysis

Survival models aim to estimate the survival function $S(t|V) = P_V(\tau > t)$ that represents the probability that an individual will survive beyond a specific time point t given the subject's covariate V. In survival analysis, the mean lifetime denoted as μ, representing the average time until an event occurs, is computed by integrating the survival function over time:

$$\hat{T} = \int_0^\infty S(t|V)\,dt \tag{1}$$

where $\hat{T}$ denotes the predicted survival time duration, and the output under attack in the victim survival models. The objective of attacks on survival models is to undermine the survival model's ranking capability and compromise survival time prediction. Specifically, the following two survival analysis metrics are monitored to assess attack effectiveness in achieving its goals.

Concordance Index: to assess concordance or agreement between the predicted survival times ranking and that of the actual survival times. A higher c-index indicates better concordance, meaning that the model's predicted rankings align more closely with the actual survival outcomes.

$$C = \frac{\sum_{i,j} \left[\mathrm{I}\left(\hat{T}_i < \hat{T}_j\right) + 0.5 \cdot \mathrm{I}\left(\hat{T}_i = \hat{T}_j\right) \right] \cdot \mathrm{I}\left(T_i < T_j\right) \cdot k_j}{\sum_{i,j} \mathrm{I}\left(T_i < T_j\right) \cdot k_i} \tag{2}$$

Mean Absolute Error (MAE): measures the accuracy of time predictions for observed data:

$$\text{MAE} = \frac{1}{N_{\text{ob}}} \sum_{i=1}^{N_{\text{ob}}} \|T_i - \hat{T}_i\| \tag{3}$$

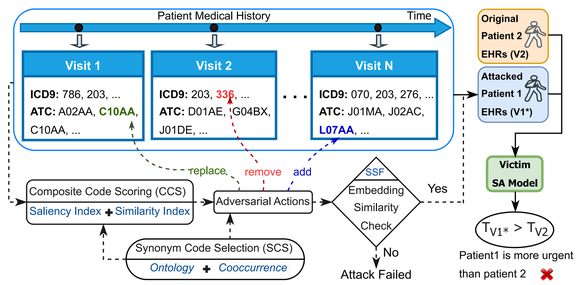

Figure 1: Overview of SurvAttack

Attack overview

SurvAttack, as a black-box adversarial attack framework, employs a **greedy search** that begins by first considering three adversarial action candidates (removal, replacement, and addition) for medical codes. To maintain clinical consistency and conformity, we have devised an **ontology-informed synonym code selection (SCS) strategy** that integrates the medical domain knowledge and statistical information, ensuring seamless alignment of any added code with the clinical semantic context. Also, we designed a **semantic similarity function (SSF)** leveraging a SOTA ontology-aware deep encoder. Using a **composite code scoring (CCS)** mechanism, SurvAttack evaluates each of the adversarial candidates. These actions are subsequently sorted in descending order based on their associated composite scores. The attack proceeds in this order, executing high-scoring adversarial candidates first until either the adversarial attack succeeds or one of the attack constraints is breached (attack fails). Finally, we propose a **dynamic SA-specific attack (DSA)** strategy, which is specifically designed for the task of survival analysis, dynamically using the defined greedy adversarial perturbation to disrupt the ranking ability of the SA model.

Ontology-informed Synonym Code Selection (SCS)

- SCS strategy incorporates both medical ontological data as clinical domain knowledge and co-occurrence probabilities as overall statistics in datasets. This meticulous integration identifies a synonym set of codes for a targeted medical code for attack, thereby maintaining both clinical conformity and statistical consistency.

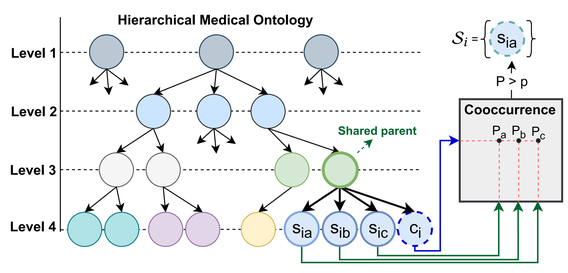

Figure 2: Ontology-informed Synonym Code Selection (SCS)

Semantic Similarity Function (SSF)

- Traditional adversarial attack studies on discrete EHR relied on perturbation count, overlooking the semantic impact of each modification. To address this, a deep semantic similarity function (SSF) is introduced.
- SSF leverages an ontology-aware transformer encoder, pre-trained for survival analysis, to generate embedding vectors and measure semantic similarity using cosine similarity.

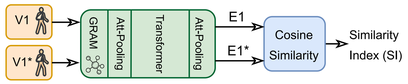

Figure 3: Semantic Similarity Function (SSF)

Composite Code Scoring (CCS)

- The attack iterates over a patient's EHR codes, scoring them based on possible adversarial actions. Two indices are computed: the saliency index (ΔF) measures the impact on the survival model's output, and the similarity index (SI) quantifies semantic preservation using the semantic similarity function.

$$\Delta F_i = F(V^*) - F(V) \qquad \text{(Saliency Index)} \tag{4}$$

$$SI_i = SSF(V, V^*) \qquad \text{(Similarity Index)} \tag{5}$$

- A composite score h is derived from ΔF and SI, prioritizing actions that maintain semantic similarity while maximizing impact.

$$h = \mathcal{H}(\Delta F, SI) = \Delta F \cdot e^{\lambda SI} \tag{6}$$

Greedy EHR Adversarial Attack

- The attack initializes V^* with V and ranks adversarial actions using the CCS strategy, executing high-scored actions first. Additions, removals, and replacements are performed iteratively while ensuring the similarity threshold is not violated.
- The replacement step selects the highest-scored synonym from $\mathcal{S}_i$ and compares removal and replacement for maximum model output change. The attack continues until the predicted survival ranking is flipped or similarity constraints prevent further modifications.

Dynamic SA-specific Attack (DSA) Strategy

- Step 1: DSA attacks all the censored subjects to reduce their expected survival times, (disrupting observed-censored pairs) and then Initialize t_{min} with the highest post-attack expected survival time of censored subjects.
- Step 2-6: Using t_{min} as the target time, the algorithm proceeds by attacking the sorted observed subjects and updates t_{min} accordingly (disrupting concordance in observed-observed pairs).

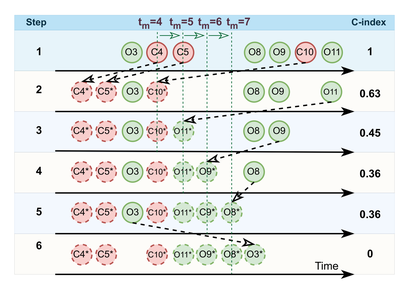

Figure 4: Semantic Similarity Function (SSF)

Results

Table 1: Baseline comparison for MAE and c-index in various cases.

Attack Method	CoxCC					DeepSurv				
	c_1	c_2	c_t	c_{cb}	MAE	c_1	c_2	c_t	c_{cb}	MAE
No Attack	0.7107	0.7107	0.7107	0.5485	1.83	0.6934	0.6934	0.6934	0.5504	2.01
Random	0.742	0.6064	0.6405	0.5369	2.00	0.7814	0.4481	0.5636	0.5051	2.79
Surv-TextBugger	0.3266	0.2722	0.1132	0.4997	2.98	0.2520	0.1946	0.08615	0.4857	3.89
Surv-WS	0.4986	0.5475	0.3292	0.5136	2.12	0.4028	0.4542	0.2215	0.5153	2.78
Surv-PWWS	0.3973	0.1808	0.098	0.5074	3.39	0.2827	0.1191	0.0738	0.4684	4.27
SurvAttack-NoSym	0.3722	0.1701	0.0918	0.4753	3.68	0.2745	0.1024	0.0721	0.4181	4.31
SurvAttack	**0.3102**	**0.1489**	**0.0841**	**0.4552**	**3.93**	**0.2057**	**0.0889**	**0.0569**	**0.4126**	**4.48**
Attack Method	**N-MTLR**					**DeepHit**				
	c_1	c_2	c_t	c_{cb}	MAE	c_1	c_2	c_t	c_{cb}	MAE
No Attack	0.5170	0.6242	0.6242	0.6242	3.49	0.5589	0.6527	0.6527	0.6527	2.06
Random	0.4997	0.7003	0.4747	0.5613	4.69	0.5056	0.7917	0.3318	0.5169	2.48
Surv-TextBugger	0.4405	0.3201	0.1918	0.0111	5.44	0.4873	0.2001	0.1551	0.078	3.08
Surv-WS	0.4844	0.5318	0.3136	0.3045	4.93	0.5012	0.2842	0.16	0.0915	2.65
Surv-PWWS	0.4880	0.5460	0.2783	0.3020	5.17	0.4509	0.2154	0.0794	0.0645	3.04
SurvAttack-NoSym	0.4270	0.2841	0.2987	0.0993	5.53	0.5026	0.2042	0.0698	0.0668	3.32
SurvAttack	**0.3926**	**0.2733**	**0.0588**	**0.0595**	**5.72**	**0.4809**	**0.1417**	**0.0688**	**0.0644**	**3.91**

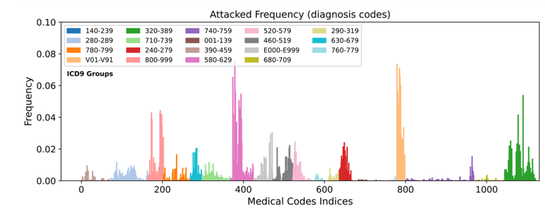

Figure 5: Frequency of adversarial attacks on Diagnosis Codes.

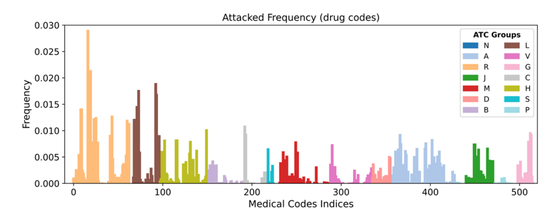

Figure 6: Frequency of adversarial attacks on drug codes.

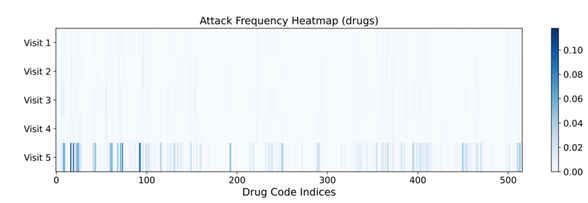

Figure 7: Attack pattern on drug codes.

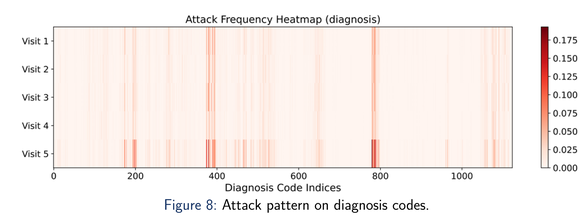

Figure 8: Attack pattern on diagnosis codes.