Event-Based Contrastive Learning for Medical Time Series

Nassim Oufattole[1]* **Hyewon Jeong**[1]* **Aparna Balagopalan**[1] **Matthew Mcdermott**[2]
Payal Chandak[1,2] **Marzyeh Ghassemi**[1] **Collin Stultz**[1,2]
[1] MIT [2] Harvard
{nassim, hyewonj, aparnab, chandak, mghassem, cmstultz}@mit.edu
matthew_mcdermott@hms.harvard.edu

Abstract

In clinical practice, one often needs to identify whether a patient is at high risk of adverse outcomes after some key medical event; e.g., the short-term risk of death after an admission for heart failure. This task, however, remains challenging due to the complexity, variability, and heterogeneity of longitudinal medical data, especially for individuals suffering from chronic diseases like heart failure. In this paper, we introduce Event-Based Contrastive Learning (EBCL) - a method for learning embeddings of heterogeneous patient data that preserves temporal information before and after key index events. We demonstrate that EBCL produces models that yield better fine-tuning performance on critical downstream tasks including 30-day readmission, 1-year mortality, and 1-week length of stay relative to other representation learning methods that do not exploit temporal information surrounding key medical events.

1 Introduction

Outcomes for patients with heart failure are highly variable, making accurate predictions challenging [1, 2]. Nevertheless, reliable outcome prediction is important for the development of effective treatment strategies [3] and for ensuring that healthcare resources are allocated appropriately, thereby ensuring that the most resources are made available to the sickest patients [4]. Prior work has leveraged data within the Electronic Health Record (EHR) for this task. Indeed, a variety of learning algorithms, including directly supervised [5, 6, 7], self-supervised [8, 9], weakly-supervised [10], generative pretraining [11, 12, 13, 14], and contrastive learning algorithms [7, 15, 16] have all been used to identify patients at the highest risk of adverse outcomes. A central feature of many predictive tasks in medicine is that one aspires to identify a patient's risk after some index event. Case in point, the ability to identify patients at high risk of death after a myocardial infarction (heart attack) forms an important part of the evaluation and care of patients with coronary artery disease. Although a number of methods use elements of the electronic record for this task, they typically do not strive to encode the temporal information surrounding the event of interest [15, 8, 9]. For example, how do patient features change pre- and post-index event? As such data often include information about the patient's response to therapy, we postulated that predictive performance could be improved, relative to existing methods, when such temporal information is explicitly modeled around the index event.

We therefore introduce a novel approach using temporal contrastive learning to emphasize the consistency of medical trajectory representations around key medical events. Our approach diverges from existing work [7, 15, 17] by imposing our specialized pre-training contrastive loss solely on data around critical events, where the most prognostic information regarding disease progression and prognosis is likely to be found. In particular, our contrastive system is trained to differentiate positive

[*] Equal contribution.

Submitted to Unifying Representations in Neural Models Workshop in NeurIPS 2023. Do not distribute.

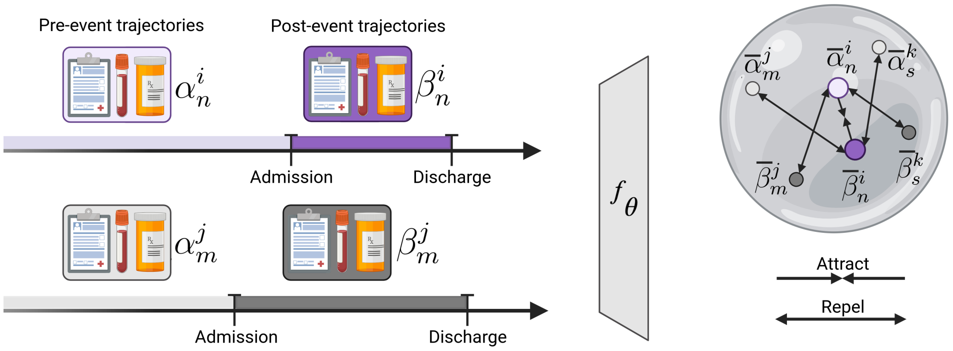


Figure 1: **Event-based contrastive learning (EBCL)** We sample a batch $D_{\mathcal{B}} = \{(\alpha_n^i, \beta_n^i) : i, n \in \mathcal{B}\}$ of pairs of pre-event and post-event time series sequences for pretraining, where i and n index the patient and the event respectively. We choose the event of interest to be an inpatient admission. Pre-event data, α_n^i, and post-event data, β_n^i, are passed separately into the transformer encoder f_θ to get $\bar{\alpha}_n^i = f_\theta(\alpha_n^i)$ and $\bar{\beta}_n^i = f_\theta(\beta_n^i)$ which is pretrained with CLIP contrastive loss. The positive pairs are pre and post-event data of the same event, $(\bar{\alpha}_n^i, \bar{\beta}_n^i)$. The negative pairs are mismatched pre-event and post-event trajectories, such as $(\bar{\alpha}_n^i, \bar{\beta}_m^j)$ where $i \neq j$ or $n \neq m$.

pairs - patient data before and after an event - from negative pairs - data from separate events (these separate events can be for either different patients or the same patient). This framework aids the model in learning consistent representations of patient history across these key events. We postulate that these learned representations can enhance our ability to capture long-term longitudinal trends in patient states and improve outcome predictions in patients with chronic diseases such as heart failure. We apply the method across several fine-tuning tasks, including 30-day readmission [18, 19, 20, 21], 1-year mortality [22, 23, 24, 25], and 1-week length of stay (LOS) prediction over a multi-center cohort of patients with heart failure. Consistently, our temporally aware contrastively pretrained model outperforms existing representation learning methods across these tasks.

Related Works Contrastive learning is a form of self-supervised learning (SSL). It utilizes automatically co-occurring sources of information ("multi-modal") [26, 27, 28, 29] or augmented versions of the same information ("multi-view") as a supervision signal [30, 31, 32, 33, 16, 27, 34, 35, 36]. Specifically, paired observations, created using multiple data modalities or augmented/mined views of the same object, are contrasted against negative observation pairs created using multi-source/view information for different objects through losses such as the InfoNCE [31] or CLIP loss [32].

A recent contrastive learning scheme called Order Contrastive Pretraining (OCP) [15] builds on prior works such as Permutation-Contrastive Learning (PCL) [7] and takes two consecutive windows of data in their original temporal sequence to construct positive pairs. Negative pairs correspond to a sequence where the temporal order of sequences is swapped. Importantly, the sequences chosen to be swapped are not chosen with respect to any particular event, while our method is event-centric.

2 Methods

2.1 Event-Based Contrastive Learning for Medical Outcome Prediction

We present a novel contrastive pretraining method called Event-Based contrastive learning (EBCL), which leverages medical time-series data around the time of a critical event.

Problem Formulation Let t_n^i denote the n^{th} clinically significant event (e.g., inpatient admission) for patient i. We note that a single patient may have more than one clinically significant event. For example, a patient with heart failure may have several hospital admissions with shortness of breath, where each admission corresponds to a clinically significant event. Pretraining data $\mathcal{D}_{\text{PT}} = \{(\alpha_n^i, \beta_n^i) : n < N_i \;\&\; i < P\}$, where α_n^i (pre-event data) corresponds to data before t_n^i

(index event) β_n^i (post-event data) corresponds to data after t_n^i until the end of the event. N_i is the number of events associated with patient i, and P is the total number of patients in the dataset.

EBCL Pretraining We encode α_n^i and β_n^i using missingness-aware triplet embedding [8] with feature, value, and time embeddings. We feed the input triplet embedding into the model, f_θ, and get pre- and post-embeddings $f_\theta(\alpha_n^i)$ and $f_\theta(\beta_n^i)$. We then compute the CLIP loss $\mathcal{L}_{\text{CLIP}}$ [32] on a batch of these embeddings. Note that (α_k^i, β_m^j) is a positive pair if $i = j$ and $k = m$.

Finetuning We finetune f_θ on the tasks defined in Section 2.2 using a dataset of time-series embedding with matched binary outcome labels: $\mathcal{D}_{\text{FT}} = \{(\alpha_n^i, \beta_n^i, y_n^i) : n < N_i \;\&\; i < P\}$. We use negative cross-entropy loss, $\mathcal{L}_{\text{CE}}$, and pass our embeddings, $f_\theta(\alpha_n^i)$ and $f_\theta(\beta_n^i)$ through a finetuning architecture depicted in Figure 3a (in appendix) to arrive at a prediction for our label.

2.2 Dataset and Tasks

We have assembled a cohort of 107,268 heart failure patients with 383,254 inpatient admissions, obtained from the electronic data warehouse of a large hospital network. We preprocess these data using a preprocessing pipeline inspired by ESGPT [37]. The dataset includes patient trajectories over a maximum span of 40 years and a maximum number of 3275 features, which includes labs, diagnoses, procedures, medications, tabular echocardiogram recordings, physical measurements, and admissions/discharges. See Appendix Table 3 for the pretraining dataset statistics.

We divide the patient trajectories into Pre-Event and Post-Event data, where the Event is an inpatient admission. We finetune and evaluate on three downstream tasks where the datasets are summarized in Table 1. Note that for the LOS task, we only use Pre-Admission data as input, as Post-Admission data would leak the LOS outcome, so $\mathcal{D}_{\text{LOS}} = \{(\alpha_n^i, y_n^i) : n < N_i \;\&\; i < P\}$. We also always restrict Post-Admission data, β_n^i, to the data prior to patient discharge, as this is the information that will be available at decision time for the mortality and readmission tasks. We partitioned our compiled dataset into training (80%), validation (10%) and testing (10%) with the split determined by individual patient instances. Additional dataset information is provided in appendix section A

Table 1: **Statistics for finetuning datasets.**

Task	# Patients	# Events	# Prevalence
30-Day Readmission	60,287	215,097	17.6%
1-Year Mortality	49,962	183,125	27.7%
1-Week LOS	107,268	383,254	54.1%

2.3 Experiments

Our transformer has two encoder layers, a 512 sequence length for pre- and post-event data, a 2,048-dimension feed-forward layer between self-attention layers, and 32-dimension token embeddings. We then perform Fusion Self-Attention [8, 38], by taking an attention-weighted average of the output embeddings of the transformer to get a single 32-dimension embedding. Finally, we have a linear projection to a 32-dimension embedding. Sequences are padded to be length 512 and attention over padded tokens is masked in the transformer and the Fusion Self-Attention layer.

To evaluate our method we perform the following experiments:

- **Supervised**: This corresponds to standard supervised training without EBCL pretraining. The model f_θ is initialized with random weights and then the model is trained in a supervised fashion for a specific task. We use the architecture in figure 3a.
- **Order Contrastive Pretraining (OCP)**[15]: OCP is a self-supervised approach that implements a pretraining task of discriminating correct and switched sequencing. We pretrain f_θ with the OCP objective where for each patient i, we take a continuous sequence of at most 512 tokens, split the sequence in half, and randomly swap the first and second halves (Figure 2). We denote this randomly sampled and swapped sequence as τ^i and display this pretraining architecture in Appendix Figure 3d. Due to slow convergence, we allow up to a maximum of 600 epochs for pretraining. For fine-tuning, we concatenate or "fuse" α_n^i and

β_n^i and pass the concatenated trajectory into the model f_θ (Figure 3b). For this experiment, α_n^i and β_n^i individually can only have a sequence length of 256 as f_θ has a max sequence length of 512. For more information see Appendix Section A.

- **Event-Based Contrastive Learning (EBCL)**: Our event-centric temporal contrastive pretrained model. We show this pretraining architecture in Appendix Figure 3c.

We finetune the pretrained contrastive learning models (OCP, EBCL) with fully connected layers to predict outcomes. Additionally, to evaluate the utility of the representations learned from pretraining, we do an additional experiment where we freeze the model weights for f_θ learned from pretraining (EBCL Frozen, OCP Frozen), and train a two-layer MLP on the output of f_θ for our finetuning task. We perform an extensive learning rate and dropout hyperparameter search for pretraining and finetuning and use a maximum of 100 epochs. We take the epoch with the highest validation set performance for pretraining and finetuning. We run 3 random seeds for pretraining and finetuning and report the mean and standard deviation of results across these seeds.

2.4 Results

We summarize the evaluation of models on the finetuning tasks in Table 2. Our proposed method, EBCL, achieves a significant improvement over the fully supervised baseline and OCP approach. The OCP method showed worse performance compared to the Supervised model on predicting 1-year mortality and 1-week LOS. We note that pretraining methods often rely on leveraging a much larger pretraining dataset than the finetuning dataset to obtain performance improvements, but in our scheme, the pretraining dataset is similar in size to our downstream tasks, so the drive for the improvement is mainly coming from solving the contrastive learning task, and not seeing more data.

Table 2: **EBCL Pretraining improves results over a supervised baseline and OCP pretraining.** Additionally, the EBCL Frozen model consistently outperforms the OCP Frozen model and even outperforms the Supervised model for 30-day Readmission prediction.

	30-Day Readmission		1-Year Mortality		1-Week LOS	
	AUC	APR	AUC	APR	AUC	APR
Supervised	64.3 ± 1.1	89.0 ± 0.5	80.2 ± 0.7	90.7 ± 0.3	86.3 ± 0.6	85.7 ± 0.5
OCP	68.1 ± 0.5	90.6 ± 0.2	78.0 ± 0.7	89.6 ± 0.3	84.1 ± 0.7	83.7 ± 0.6
OCP Frozen	59.1 ± 0.5	86.2 ± 0.4	65.9 ± 0.4	81.9 ± 0.2	59.9 ± 0.1	55.3 ± 0.3
EBCL	**70.4 ± 0.1**	**91.4 ± 0.1**	**82.3 ± 0.1**	**91.8 ± 0.0**	**87.2 ± 0.2**	**86.9 ± 0.3**
EBCL Frozen	68.2 ± 0.3	90.6 ± 0.1	78.4 ± 0.1	90.0 ± 0.1	79.7 ± 0.1	79.2 ± 0.1

Additionally, we examine how freezing the OCP and EBCL pretrained weights affects performance relative to the supervised baseline. EBCL Frozen significantly outperforms OCP Frozen for all tasks, providing further evidence that EBCL is a superior pretraining scheme for these downstream tasks. Furthermore, EBCL Frozen outperforms Supervised for readmission and is comparable for 1-year mortality, indicating EBCL learns a useful patient representation around clinical events. There is a gap between the supervised baseline and EBCL for the LOS task, but the LOS task only uses pre-event data, which suggests that the post-data EBCL embedding may be more useful for downstream tasks than the pre-data EBCL embedding. Validating this will be a subject of future experiments. Overall these results indicate that EBCL provides a useful pretraining scheme for our downstream tasks.

3 Conclusion

In this paper, we propose EBCL, a novel pretraining scheme for medical time series data, to learn temporal representations around clinically significant events. We show that EBCL outperforms supervised baselines relative to standard supervised training and a related pretraining method, OCP, which is not based on clinically significant events.

Limitations and Future Works. Given this investigation is at an early stage, we did not conduct an extensive comparison of contrastive loss schemes for pre-event and post-event data, of different model architectures, or of results on different datasets. An important direction for future work is addressing these limitations and evaluating other transformer pretraining methods. While further experiments with different model architectures are important, our current results with several pretraining methods are encouraging.

References

[1] A Mosterd. The prognosis of heart failure in the general population. the rotterdam study. *European Heart Journal*, 22(15):1318–1327, August 2001.

[2] Douglas D. Schocken, Martha I. Arrieta, Paul E. Leaverton, and Eric A. Ross. Prevalence and mortality rate of congestive heart failure in the united states. *Journal of the American College of Cardiology*, 20(2):301–306, August 1992.

[3] Kazem Rahimi, Derrick Bennett, Nathalie Conrad, Timothy M Williams, Joyee Basu, Jeremy Dwight, Mark Woodward, Anushka Patel, John McMurray, and Stephen MacMahon. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC: Heart Failure*, 2(5):440–446, 2014.

[4] Stephen F Jencks, Mark V Williams, and Eric A Coleman. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428, 2009.

[5] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357*, 2021.

[6] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[7] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469. PMLR, 20–22 Apr 2017.

[8] Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

[9] Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G Krishnan. Duett: Dual event time transformer for electronic health records. *arXiv preprint arXiv:2304.13017*, 2023.

[10] Matthew McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive ehr timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 257–278, 2021.

[11] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[12] Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

[13] Seunghyun Lee, Da Young Lee, Sujeong Im, Nan Hee Kim, and Sung-Min Park. Clinical decision transformer: Intended treatment recommendation through goal prompting. *arXiv preprint arXiv:2302.00612*, 2023.

[14] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[15] Monica N Agrawal, Hunter Lang, Michael Offin, Lior Gazit, and David Sontag. Leveraging time irreversibility with order-contrastive pre-training. In *International Conference on Artificial Intelligence and Statistics*, pages 2330–2353. PMLR, 2022.

[16] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022.

[17] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. TCLR: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, jun 2022.

[18] Adrian F Hernandez, Melissa A Greiner, Gregg C Fonarow, Bradley G Hammill, Paul A Heidenreich, Clyde W Yancy, Eric D Peterson, and Lesley H Curtis. Relationship between early physician follow-up and 30-day readmission among medicare beneficiaries hospitalized for heart failure. *Jama*, 303(17):1716–1722, 2010.

[19] Ambarish Pandey, Harsh Golwala, Haolin Xu, Adam D DeVore, Roland Matsouaka, Michael Pencina, Dharam J Kumbhani, Adrian F Hernandez, Deepak L Bhatt, Paul A Heidenreich, et al. Association of 30-day readmission metric for heart failure under the hospital readmissions reduction program with quality of care and outcomes. *JACC: Heart Failure*, 4(12):935–946, 2016.

[20] Elizabeth H Bradley, Leslie Curry, Leora I Horwitz, Heather Sipsma, Yongfei Wang, Mary Norine Walsh, Don Goldmann, Neal White, Ileana L Piña, and Harlan M Krumholz. Hospital strategies associated with 30-day readmission rates for patients with heart failure. *Circulation: Cardiovascular Quality and Outcomes*, 6(4):444–450, 2013.

[21] Ying-Chang Tung, Shing-Hsien Chou, Kuan-Liang Liu, I-Chang Hsieh, Lung-Sheng Wu, Chia-Pin Lin, Ming-Shien Wen, and Pao-Hsien Chu. Worse prognosis in heart failure patients with 30-day readmission. *Acta Cardiologica Sinica*, 32(6):698, 2016.

[22] Devika Subramanian, Venkataraman Subramanian, Anita Deswal, and Douglas L Mann. New predictive models of heart failure mortality using time-series measurements and ensemble models. *Circulation: Heart Failure*, 4(4):456–462, 2011.

[23] John A Spertus, Martha J Radford, Nathan R Every, Edward F Ellerbeck, Eric D Peterson, and Harlan M Krumholz. Challenges and opportunities in quantifying the quality of care for acute myocardial infarction: summary from the acute myocardial infarction working group of the american heart association/american college of cardiology first scientific forum on quality of care and outcomes research in cardiovascular disease and stroke. *Circulation*, 107(12):1681–1691, 2003.

[24] Ankur Gupta, Larry A Allen, Deepak L Bhatt, Margueritte Cox, Adam D DeVore, Paul A Heidenreich, Adrian F Hernandez, Eric D Peterson, Roland A Matsouaka, Clyde W Yancy, et al. Association of the hospital readmissions reduction program implementation with readmission and mortality outcomes in heart failure. *JAMA cardiology*, 3(1):44–53, 2018.

[25] Wouter Ouwerkerk, Adriaan A Voors, and Aeilko H Zwinderman. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *JACC: Heart Failure*, 2(5):429–436, 2014.

[26] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[27] Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Sequential multi-dimensional self-supervised learning for clinical time series. 2023.

[28] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[29] Lars Heiliger, Anjany Sekuboyina, Bjoern Menze, Jan Egger, and Jens Kleesiek. Beyond medical imaging-a review of multimodal deep learning in radiology. *TechRxiv*, (19103432), 2022.

[30] Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[33] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[34] Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pages 338–353. PMLR, 2022.

[35] Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pages 156–167. PMLR, 2021.

[36] Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

[37] Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac Kohane. Event stream gpt: A data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *arXiv preprint arXiv:2306.11547*, 2023.

[38] Colin Raffel and Daniel P. W. Ellis. Feed-forward networks with attention can solve some long-term memory problems, 2016.

A Dataset

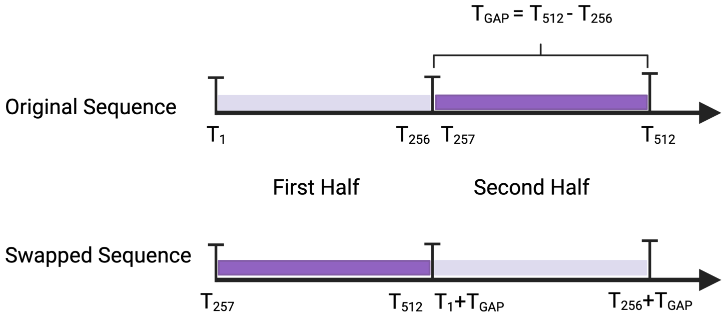
In our heart failure cohort, we include patients who have at least one inpatient admission event with at least 16 data points for both pre-admission and post-admission data. See Table 3 for pretraining dataset statistics.

Table 3: **Statistics for pretraining datasets.**

Task	# Patients	# Events
OCP Pretraining	107,268	✗
EBCL Pretraining	107,268	383,254

We use the triplet embedding strategy from Tipirneni et al [8] for modeling sequential EHR data, and this allows flexibility in how dates are encoded. For all non-OCP experiments, specifically the EBCL and supervised experiments, we encode dates as the relative time from the inpatient admission event. Since relative dates are used, we normalize the OCP and non-OCP relative dates by dividing by the standard deviation of all inpatient admission times in the dataset.

Figure 2: **Illustration of datetime encoding used for all OCP experiments.** We show the date encoding used for an original sequence and a swapped sequence in a scenario where the input trajectory has 512 data points.



OCP Data Processing We prepare a continuous trajectory of 512 consecutive data points. This trajectory might either be maintained in its original order or have its two halves swapped (Figure 2. For the case where we keep the original ordering, the first data point, T_1, is set to time 0, with subsequent data points indicating the time elapsed since T_1. Alternatively, if we swap the trajectory, the first data point of the latter half, T_{257}, becomes time 0. Other data points then denote the time difference from T_{257}. We further adjust the dates of the initial half by adding the time gap, $T_{\text{GAP}} = T_{512} - T_{256}$. This adjustment ensures that the gap between T_{256} and T_{257} remains unaltered, regardless of whether the sequence is swapped or not. This ensures that the time between the first and second half (time from T_{256} to T_{257}) can not be leveraged to detect swaps. Moreover, both the swapped and unaltered sequences ascend from zero, meaning the OCP model can't solely rely on sign or monotonicity. Instead, it must be discerned from the data patterns whether a swap has taken place.

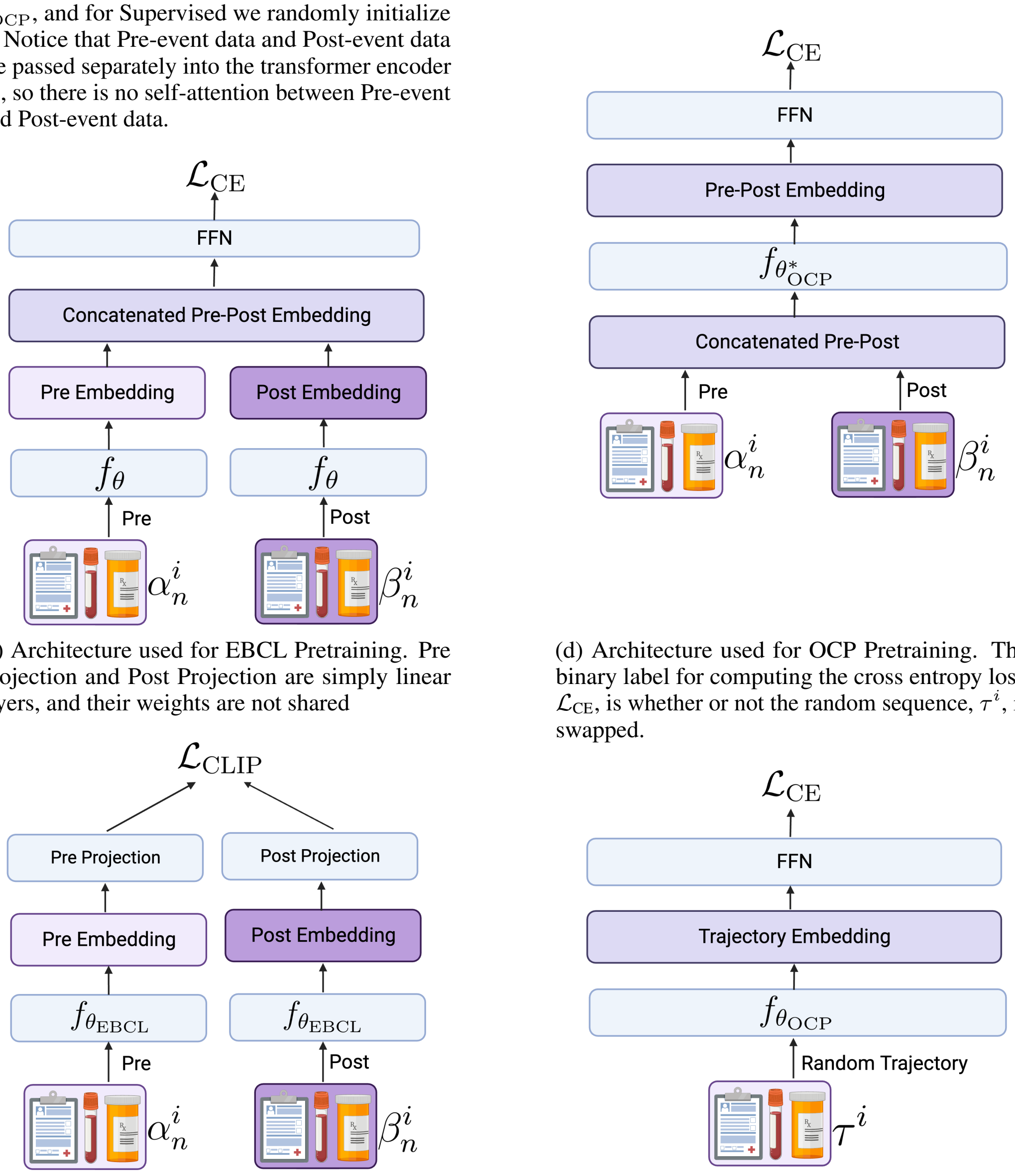
B Architectures

Figure 3 summarizes the model architecture used for pre-training and finetuning for EBCL, OCP, and Fully Supervised Experiments.

C OCP Late Fusion

For our OCP finetuning experiments, we used a standard architecture with pre and post data concatenated before being passed to f_θ (Figure 3b). This more standard architecture differs from what is

Figure 3: **Model Architectures**. FFN refers to a one hidden layer feed-forward network with a ReLU activation.



(a) This architecture is used for EBCL, OCP Late-Fusion, and Supervised experiments. For EBCL we initialize θ with the final weights from EBCL pretraining, θ^*_{EBCL}, for OCP Late-Fusion we initialize with the final weights from OCP pretraining, θ^*_{OCP}, and for Supervised we randomly initialize θ. Notice that Pre-event data and Post-event data are passed separately into the transformer encoder f_θ, so there is no self-attention between Pre-event and Post-event data.

(b) Architecture used for OCP finetuning. The pre and post data is concatenated before being fed to the transformer allowing for attention between pre and post data. We initialize f_θ with the OCP pretrained weights θ^*_{OCP}.

(c) Architecture used for EBCL Pretraining. Pre Projection and Post Projection are simply linear layers, and their weights are not shared

(d) Architecture used for OCP Pretraining. The binary label for computing the cross entropy loss, $\mathcal{L}_{\mathrm{CE}}$, is whether or not the random sequence, τ^i, is swapped.

used for the EBCL and Supervised experiments where we pass the pre and post data to f_θ separately (Figure 3a). For a fair comparison, we try OCP Late fusion (**OCP LF**), using the same finetuning architecture used for EBCL and Supervised experiments (Figure 3a), but initializing f_θ with the learned OCP pretraining weights θ^*_{OCP}.

Table 4: Performance of all models on finetuning tasks.

	30-Day Readmission		1-Year Mortality		1-Week LOS	
	AUC	APR	AUC	APR	AUC	APR
Supervised	64.3 ± 1.1	89.0 ± 0.5	80.2 ± 0.7	90.7 ± 0.3	86.3 ± 0.6	85.7 ± 0.5
OCP LF	64.6 ± 1.1	89.1 ± 0.6	79.1 ± 0.2	90.3 ± 0.1	84.1 ± 0.7	83.7 ± 0.6
OCP LF Frozen	61.5 ± 0.2	87.4 ± 0.1	67.1 ± 0.3	82.8 ± 0.2	59.9 ± 0.1	55.3 ± 0.3
OCP	68.1 ± 0.5	90.6 ± 0.2	78.0 ± 0.7	89.6 ± 0.3	84.1 ± 0.7	83.7 ± 0.6
OCP Frozen	59.1 ± 0.5	86.2 ± 0.4	65.9 ± 0.4	81.9 ± 0.2	59.9 ± 0.1	55.3 ± 0.3
EBCL	**70.4 ± 0.1**	**91.4 ± 0.1**	**82.3 ± 0.1**	**91.8 ± 0.0**	**87.2 ± 0.2**	**86.9 ± 0.3**
EBCL Frozen	**68.2 ± 0.3**	**90.6 ± 0.1**	78.4 ± 0.1	90.0 ± 0.1	79.7 ± 0.1	79.2 ± 0.1

We observe that OCP LF was consistently outperformed by EBCL, providing further evidence that EBCL is a superior pretraining task.