CXR-TFT: Multi-Modal Temporal Fusion Transformer for Predicting Chest X-ray Trajectories

Mehak Arora[1*], Ayman Ali[1], Kaiyuan Wu[1], Carolyn Davis[2], Takashi Shimazui[2], Mahmoud Alwakeel[1], Victor Moas[1], Philip Yang[2], Annette Esper[2], and Rishikesan Kamaleswaran[1]

[1] Duke University, Durham NC 27708 {mehak.arora, ayman.ali, vincent.wu, mahmoud.alwakeel, victor.moas, r.kamaleswaran}@duke.edu
[2] Emory University, Atlanta GA 30322
cmydavis@gmail.com,{tshima2,philip.yang,aesper}@emory.edu

Abstract. In intensive care units (ICUs), patients with complex clinical conditions require vigilant monitoring and prompt interventions. Chest X-rays (CXRs) are a vital diagnostic tool, providing insights into clinical trajectories, but their irregular acquisition limits their utility. Existing tools for CXR interpretation are constrained by cross-sectional analysis, failing to capture temporal dynamics. To address this, we introduce CXR-TFT, a novel multi-modal framework that integrates temporally sparse CXR imaging and radiology reports with high-frequency clinical data—such as vital signs, laboratory values, and respiratory flow sheets—to predict the trajectory of CXR findings in critically ill patients. CXR-TFT leverages latent embeddings from a vision encoder that are temporally aligned with hourly clinical data through interpolation. A transformer is trained to predict CXR embeddings at each hour, conditioned on previous CXR embeddings and clinical measurements. In a retrospective study of 20,000 ICU patients, CXR-TFT demonstrated 95% accuracy in predicting abnormal CXR findings 12 hours before they became radiographically evident, indicating that clinical data contains valuable respiratory state progression information. By providing distinctive temporal resolution in prognostic CXR analysis, CXR-TFT offers actionable predictions with the potential to improve the management of time-sensitive critical conditions, where early intervention is crucial but timely diagnosis is challenging.

Keywords: Clinical Trajectories · Multi-modal Machine Learning · Irregularly Sampled Time Series

1 Introduction

Intensive Care Unit (ICU) patients generally have complex and diverse clinical pathologies that require careful monitoring and timely intervention. Portable

[*] Corresponding Author: mehak.arora@duke.edu

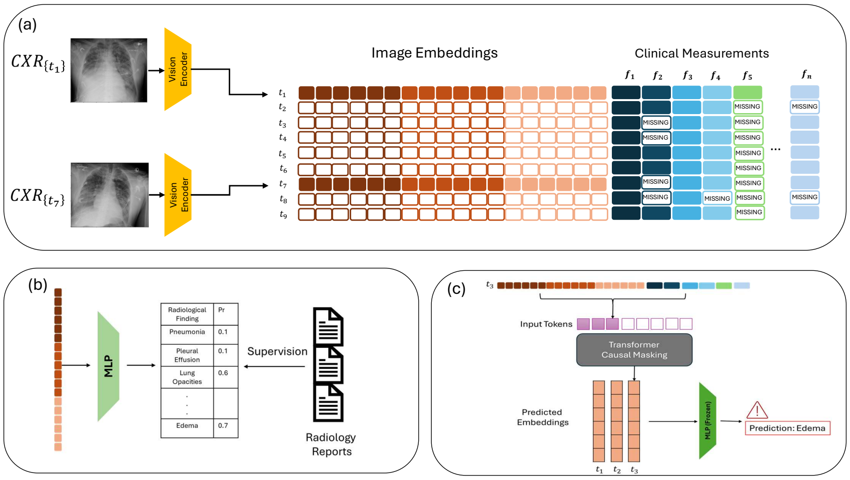


Fig. 1: The CXR-TFT Framework.(a) Sparsely recorded CXR images and irregularly sampled clinical measurements are concatenated at the input to the transformer model, (b) A Multi-layer Perceptron is trained to detect radiographic findings vision encoder embeddings, with ground truth supervision from radiology reports, (c) CXR-TFT estimates future CXR embeddings, which can predict the likelihood of radiographic findings before they are seen on subsequent CXRs.

chest radiographs (CXRs) are the most requested ICU imaging for many reasons: they are rapid to obtain, can be done at the bedside (critical for unstable patients), are used to evaluate support devices and lines, and can provide important diagnostic information, particularly for pulmonary pathology [20]. Many conditions are first diagnosed with CXRs, such as consolidation indicative of pneumonia, new pleural effusions in the setting of volume overload, or pulmonary edema [16]. These arise as complications and carry significant morbidity and mortality, making early recognition and intervention critical [3,14].

Many contemporary machine learning models that are applied in the ICU setting—for tasks like cohort phenotyping or outcome prediction—either only leverage radiology reads of CXRs or do not use imaging data altogether [11,19]. However, CXRs contain valuable information that influences clinical decision making, and radiology reports are often delayed and may not convey all information pertinent to the patient's comprehensive health state. Independently, there has been significant research on using machine learning for CXR interpretation [1] as well as CXR generation[4]. Recent foundational medical imaging models [25,5,22] have successfully learned rich CXR representations, with longitudinal approaches outperforming cross-sectional analysis [9,2,15]. However, there is a crucial need to better integrate longitudinal imaging information with other modalities of ICU data for building better clinical decision support tools.

In this study, we leverage CXR foundational models to predict ICU patient trajectories, hypothesizing that the most likely CXR could be estimated at any point during a patient's ICU stay using the previously recorded CXR and all past clinical measurements. To accomplish this, we develop CXR-TFT (Chest X-ray Temporal Fusion Transformer), a transformer-based model that integrates hourly clinical measurements—such as lab values, vital signs, and ventilator parameters—with previous CXR embeddings to predict the most probable CXR representation *in the latent space*. The latent embedding space of a vision-language model is continuous[17] and imbued with semantic meaning [25]. Operating in this space allows for interpolation between embeddings, helping us overcome the challenge of temporally aligning information from multi-modal, irregularly-sampled time series and forming the key technical contribution of this work. Predicting future pulmonary pathology could potentially shorten the time to clinical intervention and improve clinical decision support.

2 Methods

2.1 The Proposed Framework

A high-level overview of our framework is shown in Figure 1. At any time t_k during a patient's stay in the ICU, given a sequence of clinical measurements $F = \{F_{t_0}, F_{t_1}, F_{t_2} ..., F_{t_{k-1}}\}$, where $F_t = [f_t^1, f_t^2, ..., f_t^n]$ is a vector of n clinical features under consideration, and given a sequence of sparsely sampled, previously recorded CXR images $I^p = \{I_{t_0}^p, (\bullet), I_{t_2}^p ..., I_{t_{k-1}}^p\}$ where $I_t^p = \text{encoder}_{\text{vision}}(\text{CXR}_t)$ is the latent embedding representation of a CXR image obtained via a pretrained vision encoder, and $(\bullet)$ represents time points with no recorded CXR scans, the proposed model learns to predict $I_{t_k}^E$, the estimated CXR embedding at time t_k. The target output sequence $I^T = \{I_{t_0}^T, I_{t_0}^T, I_{t_2}^T ..., I_{t_{k-1}}^T\}$ used to train the model is obtained by linear interpolation in the embedding space between two recorded CXRs. Concretely, if a CXR scan was performed at time t_{k_1}, and the next CXR scan was performed at t_{k_2}, then the target sequence is defined by Equation 1.

$$I_{t_{k'}}^T = \begin{cases} I_{k_1}^T = \text{encoder}_{\text{vision}}(\text{CXR}_{t_{k_1}}), & \text{if } k' = k_1 \\ I_{k_2}^T = \text{encoder}_{\text{vision}}(\text{CXR}_{t_{k_2}}), & \text{if } k' = k_2 \\ \frac{I_{k_2}^T - I_{k_1}^T}{k_2 - k_1} \times (k' - k_1) + I_{k_1}^T, & \text{if } k_1 < k' < k_2 \end{cases} \tag{1}$$

2.2 Dataset Preparation

This is a single-center retrospective cohort study at an academic institution. Included were all adult patients admitted to any ICU between January 2015 to December 2021 who had more than one CXR performed during their hospitalization. A total of 17,690 patients met criteria, for which we extracted all single-view anteroposterior (AP) frontal chest radiographic images and their corresponding radiology reports. We also extracted demographic information and clinical measurements like vitals, laboratory values, ventilator flowsheet information across the entire ICU length of stay.

2.3 Data Preprocessing

Clinical Measurements All clinical measurements from the Electronic Medical Record (EMR) were organized into hourly bins. Multiple hourly recordings were validated against physiologically possible bounds (determined through clinician consultation), with out-of-range values discarded and the remaining values averaged. Numerical values were min-max normalized using healthy patient reference ranges. Missing clinical measurements were handled with forward-fill imputation. If no recorded value existed, missing values were imputed using the median of the normal (healthy) range of values. Categorical variables (gender, ICU type, etc.) were one-hot encoded. This processing resulted in a clinical feature vector $F_t = [f_t^1, f_t^2, ..., f_t^n]$ with $n = 82$.

Image Encoding BioMedCLIP [25], a vision language model trained to align radiology reports with corresponding image embeddings, was used to extract the latent space representation $I_{t_k} \in \mathcal{R}^{512}$ of a chest X-ray image at time t_k. Data preceding the first recorded CXR and following the last recorded CXR was excluded for training and evaluation. To facilitate training, missing values, $(\bullet)$ in the previous CXR sequence $I^p = \{I^p_{t_0}, (\bullet), I^p_{t_2} ..., I^p_{t_{k-1}}\}$ were handled using forward-fill imputation.

Radiology Reports Radiology reports provided supervision for training a downstream classifier to predict radiological findings from image embeddings. We derived 10 finding classes using the CheXPert labeler [12]: 'No Finding', 'Cardiomegaly', 'Lung Opacity', 'Edema', 'Consolidation', 'Pneumonia', 'Atelectasis', 'Pneumothorax', 'Pleural Effusion', and 'Pleural Other'.

2.4 Training CXR-TFT

The input data to the transformer model at time t_k, $X_{t_k} = [F^T_{t_k}, I^{p\,T}_{t_k}]$, was a 594×1 vector formed by the concatenation of current clinical features and the latent embedding of the previously recorded CXR. We trained an encoder-decoder transformer model [21] with a pre-norm architecture, an initial learning rate of $5e - 4$, and the AdamW optimizer with a weight decay of 0.01. Gradient clipping was used to prevent exploding gradients. The model was trained for 100 epochs with an early stopping patience of 10 epochs based on validation loss. We used a batch size of 32 and a cosine learning rate scheduler with warmup for the first 10% of training steps. To prevent overfitting, we applied dropout with a rate of 0.1 throughout the network. The mean squared error (MSE) loss between the target CXR embeddings and the decoder outputs was used as the primary optimization objective. The code to the complete data processing and training setup can be found at our Github Repository.

2.5 Classifier Regularization

We trained a lightweight multilayer perceptron (MLP) to predict key radiological findings using labels from radiology reports (Section 2.3). The MLP used BioMedCLIP vision encoder embeddings as input and was trained on the MIMIC-CXR dataset [13] (over 200,000 CXR images with reports).

The cross-entropy loss between the target labels and the predicted labels (obtained by passing the embeddings generated by the transformer through the pretrained MLP) was added to the training objective (Equation 2). This regularization encourages predicted trajectories to accurately forecast abnormal findings. For N training samples and C classes radiological findings, each with sequence length T_i where $i \in \{0, 1, \ldots, N\}$, the training objective is given by Equation 2, where $y_{i,c,t} = \mathrm{MLP}(I_t^T)$ and $p_{i,c,t} = \mathrm{MLP}(I_t^p)$, MLP is the frozen, radiological-finding classifier, θ are the parameters of our model. For this study, $C = 10$ and $\alpha = 0.5$.

$$
\begin{aligned}
\mathcal{L}_{MSE}(\theta) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}(1-\alpha)\left\|I_t^E - I_t^T\right\|_2^2 \\
\mathcal{L}_{\mathrm{BCE}}(\theta) &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{c=1}^{C}\left[y_{i,c,t}\log(p_{i,c,t}) + (1-y_{i,c,t})\log(1-p_{i,c,t})\right] \\
\mathcal{L}(\theta) &= (1-\alpha)L_{MSE}(\theta) + \alpha L_{BCE}(\theta)
\end{aligned}
\tag{2}
$$

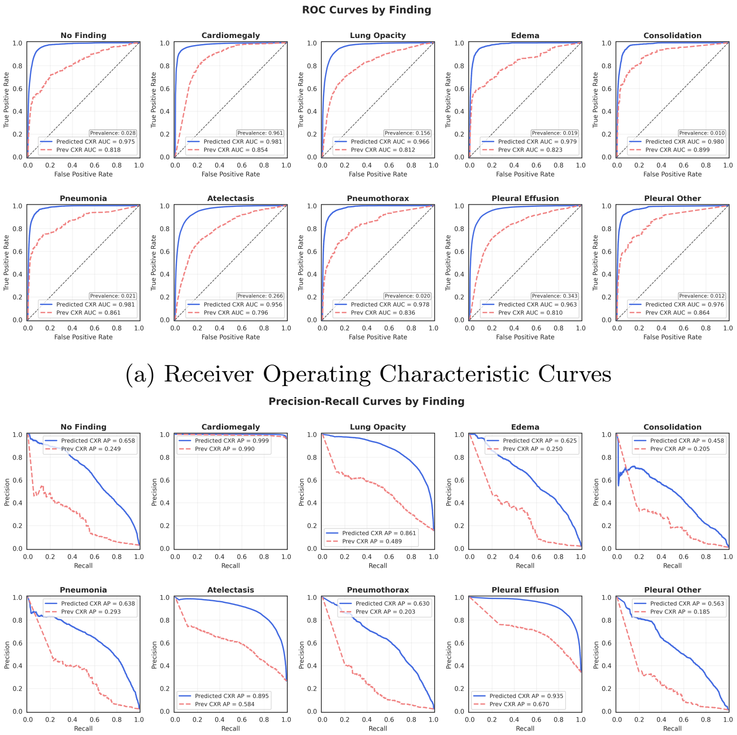
3 Results on Predicting Radiographic Findings

The radiographic-findings classifier (Section 2.5) was used to calculate the likelihood of abnormal findings in CXR embeddings generated by CXR-TFT. Following [15], previously recorded CXR findings formed the baseline, highlighting our model's utility in predicting new developments. The accuracy, precision, and

Table 1: Model Performance Metrics Across Time Horizons

Finding	Current Prediction						Future Prediction											
	Precision		Recall		Accuracy		12-hours in advance						24-hours in advance					
	M	B	M	B	M	B	Precision		Recall		Acc		Precision		Recall		Acc	
							M	B	M	B	M	B	M	B	M	B	M	B
No Finding	0.730	0.308	0.498	0.438	0.981	0.957	0.617	0.196	0.405	0.279	0.976	0.948	0.528	0.157	0.328	0.220	0.973	0.945
Cardiomegaly	0.985	0.978	0.992	0.979	0.978	0.959	0.982	0.973	0.990	0.975	0.973	0.951	0.979	0.971	0.987	0.973	0.967	0.947
Lung Opacity	0.786	0.528	0.767	0.556	0.931	0.853	0.721	0.417	0.676	0.431	0.907	0.815	0.642	0.374	0.574	0.378	0.880	0.798
Edema	0.682	0.329	0.476	0.488	0.986	0.972	0.555	0.207	0.347	0.302	0.983	0.965	0.460	0.174	0.256	0.242	0.979	0.963
Consolidation	0.588	0.263	0.362	0.424	0.991	0.982	0.513	0.132	0.275	0.212	0.990	0.978	0.360	0.079	0.175	0.127	0.989	0.976
Pneumonia	0.703	0.368	0.478	0.488	0.985	0.971	0.615	0.233	0.365	0.304	0.982	0.964	0.522	0.188	0.280	0.237	0.979	0.961
Atelectasis	0.837	0.592	0.791	0.603	0.904	0.784	0.769	0.494	0.723	0.502	0.869	0.731	0.700	0.453	0.651	0.459	0.833	0.709
Pneumothorax	0.667	0.247	0.506	0.393	0.985	0.964	0.569	0.126	0.378	0.191	0.981	0.955	0.439	0.102	0.257	0.146	0.976	0.953
Pleural Effusion	0.874	0.655	0.846	0.705	0.905	0.771	0.830	0.582	0.780	0.620	0.869	0.715	0.779	0.551	0.712	0.576	0.829	0.688
Pleural Other	0.579	0.256	0.488	0.388	0.989	0.979	0.490	0.117	0.353	0.170	0.987	0.973	0.344	0.078	0.211	0.106	0.984	0.972
Average	**0.743**	**0.452**	**0.620**	**0.546**	**0.964**	**0.919**	**0.666**	**0.348**	**0.529**	**0.399**	**0.952**	**0.900**	**0.575**	**0.313**	**0.443**	**0.346**	**0.939**	**0.891**

Note: M = CXR-TFT Model, B = Baseline, Acc = Accuracy.



(a) Receiver Operating Characteristic Curves

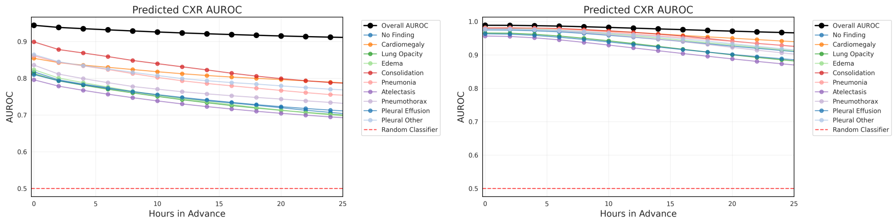
(b) Precision Recall Curves

Fig. 2: Performance comparison of detecting radiographic findings on embeddings predicted by the transformer model, and baseline of the previously recorded CXR. Figure (a) denotes the prevalence of each class in the test set.

recall are reported in Table 1. The “Current Prediction” results evaluate model performance by comparing predicted labels at each hour, with labels derived from interpolated target CXR trajectories. The “Future Prediction” results assess the model by comparing predicted labels with ground truth labels from radiology reports corresponding to the subsequent CXR for time-points with a recorded CXR and without a prior CXR recorded within the lead time frame. The Receiver-Operating Characteristic Curves (ROC) and Precision-Recall Curves (PRC) for “Current Predictions“ are shown in Figure 2, and changes in the area under the ROC curves (AUROC) for varying prediction lead-times is shown in Figure 3.

4 Discussions

With CXR-TFT, we demonstrate that modeling CXR trajectories in the vision-language latent space enables temporally-aligned integration of CXR and clinical data, accurately predicting abnormal findings 12-24 hours before they appear on



(a) Baseline: Previous CXR Embeddings (b) Predicted CXR Embeddings

Fig. 3: Temporal AUROC trends for early prediction of radiological findings with prediction horizon measured in hours before the next recorded CXR.

subsequent CXRs. This has critical clinical implications—predicting the develpement of abnormal pathology can prompt early diagnostic imaging and subsequent clinical intervention. For example, predicting pneumonia before overt clinical indication may enable earlier antibiotics and reduced complications.

Most previous radiological trajectory research is limited to broad categorizations (worsening, stable, improving) [8] or predicting severity outcomes (mortality, ICU readmission) [7], offering limited clinical utility as their predictions rarely influence real-time patient management. In contrast, our approach generates actionable, bedside predictions reflecting important physiological changes at a higher temporal resolution than the relatively sparse chest X-rays alone. Recent research on multi-modal CXR models [10,24] train unimodal encoders for static tasks like phenotyping or mortality prediction, while CXR-TFT predicts hourly latent embeddings in a seq-to-seq manner, addressing the temporal alignment challenge. Other studies address data asynchronicity by generating CXRs to fill in temporal gaps [23]. The closest work to our research is the diffusion-based CXR generation model proposed by Kyung et al. [15], which is conditioned on the most recent CXR and clinical data. CXR-TFT diverges in three key aspects: (1) predicting future CXRs in the latent embedding space rather than pixel space, improving efficiency while eliminating hallucination risks; (2) incorporating comprehensive contextual information from clinical measurements and previous CXR embeddings from admission until prediction time; (3) achieving finer-grained monitoring capabilities by hourly temporal alignment between CXR embeddings and clinical measurements.

This study has limitations. Since clinical studies show that restrictive, clinically-indicated imaging achieves similar outcomes to daily CXRs [20,6], and most ICUs don't perform routine daily CXRs, assessing direct clinical impact requires prospective study. Next, we used a single institution for our cohort, which limits the generalization of our findings. Lastly, our work focuses on a single approach to the sequence-to-sequence task, but could be improved by exploring alternative model architectures, different multi-modal fusion strategies [18], and more sophisticated embedding interpolation techniques.

5 Conclusion

This work demonstrates that multimodal models using clinical time-series and pretrained vision-language embeddings can successfully predict future radiological findings. Future directions of research include further retrospective and prospective clinical studies to validate findings, exploring different model architectures, and expansion to include other data and imaging modalities.

Acknowledgments. This study was funded by the National Institutes of Health (Award Numbers R01GM139967, R21GM148931).

Disclosure of Interests. The authors have no competing interests to disclose.

References

1. Ahmad, H.K., Milne, M.R., Buchlak, Q.D., Ektas, N., Sanderson, G., Chamtie, H., Karunasena, S., Chiang, J., Holt, X., Tang, C.H., et al.: Machine learning augmented interpretation of chest x-rays: a systematic review. Diagnostics **13**(4), 743 (2023)
2. 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)
3. Bellani, G., Pham, T., Laffey, J.G.: Missed or delayed diagnosis of ards: a common and serious problem. Intensive care medicine **46**, 1180–1183 (2020)
4. Bluethgen, C., Chambon, P., Delbrouck, J.B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S.: A vision–language foundation model for the generation of realistic chest x-ray images. Nature Biomedical Engineering pp. 1–13 (2024)
5. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
6. Clec'h, C., Simon, P., Hamdi, A., Hamza, L., Karoubi, P., Fosse, J.P., Gonzalez, F., Vincent, F., Cohen, Y.: Are daily routine chest radiographs useful in critically ill, mechanically ventilated patients? a randomized study. Intensive care medicine **34**, 264–270 (2008)
7. Duanmu, H., Ren, T., Li, H., Mehta, N., Singer, A.J., Levsky, J.M., Lipton, M.L., Duong, T.Q.: Deep learning of longitudinal chest x-ray and clinical variables predicts duration on ventilator and mortality in covid-19 patients. Biomedical engineering online **21**(1), 77 (2022)
8. Gourdeau, D., Potvin, O., Archambault, P., Chartrand-Lefebvre, C., Dieumegarde, L., Forghani, R., Gagné, C., Hains, A., Hornstein, D., Le, H., et al.: Tracking and predicting covid-19 radiological trajectory on chest x-rays using deep learning. Scientific reports **12**(1), 5616 (2022)

9. Gu, Y., Yang, J., Usuyama, N., Li, C., Zhang, S., Lungren, M.P., Gao, J., Poon, H.: Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys (2023), https://arxiv.org/abs/2310.10765
10. Guerra-Manzanares, A., Shamout, F.E.: Mind: Modality-informed knowledge distillation framework for multimodal clinical prediction tasks. arXiv preprint arXiv:2502.01158 (2025)
11. Gutierrez, G.: Artificial intelligence in the intensive care unit. Critical Care **24**, 1–9 (2020)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
13. Johnson, A.E., 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**(1), 317 (2019)
14. König, I.R., Fuchs, O., Hansen, G., von Mutius, E., Kopp, M.V.: What is precision medicine? European respiratory journal **50**(4) (2017)
15. Kyung, D., Kim, J., Kim, T., Choi, E.: Towards predicting temporal changes in a patient's chest x-ray images based on electronic health records (2024)
16. Laroia, A.T., Donnelly, E.F., Henry, T.S., Berry, M.F., Boiselle, P.M., Colletti, P.M., Kuzniewski, C.T., Maldonado, F., Olsen, K.M., Raptis, C.A., et al.: Acr appropriateness criteria® intensive care unit patients. Journal of the American College of Radiology **18**(5), S62–S72 (2021)
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
18. Rasekh, A., Heidari, R., Rezaie, A.H.H.M., Sedeh, P.S., Ahmadi, Z., Mitra, P., Nejdl, W.: Towards precision healthcare: Robust fusion of time series and image data. arXiv preprint arXiv:2405.15442 (2024)
19. van de Sande, D., van Genderen, M.E., Huiskens, J., Gommers, D., van Bommel, J.: Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive care medicine **47**, 750–760 (2021)
20. Toy, D., Siegel, M.D., Rubinowitz, A.N.: Imaging in the intensive care unit. In: Seminars in Respiratory and Critical Care Medicine. vol. 43, pp. 899–923. Thieme Medical Publishers, Inc. (2022)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
22. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.H., Kiraly, A., Kazemzadeh, S., Melamed, Z., et al.: Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. In: arXiv preprint arXiv:2308.01317 (2023)
23. Yao, W., Liu, C., Yin, K., Cheung, W., Qin, J.: Addressing asynchronicity in clinical multimodal fusion via individualized chest x-ray generation. Advances in Neural Information Processing Systems **37**, 29001–29028 (2024)
24. Yao, W., Yin, K., Cheung, W.K., Liu, J., Qin, J.: Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 16416–16424 (2024)

25. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)