Supplementary: Reinforced Sequential Decision-Making for Sepsis Treatment: The POSNEGDM Framework with Mortality Classifier and Transformer

Anonymous Author(s)
Affiliation
Address
email

1 Sepsis Data Description

The MIMIC-III database Johnson et al. [2016] is a rich source of de-identified clinical data from Beth Israel Deaconess Medical Center in Boston, available to global researchers under a data use agreement. Spanning over a decade, it covers 53,423 adult and 7,870 neonate admissions, involving 38,597 adult patients with a median age of 65.8 years and an 11.5% in-hospital mortality rate. The database includes detailed patient demographics, vital signs, lab results, treatments, outcomes, and provider notes.

ICU stays last a median of 2.1 days, hospital stays average 6.9 days, with an average of 4,579 observations and 380 lab measurements per admission. Widely used in critical care research, MIMIC-III provides a valuable dataset for studying sepsis and critical illnesses.

Sepsis MIMIC-III is a subset of the MIMIC-III database focusing on ICU patients diagnosed with sepsis. It's valuable for sepsis detection, treatment modeling, and research. The dataset contains 19,614 sepsis-specific treatment trajectories, with 30% for testing and 70% for training.

In this dataset, positive trajectories indicate patient survival and negative ones represent fatalities. Both the training and testing sets have an approximate 9.5% mortality rate.

2 Proposed Network

In this section, we first delineate the algorithm's components, which include the mortality classifier and the transformer-based dualsight network. Subsequently, we present the complete POSNEGDM framework, with a particular focus on the novel training objectives.

2.1 Mortality Classifier

We designed a mortality classifier to predict patient survival likelihood. It takes the current patient state as input, utilizing a five-layer fully-connected architecture (size 64). Training details are in Table 1. Dealing with a highly imbalanced dataset, we used Borderline SMOTE Han et al. [2005] to upsample the minority class, mitigating the imbalance issue.

Table 1: Hyperparameters for training the mortality classifier

Parameter	Learning Rate	Weight Decay	Optimizer	Dropout
Value	1e-3	1e-5	Adam	0.2

Submitted to the Workshop on Deep Generative Models for Health at NeurIPS 2023. Do not distribute.

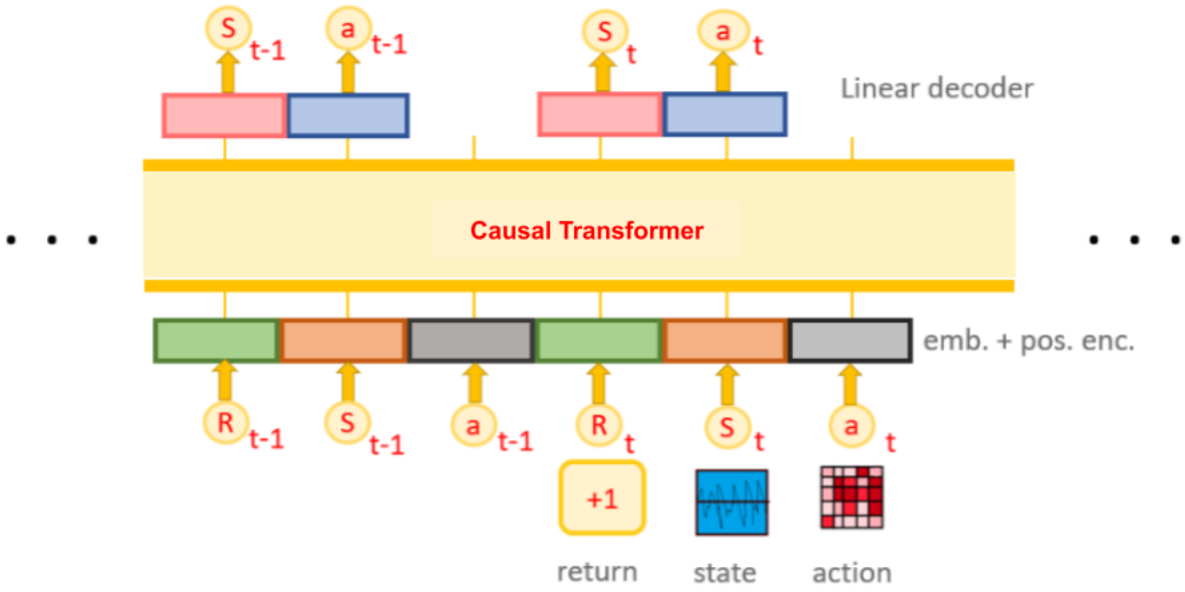


Figure 1: The DUALSIGHT decision maker takes in states, actions, and returns as input, which are first embedded into linear representations that are specific to each modality. The positional episodic timestep encoding is added to the input to help the model understand the order of events. The tokens are then fed into the GPT architecture, which uses a self-attention mechanism to predict actions and next states. The causal mask ensures that the model can only attend to previous tokens, preserving the causality of the system.

$$L_{survival}(MC, \text{DUALSIGHT}, \tau) = \mathbb{E}_{\tau \sim P_{data}(\cdot)}[\log(1 - MC(\text{DUALSIGHT}(\tau)))] \quad (1)$$

The classifier aims for high prediction accuracy in treatment trajectory outcomes. It has empirically shown an impressive 96.7% accuracy on the test dataset. It serves to assess treatment decision quality from different algorithms and offers feedback for training the transformer-based decision maker, as illustrated in the loss function (Equation (1)).

2.2 DUALSIGHT Decision Maker

$$L_{total} = \alpha L_{action} + \beta L_{state} + \gamma L_{survival} \quad (2)$$

The DUALSIGHT decision maker is trained on MIMIC-III dataset (Section 1). Actions in this RL setup represent medical interventions within a 4-hour window, categorized into a 5×5 discrete space, with 0 for no drug use and four quartiles for varying dosages. The reward function assigns +1 for positive trajectories (patient survival) and -1 for negative trajectories (patient demise) at the final time step. The agent's goal is to maximize cumulative rewards during treatment.

We adopt the Decision Transformer (DT) network structure Chen et al. [2021] for our decision maker. The Transformer model, known for its effectiveness in sequence tasks as seen in models like GPT Brown et al. [2020] and BERT Kenton and Toutanova [2019], is the foundation of DT. DT, with a GPT-like architecture, has outperformed state-of-the-art offline RL benchmarks.

In Figure 1, our decision maker takes a sequence of past states, actions, and returns as input. Returns aggregate future rewards from time step t until the episode ends, representing expected outcomes. This input sequence undergoes processing with modality-specific linear embeddings and positional episodic timestep encoding to maintain event order. Then, tokens enter a GPT architecture, predicting output in an autoregressive manner using causal self-attention. This self-attention module calculates weighted sums of input states based on similarity, effectively capturing dependencies and relationships among states.

A crucial difference between our architecture and DT is our decision maker's ability to anticipate not only the immediate action but also the subsequent state, providing dual insights. Predicted next states help the mortality classifier assess patient survival likelihood, enhancing decision maker performance through feedback. This unique configuration significantly improves our decision maker's effectiveness.

3 Ablation studies and Additional experiments

3.1 Ablation studies

Table 2: Importance of the action prediction loss (L_{action}). Here, $\beta = 0.1$ and $\gamma = 1$.

Loss importance (α)	Action Prediction Accuracy %	Mortality %	
		Step by Step	Complete Traj.
0.0	4.6	6.12	0.36
0.1	91.6	5.61	0.65
0.3	94	6.07	0.63
0.5	93.6	3.33	0.18
0.8	92.8	3.06	0.40
1.0	**94.6**	**2.61**	**0.18**

Table 3: Importance of the survival loss ($L_{survival}$). Here, $\alpha = 1$, $\beta = 0.1$.

Loss importance (γ)	Action Prediction Accuracy %	Mortality %	
		Step by Step	Complete Traj.
0.0	**94.7**	60.41	10.29
0.1	94.3	5.21	0.49
0.3	93.9	3.95	0.45
0.5	93.4	**2.32**	0.36
0.8	93.9	2.41	**0.18**
1.0	94.6	2.61	**0.18**

In this section, we perform ablation studies to assess the impact of the three objective terms in our overall loss function (Equation (2)), as shown in Tables 2 to 3.

We evaluate action prediction accuracy on positive data and mortality. Action prediction accuracy measures our model's ability to mimic expert actions, while mortality reflects how well the model guides patient states towards survival. Given our limited access to model internals, we calculate mortality using two methods: Step-by-step and Complete-trajectory.

In the Step-by-step method, we input states from a test trajectory and check if the next predicted state is alive for each state in the trajectory. The Complete-trajectory approach begins with the initial test state and lets the model generate the next ten actions and states. If any state results in a death prediction, we consider the patient deceased in that trajectory. These mortality figures encompass both positive and negative test trajectories (total mortality).

Table 2 demonstrates the impact of action prediction loss on POSNEGDM system performance. Removing L_{action} ($\alpha = 0$) significantly reduces action prediction accuracy on expert data, while increasing α improves overall performance, as evident in total mortality.

Table 4 illustrates the influence of subsequent state prediction loss on the POSNEGDM system's performance. Without L_{state} ($\beta = 0$), subsequent state prediction and mortality estimation from the classifier become unattainable. Moreover, total mortality rises with increasing β.

Table 3 assesses the influence of survival loss on POSNEGDM system performance. Excluding $L_{survival}$ leads to a significant increase in mortality rates, emphasizing its importance. Additionally, we observe that $\gamma = 0.8$ performs comparably to $\gamma = 1.0$ and, in fact, yields better mortality outcomes than the parameter configuration in Performance Comparison Table in the main paper.

In summary, these results emphasize the sensitivity of mortality rates to different hyperparameter settings. It underscores the necessity of careful hyperparameter tuning to optimize performance

Table 4: Importance of subsequent state prediction loss (L_{state}). Here, $\alpha = 1$ and $\gamma = 1$.

Loss importance (β)	Action Prediction Accuracy %	Mortality %	
		Step by Step	Complete Traj.
0.0	94.7	N\A	N\A
0.1	94.6	**2.61**	**0.18**
0.3	92.9	3.57	0.36
0.5	93.5	6.33	0.40
0.8	92.9	7.54	0.20
1.0	92.8	10.02	0.67

with regard to mortality rates. We have chosen $\alpha = 1, \beta = 0.1, \gamma = 1$ (as shown in Experiments Section in the main paper), which produces superior results (Performance Comparison Table in the main paper). While an exhaustive grid search might uncover better hyperparameters, it would be time-intensive. Importantly, it is essential to note that higher action prediction accuracy does not necessarily translate to lower mortality rates. Our work introduces mortality as a critical factor in decision-making, which had not been explicitly addressed previously, to the best of our knowledge.

3.2 Additional experiments

Figure 2 displays 2D histograms illustrating the aggregated actions recommended by the physician (Ground Truth), POSNEGDM, and Behavioral Cloning (BC). These histograms facilitate qualitative analysis akin to the approach in Raghu et al. [2017]. The bins along the horizontal and vertical axes represent Vasopressor and IV fluid dosages prescribed by the policy, respectively. An action value of 0 denotes no drug administration, with increasing values indicating higher dosages, categorized into quartiles. Each grid cell represents a specific action, with color denoting its frequency. Notably, POSNEGDM closely mirrors the ground truth, particularly in positive cases, outperforming BC. Decision Transformer results are excluded, as it deviates from the ground truth even more than Behavioral Cloning.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

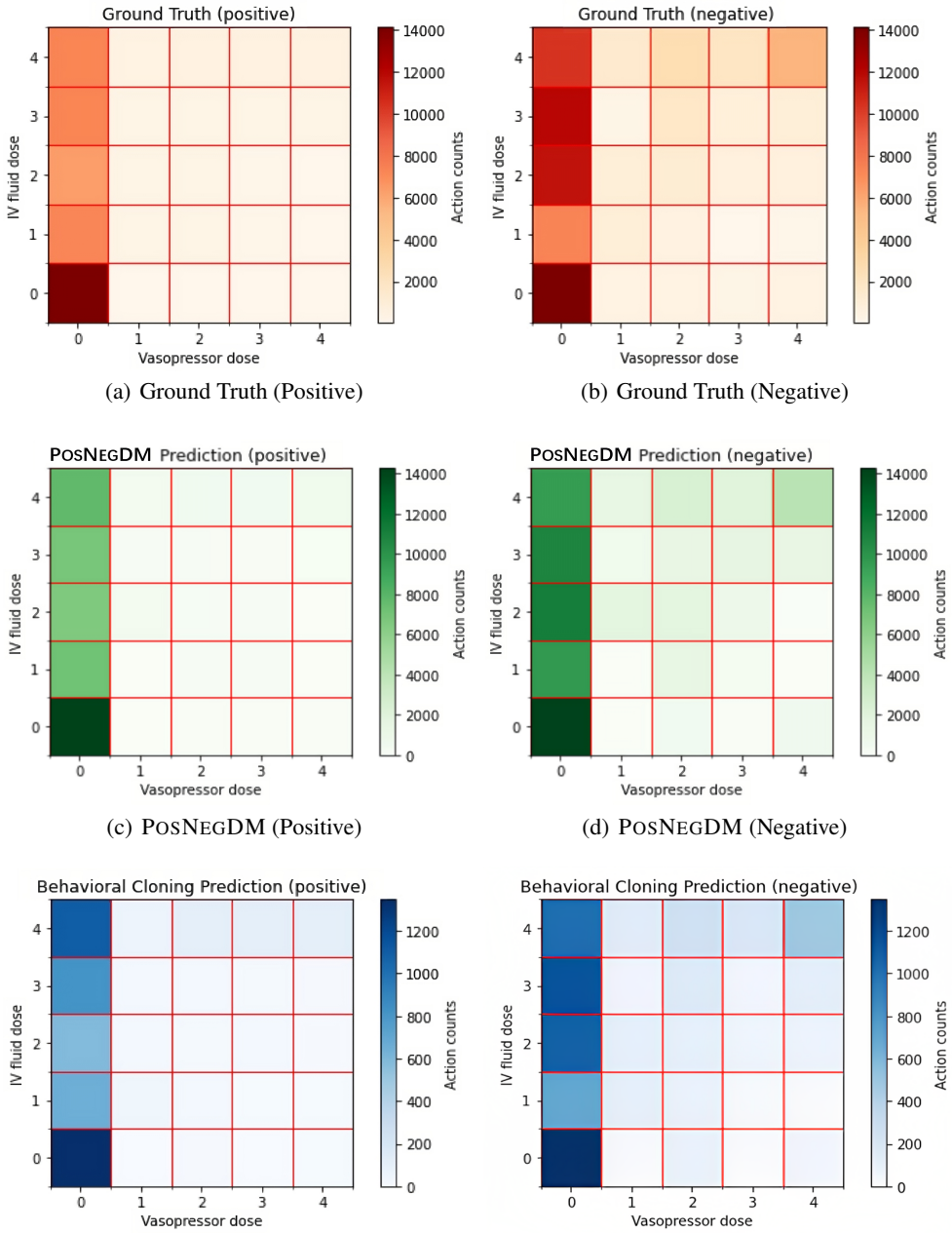
Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

Alistair EW 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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.



(a) Ground Truth (Positive)

(b) Ground Truth (Negative)

(c) PosNegDM (Positive)

(d) PosNegDM (Negative)

(e) BC (Positive)

(f) BC (Negative)

Figure 2: The three rows in the visualization represent the policies as provided by physicians, PosNegDM, and Behavioral Cloning (BC) respectively, each applied to both positive and negative test data. The axis labels correspond to the discretized action space, where '0' signifies no drug administration, and '4' indicates the maximum dosage of a particular drug. Each grid cell represents a specific action, with its color indicating the frequency of its occurrence.