XAI for Better Exploitation of Text in Medical Decision Support

Ajay Madhavan Ravichandran[1] **Julianna Grune**[1] **Nils Feldhus**[1]
Aljoscha Burchardt[1] **Sebastian Möller**[1,2] **Roland Roller**[1]
[1]German Research Center for Artificial Intelligence (DFKI)
[2]Technische Universität Berlin
{firstname.lastname}@dfki.de

Abstract

In electronic health records, text data is considered a valuable resource as it complements a medical history and may contain information that cannot be easily included in tables. But why does the inclusion of clinical texts as additional input into multimodal models, not always significantly improve the performance of medical decision-support systems? Explainable AI (XAI) might provide the answer. We examine which information in text and structured data influences the performance of models in the context of multimodal decision support for biomedical tasks. Using data from an intensive care unit and targeting a mortality prediction task, we compare information that has been considered relevant by XAI methods to the opinion of a physician.

1 Introduction

Electronic health records often contain factual information in short, tabular form, including laboratory values, diagnoses, gender, and age. They also include longer texts in various forms written for many different purposes. Depending on the origin and context of the data, the text could be a clinical or nursing note, a discharge summary, or a radiology report, to name a few. The text might provide a high-level interpretation of the current patient situation, taking different kinds and sources of information into account. The text might refer directly to some given structured facts in the database (e.g., a lab value is above borderline) but might also consider additional information such as general impressions, assumptions, and information gathered directly from the patient or other medical personnel (e.g., the patient is not very adherent). For this reason, the texts are generally considered valuable resources in the clinical routine. In the context of machine learning for healthcare, however, the inclusion of such texts has shown in various setups only marginal effects (Khadanga et al., 2019; Yang and Wu, 2021), although one would assume that the additional information and complementary perspective should improve a system's performance.

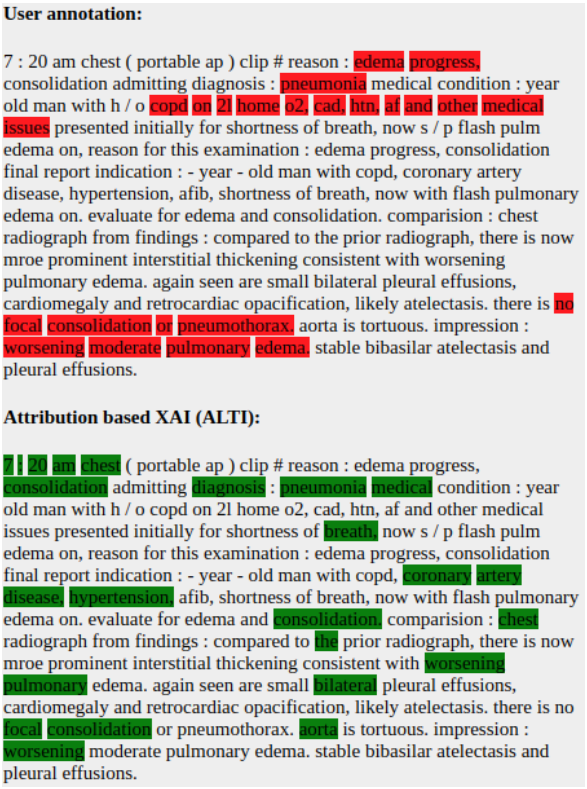

Figure 1: Comparison of human annotation regarding relevant tokens for in-hospital mortality, versus XAI

Many papers in this area deal with multi-modal data, integrating, for instance, image and text, or structured and unstructured data into one model. MIMIC-III (Johnson et al., 2016) is a popular dataset in this context, as it can be easily accessed by researchers. It contains data from an intensive care unit (ICU) of a US hospital, including patient demographics, time series data, or text, such as nursing notes, discharge summaries, or social worker notes. However, while many approaches in other domains do achieve a boost in performance using multimodal (text) data, the performance dif-

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 506–513

ference between unimodal and multi-modal models in the medical context can be modest (Deznabi et al., 2021). In this work, we explore which information is valuable for multi-modal machine learning using MIMIC data. More precisely, we re-implement two multi-model (MM) approaches for the task of in-hospital mortality prediction. We then introduce an XAI approach for the given MM approaches and examine the attributed information according to their faithfulness. Finally, we investigate if the attributions are plausible from a physician's perspective.

2 Related Work

Recent years have seen a surge in leveraging deep learning approaches utilizing diverse clinical data sources for clinical outcome predictions. These include textual clinical notes, longitudinal data, and demographic data. Unimodal approaches like CNNs (Rocheteau et al., 2021), LSTMs (Choi et al., 2016), and BERT (Naik et al., 2022) have laid the groundwork. Later expanded to multimodal approaches such as additive fusion (Khadanga et al., 2019; Deznabi et al., 2021) to more sophisticated cross attention fusion (Zhang et al., 2022; Qiao et al., 2019). Yang and Wu (2021) and Deznabi et al. (2021) implemented additive and gated fusion-based multimodal models for tasks like diagnosis prediction, acute respiratory failure prediction, and in-hospital mortality prediction. We extend their work by applying explainability methods to models and evaluating the quality of explanations.

Explainable AI (XAI) enhances transparency and trust in healthcare applications, especially within medical decision support systems (Markus et al., 2021) and clinical NLP (Roller et al., 2022a). Notably, Naylor et al. (2021) compared the faithfulness of various explanation methods for models like BERT in mortality prediction. Additionally, DeYoung et al. (2020) introduced a benchmark with human annotations to evaluate NLP models explainability for faithfulness and plausibility. However, previous research has mainly focused on quantitative evaluations of explainability methods for uni models. This study addresses this gap by quantitatively evaluating XAI in multimodal models.

3 Method

3.1 Data

We use the Medical Information Mart for Intensive Care (MIMIC-III) dataset (Johnson et al., 2016) in our experiments. MIMIC comprises authentic electronic health record (EHR) data, including vital signs, laboratory measurements, and clinical notes (free text), from ICU patients. One of its tasks involves predicting patient mortality risk in the intensive care unit (ICU) based on the first 48 hours of patient stay. Mortality, in this context, refers to the likelihood of a patient dying while receiving intensive care.

For our cohort selection and setup, we mostly follow Harutyunyan et al. (2019) and Yang and Wu (2021) and focus on patients aged 18 years and older with ICU stays lasting 48 hours or more, accompanied by clinical notes. The original cohort of Harutyunyan et al. (2019) includes 17 different features that undergo different pre-processing steps, such as inserting missing information by previous or plausible default values and converting them into time series data. As we are particularly interested in text data, we extend the original cohort by two different sources of text, namely nursing notes and admission notes.[1]

The final data consists of three different modalities: a) text, consisting of either nursing notes or admission notes; b) time series data, such as heart rate, blood pressure, or glucose; and c) time-invariant data, such as age or ethnicity. While some time series features are numeric, others are categorical (e.g., Glasgow coma scale eye-opening), which are converted into several binary features during pre-processing following the approach of Harutyunyan et al. (2019). More details about data imputation and a synthetic example of the data are added to the Appendix B.

3.2 Multimodal Models

In this study, we employ diverse architectures to encode information from different modalities into latent vectors. Specifically, we use LSTMs to process time series data, linear layers to handle time-invariant data, and transformer models for textual data. To integrate all the encoded information effectively, we use two fusion approaches: The gated fusion approach proposed by Yang and Wu (2021) and the concatenation fusion approach introduced by Deznabi et al. (2021). In the gated fusion approach, a gated attention mechanism is applied over the encoded vectors to generate a fused representation that incorporates context from all the encoded

[1] Explanation of this terminology can be found in Appendix A.

vectors. Conversely, in the concatenated fusion approach, all the encoded vectors are concatenated into a single vector to produce a fused representation. Figure 2 depicts a simplified overview of the multimodal architectures. Subsequently, the fused vector from both fusion approaches is projected into a fully connected layer for prediction.

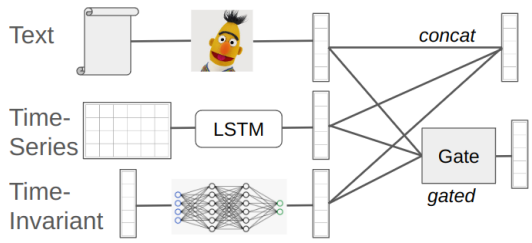

Figure 2: Combining modalities using concatenation or gated fusion.

Both approaches use a pre-trained ClinicalBERT (Huang et al., 2019) to encode the clinical notes (nursing, radiology, others, etc.). Both approaches use the average embedding over all the clinical notes as the encoded textual feature.

3.3 Multi-Modal XAI

To identify which information is crucial for successful predictions in our multimodal setup, we integrate XAI techniques using state-of-the-art methods based on gradient and attention. For pinpointing significant information in time series data processed by the LSTM, we employ Integrated Gradients (IG) (Sundararajan et al., 2017). For the textual data fed into the BERT model, we use the attention vector norm (Kobayashi et al., 2020) and layer-wise Token-to-token Interaction (ALTI) (Ferrando et al., 2022). These methods have shown promising results in explaining transformer-based models. They let us identify relevant tokens in the texts and pertinent features in the time series data, which we can then compare with annotations provided by medical professionals.

4 Experimental Results

Our first experiment concerns the reproduction of multimodal and unimodal methods and application to the in-hospital mortality task. For the evaluation, we follow a similar methodology to related work (§2), utilizing ROC (Area Under the Receiver Operating Characteristic curve) and AUPR (Area Under the Precision-Recall curve).

In the second experiment, we apply XAI to the models and examine which information is considered by the model as valuable for the prediction. Following Jacovi and Goldberg (2020), we explore faithfulness by replacing the top X attributed token or time point of the time series with a mask token or zero and observe the drop in model performance.

Finally, we conduct a plausibility test, as suggested by DeYoung et al. (2020). Here, we directly compare the attributions on text and structured data to the relevant information according to a physician's perspective. Only annotation of text data is quantitatively analyzed based on the overlap between annotated tokens and attributed tokens, such overlap matching is not possible for time-series data. As we are particularly interested in examining the benefit of text data, we randomly select 100 patient cases in which a multi-modal approach predicts a higher probability score for mortality than the unimodal LSTM approach. Likewise, we randomly select 100 cases in which the multi-modal approach predicts a lower probability score for mortality. For those cases, we assume that text data provided additional information to make a stronger prediction assumption (independent of whether the prediction is correct or not).

A final-year medical student annotated these 200 cases. The student was asked to identify parts of the text and important time-series features that support the outcome of mortality or survival. In addition, the student was asked to provide their estimation of the patient's survival and whether the text was useful in solving the task.

4.1 Results

Model performance: Unimodal vs. multi-modal Table 1 presents the results of the two multimodal approaches in comparison to the unimodal models for both text types. The first observation is that LSTM provides stronger results compared to the two BERT approaches, and all multimodal approaches outperform the unimodal models. This slight performance gain is particularly visible when using nursing notes in comparison to admission notes. Moreover, the more complex gated mechanism shows a slight benefit over the concatenation. Overall, the presented results are comparable to what has been reported already in other related work (Khadanga et al., 2019; Lyu et al., 2022).

We can conclude that for the given data and the given problem, structured (time series) data seems to have a stronger influence on the model performance, and adding both 'worlds' can lead to further, but rather minor, improvements. However, an ad-

Table 1: Performance of multimodal (MM) approaches in comparison to the unimodal models in predicting in-hospital mortality according to ROC and AUPR.

Model	ROC	AUPR
BERT (nursing)	0.80	0.37
BERT (admission)	0.74	0.30
LSTM	0.80	0.42
ConcatMM (nursing)	0.81	**0.44**
ConcatMM (admission)	0.81	0.37
GatedMM (nursing)	**0.83**	0.43
GatedMM (admission)	0.82	0.39
LSTM w/o height+weight	0.78	0.38

ditional analysis of the data reveals that the two features *height* and *weight* are often missing and imputed with default values. For this reason, we removed those two features from the original data and trained an LSTM. Without those two features, however, the model suffers a drop in performance.

Explanation faithfulness test Table 2 presents the faithfulness test, in which we examine which information influences the models' prediction. To do so, we replace the top-5 (top-10 and top-15) strongest (XAI) attributed tokens or time points of the time-series data and compare this to a random replacement of the same amount of information. The table shows that removing the attributed tokens leads to a stronger drop in performance, compared to the random removal. This indicates that the model relies on information (and particularly text tokens), which are useful for the mortality prediction task.

Table 2: ROC Performance after replacement of top-X text tokens or time point of time-series data. The table compares a random replacement against the replacement of attributed information (XAI). The table compares BERT (admission) with ROC=0.80 for text and MM with ROC=0.83.

Modality	Top	Attribution	Random
Text	5	0.769 (0.031)	0.801 (0.000)
	10	0.744 (0.056)	0.800 (0.000)
	15	0.734 (0.066)	0.799 (0.001)
Struct.	5	**0.664** (0.166)	0.726 (0.104)
	10	**0.595** (0.235)	0.674 (0.156)
	15	**0.585** (0.245)	0.632 (0.198)

Regarding the attributed tokens in the text data, we found the following patterns: First, highly attributed information is often spread widely across the document. In many cases, attributed tokens in a document include medical conditions such as symptoms or diseases (e.g., pain, cirrhosis, pneumonia), in some others also body parts such as heart or lung and sometimes medications. However, many other seemingly irrelevant tokens are highlighted, such as the word *patient* or a specific time mentioned in the text. Finally, even though information tends to be spread across the document, the attribution also covers sequences of words, such as the patient's age ('53 y. o. man'), negations ('denies pain'), and other connected information ('chest pain,' 'renal failure').

When looking closer at the attributed time-series data, the following five features play a particularly important role in the model's performance drop: Glasgow Coma Scale (total), blood pressure (mean), Glasgow Coma Scale (motor response), oxygen saturation, and Glasgow Coma Scale (verbal response).

Explanation plausibility evaluation For the 200 patient cases that a physician annotated, we first conducted a manual analysis to find differences and similarities to the attributed tokens. Figure 1 depicts an example text with human and machine (XAI) annotation. In general, the annotations show that, in many cases, a few larger chunks of text sequences were annotated. Moreover, even though severe conditions seem to be mentioned multiple times in the documents (redundancy), the physician often annotated each condition just once – the explanations, however, also highlight the same condition in multiple parts of the document. Moreover, the physician annotated some measurements of values as relevant, whereas XAI never detected anything comparable – although it considers, in some cases, age and gender as useful. On a time-series data, the physician considers similar information useful compared to XAI, namely the Glasgow Coma Scale (eye-opening), the Glasgow Coma Scale (motor response), the Glasgow Coma Scale (total), oxygen saturation, and respiratory rate.

Table 3: Plausibility evaluation measuring agreement with human-annotated of the clinical text (nursing and admission) for mortality prediction. The table shows the lenient-f1 scores obtained by measuring the overlap between the annotated token and the attributed token.

Model	Precision	Recall	Lenient-F1
BERT (nursing)	**0.141**	**0.204**	**0.166**
BERT (admission)	0.064	0.090	0.075
ConcatMM (nursing)	0.102	0.159	0.124
GatedMM (admission)	0.110	0.168	0.133

Second, we quantitatively evaluated plausibility

by measuring the lenient-F1 score for the overlap between annotated and attributed text. Since our main focus is textual data, we did not create annotations for time points, making such overlap evaluation impossible for time series data. Table 3 shows that the BERT model attributions align more closely with human annotations for nursing notes, while multimodal models exhibit lower agreement with human annotations. However, the overall agreement, as measured by the lenient-F1 score, is very low. This low agreement is likely because the models struggle to differentiate between acute conditions (e.g., active bleeding, signs of severe infection) and pre-existing conditions (e.g., pneumonia, diabetes mellitus), missing out on the negation of medical conditions by attending only to pathology (e.g., 'no melena' is annotated by physician but the model attribution identifies only "melena").

5 Discussion

The initial results align with findings from related work: text data is a valuable resource for improving predictions, but its benefit varies depending on the task and the text source. For instance, nursing notes led to higher results than admission notes, despite the fact that nursing notes were often truncated due to BERT's restricted input length. Given the redundancy in clinical texts, it may be beneficial to compress larger texts into shorter documents to accommodate additional text sources.

Another notable finding is the performance drop when removing height and weight, two features that are often missing and filled with default values. Our medical expert confirmed that *height* and *weight* do not influence the given task, which may reduce overall trust in our model. However, it is not unusual for machine learning models to consider seemingly irrelevant information as useful. For example, in Roller et al. (2022b), a nephrology outcome prediction model found the number of lab measurements in the last month to be very useful, which may indirectly indicate a patient's deteriorating condition. In our case, the model's reliance on height and weight might be justified by the context in which these features are used. For instance, weight may be measured over time to monitor fluid balance. Thus, the model might be capturing an important dependency that is not immediately apparent.

In the second experiment evaluating the faithfulness of the attributions, we observed a significant drop in model performance when the top-attributed information was replaced in the input, compared to a random replacement. This stronger decline in performance was particularly pronounced when time-series data was replaced, indicating that time-series information plays a crucial role in the model's performance for the given task. Conversely, it also shows that some tokens, such as medical conditions that are mentioned in the text, have a positive influence on the model.

In the third experiment, comparing human and XAI annotations of texts suggests that systems can extract relevant information (pre-existing conditions are identified more often than acute conditions). On the other hand, the extracted information is not always humanly plausible. The comparison of human and XAI annotated time-series features showed that both the physician and the model consider similar features useful for the given prediction task. However, multimodal quantitative analysis of plausibility remains a bottleneck that should be addressed in future work.

6 Conclusion

We analyzed the relevance of text and structured data in the context of a multimodal decision support system for in-hospital mortality. We found that the source of text influences the model performance (nursing vs admission notes). Moreover, sparse information (e.g., patient height and weight) can benefit the performance of models, although such information does seem irrelevant from an expert's perspective.

In our experiments, we found that the model performance drops considerably when structured information (time series) is replaced in the input compared to textual inputs. In general text data could provide additional context in a multimodal setup, but its benefit depends on the task (other tasks might lead to more benefits) as our results showed only a marginal boost in performance compared to unimodal models.

Finally, our comparison between human and XAI annotations of the texts indicates that the models can extract relevant information but not always. It seems that for multimodal data such as text and time series, quantitative analysis of plausibility is a bottleneck, and it should be addressed in future work.

Limitations

Our approach has clear limitations in terms of applied models (for instance, a multimodal LLM could have been tested) as well as additional XAI methods (e.g. LIME or SHAP). Moreover, in order to gain more insights into the human perspective, a large-scale annotation from a human perspective is necessary, considering additional human annotators, patient cases, and datasets.

Ethical Considerations

Although we build multimodal machine learning models for healthcare with the intention of creating a positive impact on society, our model is trained and tested only on retrospective and anonymized data. In this way, we do not influence patient outcomes.

Acknowledgments

The project has received funding from the German Federal Ministry of Education and Research (BMBF) through the project PRIMA-AI (01GP2202C).

References

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 301–318, Northeastern University, Boston, MA, USA. PMLR.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. 2021. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4026–4031, Online. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for ICU management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437, Hong Kong, China. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. 2022. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2022:719–728.

Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States. Association for Computational Linguistics.

Mitchell Naylor, Christi French, Samantha Terker, and Uday Kamath. 2021. Quantifying explainability

in NLP and analyzing algorithms for performance-explainability tradeoff. *Interpretable ML in Healthcare workshop at ICML 2021*.

Zhi Qiao, X. Wu, Shen Ge, and Wei Fan. 2019. Mnn: Multimodal attentional neural networks for diagnosis prediction. In *International Joint Conference on Artificial Intelligence*.

Emma Rocheteau, Pietro Liò, and Stephanie Hyland. 2021. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 58–68, New York, NY, USA. Association for Computing Machinery.

Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke, and Bilgin Osmanodja. 2022a. An annotated corpus of textual explanations for clinical decision support. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2317–2326, Marseille, France. European Language Resources Association.

Roland Roller, Manuel Mayrdorfer, Wiebke Duettmann, Marcel G Naik, Danilo Schmidt, Fabian Halleck, Patrik Hummel, Aljoscha Burchardt, Sebastian Möller, Peter Dabrock, Bilgin Osmanodja, and Klemens Budde. 2022b. Evaluation of a clinical decision support system for detection of patients at risk after kidney transplantation. *Frontiers in Public Health*, 10.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, and Alexander Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 881–893. Association for Computational Linguistics.

Bo Yang and Lijun Wu. 2021. How to leverage the multimodal EHR data for better medical prediction? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4029–4038, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ying Zhang, Baohang Zhou, Kehui Song, Xuhui Sui, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. 2022. PM^{2}F^{2}N: Patient multi-view multi-modal feature fusion networks for clinical outcome prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1985–1994, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Nursing and admission notes

Nursing notes are small text snippets written by medical personnel during a patient's stay in the ICU. They describe general observations, current medical conditions, and treatment. As we target the patient situation within the first 48 hours, we concatenate all nursing notes from that time into one document.

Following the work of van Aken et al. (2021), we simulate patient textual information at the time of admission by extracting the chief complaint, present illness, medications, allergies, physical exam, and family and social history from discharge summaries. We refer to this as admission notes.

B Imputation value and synthetic sample

Table 4: Shows the selected time-series values and their corresponding impute values (plausible).

Variable	Impute value
Capillary refill rate	0
Diastolic blood pressure	59.0
Fraction inspired oxygen	0.21
Glascow coma scale eye opening	4 spontaneously
Glascow coma scale motor response	6 obeys commands
Glascow coma scale total	15
Glascow coma scale verbal response	5 oriented
Glucose	128.0
Heart Rate	86
Height	170.0
Mean blood pressure	77.0
Oxygen saturation	98.0
Respiratory rate	19
Systolic blood pressure	118.0
Temperature	36.6
Weight	81.0
pH	7.4

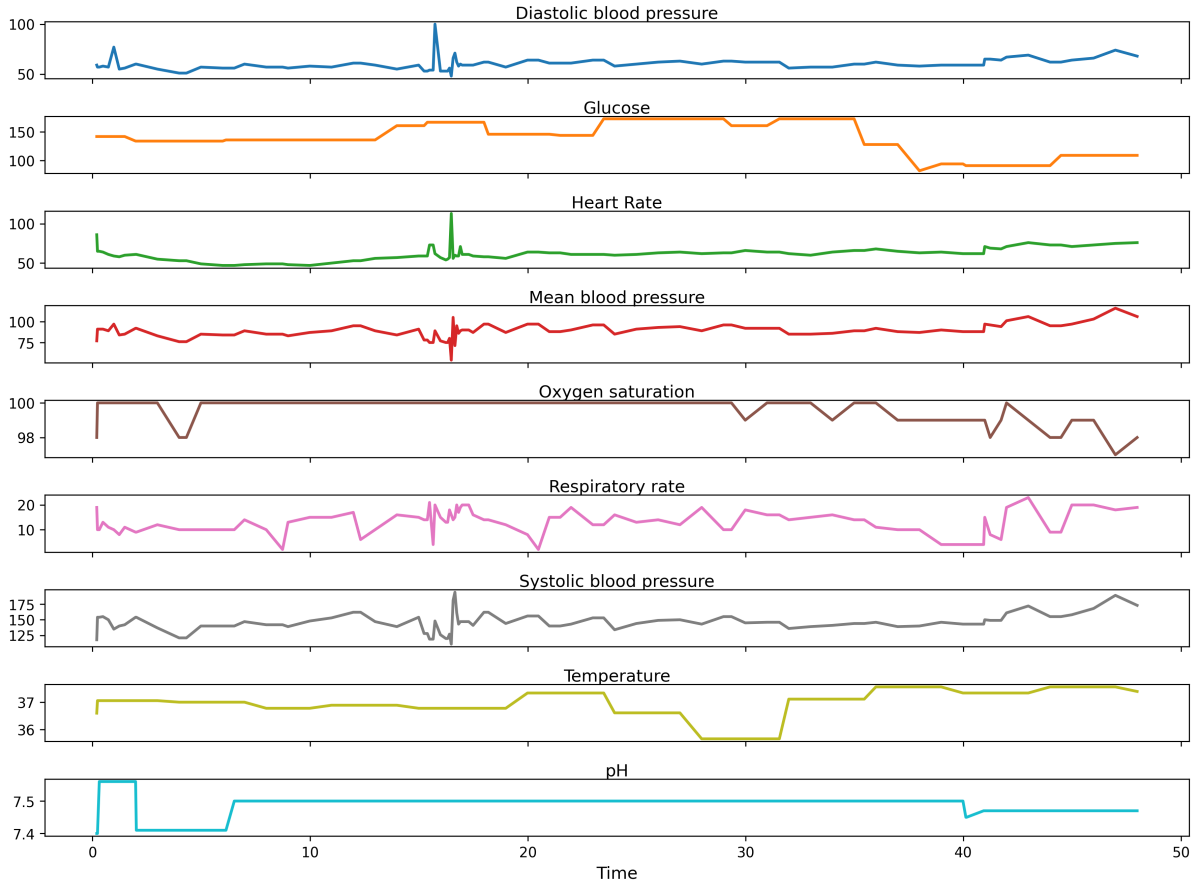

Figure 3: A synthetic sample of a patient's time-series in the MIMIC-III dataset.

Gender: Female **Age:** 32 **Ethnicity:** Hispanic

32-year-old female with a history of asthma since childhood. Admitted for severe exacerbation with respiratory distress. Received multiple nebulizations and systemic corticosteroids. Developed hypoxia overnight, required intubation and transfer to ICU for mechanical ventilation. Blood gas analysis showed severe respiratory acidosis. Managed with lung protective ventilation strategy and continuous monitoring. Family notified and involved in care decisions.

Figure 4: A synthetic sample of a patient's clinical text in the MIMIC-III dataset.