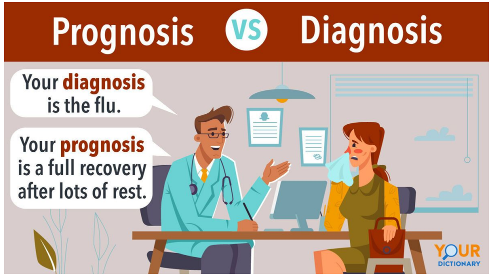
# Time-to-Event Pretraining for 3D Medical Imaging

Zepeng Huo*, Jason Alan Fries*, Alejandro Lozano*,
Jeya Maria Jose Valanarasu, Ethan Steinberg, Louis Blankemeier,
Akshay S. Chaudhari, Curtis Langlotz, Nigam H. Shah

Stanford Medicine School of Medicine
Stanford Medicine Clinical Excellence Research Center
BMIR Stanford Center for Biomedical Informatics Research
Stanford AIMI Center for Artificial Intelligence in Medicine and Imaging
HAI Stanford University Human-Centered Artificial Intelligence

ICLR 2025

# Challenges and Motivations




What's the motivation:

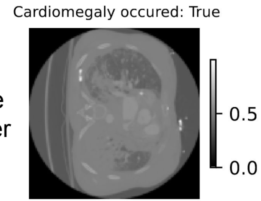
- Medical vision foundation models (VFMs) should help to better clinical decision making
  - In medicine, **<u>prognosis</u>** can help long-term treatment planning, in addition to **<u>diagnosis</u>**

Clinical use case:

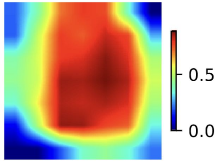
Use CT scan to predict probability of cardiomegaly (enlarged heart) in the next 12 month

E.g. VFMs should learn to identify biomarkers through self-supervised pretraining

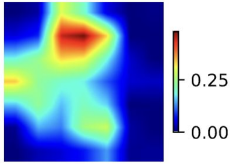
* Example from paper CT input




Use GradCAM to probe models




✅Ideal

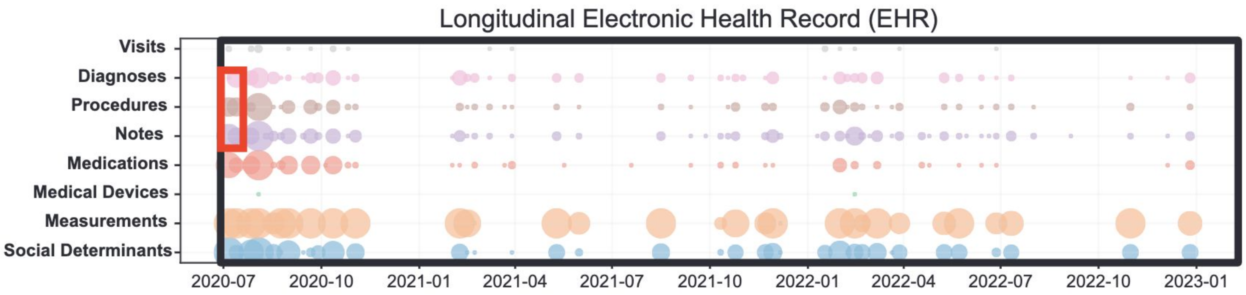


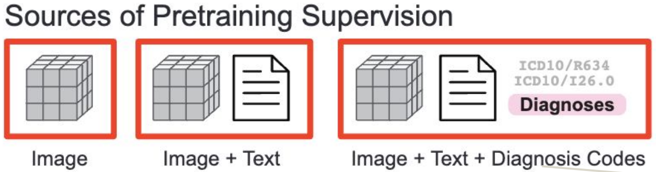

❌ Not ideal

Why fail?

# Missing context problem

Longitudinal information from EHR is largely ignored by current pretraining methods
=> only local information (**redbox** below)




Leads to *missing context problem*: no long-term temporality

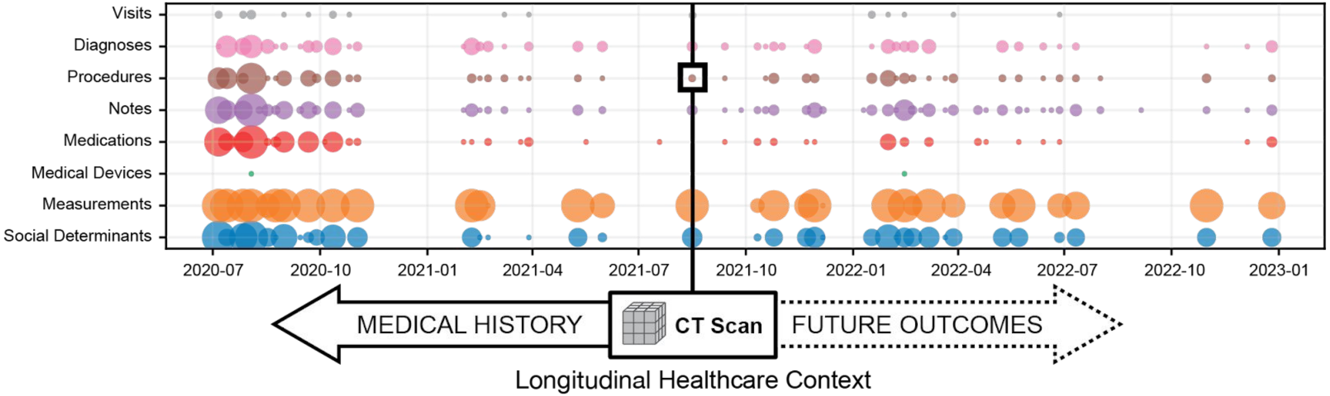
Self-supervised learning within image

Contemporaneously paired notes

Short time horizon from EHR timeline

# So we propose

- Medical imaging models should improve ***prognosis*** which is important in medicine
- EHR is the great source of pretraining information with *long-term temporality*



# Methodology: Time-to-event pretraining

## Future-guided pretraining is a great way to learn **prognosis** signals

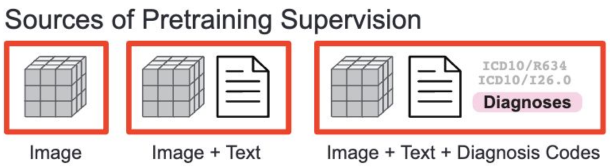
- How to use future information to guide pretraining?

## Materialize the pretraining under Time-to-event framework (a.k.a survival models)

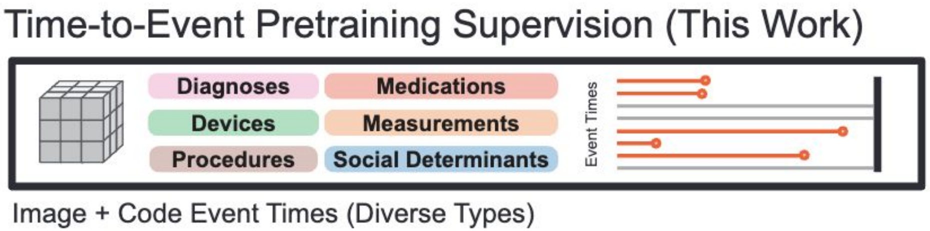
- Naturally handles *censorship* (bias from loss of follow up)
- Not only *what* will happen but *when* (vs. autoregressive next-token loss)
- Data efficiency (squeeze more from weak supervision signals)

This work

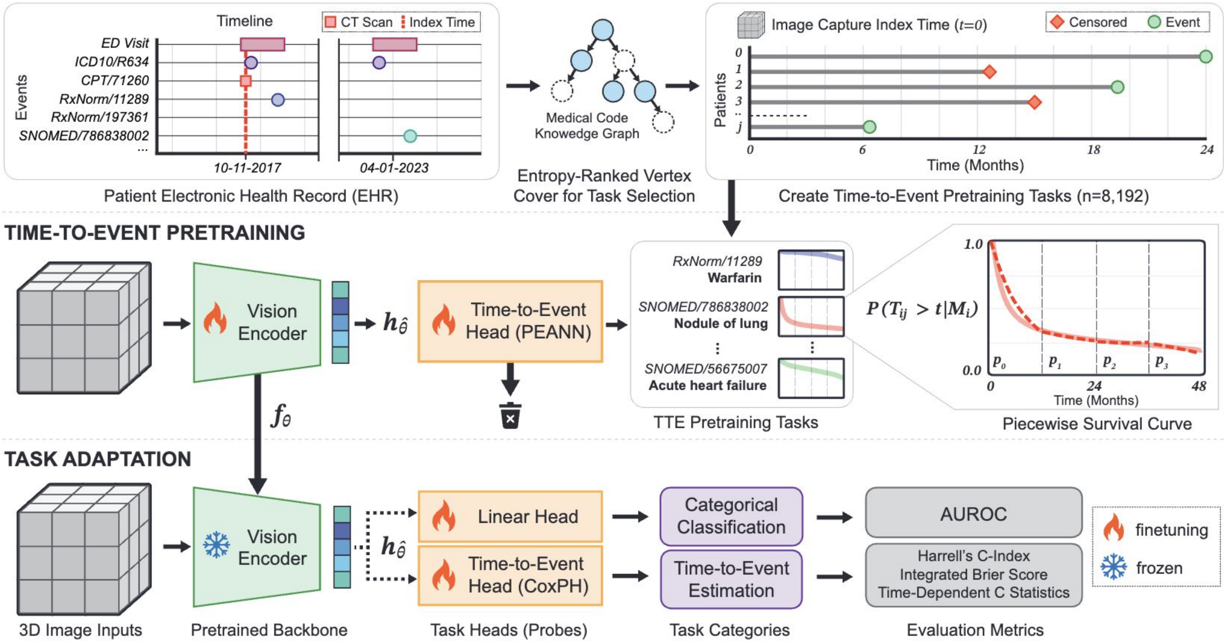
Current methods




VS.



# Pipeline overview




Step 1:
Transform EHR timelines into TTE tasks (8,192), resulting in event time distributions with censorship.

Step 2:
Pretrain 3D vision encoders to extract imaging biomarkers to learn each task's survival function.

Step 3:
Linear probe 3D vision encoders for different downstream tasks via different adaptations.

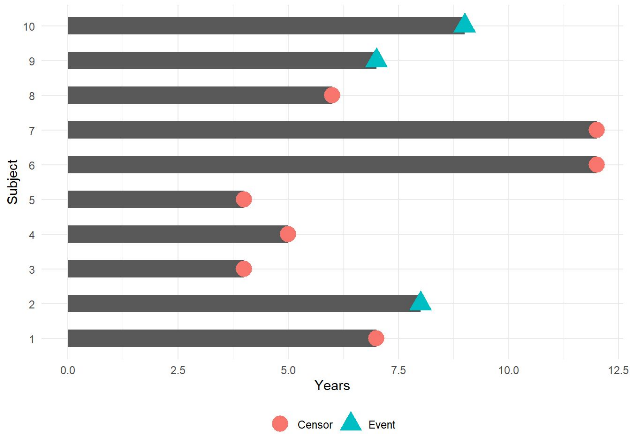
# Time-to-event loss materialization

Model the **time until an event occurs** (e.g. death) while account for censorship

## Censorship

Event of interest is not observed by the end of study

$\boxed{(X_i, T_i)}$ **Biased** $(X_i, T_i, \delta_i)$ $\delta_i = \begin{cases} 1 & \text{event observed} \\ 0 & \text{censored} \end{cases}$




Proposal of using Deep Piecewise Exponential Model

Different distribution parameters across different **timepieces**

✅ provide additional flexibility than hazard portionality

No need to make sure **uncensored** patients in a batch

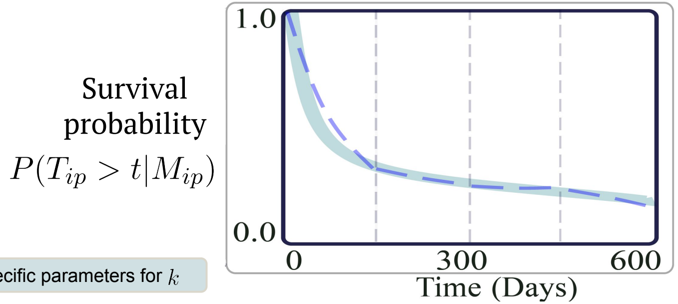
✅ more efficient than Cox-PH

**Our Goal** Estimate the probability distribution of event times $P(T_i = t)$ accounting for censoring

# Time-to-event loss materialization

## Deep Piecewise Exponential Model

- Partition time into **pieces** for more expressive risk modeling
- For **piece** *p*, interval start and end time [ $S_p$ , $E_p$)




For an **image** $i$, **task** $k$, **piece** $p$

Piece-specific linear projection

Task-specific parameters for $k$

**Hazard function:**

$$\lambda_{ip} = exp(W_p M_{ipk} \cdot \hat{\beta}_k + b_k)$$

Image $i$ feature embedding

**Survival function:**

$$S_i(t) = \prod_{p=1}^{P} \exp\left[-\lambda_{ip}(\min(t, E_p) - S_p)\mathbb{I}(t \geq S_p)\right]$$

min of observed time or censorship

Indication function for event happening in a timepiece

**Loss function:**

$$\mathcal{L}_i = [S_i(t)]^{1-\Delta_i}[f_i(t)]^{\Delta_i}$$

**(* omitting summation of task *k*, timepiece *p* for succinctness)**

Survival probability until *t*

Probability density function of hazard at *t*

# Experiment setup

## Dataset

- Pretraining dataset → Test dataset

| Pretraining dataset | Test dataset |
|---|---|
| INSPECT [1] train split<br>- Average of 5 years follow-up data post-CT scan | INSPECT test split<br>RSPECT [2] |

| | INSPECT | RSPECT |
|---|---|---|
| # Patients | 19,402 | 7,279 |
| # Train | 18,945 | 5,823 |
| # Valid | 1,089 | 364 |
| # Test | 3,214 | 1,092 |
| Imaging | ✓ | ✓ |
| EHR | ✓ | ✗ |
| TTE Tasks | ✓ | ✗ |
| Diag. Tasks | ✓ | ✓ |
| Scan Type | Chest CT | Chest CT |

## Evaluation

- 4 Binary tasks *(AUROC)*
  - Mortality, readmission, pulmonary hypertension, pulmonary embolism
- 5 time-to-event tasks *(C Statistics and variations)*
  - Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion

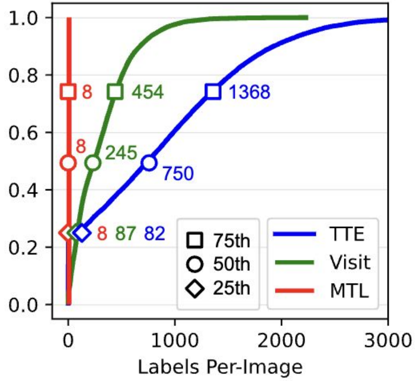
[1] Huang, Shih-Cheng, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis P. Langlotz, Serena Yeung, Nigam H. Shah, and Jason A. Fries. "Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis." *arXiv preprint arXiv:2311.10798* (2023).
[2] Colak, Errol, et al. "The RSNA pulmonary embolism CT dataset." Radiology: Artificial Intelligence 3.2 (2021): e200254.

# Experimental Findings

- **Architecture agnostic**
  - SwinUNETR ✅ ResNet ✅ DenseNet ✅
    - All boosted performance across 8 benchmark tasks over baselines
      - AUROC **23.7%** ↑
      - Harrell's C-index **29.4%** ↑
      - Integrated Brier Score **54%** ↓

- **Effectiveness of TTE pretraining**:
  - Better than image-only pretraining (MAE) and local EHR or report information pretraining (Merlin [1])

- **High data efficiency**
  - 3x labels per image at 50 percentile



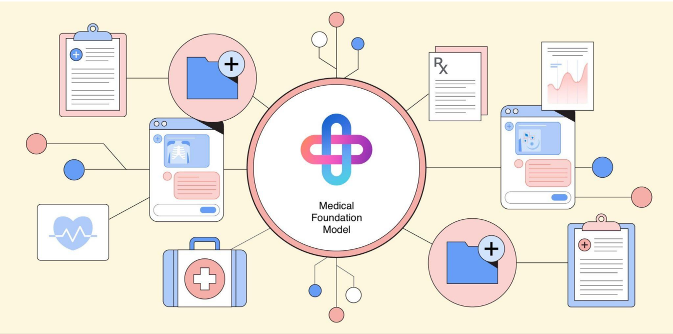

- **Without sacrificing** diagnosis performance!
  - Tested on out-of-distribution RSPECT data

[1] Blankemeier, Louis, et al. "Merlin: A vision language foundation model for 3d computed tomography." Research Square (2024): rs-3.

# Future work and acknowledgement

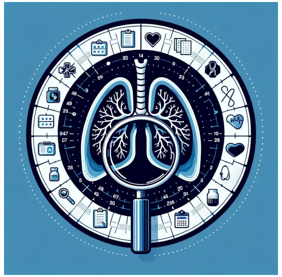
## Future work

- Multimodality backbone
  - Beyond image-only encoder
  - Link **<u>any modality</u>** to future event through TTE
    - E.g. genomics, wearable, reports, etc.
  - Comprehensive view of health trajectory




## Data download:

- http://inspect.stanford.edu



## Acknowledgment:

NIH National Heart, Lung, and Blood Institute

Stanford Medicine | School of Medicine

Stanford Medicine | Clinical Excellence Research Center

BMIR Stanford Center for Biomedical Informatics Research

Stanford AIMI Center for Artificial Intelligence in Medicine and Imaging

HAI Stanford University Human-Centered Artificial Intelligence