On the Importance of Clinical Notes in Multi-modal Learning for EHR Data

Severin Husmann  
Hugo Yèche  
Gunnar Rätsch *  
Rita Kuznetsova *

{shusmann,hyeche,raetsch, mkuznetsova}@ethz.ch  
Department of Computer Science, ETH Zürich

Abstract

Understanding deep learning model behavior is critical to accepting machine learning-based decision support systems in the medical community. Previous research has shown that jointly using clinical notes with electronic health record (EHR) data improved predictive performance for patient monitoring in the intensive care unit (ICU). In this work, we explore the underlying reasons for these improvements. While relying on a basic attention-based model to allow for interpretability, we first confirm that performance significantly improves over state-of-the-art EHR data models when combining EHR data and clinical notes. We then provide an analysis showing improvements arise almost exclusively from a subset of notes containing broader context on patient state rather than clinician notes. We believe such findings highlight deep learning models for EHR data to be more limited by partially-descriptive data than by modeling choice, motivating a more data-centric approach in the field.

1 Introduction

Recently, researchers have explored using deep neural networks for clinical decision support, particularly for Intensive Care Unit (ICU) prediction tasks. Most existing literature focused on using either available electronic health records [1, 2, 3, 4, 5, 6, 7] (EHR) or clinical notes [8, 9, 10, 11] individually. However, more recent research shows that combining both data types in a cross-modal fashion improves performance [12, 13] on a popular suite of MIMIC-III [14] benchmark tasks [15]. Ultimately, the goal of such improvement in the field can help better plan the resource usage in ICU [12, 16]. Though, as for any real-world application, acceptance by clinicians will require an understanding of the behavior of these new cross-modal models. Other studies provided analyses for uni-modal models using EHR data [4] or clinical notes [8]. To the best of our knowledge, similar work investigating the cross-modal interactions when combining both modalities to monitor patients in ICU does not exist.

We aim to fill that gap by exploring the underlying reasons for performance improvements obtained from jointly using EHR data and clinical notes. Like Horn et al. [4], which analyzed the performances of EHR-only models, we rely on a transformer-based [17] model, as attention weights provide a direct explanation of the deep learning model behavior. Due to the multi-modal nature of the data, we adopt a multi-modal extension [18] to this original transformer architecture. In this work, first, we

*Co-supervised

confirm that the most basic cross-modal transformer model combining EHR data and clinical notes performs significantly better than state-of-the-art uni-modal approaches on all classification tasks from the MIMIC-III benchmark [15]. Second, by performing a two-step analysis, we show that the performance gain compared to EHR-only models arises exclusively from a subset of nurses’ notes and radiology reports. These notes provide additional context on patients’ states. Moreover, we also show that using clinicians’ notes, which summarize EHR contents rather than containing additional information, does not provide any performance gain. Thereby, we refute the hypothesis that current architectures poorly represent EHR information. We believe our findings motivate more data-centric approaches in the field, focusing on leveraging more descriptive data to model patients’ states in the ICU.

2 Related work

Deep Learning for ICU patient monitoring. As previously mentioned, we can categorize existing research using deep learning to monitor patients’ evolution in the ICU according to the modality used. First, the most common line of work relies on the EHR, viewed as a multivariate time series. A large group of related literature tackles this problem with various architectures in both supervised [12, 13, 4, 5, 6] and unsupervised fashion [19, 20]. Other researchers focus exclusively on clinical notes [8, 9, 10, 11]. At last, motivating our work, more recent research explored fusing both modalities [12, 13, 21, 22], yielding significant improvement over previous uni-modal approaches. Unfortunately, in this last body of work, the authors work with independent representations for each modality making it difficult to retrieve cross-modal interaction, thus limiting interpretability.

Interpretability in deep learning. Widely used methods in deep learning interpretability rely on saliency-map-based methods like LIME [23], GRAD-CAM [24], and SHAP [25]. Though, these methods are limited [26] and are not necessarily compatible with sequence data. Thus, we base our analysis on attention-based methods, a more common approach for sequence modeling. Various papers have investigated the use of attention weights as a potential post hoc interpretability mechanism [27, 28, 29, 30]. In addition, other literature further developed methods beyond attention weights, such as attention flow and rollout [31]. In the case of attention flow, [32] proved that attention flows are layer-wise Shapley values. Attention rollout is often employed in computer vision to provide attention heatmaps over the input [33, 34]. In the context of ICU data, similarly to us, Horn et al. [4] also based their uni-modal model behavior analysis on attention weights.

3 Method

In this section, we introduce our interpretable model to handle both clinical notes and EHR data. We also describe hyperparameter choice and data pre-processing to ensure a better analysis of clinical notes’ importance.

Architecture. We consider two modalities $X_{EHR} \in \mathbb{R}^{T_{EHR} \times d_{EHR}}$ and $X_{CN} \in \mathbb{R}^{T_{CN} \times d_{CN}}$, where $T_{EHR}$ and $T_{CN}$ denote the length of the EHR data and clinical notes during the given patient’s stay, and $d_{EHR}$ and $d_{CN}$ the input feature dimensionality. To obtain $X_{CN}$ we use the pre-trained Clinical BERT model [35] as a text feature extractor, $d_{CN} = 769$. Note that $d_{CN}$ is greater than the original Clinical BERT output dimension by one. This is because clinical notes are not regularly spaced, which cannot be accounted for by positional encoding. Thus, we add the entry time of notes as the last feature. The dimensionality of EHR data is $d_{EHR} = 42$ after one-hot encoding of categorical variables. Following the original paper [18], we pass the input modalities through a 1D convolutional layer to embed the text series and the time series into the same latent space. Then we add positional encoding to each modality. Since we make predictions at $X_{EHR}$ frequency, we always obtain queries from $X_{EHR}$ and keys and values from $X_{EHR}$ or $X_{CN}$ in the self-attention and cross-attention layers, respectively. Figure 1 shows the employed architecture.
**Tasks definition.** Previous literature jointly using EHR data and clinical notes \[12, 13\] relied on MIMIC-III data \[36\]. Like other existing datasets, MIMIC-III consists of physiological measurements and information about laboratory tests. However, contrary to others, in MIMIC-III, a significant number of patients have clinical notes written by various medical personnel allowing for a multi-modal approach. Thus, for our experiments, we follow previous research by analyzing clinical notes’ importance on Harutyunyan et al. \[15\]’s benchmark tasks. In particular, we focus our work on three tasks. First, *decompensation*, an hourly binary prediction task on whether the patient dies in the next 24 hours. Second, *in-hospital mortality* (IHM), a similar task with a single binary prediction task after 48 hours of observed ICU data on whether the patient dies in the ICU. Finally, *phenotyping*, a multi-class classification problem where one predicts which acute care conditions were present during a patient’s stay out of 25 care conditions. The benchmark contains another task, hourly predicting the remaining length of stay. We do not consider this task due to its lower clinical relevance and the poor performance of all existing deep learning models on the task. We further discuss the data and its processing in Appendix A.

**Hyperparameter tuning.** Increasing the number of heads or layers in transformers usually improves performance but also makes the models’ analysis more complex. Since we already observed significant improvement over state-of-the-art uni-modal models using a single layer and head, we fixed these two hyperparameters to 1 in all experiments. Doing so allowed us to preserve a more explainable model at the cost of a slight loss in performance. All other hyperparameters were selected using a grid search on the validation performance of *decompensation*. For the convolutional layer, we use 64 filters with a size equal to the input space. In the transformer, we use a latent space of size 64 everywhere and a dropout rate of 0.2. We trained all models with a batch size of 16 patient stays, a learning rate of $1 \times 10^{-5}$, and Adam optimizer \[37\].

**Data processing.** As it is common for MIMIC-III benchmark data \[15\], we used forward imputation and standard scaling as pre-processing steps. More importantly, we mask out the last note in all tasks to avoid label leaks through direct death mentions. This is a simple way to reduce information leaking. Though, it is possible to tackle this more in-depth by filtering samples using keyword search as \[11\]. As for the choice of hyperparameters, masking the last note might reduce absolute performance compared to previous research not following the same procedure \[12, 13\], but ensure a fairer analysis of the role of the clinical notes.

4 Experiments

In this section, we first show that even with hard constraints on model complexity, jointly using EHR data and clinical notes outperforms existing uni-modal models on the considered tasks. We then perform an ablation study on note types and analyze cross-modal attention weights highlighting the importance of either “nurse” or “radiology” notes for cross-modal model decision-making. All results reported are from the test set and obtained over 5 random seeds.

**Overall performance.** As shown in Table 1 while only employing a single cross-attention layer with a single head and not attending to the last note, using a joint model surpasses previous literature’s performance utilizing either EHR data \[15\] or clinical notes \[12\]. This observation confirms prior findings on the addition brought by clinical notes in patient monitoring \[12, 13\] and justifies the need to understand the underlying reason for such improvement. Interestingly, we also observe that our model performs on par with the best cross-modal model on both *decompensation* and *phenotyping* tasks despite our constraints.

**Note type ablation.** To understand which note types are important to the decision-making in our cross-modal model, we carry out ablation studies where we cumulatively add note types by increasing or by decreasing frequency (see Appendix B). Figure 2 shows the ablation studies for the *decompensation* and *IHM* tasks by increasing frequency. We draw two conclusions. On the one hand, adding doctors’ (“Physicians”) notes does not improve performance. This suggests that current representations obtained from the EHR data alone are sufficient. On the other hand, adding nurses’ notes and radiology reports for *IHM* significantly improves AUPRC on both tasks. Similar results can be found for *phenotyping* in Appendix B. These findings show that, when using clinical notes, deep learning models benefit from additional textual information rather than a possibly better representation of similar information in doctors’ notes summarizing the patient’s state.
Table 1: Performances on MIMIC-III Benchmark [15] for uni-modal (top, middle) and cross-modal models (bottom). Reported errors, if provided, are 95% confidence intervals on the mean.

<table>
<thead>
<tr>
<th>Task</th>
<th>Decompensation</th>
<th>IHM</th>
<th>Phenotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metric</td>
<td>AUPRC</td>
<td>AUROC</td>
</tr>
<tr>
<td>EHR-Only LSTM [15]</td>
<td>33.3 ± 1.0</td>
<td>90.6 ± 0.3</td>
<td>51.5 ± 5.5</td>
</tr>
<tr>
<td>EHR-Only Transformer (Ours)</td>
<td>31.7 ± 0.4</td>
<td>90.3 ± 0.2</td>
<td>49.8 ± 0.9</td>
</tr>
<tr>
<td>Text-only LSTM [12]</td>
<td>8.1 ± 0.7</td>
<td>79.3 ± 0.7</td>
<td>30.3 ± 1.3</td>
</tr>
<tr>
<td>Text-only Transformer (Ours)</td>
<td>19.2 ± 0.6</td>
<td>83.3 ± 0.3</td>
<td>34.1 ± 0.6</td>
</tr>
<tr>
<td>Cross-Modal LSTM [12]</td>
<td>34.5 ± 0.7</td>
<td>90.7 ± 0.7</td>
<td>52.5 ± 1.3</td>
</tr>
<tr>
<td>Cross-Modal Ensemble [13]</td>
<td>40.4 ± n.a.</td>
<td>92.0 ± n.a.</td>
<td>58.2 ± n.a.</td>
</tr>
<tr>
<td>Cross-Modal Transformer (Ours)</td>
<td>39.7 ± 0.6</td>
<td>92.2 ± 0.2</td>
<td>52.7 ± 1.0</td>
</tr>
</tbody>
</table>

Figure 2: Area under the precision-recall curve performance of the cross-modal transformer when adding different note types by increasing frequency. (left) Decompensation. (right) In-hospital mortality.

**Cross-modal attention weights analysis.** For all tasks, we observed that notes among the two most frequent types, nurses’ and radiology notes, lead to the greatest improvements. Thus, one could argue that the note frequency explains the increase in performance over the note type itself. To answer this question, we perform an analysis of cross-modal attention weights. As shown with an illustrative example in Figure 3, attention weights are very salient, making each EHR timestep attend to only a subset of notes. In addition, also shown in Figure 3, highly attended notes, such as the 7th note in the example, also impact the model’s final prediction. Indeed, when it obtains this note at hour 45, the cross-modal model’s predictions diverge from the EHR-only model ones. This high saliency of attention weights refutes the hypothesis that the nurses’ and radiology notes are the most relevant types due to their high frequency. Indeed, in practice, only a subset of these notes are attended to independently of the notes type frequency. We provide additional examples to support this claim in Appendix C.

Figure 3: Illustrative example of model’s behaviour for decompensation. (left) Predictions for EHR-only (blue) and cross-modal (orange) models over time compared to ground truth labels (green). (right) Cross-attention over time for the cross-modal model where EHR timesteps (x-axis) attends to clinical notes (y-axis). Both figures represent patient 11555.
5 Conclusion

In this paper, we aim to understand the underlying reason behind the significant improvement brought by jointly using EHR data and clinical notes in a multi-modal fashion in ICU-related tasks. To answer this question, we analyze the performance impact of the available note types in MIMIC-III and the cross-modal attention weights in our model. Based on our experimental results, we conclude that improvements yielded by adding clinical notes come from notes containing additional context on patient states, such as nurses’ notes and radiology reports. We also find that doctors’ notes, which often summarize EHR content, do not improve performance, suggesting that representations learned from EHR data alone preserve the input information. We believe these findings show how EHR data alone is only a narrow view of patient states in the ICU, motivating more data-centric approaches in the field.
References


A Dataset

A.1 Pre-processing

We first run the pre-processing step provided by Harutyunyan et al. [15] for EHR data. Next, we extract all patients with notes. Echo, electrocardiogram (ECG), and discharge summary notes have no CHARTTIME, the exact time at which the note is entered in the system. However, these notes might contain valuable medical information. Thus, we set their CHARTTIME to the end of the day of their CHARTDATE, which contains only the day on which the note was entered. It ensures no temporal leak of information. Following previous work [12], we then exclude patients without any notes. From the benchmark task with 6281 ICU stays in the test set, we exclude 27 patients and end up with 6254 ICU stays. For some of these stays, only phenotyping task labels exist, hence, for the decompensation task, we end up with 6215 labeled ICU stays in the test set. We follow the same approach for the training set, validation set, and test set.

A.2 Statistics

For further details, we provide in Table 2 the test set statistics for the decompensation and IHM tasks. In addition, Table 3 shows the note type counts in the training and test set. We ran the ablation studies according to the note type frequency in the training set.

Table 2: Number of negative and positive labels for the IHM task and the decompensation task test sets.

<table>
<thead>
<tr>
<th>(a) Decompensation task test set</th>
<th>(b) IHM task test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>Number of labels</td>
</tr>
<tr>
<td>0</td>
<td>512862</td>
</tr>
<tr>
<td>1</td>
<td>9427</td>
</tr>
<tr>
<td>Total</td>
<td>522289</td>
</tr>
</tbody>
</table>

Table 3: Number of notes per note type in the test and training cohort.

<table>
<thead>
<tr>
<th>(a) Note type distribution in test set</th>
<th>(b) Note type distribution in training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note type</td>
<td>Number of notes</td>
</tr>
<tr>
<td>Nursing</td>
<td>66206</td>
</tr>
<tr>
<td>Radiology</td>
<td>39744</td>
</tr>
<tr>
<td>ECG</td>
<td>16031</td>
</tr>
<tr>
<td>Physician</td>
<td>14261</td>
</tr>
<tr>
<td>Discharge summary</td>
<td>6809</td>
</tr>
<tr>
<td>Echo</td>
<td>3762</td>
</tr>
<tr>
<td>Respiratory</td>
<td>3036</td>
</tr>
<tr>
<td>Nutrition</td>
<td>885</td>
</tr>
<tr>
<td>General</td>
<td>812</td>
</tr>
<tr>
<td>Rehab Services</td>
<td>562</td>
</tr>
<tr>
<td>Social Work</td>
<td>263</td>
</tr>
<tr>
<td>Case Management</td>
<td>102</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>13</td>
</tr>
<tr>
<td>Consult</td>
<td>12</td>
</tr>
</tbody>
</table>

B Ablation Studies

In this section, we provide further results regarding notes type ablations.
Table 4: Note type ablation study results when adding clinical notes by increasing frequency.

<table>
<thead>
<tr>
<th>Task</th>
<th>Decomposition</th>
<th>IHM</th>
<th>Phenotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUPRC</td>
<td>AUROC</td>
<td>AUPRC</td>
</tr>
<tr>
<td>ICU + Other</td>
<td>29.7 ± 1.1</td>
<td>90.0 ± 0.1</td>
<td>49.2 ± 0.5</td>
</tr>
<tr>
<td>+ Respiratory</td>
<td>30.9 ± 0.6</td>
<td>90.1 ± 0.1</td>
<td>48.6 ± 0.5</td>
</tr>
<tr>
<td>+ Echo</td>
<td>32.0 ± 0.9</td>
<td>90.2 ± 0.1</td>
<td>49.3 ± 0.3</td>
</tr>
<tr>
<td>+ Discharge</td>
<td>31.5 ± 0.9</td>
<td>90.1 ± 0.1</td>
<td>49.5 ± 0.4</td>
</tr>
<tr>
<td>+ ECG</td>
<td>31.4 ± 0.7</td>
<td>90.1 ± 0.1</td>
<td>49.8 ± 0.4</td>
</tr>
<tr>
<td>+ Physician</td>
<td>31.4 ± 0.6</td>
<td>90.0 ± 0.2</td>
<td>49.5 ± 0.6</td>
</tr>
<tr>
<td>+ Radiology</td>
<td>32.5 ± 0.7</td>
<td>90.6 ± 0.2</td>
<td>51.4 ± 0.4</td>
</tr>
<tr>
<td>+ Nurse</td>
<td>39.9 ± 0.7</td>
<td>92.3 ± 0.2</td>
<td>52.7 ± 1.0</td>
</tr>
</tbody>
</table>

B.1 Increasing Frequency

We first summarize results from the ablation by increasing frequency from Figure 2 in Table 4. We find, as previously mentioned, that for all tasks, nurses’ or radiology notes contribute to improved performance. For the phenotyping task, we see that also other note types help. It may be because these types contain information for specific illnesses not necessarily present in the other notes, thus also from some additional context on patients’ states. Nonetheless, the main contribution to performance remains from radiology reports.

Figure 4: Micro-averaged AUROC note type ablation study results for the phenotyping task. (left) By increasing note type frequency. (right) by decreasing note type frequency.

B.2 Decreasing Frequency

We also provide an ablation study on note types by decreasing frequency, as a sanity check, in Figure 5 and Table 5. We draw the same conclusions on note types importance.

Figure 5: Note type ablation study results by adding types by decreasing frequency. (left) Decompensation. (right) In-hospital mortality.
Table 5: Note type ablation study results when adding clinical notes by decreasing frequency.

<table>
<thead>
<tr>
<th>Model</th>
<th>Decompensation AUPRC</th>
<th>Decompensation AUROC</th>
<th>IHM AUPRC</th>
<th>IHM AUROC</th>
<th>Phenotyping AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU + Nurse</td>
<td>39.2 ± 0.4</td>
<td>92.3 ± 0.1</td>
<td>52.5 ± 0.7</td>
<td>86.6 ± 0.2</td>
<td>79.2 ± 0.2</td>
</tr>
<tr>
<td>+ Radiology</td>
<td>39.5 ± 0.7</td>
<td>92.4 ± 0.1</td>
<td>52.6 ± 1.6</td>
<td>86.8 ± 0.9</td>
<td>81.4 ± 0.2</td>
</tr>
<tr>
<td>+ Physician</td>
<td>40.1 ± 0.7</td>
<td>92.4 ± 0.2</td>
<td>52.3 ± 1.5</td>
<td>86.8 ± 0.8</td>
<td>81.5 ± 0.2</td>
</tr>
<tr>
<td>+ ECG</td>
<td>39.4 ± 0.8</td>
<td>92.3 ± 0.2</td>
<td>52.7 ± 1.6</td>
<td>87.1 ± 0.8</td>
<td>82.2 ± 0.1</td>
</tr>
<tr>
<td>+ Discharge</td>
<td>39.6 ± 0.9</td>
<td>92.3 ± 0.3</td>
<td>52.7 ± 1.5</td>
<td>87.1 ± 0.8</td>
<td>82.3 ± 0.1</td>
</tr>
<tr>
<td>+ Echo</td>
<td>39.5 ± 0.8</td>
<td>92.2 ± 0.2</td>
<td>52.8 ± 1.3</td>
<td>87.1 ± 0.7</td>
<td>82.7 ± 0.1</td>
</tr>
<tr>
<td>+ Respiratory</td>
<td>39.5 ± 0.8</td>
<td>92.2 ± 0.2</td>
<td>52.7 ± 1.1</td>
<td>87.1 ± 0.7</td>
<td>82.6 ± 0.2</td>
</tr>
<tr>
<td>+ Other</td>
<td>39.7 ± 0.6</td>
<td>92.2 ± 0.2</td>
<td>52.7 ± 1.0</td>
<td>87.1 ± 0.6</td>
<td>82.6 ± 0.1</td>
</tr>
</tbody>
</table>

C Interpretability results

In this section, we describe our attempt to analyze notes importance more in-depth through attention rollout and chunk level representations. We also provide a similar analysis for another illustrative example.

C.1 Deeper analysis on main manuscript example

In the example from the main manuscript, we found that the 7th note was strongly attended to in Figure 3. To further understand, what lead this note to significantly change the model prediction toward a imminent death, we ran attention rollout [31] and summed over the words [27] (see Figure 6). Additionally, we normalized by the number of words in each chunk when a note does not fit a single Clinical BERT pass (i.e., it is longer than 128 tokens). One interesting observation is how dnr is significant in the overall note representation. Indeed this stands for “do no resucitate”, indicating a clear change of treatment in case of cardiac or respiratory arrest and thus a more probable death.

Figure 6: Attention rollout of Clinical BERT on 7th note from patient 11555

Chunk level representation of notes. In our model, if a note is longer than 128 tokens on which Clinical BERT was trained, we passed the chunks separately through Clinical BERT and averaged the global token embedding. For an additional level of interpretability, we use the same model architecture and train using all embedded chunks of the same note with the same time appended. For the identical patient, we see that chunk 12 is attended early on, and the 25th and 27th chunks are the strongest attended to when the prediction gap increases. Figure 8 shows the attention rollout results on the chunks separately. Once again in the most attended chunk (see Figure 8c), we see the nurses mentions to “please allow natural death”.

C.2 Other Illustrative Example

To further illustrate our conclusions, we also present a second example corresponding to the patient with ID 12767. Figure 9 shows the predictions and the corresponding cross-attention from ICU
time series to text series over time. We observe that at the 63rd timestep when the multi-modal
transformer’s prediction deviates from the EHR-only transformer’s prediction it attends to the newly
available 8th note.

As in the previous example, we compute the attention rollout from this particular note in Figure 9.
This short note states that the patient is most likely to be made DNR/DNI. We see that besides the
DNR status, attention is seemingly strongly placed also on unresponsive and tylenol. Interestingly,
we observe that the last note is highly attended to. Indeed, we see that the note already discusses the
patient’s death, highlighting the existing leakage of information. Hence, it is clear why the model is
above 90% certain that the patient will die within the next 24 hours. We show this last note attention
rollout in Figure 11. Interestingly, not necessarily expired is focused in contrast to mortem. We
believe this is because the former is a clinical term, thus, more frequent in the corpus used to train
clinical BERT.
As for the previous examples, we also explore the chunks of the notes directly. Figure 12 shows the ICU to nurse note chunks cross-attention over time. For this patient the 8th, 17th, 23rd, 25th, and 33rd chunks are strongly attended to over time. In Figure 13 we present the corresponding attention rollout results. In chunk 7, we find medically relevant observations like obtundation or decerebrate like posturing attended to. Obtundation describes the status when the patient has a reduced level of consciousness and alertness. Decerebrate posture is a body pose indicating severe brain damage. In the other chunks, the DNR status is again highlighted. Though, the cross-modal network actually does not have a higher likelihood of death for this patient at the time of this note. Hence, it might not have been able to learn about such medical knowledge yet. Interestingly the last note chunk, which strongly indicates that death is near, shows attention on the namepatterns and not the patient being placed on comfort measures and that they were extubated (cf. Figure 13e). Since we did not train end-to-end due to resource constraints, it is possible that the model was unable to capture the expected humanly relevant rationales correctly and represented this observation differently instead.
nursing / other pt expired mr [ * * known lastname 5307 * * ] had been made cmo yesterday. comfort maintained today using morphine. pt repositioned vs monitored and family updated q2-3 hours on pt condition; their desire was to be updated if there was a change. rn shaving patient at 1615 and his hr dropped from 115 to the 60’s. and rr decr to 9 from rate of 30. family called immediately. ** first name11 ( namepattern1 ) 2294 * * | ** initials ( namepattern4 ) | ** last name ( namepattern4 ) 2295 * * | ** name ( ni ) 2489 * * | ** name ( ni ) 5307 * *. rn spoke with ** ( name ( ni ) 2489 * * | ** name ( ni ) 5307 * *. pt son and told him the family should come in. bp dropped to 72/20. hr and rr continued to decr. rn called son again and told him that they would not make in in time, that pt was dying and rn would stay with him in the room. pt expired at 161. ** first name ( titles ) * * | ** last name ( titles ) 64 * *. family to arrive within an hour. neurosurgeon resident called. post-mortem care to be done and pt will be transported to morgue after family has arrived.

Figure 11: Attention rollout of Clinical BERT of last note from the patient with ID 12767.

Figure 12: Chunk cross-attention over time for the cross-modal model where EHR timesteps (x-axis) attends to clinical notes (y-axis) for the patient with ID 12767.
nursing / other night note : mcu | neuro : unchanged through night
responds to voice intermittently , able to hold one thumb up on
command w / r hand ; wiggle toes sl | on r | very weak | no movement
on left side | neuro and neuro sicu in to eval pt | pupils [* * * 1 - 21 * ]
and sluggish | periods of deep obtundation to more wake | almost
decerebrate | like posturing on r | nd s in while this was occurring | plan
for f / u el scan today

(a) Attention rollout for 8th chunk

doctors | son in particular seemed upset and eager for info | neuro [* *
doctor first name * * ] team member spoke with family | no real
change in neuro exam | stable post parietal bleed | p - will wean nspnews
as we can once expressed kicks in | goal bp 150 | for now will leave
intubated until neuro exam brightens | check culture results | watch | s and o s | labs | can start tube feeds in am if still intubated | monitor
neuro exam for changes | keep family advised as to all

di : foley patent draining mod amt ' s of clear yellow urine | gu : ngt
clamped | npo | no stool | nd soft distended | hypoactive bowel sounds |
neuro : no chg in neuro status | cont to be unresponsive | does not open
eyes | moving right arm and right leg | no movement on left side | pupils
2mm react briskly | iv fluid cont at 40 / hr | pt waiting till morning to
most likely be made dnr / di

(b) Attention rollout for 17th chunk

(c) Attention rollout for 23rd chunk

defined on left | wife , son , and daughter want to uphold wishes of pt
which is not to have non quality life | as his mother did for 6 yrs in a
hosp home | s / p stroke | ferr order obtained | family waiting for last
daughter to arrive from | * * name ( ni ) 794 * * | late tonight | family
wishes to give pt a few days to see if any improvements occur | di |
* * last name ( stitle ) 530

(d) Attention rollout for 25th chunk

(e) Attention rollout for 33rd chunk

Figure 13: Attention rollout over Clinical BERT from nurse’s note chunks for the patient with ID
12767