Contrastive Learning of Temporal Distinctiveness for Survival Analysis in Electronic Health Records

Mohsen Nayebi Kerdabai, Arya Hadizadeh Moghaddam, Bin Liu, Mei Liu, Zijun Yao

CIKM 2023

Outline

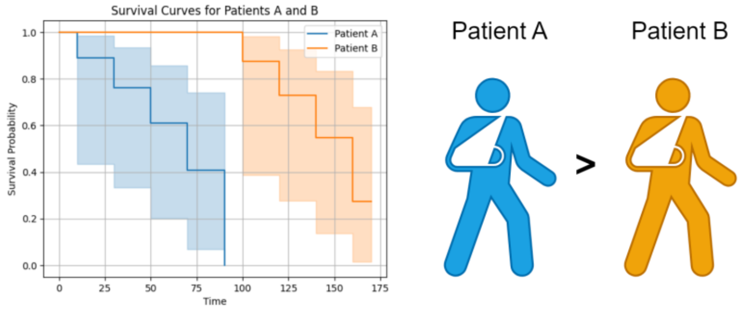
- **Background and Motivation**
- **Proposed Method**
- **Evaluation**
- **Conclusion**

Background and Motivation

- **What is Survival Analysis (SA)?**
  - Methods used to analyze **time-to-event data => Time Modeling**
  - Applied in various fields such as healthcare, economics, engineering, etc.

- **Goals**
  - Rank patients regarding their predicted survival time/chance.
  - Time prediction ability

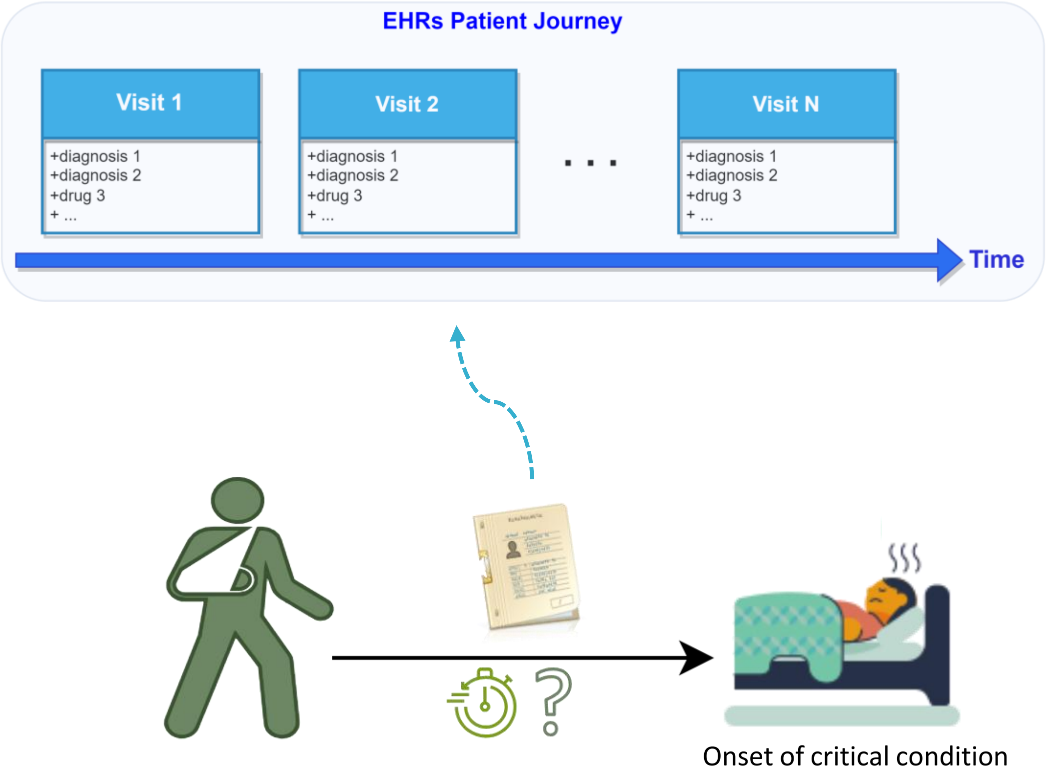
- **Semi-supervised Learning**
  - Labeled Data : **Observed**
  - Unlabeled Data : **Censored** (a key challenge)



- **Challenges and Opportunities:**
  - Traditionally, ranking and survival likelihood objectives are used.
  - Learning discriminative patient embedding based on survival time difference was not explored in the literature

Background and Motivation

- Learning the longitudinal and rich Electronic Health Records (EHRs) for comprehensive patient profiling

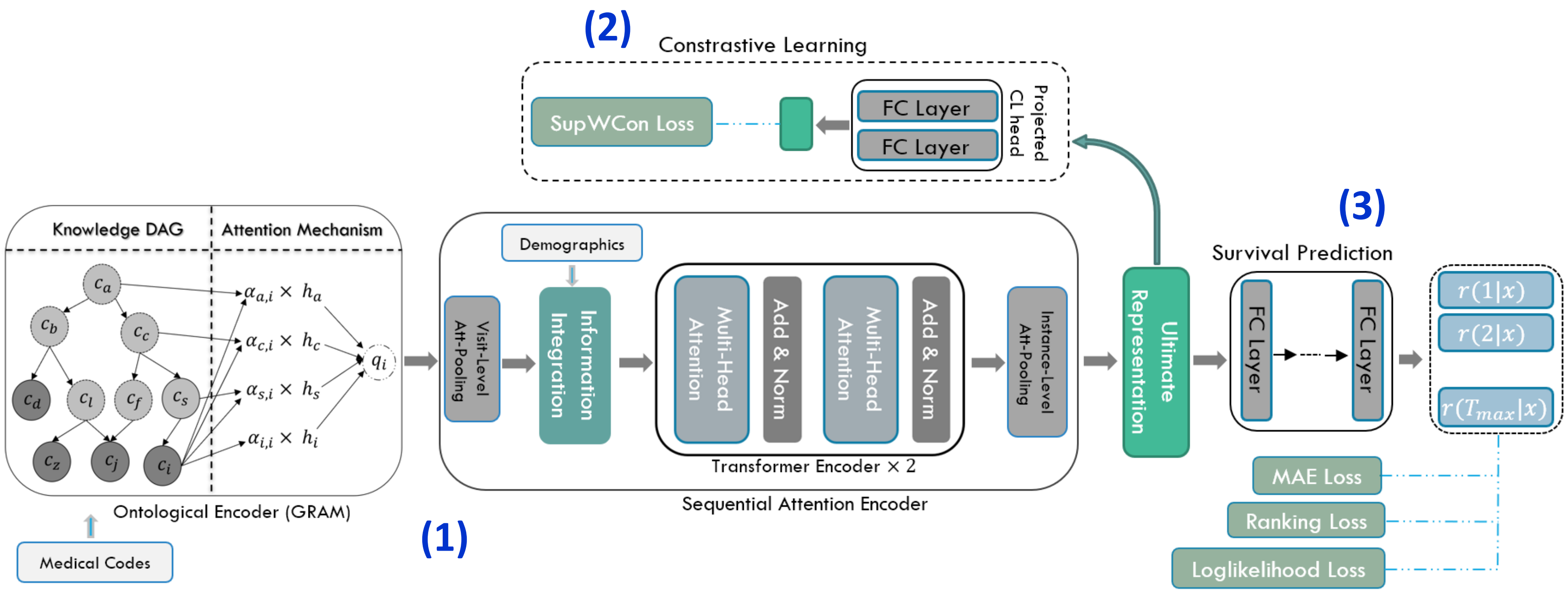


Outline

- Background and Motivation
- **Proposed Method**
- Evaluation
- Conclusion

Proposed Method

- We present **Ontology-aware Temporality-based Contrastive Survival (OTCSurv)** Analysis to predict time-to-event using patient history in EHRs.
- Three main components:
  1. **Sequential Attention-based Ontological Encoder**– encoding patient history with multiple attention-based components as well as domain knowledge from ontology
  2. **Contrastive Component** – utilizes survival durations to derive hardness-aware temporal distinctiveness
  3. **Survival Prediction** – outcomes necessary survival probabilities for each discrete time interval.



Proposed Method - Ontological Encoder

Ontology based EHR code encoding using GRAM*

- For each code (EHRs feature) find its ancestors in the ontology
- Assign each of them a vector $h_j \in \mathbb{R}^{d_c}$
- Use attention mechanism to generate attention weights for each of them$\alpha_{ij} \in \mathbb{R}^{+}$
- Use the attention weights to produce a convex combination of the initial embeddings as a final vector $q_i \in \mathbb{R}^{d_c}$

attention mechanism

$$f(h_i, h_k) = \omega_\alpha^T \tanh(W_\alpha Concat(h_i; h_k) + b_\alpha)$$

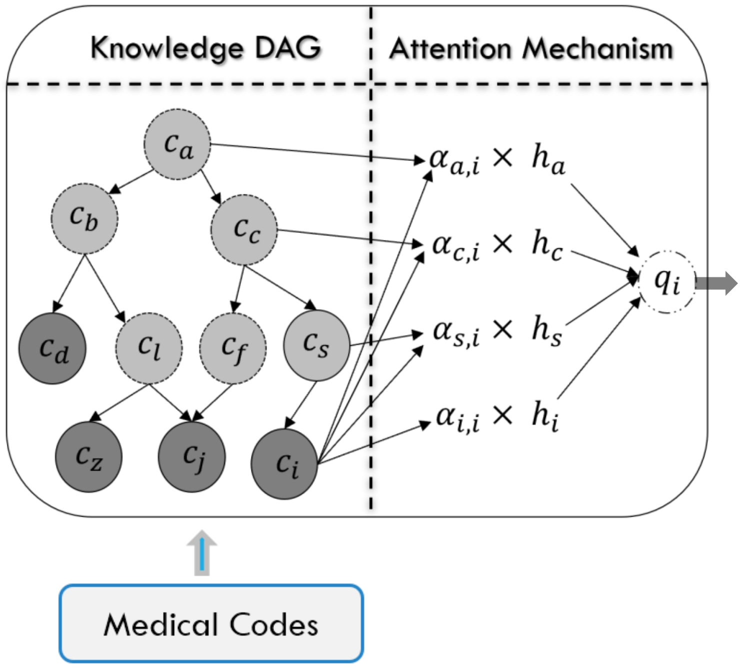
$f(.) \Rightarrow MLP$ operator

$$\alpha_{ij} = \frac{\exp(f(h_i, h_j))}{\sum_{k \in A(i)} \exp(f(h_i, h_k))}$$

Softmax function

$$q_i = \sum_{j \in A(i)} \alpha_{ij} h_j, \quad \sum_{j \in A(i)} \alpha_{ij} = 1, \quad \alpha_{ij} \geq 0 \ \ for \ \ j \in A(i)$$

Convex combination



* Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning, SIGKDD2017

Proposed Method - Sequential Attention Encoder

- **Attention-pooling**
  - **Visit-level :** compress and combine the information of different codes of a visit using the attention mechanism and produce the visit representation
  - **Instance-level -** compress and combine the information of different visits of an instance using the attention mechanism and produce the instance representation

$$e^n = l(v_n) = W_2\sigma(v_n W_1 + \beta_1), \quad 1 \le n \le N$$

$$\alpha^n = \mathrm{softmax}(e^n)$$

$$p^n = \sum_{i=1}^{M} \alpha_m^n q_m^n$$

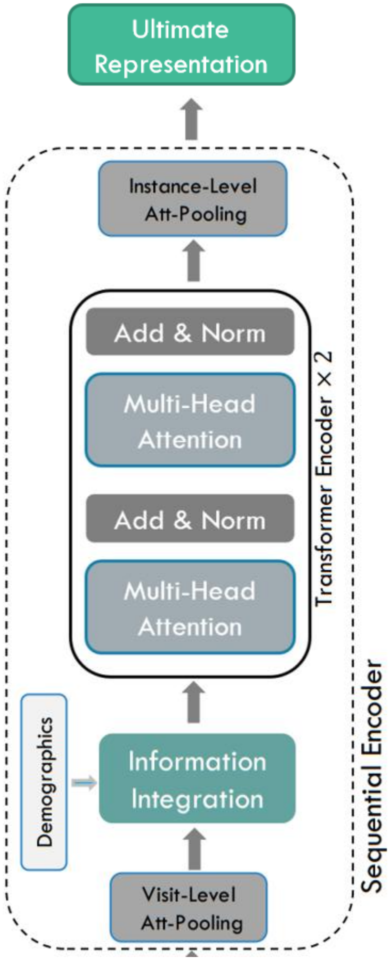
Information Integration: simple concatenation

- **Transformer encoder –** Using self-attention mechanism extract the interactions of medical visits to produce a rich representation of a patient

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

$$\mathrm{SelfAttention}(Q, K, V) = \mathrm{softmax}\left(QK^T / \sqrt{d_k}\right) V$$

$$U = Concat(head_1, head_2, \ldots, head_h)W^O$$

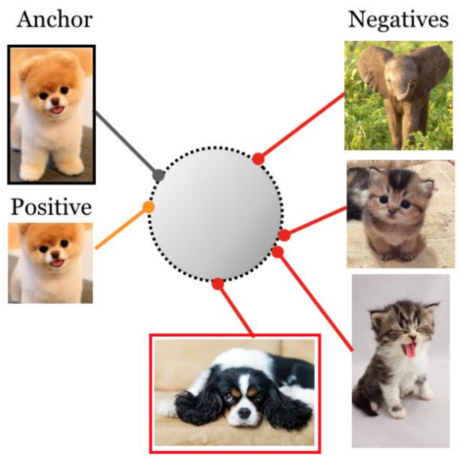


Proposed Method – Contrastive Component

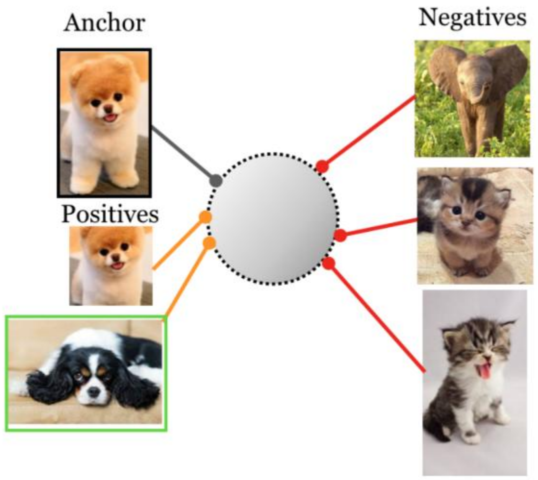
- **What is contrastive learning?**
  - **Representation Learning**
  - **Contrasting positive samples with negative samples.**

- **Originally, self-supervised, but later, supervised and semi-supervised**

- **Objective: bring positive samples closer together in the embedding space while pushing negative samples farther apart.**



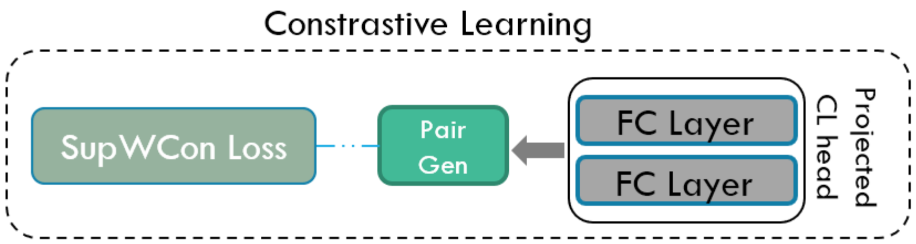
Self Supervised Contrastive [1]



Supervised Contrastive [2]

[1] A Simple Framework for Contrastive Learning of Visual Representations, Chen *et all.*, PMLR 2020
[2] Supervised Contrastive, Khosla *et all.*, Learning, NeurIPS 2020

Proposed Method – Contrastive Component



- **Nonlinear Transformation**

  FNN with a nonlinear AF--> transfers the representation to a different latent space

- **Pair Generator**

  Negative and positive pair generation criteria

- **Supervised Weighted Contrastive (SupWCon) loss**
  - Constant temperature parameter for + pairs $\tau \in \mathbb{R}^+$
  - Adaptive temperature parameter for - pairs $\tau_{ia} \in \mathbb{R}^+$
    - The inverse of their difference in survival duration
    - Encourage our model to regulate the amount of dissimilarity.

+ pair similarity

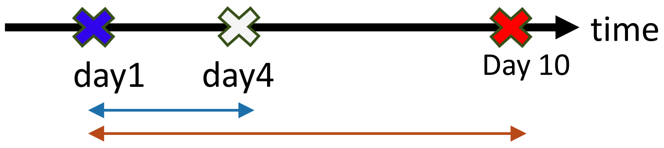
$$L^{\text{SupWCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau_{ia}\right)}$$

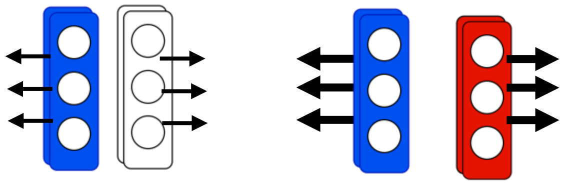
pair similarity

the set of instances that make a positive pair with anchor

$$\tau_{i,a} = |T_i - T_a|^{-1}$$

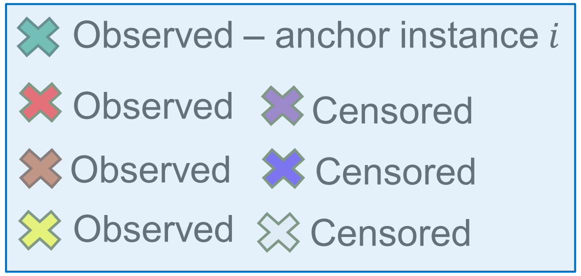
hardness − aware

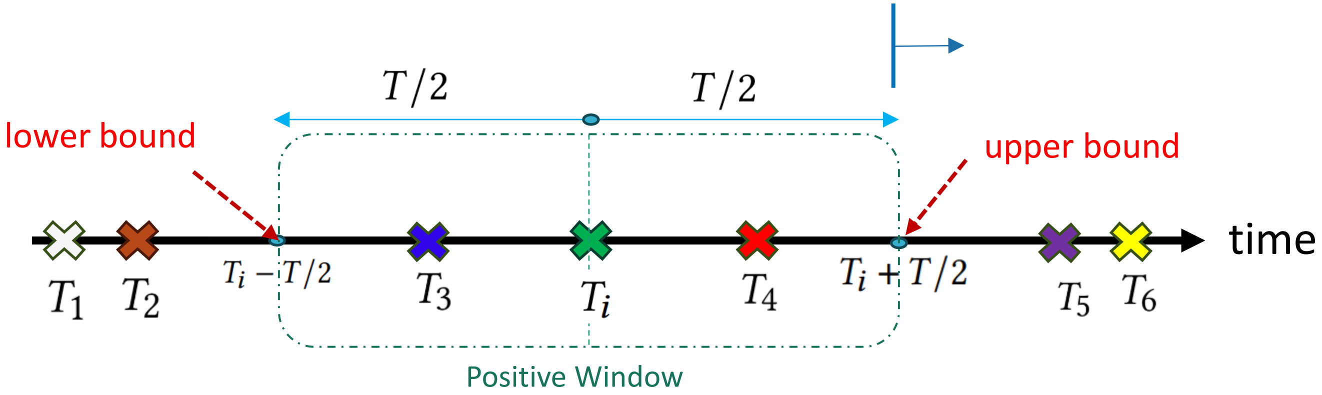




Proposed Method – Contrastive Component

Pair generation for contrastive learning

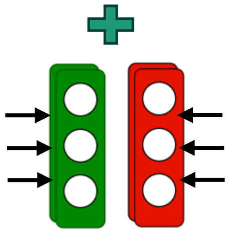




$$T_i - T/2 \leq t < T_i + T/2$$

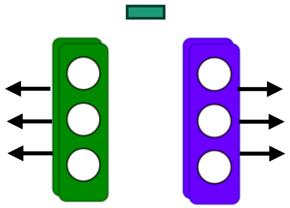
Positive pairs criteria:

- Any observed subject with survival time inside the positive window.
    - Positive pairs example : (T_i, T_4)



Negative pairs criteria :

- Any observed subject with survival time outside the positive window:
    - Pair example: (T_i, T_2), (T_i, T_6)



- Any censored subject with survival time larger than the *upper bound* of the positive window:
    - Pair example: (T_i, T_5)

Proposed Method – Other Loss Components

1. Loglikelihood Loss

Maximize the summation the survival probabilities before T

Minimize the summation the survival probabilities after T

$$L_{\text{ob}}^{\text{Loglikelihood}} = -\sum^{T-1} \log \hat{S}(t|x) - \sum_{t=T}^{T_{\max}} \log(1-\hat{S}(t|x))$$

$$L_{\text{cen}}^{\text{Loglikelihood}} = -\sum_{t=1}^{T} \hat{S}(t|x)$$

$$L^{\text{Loglikelihood}} = L_{\text{cen}}^{\text{Loglikelihood}} + L_{\text{ob}}^{\text{Loglikelihood}}$$

2. Pairwise Ranking Loss

- Based on concordance
- penalize the discordant pairs

$$L^{\text{Ranking}} = \max(0, (T_j - T_i) - (\hat{T}_j - \hat{T}_i))$$

penalize the discordant pairs

3. Mean Squared Error (MSE) Loss

penalizes wrong predicted survival duration times for only observed data points

$$\text{MSE} = \frac{1}{N_{\text{ob}}} \sum_{i=1}^{N_{\text{ob}}} (T_i - \hat{T}_i)^2$$

Outline

- Background and Motivation
- Proposed Method
- **Evaluation**
- Conclusion

Evaluation

- **EHRs Dataset**
  - A real-world EHR dataset from the University of Kansas Medical Center (KUMC) gathered for studying Acute Kidney Injury (AKI)
- **Medical Ontologies:**
  - Drug codes: Anatomical Therapeutic Chemical (ATC) classification
    Diagnosis codes : The International Classification of Diseases (ICD-9)
- **Evaluation Metrics**
  - The time-dependent discrimination index C^{td}
    - Extension of Harrell's concordance index (c-index)
    - Evaluating ranking ability
  - The Mean Absolute Error (MAE)
    - Evaluating time prediction ability

Table 1: Dataset Statistics

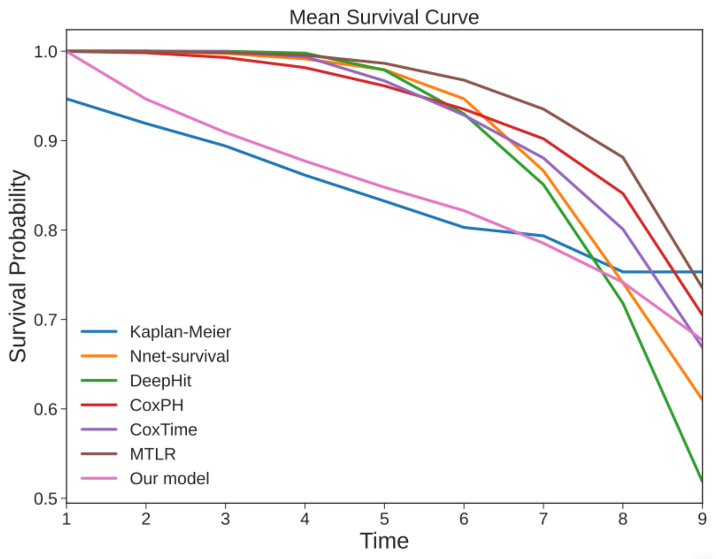
Number of patinets	56779
Number of censored patients	47773 (84.1%)
Number of observed patients	9006 (15.8%)
The average age of patients	59.47
Sex of patients distribution	(52% M, 48% F)
Average number of medical codes in a visit	13.29
Average number of medical codes in a patient	66.46

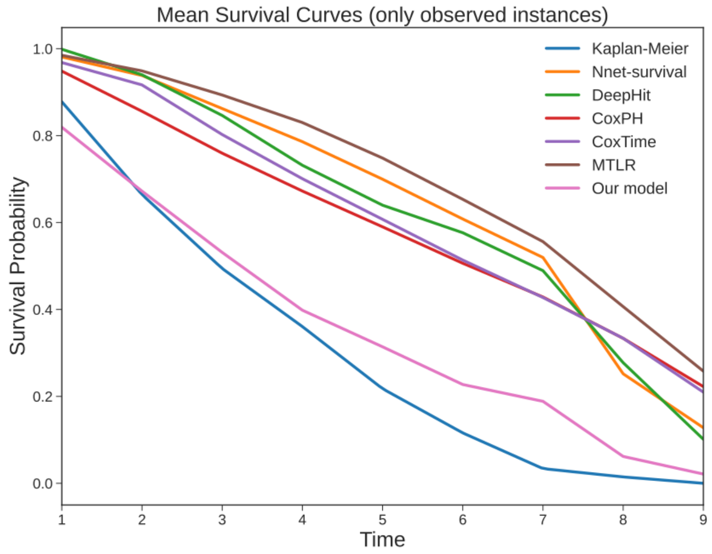
Evaluation - Results

- **Baseline Comparison**
  - OTCSurv outperforms other baselines regarding both evaluation metrics

Model Name	C^{td}	MAE
Nnet-survival	0.6332	3.165
MTLR (N-MTLR)	0.6712	4.081
DeepHit	0.6929	3.012
CoxTime	0.6912	2.980
DeepSurv (CoxPH)	0.6898	3.007
Our model	**0.6990**	**1.890**

- **Survival Curves Comparison**
  - OTCSurv's mean survival curve is the closest to the Kaplan-Meier curve => accurately captures the survival behavior and provides survival predictions that are more consistent with the actual outcomes.





Results - Ablation

- We experiment with different combinations of loss components.
- **SupWCon** is effective on improving **both** evaluation metrics.
- With all four loss components, we achieve the best **trade-off** in the performance.
- We utilized **a weighted summation** of these four losses as follows:
$$L^{\text{Total}} = \lambda_1 L^{\text{Loglikelihood}} + \lambda_2 L^{\text{Ranking}} + \lambda_3 L^{\text{SupWCon}} + \lambda_4 L^{\text{MSE}}$$
- Ontological Encoder and attention-poolings also play a significant role in improving the final results

Loss functions	C^{td}	MAE
$L^{\text{Loglikelihood}}$	0.6647	2.30
L^{Ranking}	0.6907	2.80
$L^{\text{Loglikelihood}}, L^{\text{SupWCon}}$	0.6808	1.90
$L^{\text{Loglikelihood}}, L^{\text{Ranking}}$	0.6951	2.43
$L^{\text{Loglikelihood}}, L^{\text{Ranking}}, L^{\text{SupWCon}}$	**0.7030**	**1.91**
$L^{\text{Loglikelihood}}, L^{\text{Ranking}}, L^{\text{SupWCon}}, L^{\text{MSE}}$	**0.6990**	**1.89**

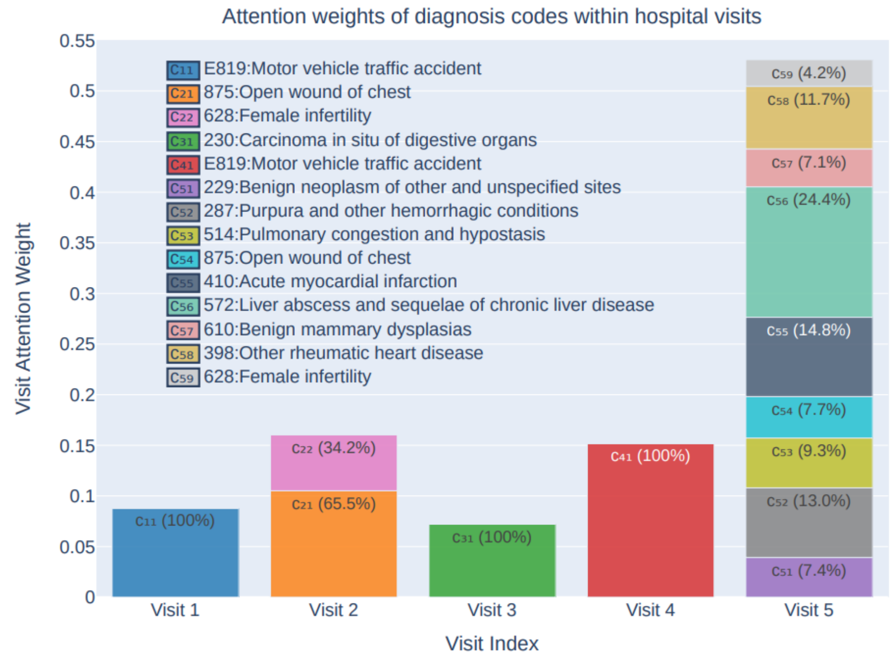
Window size	C^{td}	MAE
w/o ontology	0.6888	2.11
w/o attention-pooling	0.6795	2.21
Full model	**0.6990**	**1.89**

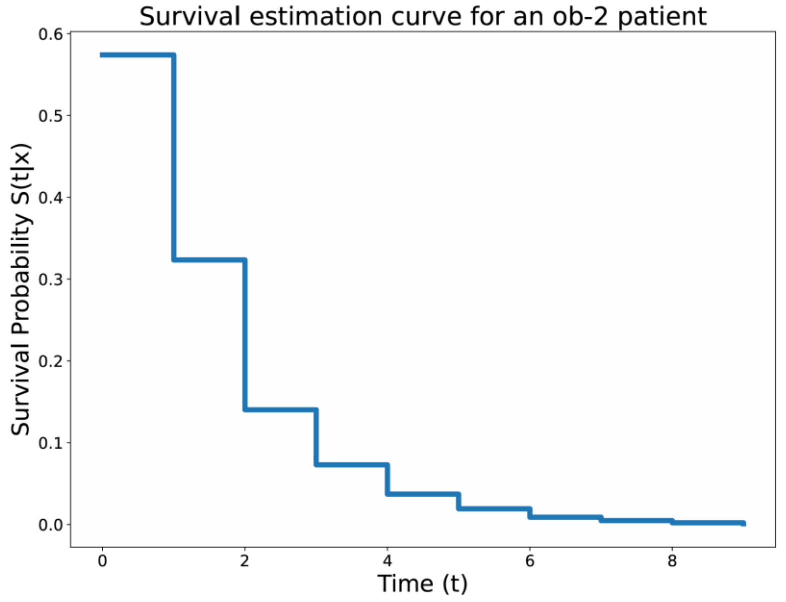
Results - Interpretability

Interpretability by Attention Analysis:

Attention weights in the **visit-level** attention-pooling
- the relative significance of each feature (code)
  - The height of each stacked bar
  - Example
    - c_{56}: "572" : liver abscess and sequelae of chronic liver disease
    - c_{55}: " 410 " : acute myocardial infarction (AMI)

Attention weights in the **instance-level** attention-pooling
- The relative importance of each of the visit
  - Each bar inside the stacked bars presents a medical code Inside the visit and its attention weight.





Outline

- Background and Motivation
- Proposed Method
- Evaluation
- **Conclusion**

Conclusion

Ontology-aware Temporality-based Contrastive Survival (OTCSurv) analysis framework

- **We design a Supervised Weighted Contrastive (SupWCon) Learning loss function**
  - Based on survival duration as its contrastive pairing criteria
  - Handles and takes into account censored data
  - Features an adaptive hardness-aware temperature parameter
- **We used a sequential attention-based ontological encoder**
  - Incorporate knowledge domain from medical ontology
  - Learn embedding with rich context using multiple attention-based components
- **We optimize survival prediction with a configuration of SupWCon, along with three more loss functions, focusing on two key goals**
  - Precisely ranking the survival chances of patients
  - Accurately predicting survival duration time

Thank you

Questions ?

mohsen.nayebi@ku.edu