A Holistic Time-Aware Classification Model for Multimodal Longitudinal Patient Data

Tobias Susetzky[1], Huaqi Qiu[1], Rickmer Braren[2], and Daniel Rueckert[1,3,4]

[1] Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany
[2] TUM University Hospital, Munich, Germany
[3] Department of Computing, Imperial College London, UK
[4] Munich Center for Machine Learning (MCML), Munich, Germany
{tobias.susetzky, huaqi.qiu, rbraren, daniel.rueckert}@tum.de

Abstract. Current prognostic and diagnostic AI models for healthcare often limit informational input capacity by being time-agnostic and focusing on single modalities, therefore lacking the holistic perspective clinicians rely on. To address this, we introduce a **T**ime-**A**ware **M**ulti**M**odal Transformer **E**ncoder (TAMME) for longitudinal medical data. Unlike most state-of-the-art models, TAMME integrates longitudinal imaging, textual, numerical, and categorical data together with temporal information. Each element is represented as the sum of embeddings for high-level categorical type, further specification of this type, time-related data, and value. This composition overcomes limitations of a closed input vocabulary, enabling generalization to novel data. Additionally, with temporal context including the delta to the preceding element, we eliminate the requirement for evenly sampled input sequences. For long-term EHRs, the model employs a novel summarization mechanism that processes sequences piecewise and prepends recent data with history representations in end-to-end training. This enables balancing recent information with historical signals via self-attention. We demonstrate TAMME's capabilities using data from 431k+ hospital stays, 73k ICU stays, and 425k Emergency Department (ED) visits from the MIMIC dataset for clinical classification tasks: prediction of triage acuity, length of stay, and readmission. We show superior performance over state-of-the-art approaches especially gained from long-term data. Overall, our approach provides versatile processing of entire patient trajectories as a whole to enhance predictive performance on clinical tasks. Code is available at github.com/go31glX57/tamme.

Keywords: Multimodal Learning · Temporal Modeling · Longitudinal Health Records · Clinical AI .

1 Introduction

In clinical decision-making, practitioners exploit the full range of information available for an individual patient, comprising medical imaging, textual reports,

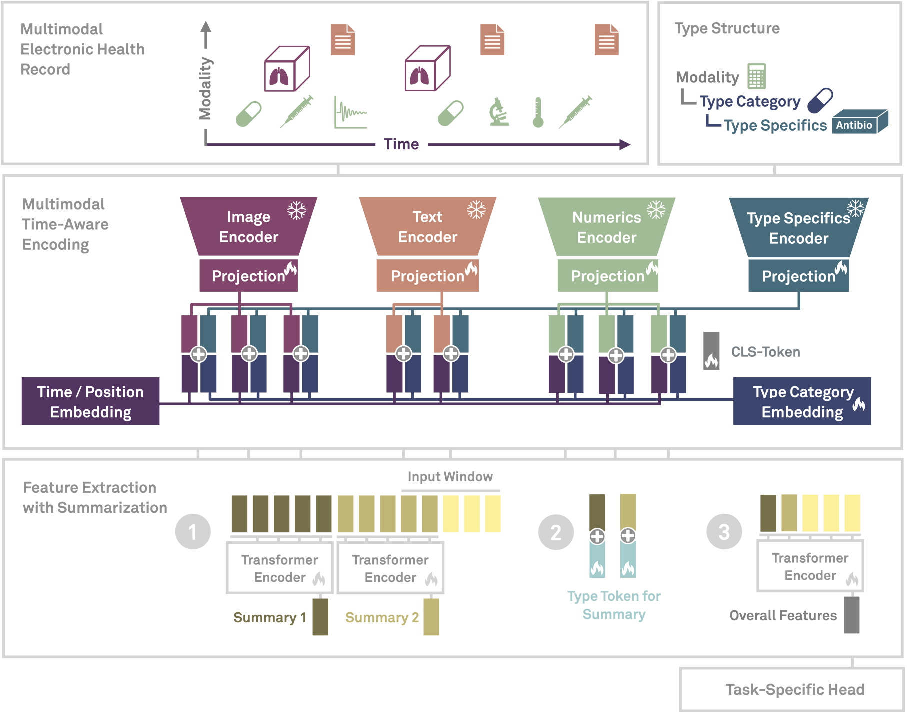

Fig. 1. TAMME is a transformer encoder that takes in a sequence of EHR elements of different modalities comprising numerics, categoricals, imaging, and text. Each element is represented as a token resulting from the sum of embeddings for value, position/time, type specifics, and type category. The latter are learned during training; numerical values are represented by Fourier Features; for other values, external pre-trained models are used as encoders. All embeddings are projected into a shared space by trainable projection modules. By this, a uniform sequence of tokens is obtained that can be processed by a regular transformer with a task-specific head. In order to cope with long sequences beyond regular capacity in an end-to-end fashion, we add a summarization mechanism. By this, features for overflowing subsequences are extracted piece-wise. Subsequently, a dedicated trainable type token is added to those summaries before pre-pendeding them to the most recent items. The transformer encoder then extracts the final features from this resulting sequence. Throughout the whole diagram, the flame icon indicates trainable components while a snowflake symbol is attached to modules that are frozen during the whole training process.

prescriptions, vitals, previously applied procedures, and more. This multimodal data is typically used to put the current health status of a patient into the context of their individual records and disease progression. In particular, incorporated temporal information for each entry allows doctors to compare items longitudinally when assessing individual relevancy. Recent AI models in healthcare also incorporate multiple modalities but are mostly limited to a few modalities [20,5] or focus on vision-language tasks [1]. Moreover, they often omit longitudinal dynamics or temporal information and isolate recent single-time events. However, in extensive real-world datasets such as MIMIC [8,9,7,11], the patient's history contains the vast majority of information (cf. Figure 3). We propose a versatile transformer-based classification model capable of processing both recent and historical data of a patient as one holistic sequence along with temporal information while weighting its elements individually. We show that this time-aware multimodal transformer encoder can facilitate multiple clinically relevant downstream tasks. Our key contributions can be summarized as follows:

1. We develop a highly-generalizable approach of representing rich EHRs including a wide variety of modalities as a sequence of uniform tokens within a joint embedding space. In particular, it allows a model to extend to novel data types such as pharmaceuticals without re-training.
2. We propose the integration of temporal information, eliminating the need for equally sampled input sequences and consequently removing the need for explicit handling of missing data.
3. We introduce a summarization mechanism to learn from arbitrary long EHRs in an end-to-end manner.
4. We demonstrate our approach on the MIMIC dataset, where the resulting model outperforms baseline approaches in five of six tasks.

2 Related Works

The clinical classification tasks our model is trained for have been widely covered in literature, in particular for the MIMIC-IV dataset. For predicting triage acuity in ED, a decision tree trained on polynomial features from vitals, medication, complaints, and metadata achieves SOTA results [23]. MeTra [12], a modified vision transformer, is the first use of a transformer for ICU survival prediction. It prepends patch projections of one image per patient with tokens for selected numerical data from a limited timespan. On MIMIC-IV, XGBoost succeeded in ICU survival prediction [16] from crafted numerical features, while MedFuse [5] combines numerical EHR elements with images to predict in-hospital mortality. ICU length of stay classification has been done using BERT-based models [19], Random Forest Classifier [6], or an LSTM model [3].The duration of hospital stay and readmission have been successfully predicted, for example, by LSTM models [3] or a so-called Hierarchical Attention Network [4], again using numerical clinical data. There is also a considerable amount of work leveraging the longitudinal data structure of EHRs, some also incorporating temporal information. MOTOR [21] and BEHRT [15] both use a sequence of categorical

EHR events timestamped with the current patient age to train a transformer for prognosis, albeit with highly different architecture and training strategies. Another proposed option for time integration is explicitly scaling the transformer's self-attention to prioritize recent over historic data [13]. There have also been experiments with both, time-aware attention scaling and time embeddings that are added to the input tokens for longitudinal imaging data [14]. Early work on multimodal sequence modeling including temporal aspects uses a transformer to process numeric and categorical ICU data together with patient metadata and history statistics [20]. We significantly go beyond this with a highly versatile input representation that can include all kinds of numeric, categorical, and meta data as well as texts and images, while still being extendable to novel data types. Further, our model incorporates all relevant relative temporal information and is capable of processing long-term history restricted in length only by the availability of computational power.

3 Methods

Modular Token Representation

We propose a novel approach for holistic token representations of multi-modal EHRs. In our setting, a temporally ordered sequence of numerical, categorical, textual, and imaging EHR elements from a single patient constitutes a sample. Each element consists of a high-level type category (e.g., drug, procedure), free text type specifics (e.g., drug name, description), temporal information, and a value (image, text, number, or none for categoricals). TAMME obtains a uniform sequence where each token representation is the sum of embeddings for type category, type specifics, time, and value (cf. Figure 1). To obtain the value embeddings, numerics are encoded as Fourier Features [22] while for images and texts, external domain-specific pre-trained models are used, namely RAD-DINO [17] and BioLORD 2023 [18]. All value embeddings remain frozen throughout the whole training process. For temporal information, we use a sinusoidal representation of patient age, time delta to the preceding EHR element, element position within the sequence, and an optional time delta to an anchor event such as the latest hospital admission. For type information, a large-scale closed vocabulary would limit extendability of a trained model to novel datasets and EHR elements such as unseen drugs. We overcome this by splitting type information into a high-level category and specifics. In close consultation with experienced clinicians, we define a minimal yet comprehensive set of 32 categories which are represented by randomly initialized and fully trainable embeddings. These categories are further specified by free-text details which are statically encoded by the same domain-specific pre-trained model as used for text values. Figure 1 includes an illustration. By this, we obtain semantically valid embeddings which are robust to most data inconsistencies in spelling, terminology, etc. and technically enable the zero-shot application of a trained model to novel data. Overall, we propose a versatile approach for joint learning from most kinds of patient data with longitudinal structure and temporal information.

Model Architecture

TAMME employs a well-known transformer architecture with minor modifications: input sequences are being encoded as described above; static embeddings for type specifics as well as for numeric, textual, and image values are projected to a shared space by linear layers that are trained end-to-end along with the other building blocks of the model. The projections are combined with the trainable type category embedding via addition, resulting in a sequence of uniform tokens. Instead of using standard positional embeddings, we leverage the temporal information of each token represented as sinusoidal frequencies and add them to the corresponding tokens. A transformer encoder is used as a feature extractor on these token sequences together with a CLS token. We attach a specific head and train the whole model separately for each one of a selection of classification tasks relevant to clinical decision making. To enable the full processing of individual EHRs far beyond the regular input length n of the model, TAMME features piecewise summarization: the remainder of an input sequence exceeding specified capacity by m tokens is processed piecewise. Each of the $s := \lceil \frac{m}{n} \rceil$ overflowing subsequences of maximal length n is being replaced by the feature vector produced by the model for this isolate subsequence. The most recent $n - s$ tokens of the input sequence are kept unchanged and prepended with the s summary tokens. An illustration is included in Figure 1. To mark a summary token as such, a learnable type embedding is applied via addition. The entire setup remains trainable in an end-to-end manner and enables the complete processing of individual patient journeys in the form of long-term EHRs.

Self-Supervised Pre-training

We also explore a variant where we first pre-train TAMME using Masked Sequence Modelling in a self-supervised fashion. Respecting imbalance among high-level type categories, we apply biased-random sampling when selecting the tokens to mask. Only the unmasked tokens are processed by the transformer encoder and a transformer decoder model is added to reconstruct the masked tokens from context. We task the model to recover the type category as well the embeddings for type specifics and value for each of the masked tokens. In a second stage, we omit the decoder and fine-tune the transformer encoder module for each of our tasks individually.

4 Experiments & Results

To investigate our approach we merge MIMIC-IV [8], MIMIC-ED [7], MIMIC-Note [9], as well as MIMIC-CXR [11,10], all at version 2.2 from PhysioNet [2]. From MIMIC-Note, we use the findings section of radiology reports and all contents of discharge summaries. We add information such as patient age, precompute static embeddings, and create subsets centered around hospital, ED, and ICU stays. All data is split randomly into training, validation, and test set using a 70-15-15 split. We ensure all data from one patient is only included in one split. The dataset contains 425M+ single events from 303k+ patients, 431k+ hospital stays, 73k+ ICU stays, and 425k+ ED visits. We train and evaluate our model on ED visits, stationary hospital stays, and ICU stays for the following classification tasks, where applicable: length of stay, survival, readmission, and triage acuity (i.e. assigning a criticality level to patients at ED, see [7]). Table 1 provides an overview. For each stay, we label a patient as survived, if there is no death charted within the respective stay nor within 24h after discharge, i.e. the end of the stay. This temporal tolerance accounts for precision shortcomings in the original charting and data collection process of MIMIC. To prevent information leakage, we exclude events from the model's input that might reveal the prediction target, such as the current discharge summary. We constrain the model's input data based on the covered time span. By default, it ranges from the onset of the current stay to the earliest of the following: prediction target event, end of stay, or a fixed maximum of 12 hours from the beginning of the stay. For readmission prediction, the 12-hour cap is omitted, allowing the model to leverage all data of the most recent stay. For all tasks, we experiment with extending this time window to historic records dating back up to one week, six months, or one year prior to the beginning of the current stay. The results detailed in Table 2 generally indicate performance improvements by historic data, albeit subject to the specific task. In comparison to training task-specific models from scratch, we also evaluate fine-tuning a task-agnostic model that has been pre-trained in a self-supervised manner (see Section 3). We assume that uniform pre-training affects performance equally for all context windows used in fine-tuning. Thus, we only use to best configuration from Table 2 to reduce computational cost. Pre-training is done on 400k+ multimodal EHR subsequences obtained by a sliding window of size 3000. At a second stage, we then fine-tune the resulting model on the exact same data and parameter set as we did when training from scratch. Further, we benchmark against the state of the art baseline. Overall, our trainings are done on 2-8 Nvidia A100 GPUs for up to 103 hours, highly depending on the length of input sequences. As summarized in Table 3, we overall observe slight gains from pre-training, while TAMME exceeds baseline performance in five of six tasks. For the triage acuity task, we follow [23] using a subsampled dataset to tackle heavy class imbalance. We still observe inferior performance of our general approach when compared to the specialized baseline. We hypothesize that short-term signals relevant for triage might be easily cluttered by long-term data. We suspect a similar signal-vs-noise tradeoff explains the minor performance drops when using an extended EHR window in some tasks. However, this negative impact from long-term information is marginal, whereas its positive impact is significant (Tasks 2, 3, 5). Overall, we discover that the ideal EHR input window varies among tasks.

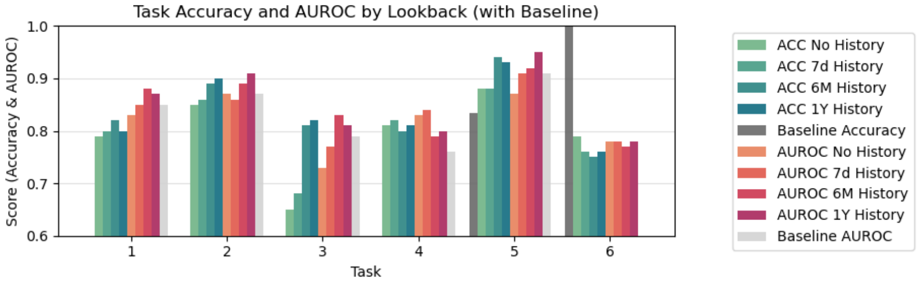

Fig. 2. Comparison of classification performance of TAMME using different extents of included EHR history versus the performance of the baseline model.

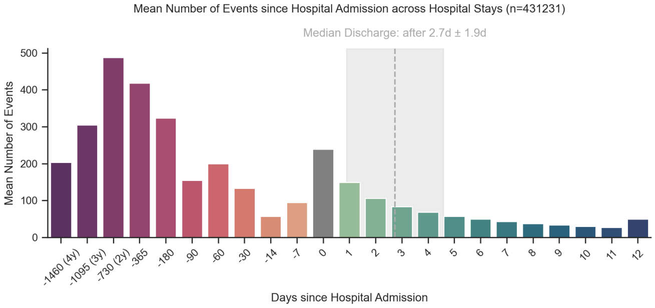

Fig. 3. The mean number of EHR entries (events) per day before and after admission, across all patients in the MIMIC dataset. For readability, historic data points are aggregated into bars covering time spans larger than one day. For instance, on average, patients have around 300 EHR entries charted 90 to 180 days before admission. Overall, this illustrates that patient history is much richer on information compared to the current stay.

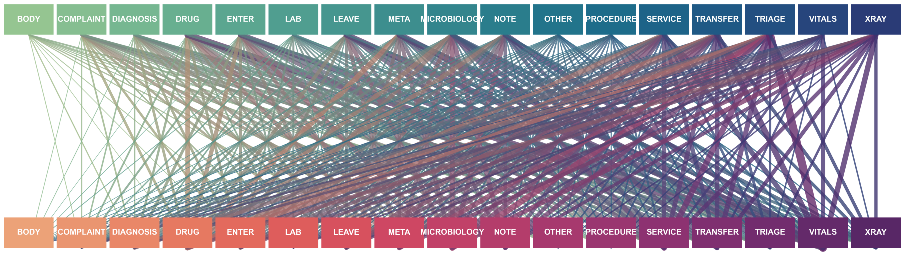

Fig. 4. Average attention among high-level type categories of a model trained for prediction of hospital length of stay. Higher line thickness indicates increased attention. We observe that, for instance, vitals appear to be crucial for interpreting most other data, especially for triage scores.

Table 1. Task targets and classes.

Task ID	Target	Classes
1	Hospital Length of Stay	short ($\leq$3d), long ($>$3d)
2	In-Hospital Survival	True, False
3	Hospital Readmission	within 30 days, after 30 days / never
4	ICU Length of Stay	short ($\leq$3d), long ($>$3d)
5	In-ICU Survival	True, False
6	ED Triage Acuity	1, 2, 3, 4, 5

Table 2. Classification performance for different extents of included historic EHRs.

Task ID	**Target**	**No History**		**7 Days**		**6 Months**		**1 Year**	
		ACC	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**
1	Hospital Length of Stay	0.79	0.83	0.80	0.85	**0.82**	**0.88**	0.80	0.87
2	In-Hospital Survival	0.85	0.87	0.86	0.86	0.89	0.89	**0.90**	**0.91**
3	Hospital Readmission	0.65	0.73	0.68	0.77	0.81	**0.83**	**0.82**	0.81
4	ICU Length of Stay	0.81	0.83	**0.82**	**0.84**	0.80	0.79	0.81	0.80
5	In-ICU Survival	0.88	0.87	0.88	0.91	**0.94**	0.92	0.93	**0.95**
6	ED Triage Acuity	**0.79**	**0.78**	0.76	0.78	0.75	0.77	0.76	0.78

5 Discussion & Conclusion

We introduced TAMME, a time-aware transformer encoder for multimodal longitudinal EHRs. Unlike existing models, TAMME is extendable, versatile, and integrates most clinically relevant data types while preserving longitudinal structure and temporal information. Using piecewise summarization, it can process long patient histories effectively. While still being limited by the availability of computational power, this enhances control on resource usage and contributes to the model's ability to distinguish between recent and historical signals. Our experiments demonstrate that leveraging extended patient histories can indeed improve predictive performance on clinical tasks, surpassing state-of-the-art baselines in five of six tasks. This highlights the importance of capturing temporal dependencies and multimodal information for targeted decision-making. Further evaluation on other datasets and additional clinical tasks, such as personalized treatment recommendations, might help to assess the model's broader applicability as well as a clinical trial, exploring the model's reasoning, and an extensive ablation study beyond the length of historic context. In conclusion, TAMME offers a scalable approach for processing multimodal longitudinal EHRs, advancing clinical predictive modeling and decision support.

Acknowledgments. This work benefited from resources provided by the joint project "Open Medical Inference", a Module 3 project of the Medical Informatics Initiative of the Federal Government, funded by the German Federal Ministry for Research, Technology, and Space (Grant Number 01ZZ2315B).

Table 3. Classification performance of our model trained from scratch vs. a fine-tuned variant based on self-supervised pre-training, all compared against the baseline. For each task, the best configuration according to Table 2 has been used.

Task ID	Target	Ours				Baseline		
		pre-trained		from scratch		result		ref
		ACC	AUC	ACC	AUC	ACC	AUC	
1	Hospital Length of Stay	0.83	**0.89**	0.82	0.88	-	0.85	[3]
2	In-Hospital Survival	0.96	**0.95**	0.90	0.91	-	0.87	[5]
3	Hospital Readmission	0.81	**0.84**	0.81	0.83	-	0.79	[3]
4	ICU Length of Stay	0.82	0.83	0.82	**0.84**	-	0.76	[3]
5	In-ICU Survival	0.94	**0.95**	0.93	**0.95**	0.83	0.91	[16]
6	ED Triage Acuity	0.83	0.81	0.79	0.78	**0.99**	-	[23]

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
2. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000). https://doi.org/10.1161/01.CIR.101.23.e215, https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215, [Online]
3. Gupta, M., Gallamoza, B., Cutrona, N., Dhakal, P., Poulain, R., Beheshti, R.: An extensive data processing pipeline for mimic-iv. In: Machine Learning for Health. pp. 311–325. PMLR (2022)
4. Harerimana, G., Kim, J.W., Jang, B.: A deep attention model to forecast the length of stay and the in-hospital mortality right on admission from icd codes and demographic data. Journal of biomedical informatics **118**, 103778 (2021)
5. Hayat, N., Geras, K.J., Shamout, F.E.: Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In: Machine Learning for Healthcare Conference. pp. 479–503. PMLR (2022)
6. Hempel, L., Sadeghi, S., Kirsten, T.: Prediction of intensive care unit length of stay in the mimic-iv dataset. Applied Sciences **13**(12), 6930 (2023)
7. Johnson, A., Bulgarelli, L., Pollard, T., Celi, L.A., Mark, R., Horng, S.: Mimic-iv-ed (version 2.2) (2023). https://doi.org/10.13026/5ntk-km72, https://doi.org/10.13026/5ntk-km72
8. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv (version 2.2) (2023). https://doi.org/10.13026/6mm1-ek67, https://doi.org/10.13026/6mm1-ek67

9. Johnson, A., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv-note: Deidentified free-text clinical notes (version 2.2) (2023). https://doi.org/10.13026/1n74-ne17, https://doi.org/10.13026/1n74-ne17
10. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**, 317 (2019). https://doi.org/10.1038/s41597-019-0322-0, https://doi.org/10.1038/s41597-019-0322-0
11. Johnson, A.E.W., Pollard, T.J., Mark, R., Berkowitz, S.J., Horng, S.: Mimic-cxr database (version 2.0.0) (2019). https://doi.org/10.13026/C2JT1Q, https://doi.org/10.13026/C2JT1Q
12. Khader, F., Kather, J.N., Müller-Franzes, G., Wang, T., Han, T., Tayebi Arasteh, S., Hamesch, K., Bressem, K., Haarburger, C., Stegmaier, J., et al.: Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. Scientific Reports **13**(1), 10666 (2023)
13. Li, T.Z., Still, J.M., Xu, K., Lee, H.H., Cai, L.Y., Krishnan, A.R., Gao, R., Khan, M.S., Antic, S., Kammer, M., et al.: Longitudinal multimodal transformer integrating imaging and latent clinical signatures from routine ehrs for pulmonary nodule classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 649–659. Springer (2023)
14. Li, T.Z., Xu, K., Gao, R., Tang, Y., Lasko, T.A., Maldonado, F., Sandler, K.L., Landman, B.A.: Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography. In: Medical Imaging 2023: Image Processing. vol. 12464, pp. 229–238. SPIE (2023)
15. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: Behrt: transformer for electronic health records. Scientific reports **10**(1), 7155 (2020)
16. Pang, K., Li, L., Ouyang, W., Liu, X., Tang, Y.: Establishment of icu mortality risk prediction models with machine learning algorithm using mimic-iv database. Diagnostics **12**(5), 1068 (2022)
17. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., Wetscherek, M.T., Codella, N., Hyland, S.L., Alvarez-Valle, J., Oktay, O.: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence (Jan 2025). https://doi.org/10.1038/s42256-024-00965-w, https://doi.org/10.1038/s42256-024-00965-w
18. Remy, F., Demuynck, K., Demeester, T.: BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. Journal of the American Medical Informatics Association p. ocae029 (02 2024). https://doi.org/10.1093/jamia/ocae029, https://doi.org/10.1093/jamia/ocae029
19. Röhr, T., Figueroa, A., Papaioannou, J.M., Fallon, C., Bressem, K., Nejdl, W., Löser, A.: Revisiting clinical outcome prediction for mimic-iv. In: Proceedings of the 6th Clinical Natural Language Processing Workshop. pp. 208–217 (2024)
20. Shickel, B., Silva, B., Ozrazgat-Baslanti, T., Ren, Y., Khezeli, K., Guan, Z., Tighe, P.J., Bihorac, A., Rashidi, P.: Multi-dimensional patient acuity estimation with longitudinal ehr tokenization and flexible transformer networks. Frontiers in Digital Health **4**, 1029191 (2022)
21. Steinberg, E., Fries, J., Xu, Y., Shah, N.: Motor: a time-to-event foundation model for structured medical records. arXiv preprint arXiv:2301.03150 (2023)

22. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
23. Yaddaden, Y., Benahmed, Y., Rioux, M.D., Kallel, M.: Machine learning-based pre-diagnosis tools in emergency departments: Predicting hospitalization, mortality and triage acuity. In: 2023 IEEE Third International Conference on Signal, Control and Communication (SCC). pp. 1–6. IEEE (2023)