Empowering Clinicians with MeDT: A Framework for Sepsis Treatment

Aamer Abdul Rahman[1,3] **Pranav Agarwal**[1,3] **Vincent Michalski**[1,2] **Rita Noumeir**[3]
Philippe Jouvet[2,4] **Samira Ebrahimi Kahou**[1,3,5]
[1]Mila [2]Université de Montréal [3]École de Technologie Supérieure
[4]CHU Sainte-Justine Hospital [5]CIFAR
{aamer.abdul-rahman.1,pranav.agarwal.1,samira.ebrahimi-kahou}@ens.etsmtl.net
vimichals@gmail.com
philippe.jouvet.med@ssss.gouv.qc.ca

Abstract

Offline reinforcement learning has shown promise for solving tasks in safety-critical settings, such as clinical decision support. Its application, however, has been limited by the need for interpretability and interactivity for clinicians. To address these challenges, we propose *medical decision transformer (MeDT)*, a novel and versatile framework based on the goal-conditioned reinforcement learning (RL) paradigm for sepsis treatment recommendation. *MeDT* is based on the decision transformer architecture, and conditions the model on expected treatment outcomes, hindsight patient acuity scores, past dosages and the patient's current and past medical state at every timestep. This allows it to consider the complete context of a patient's medical history, enabling more informed decision-making. By conditioning the policy's generation of actions on user-specified goals at every timestep, *MeDT* enables clinician interactability while avoiding the problem of sparse rewards. Using data from the MIMIC-III dataset, we show that *MeDT* produces interventions that outperform or are competitive with existing methods while enabling a more interpretable, personalized and clinician-directed approach. For future research, we release our code at `https://aamer98.github.io/medical_decision_transformer/`.

1 Introduction

Healthcare tasks can be seen as a form of sequential decision-making process, where clinicians aim to optimize a patient's health by selecting appropriate medical interventions, considering the patient's historical health data and prior treatments. Clinical decision support systems [24] can help healthcare professionals in making more informed decisions. Particularly in intensive care units (ICUs), where clinicians face challenges in choosing optimal medication dosages due to the complex nature and rapid progression of diseases. This is where RL comes in as a promising solution for developing policies that recommend optimal treatment strategies [19, 10, 9, 21, 6]. Given the risks associated with direct interaction with the environment in safety-critical applications [5], we use offline RL. However, current approaches are limited by issues such as credit assignment from sparse reward functions [14, 11], limited context lengths [18] and a lack of interpretable models [20].

Recent research in RL [17, 16, 7, 25] is shifting towards attention-based networks [15] like Transformers [26, 1] which can effectively model long contexts and can be trained in a parallelizable manner [1], countering most of the challenges of recurrent neural networks (RNNs) [27]. Chen et al. [3] proposed the decision transformer (DT), a transformer based RL policy learning model that has proven effective for offline RL [4, 28, 13]. DT addresses the problem of sparse or distracting rewards

Goal-Conditioned Reinforcement Learning Workshop in the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

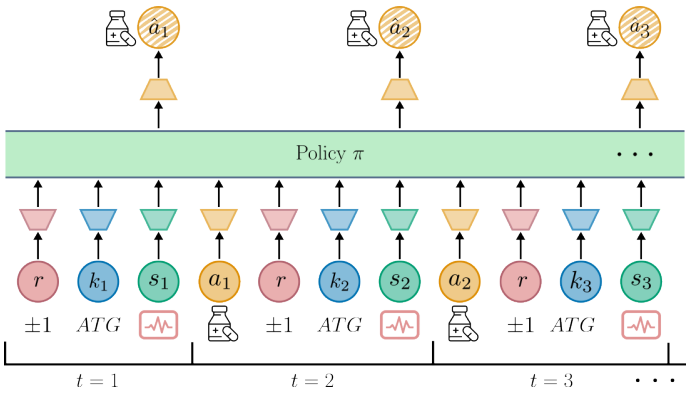

Figure 1: Architecture of the proposed MeDT framework.

by leveraging self-attention for credit assignment [3], which incorporates contextual information into the learning process. Furthermore, the transformer's ability to model long sequences enables it to consider the patient's history of states and medications to make predictions [26].

Building on this idea, we propose an offline RL framework where treatment dosage recommendation is framed as a sequence modeling problem. The proposed framework called the *medical decision transformer (MeDT)*, shown in Fig. 1, is based on the DT architecture and recommends optimal treatment dosages by autoregressively modeling a patient's state while conditioning on hindsight return information. To provide the policy with more informative and goal-directed input, we also condition *MeDT* on hindsight patient acuity scores [12] at every time step. This enhances interpretability of the conditioning, facilitating interaction of clinicians with the model.

Our contributions are threefold. 1) We propose *MeDT*, a transformer-based policy network that autoregressively models the full context of a patient's clinical history and recommends optimal medication dosages. 2) To alleviate the burden of sparse rewards, we condition *MeDT* on hindsight patient acuity scores as a goal at every timestep. We modify these scores to make them more interactable for clinicians. 3) In addition to fitted Q-evaluation (FQE) to evaluate learnt policies, we leverage a transformer network, the *state predictor*, to serve as an approximate model to capture the evolution of a patient's clinical state in response to treatment. This model enables autoregressive inference of *MeDT* and also serves as an evaluation framework of models used for clinical dosage recommendation.

2 Methodology

Our proposed *MeDT* architecture is inspired by the Upside-Down RL approach [23] for learning policies, which maps rewards to corresponding actions. Specifically, our work builds on the idea of modeling RL as a sequence modeling problem, as first proposed in [3], and further developed in related works [13, 29, 7].

We frame our problem as a Markov decision process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{S}')$, where $\mathcal{S}$ represents the patient states, $\mathcal{A}$ is the set of possible dosage recommendations, $\mathcal{P}$ is the state transition function, $\mathcal{R}$ is the reward function and $\mathcal{S}'$ is the next patient state. While this framework is well-suited for addressing RL problems, direct interaction with the environment can be risky in safety-critical applications like ours. To mitigate this risk, we use offline RL, a subcategory of RL that learns an optimal policy using a fixed dataset.

MeDT is trained to model trajectories such that the transformer is conditioned on future desired returns to generate treatment dosage recommendations (Fig. 1). Specifically, during training, we use returns-to-go (RTG) $r_t = \sum_{t'=t}^{T} R_{t'}$, which represents the observed treatment outcome (death or survival), to condition the model, while it is fixed to $+1$ for survival during evaluation. In addition, we propose to condition *MeDT* on future patient acuity scores (SAPS2), or acuity-to-go (ATG), where the acuity score provides an indication of the severity of illness of the patient in the ICU, based on the status of the patient's physiological systems. This formulation allows clinicians to input desired acuity scores for the next state upon monitoring the current physiological state of the

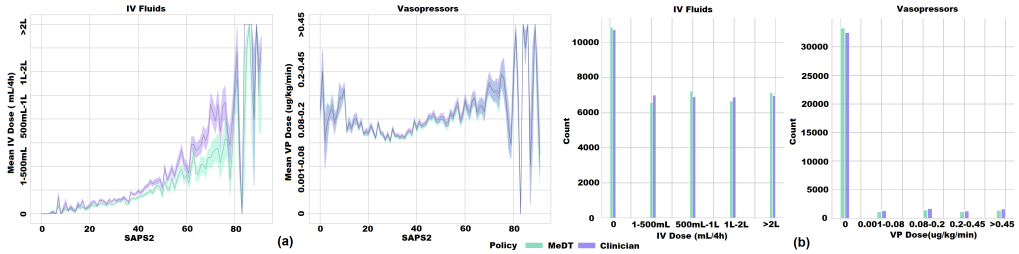

Figure 2: (a) Dosage recommended by MeDT and clinician policy for different SAPS2 scores. (b) Distribution of IV fluids and VPs given by the MeDT and clinician policies.

patient, providing additional context for the policy to generate optimal actions. This leads to more information-dense conditioning, allowing clinicians to interact with the model and guide the policy's generation of treatment dosages.

To enable clinicians to provide more detailed inputs, we break down the SAPS2 score into constituent scores that correspond to specific organ systems [12]. Following the definitions provided in [22], we define split scores $k = (kc, kr, kn, kl, kh, km, ko)$ to represent the status of the cardiovascular, respiratory, neurological, renal, hepatic, haematologic and other systems, respectively. This enhances the usability of the model for clinicians, enabling efficient interaction with the model for future dosage recommendations, considering the current state of the patient's organs. Using these scores, the treatment progress over T time steps forms a trajectory

$$\tau = ((r_1, k_1, s_1, a_1), (r_2, k_2, s_2, a_2), \ldots, (r_T, k_T, s_T, a_T)) \,. \tag{1}$$

We train our policy, a causal transformer network, to predict ground-truth dosages that were administered by the clinician given the patient's past and current state via teacher forcing, while ignoring future information via masking (Fig. 1). MeDT aims to learn an optimal policy distribution $P_\pi(a_t | s_{\leq t}, r_{\leq t}, k_{\leq t}, a_{<t})$, following the model architecture and hyperparameters used in [3]. We use an encoder with a linear layer and a normalization layer for each type of input (i.e. RTG, ATG, state, action) to project raw inputs into token embeddings. To capture temporal dynamics of the patient's state, we use learned position embeddings for each timestep which are added to the token embeddings. Finally, the resulting embeddings are fed into a causal transformer, which autoregressively predicts the next action tokens.

Evaluation

In this work, we train and evaluate the performance of *MeDT* on a cohort of septic patients. The cohort data is obtained from the medical information mart for intensive care (MIMIC-III) dataset [8], which includes 19,633 patients, with a mortality rate of 9%. To preprocess the data, we follow the pipeline defined by Killian et al. [9]. We extract physiological measurements of patients recorded over 4-hour intervals and impute missing values using K-nearest neighbour. Multiple observations within each 4-hour window are averaged. The patient state consists of 5 static demographic and 38 time-varying continuous variables such as lab measurements. We focus on the administration of two drugs: vasopressorss (VPs) and intravenous (IV) fluids. The administration of each drug for patients is sampled at 4-hour intervals. We discretized the dosages for each drug into 5 bins, resulting in a combinatorial action space of 25 possible treatment administrations.

In online RL, policies are assessed by having them interact with the environment. However, healthcare involves patients, where employing this evaluation method is unsafe. As a stand-in for the simulator during inference, we propose to additionally learn an approximate model (state predictor) of $P_\theta(s_t | a_{<t}, s_{<t})$ with a similar architecture as the policy model. During inference, this model allows autoregressive generation of a sequence of actions by predicting how the patient state evolves as a result of those actions (Alg. 1). Figure 4 visualizes this rollout procedure.

Additionally, we utilize FQE to produce the estimated Q-value for a given policy (Fig. 5). FQE takes as input a policy π and a set of transitions $\{s_t, a_t, s_{t+1}, r_{t+1}\}_{t=1}^n$. At each step k, the algorithm computes the target $y_t = r_t + \gamma Q_{k-1}(s_{t+1}, \pi(s_{t+1}))$, solving the equation $Q_k = \operatorname{argmin}_{f \in F} \sum_{i=1}^n (f(s_t, a_t) - y_t)^2$. This yields a neural network labeled as Q_π, which

Table 1: Estimated final patient acuity scores (averaged over 2898 patients) for BCQ: batch constrained Q-learning, BC: behaviour cloning, DT: decision transformer and MeDT: medical decision transformer.

Models	Overall ↓	Low ↓	Mid ↓	High ↓
BCQ	42.10±0.03	41.78±0.07	42.38±0.03	41.59±0.15
BC	40.50±0.03	40.33±0.07	40.56±0.03	40.29±0.12
DT	40.38±0.03	40.16±0.06	40.49±0.03	**40.06±0.12**
MeDT	**40.31±0.03**	**40.05±0.06**	**40.40±0.03**	40.35±0.14

serves the purpose of estimating the value associated with any given state-action pair (s, a) within the dataset D, as dictated by the policy π. To gauge a policy's effectiveness, the average value of the initial state is computed.

3 Results and Discussion

Quantitative Analysis. We evaluate MeDT using autoregressive inference (Fig. 4) with a state predictor (Algorithm 1), in comparison to various baselines, including BCQ [9], transformer-based BC, and DT (Table 1). To assess patient outcomes, we calculate SAPS2 scores with the state predictor and conduct off-policy evaluation (OPE) using FQE. The policies are run in the loop with state predictor for only ten timesteps, to avoid accumulation of errors resulting from the autoregressive nature of evaluation. From Table 1, we observe that the proposed MeDT policy resulted in more stable estimated patient states relative to the baselines. Similarly, Fig. 5 shows that MeDT marginally outperforms BCQ, resulting in higher estimated Q-values. This suggests that the proposed goal conditioning has the intended effect.

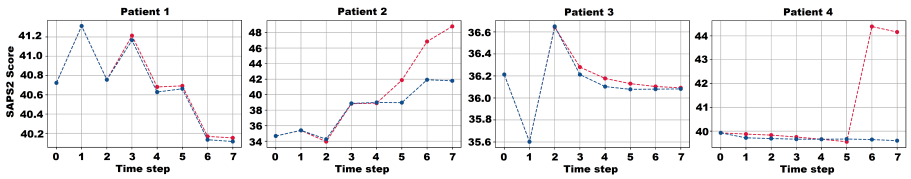

Figure 3: Visualization of 4 patient trajectories computed by the state predictor following treatment recommendation from DT (red) and MeDT (blue).

Qualitative analysis. We qualitatively evaluate the policy of MeDT against the clinician's policy. To ensure accurate analysis, we use ground-truth trajectories as input sequences instead of relying on autoregressive inference, which may lead to compounding errors. In Fig. 2a, we conduct a comparative analysis of the mean dose of vasopressors and IV fluids recommended by the MeDT policy and the clinician policy, for patient states with varying SAPS2 scores. Our results show that the MeDT policy generally aligns with the clinician's treatment strategy but recommends lower doses of IV fluids on average. Both policies exhibit a similar trend of increasing medication doses with worsening patient condition, for both vasopressors and IV fluids. It is important to note that the MeDT policy differs from previous works in that it does not recommend minimal dosages for patients with high SAPS2 scores [19]. Fig. 2b presents the dosage distribution of IV fluids and vasopressors recommended by both the MeDT and clinician policies. Our analysis reveals that the MeDT policy uses more zero dosage instances for both IV fluids and vasopressors, compared to the clinician policy.

In Fig. 3, we visualize the trajectories of multiple patients computed by the state predictor, following treatment actions recommended by both the DT and *MeDT* policies. The impact of ATG conditioning on patient health is evident, as *MeDT* leads to more stable trajectories, demonstrating the potential of our framework to generate targeted and improved treatment recommendations by considering both the hindsight returns and ATG at each timestep.

Our experimental results demonstrate the potential of *MeDT* to bolster clinical decision support systems by providing clinicians with an interpretable and interactive intervention support system.

References

[1] Pranav Agarwal, Aamer Abdul Rahman, Pierre-Luc St-Charles, Simon JD Prince, and Samira Ebrahimi Kahou. Transformers in reinforcement learning: A survey. *arXiv preprint arXiv:2307.05979*, 2023.

[2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[3] 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.

[4] Hiroki Furuta, Yutaka Matsuo, and Shixiang Shane Gu. Generalized decision transformer for offline hindsight information matching. *arXiv preprint arXiv:2111.10364*, 2021.

[5] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

[6] Yong Huang, Rui Cao, and Amir Rahmani. Reinforcement learning for sepsis treatment: A continuous action space solution. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 631–647. PMLR, 05–06 Aug 2022.

[7] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

[8] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[9] Taylor W. Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare. In *ML4H@NeurIPS*, 2020.

[10] Matthieu Komorowski, Leo Anthony Celi, Omar Badawi, Anthony C. Gordon, and Aldo A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24:1716–1720, 2018.

[11] Flemming Kondrup, Thomas Jiralerspong, Elaine Lau, Nathan de Lara, Jacob Shkrob, My Duc Tran, Doina Precup, and Sumana Basu. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15696–15702, 2023.

[12] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[13] Linghui Meng, Muning Wen, Yaodong Yang, Chenyang Le, Xiyun Li, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, and Bo Xu. Offline pre-trained multi-agent decision transformer: One big sequence model conquers all starcraftii tasks. *arXiv preprint arXiv:2112.02845*, 2021.

[14] A. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, 1999.

[15] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[16] Emilio Parisotto and Ruslan Salakhutdinov. Efficient transformers in reinforcement learning using actor-learner distillation. *arXiv preprint arXiv:2104.01655*, 2021.

[17] Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning. In *International conference on machine learning*, pages 7487–7498. PMLR, 2020.

[18] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[19] Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *ArXiv*, abs/1711.09602, 2017.

[20] Elsa Riachi, Muhammad Mamdani, Michael Fralick, and Frank Rudzicz. Challenges for reinforcement learning in healthcare. *arXiv preprint arXiv:2103.05612*, 2021.

[21] Suchi Saria. Individualized sepsis treatment using reinforcement learning. *Nature Medicine*, 24:1641 – 1642, 2018.

[22] Luregn J. Schlapbach, Scott L. Weiss, Melania M. Bembea, Joseph A. Carcillo, Francis Leclerc, Stephane Leteurtre, Pierre Tissieres, James L. Wynn, Jerry Zimmerman, Jacques Lacroix, and Marie E Steiner. Scoring systems for organ dysfunction and multiple organ dysfunction: The podium consensus conference. January 2022.

[23] Juergen Schmidhuber. Reinforcement learning upside down: Don't predict rewards–just map them to actions. *arXiv preprint arXiv:1912.02875*, 2019.

[24] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

[25] Tianxin Tao, Daniele Reda, and Michiel van de Panne. Evaluating vision transformer methods for deep reinforcement learning from pixels. *arXiv preprint arXiv:2204.04905*, 2022.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[27] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

[28] Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *International Conference on Machine Learning*, pages 24631–24645. PMLR, 2022.

[29] Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In *International Conference on Machine Learning*, pages 27042–27059. PMLR, 2022.

Appendix

3.1 Training

The transformer policy is trained on mini-batches of fixed context length, which are randomly sampled from a dataset of offline patient trajectories. In our case, we choose a context length of 20, which is the longest patient trajectory in the dataset following preprocessing. For trajectories shorter than this length, we use zero padding to adjust them. During training, we use teacher-forcing, where the ground-truth sequence is provided as input to the model. At each timestep, the ATG (k_t) is set to the actual acuity scores of the state at the next time step in the sequence. The prediction head of the policy

model, associated with the input token s_t is trained to predict the corresponding discrete treatment action a_t using cross-entropy loss. The losses over each timestep are then averaged. Additionally, the state estimator is trained to predict the patient's state following the treatment actions. The prediction head of the state predictor model, corresponding to the input token a_t is trained to estimate the continuous state s_{t+1} using mean square error loss. The models are each trained on a single NVIDIA V100 GPU. The results of the experiments are averaged over 5 seeds.

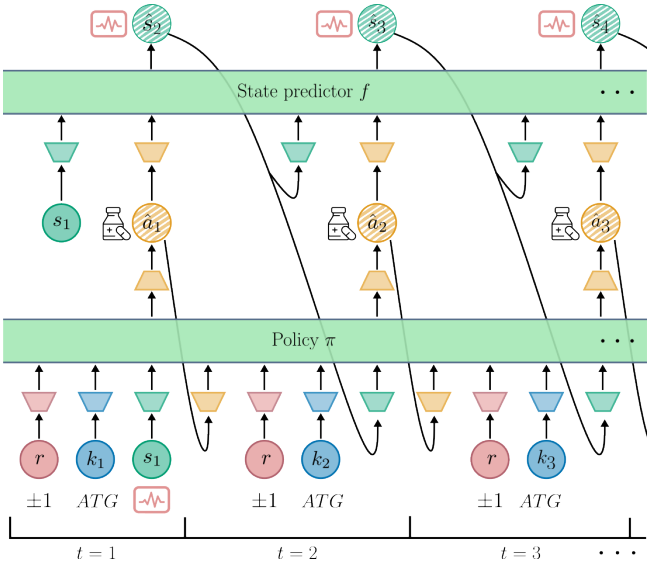

Figure 4: Autoregressive evaluation pipeline: At each timestep t, the pretrained state predictor attends to past recommended doses and predicted patient states, and outputs state prediction $\hat{s}_{t+1}$. Both dosage recommendations $\hat{a}_{t+1}$ and predicted states are fed back to MeDT to simulate treatment trajectories.

Algorithm 1 Evaluation Loop

Input: Initial patient state s_0
Output: Acuity score $g_1, \ldots, g_T$
Set target return $r_T = 1$
Initialize state $s_1 = s_0$, target return $r_{1:T} = r_T$ and action sequence $a_{0:t-1} = \{\}$
for $t = 1, 2, \ldots, T$ **do**
    Select desired Acuity To Go k_t
    Select action $a_t = \text{MeDT}(r_{1:t}, k_{1:t}, s_{1:t}, a_{0:t-1})$
    Append a_t to the sequence of actions: $a_{1:t} = a_{0:t-1} + [a_t]$
    Estimate new state: $s_{t+1} = \text{state\_estimator}(s_{1:t}, a_{1:t})$
    Evaluate acuity score g_{t+1} for state s_{t+1}
    Append s_t to the sequence of states: $s_{1:t} = s_{1:t-1} + [s_t]$
end for
return acuity score $g_1, \ldots, g_T$

Results and analysis

Interpretability. To enhance the interpretability and reliability of our *MeDT* model for external users, we analyse attention maps of different layers (second, fourth and sixth), as early, middle and last, respectively (Fig. 6). During inference, the input sequence is passed through the pretrained model and attention scores are obtained for each layer by averaging across multiple heads. As causal transformers attend only to current and past timesteps, the upper right part of the attention map is masked.

In the early layers of DT, attention is primarily focused on the patient's state. However in *MeDT*, with the addition of ATG, this attention is distributed across different features, with more emphasis

on the ATG constituents. In the middle and last layers, we observe that DT prioritizes the initial returns, while *MeDT* emphasizes the initial returns and the ATG scores, with greater emphasis on the cardiovascular acuity score. Such attention visualizations enable clinicians to understand the model's reasoning behind dosage prediction, making transformer-based models more interpretable.

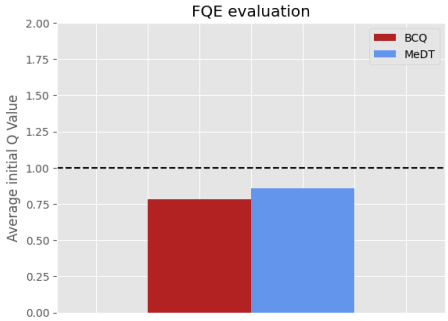

Figure 5: Mean initial Q-values for BCQ and MeDT. The maximum expected return is represented by the dotted line.

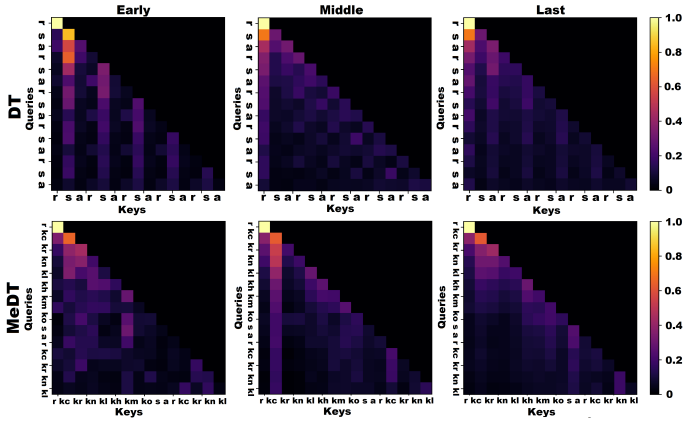

Figure 6: Visualization of attention maps of DT and MeDT at inference.

Limitations and future work. The MIMIC-III dataset has some limitations, as it only represents a specific geographic area, which could result in an over-representation of certain patient populations and an under-representation of others. Consequently, using the state predictor for evaluation may introduce biases inherent in the dataset on which it was trained. To mitigate these potential biases, we will investigate causal representation learning and pretraining techniques for that enhance model robustness. Moreover, further work can be carried out to enhance interpretability following the work of Chefer et al. [2], which proposes a more robust method of propagating relevancy through transformer layers. Despite these limitations, MeDT provides a general framework to harness the vast amount of data found in large-scale electronic health records (EHRs) from different modalities. As a result, researchers can explore the scalability of the transformer architecture to develop treatment recommendations for various medical conditions in the future.