

We thank the reviewers for their time, suggestions and questions that we believe will improve the quality of the paper. Below we summarize our overall response to the reviewer’s questions and comments.

- We will add a discussion on the relationship between the key metrics, NEWS and MAP, and the cost function, as shown in Figure 1 of the attachment. The cost function can indeed capture dangerous states that the reward function overlooks, addressing the issue of latent variables that cannot be incorporated into the reward function.
- We will include an off-policy evaluation as suggested by reviewer guwJ. Referring to [1], we will evaluate our method and others using multiple evaluation metrics under the same reward function and behavior policy, as shown in Table 1 of the attachment. Our policy performs well in RMSEIV、WISb、WIS_t、WIS_{bt}. For the specific meanings of these metrics, please refer to [1].
- We will add a sensitivity analysis of the generative world model to CDT+CT, as suggested by reviewer sHtP. As the target reward increases, the generated world model exhibits more aggressive behavior, which can improve the performance of the estimated policy, but there is an upper limit to this effect.
- We will clarify the meaning and calculation process of the evaluation metric ω , as suggested by reviewers sHtP and cbZJ. For details, please refer to our responses to reviewer sHtP (Response2-1) and cbZJ (Response5-3).

References:

[1] Luo, Zhiyao, et al. "Position: Reinforcement Learning in Dynamic Treatment Regimes Needs Critical Reexamination." Forty-first International Conference on Machine Learning.

Review1- Reviewer R5ke

Weaknesses:

- It seems like the system (which is the patient) is not feasible to model as evolving according to an Markov process on the observed state at each time, but instead a POMDP with a high dimensional latent state.

Thank you for your suggestion. We agree that the complexity of the patient system should be described as a POMDP, as the MDP model indeed oversimplifies the system. We constructed a non-Markovian model by incorporating historical state sequences into the RL decision-making process, with the goal of capturing potential states from history, which aligns with your suggestion. To ensure a more rigorous presentation, we have reformulated it as a CPOMDP framework in Section 3 Problem Formulation, defined as

$(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{Z}, \mathcal{R}, \mathcal{C}, \gamma, \kappa, \rho_0)$, where \mathcal{O} is the set of observations o ,

$\mathcal{P}(s'|s, a) = Pr(s_{t+1} = s' | s_t = s, a_t = a)$ is the transition probability, and

$\mathcal{Z}(o|s', a) = Pr(o_{t+1} = o | s_{t+1} = s', a_t = a)$ is the observation probability.

- Clarity and presentation can be improved. In Equation (4), what is τ^\wedge , this was never defined for being such a key component of the procedure. Similar issues in Eq (8) related to clarity.

We apologize for any inconvenience caused in your reading. The explanations for the two

points mentioned above are as follows:

1. In Equation (4), τ^\wedge is the trajectory which are sampled from the executing policy (training policy) $\mathcal{M}^\wedge\{\hat{\zeta}_\theta\}$. In an online environment, we can use the execution policy to interact with the environment and generate some trajectories τ^\wedge .
2. In Equation (8), $x_t = \{h_t \cup a_t\} = \{\hat{R}_{-K:t}, s_{-K:t}, a_{-K:t}\}$ is the input which includes the reward R , states s and action a from the preceding K timesteps, where K is the context length. The input is encoded by linear layers and pass through the casual transformer to predict hidden tokens g_t . Next, we use the hidden tokens as input and employ two linear layers (L_μ & L_φ) to predict the current reward r_{t-1} and the next state s_t , with the objective of minimizing the mean squared error for each linear layer, defined as:

$$\min_{\varphi, \mu} \mathbb{E}_{s_t, r_{t-1} \in x_t \sim \mathcal{D}_e} [\{(s_t - \ell_\varphi(g_t))^2 + (r_{t-1} - \ell_\mu(g_t))^2\}]$$

In addition, we will improve the overall clarity and presentation of the paper to avoid such issues.

Questions:

- In explaining the main Equation (4), the blurb in lines 162-164 needs some attention. Specifically, what does "MDP obtained after augmenting M with cost function C_θ , using the executing policy ..." mean exactly?
The ICRL approach involves two policies: one is the expert policy π_e , and the other is the policy being trained (execution policy $\pi_{M^{\{\zeta_\theta\}}}$). This sentence defines the Markov Decision Process (MDP) model for the latter. $M^{\{\zeta_\theta\}}$ represents the MDP that results from adding the cost function C_θ to the original MDP M . The executing policy for this augmented MDP is denoted as $\pi_{M^{\{\zeta_\theta\}}}$.
- In Eq (8) what does $s_t, r_{t-1} \in x_t \sim \mathcal{D}_e$ represent? It looks like x_t is a representation generated by the transformer, and s_t, r_t are from the data. Is there a problem in the notation?
Sorry, there is indeed an issue here. We have redefined the data generated by the transformer as g_t .
- Is the transformer representations being trained from the gradients of both the policy model as well as the world model? If so, how do they interact? If not, consider adding stop gradients at the relevant places in Eq (7) and/or (8) to make this explicit.
Yes, in the model-based offline RL framework, the transformer is trained together with the policy model and the world model. The objectives of the policy model and the world model are to generate actions, rewards, and the next state, respectively. The transformer structure is used to extract historical information for the policy and world models. When training model-based offline RL, the goal is to simultaneously minimize Equations 7 and 8. During this process, the transformer is also trained alongside the objectives until convergence.

Review2- Reviewer guwJ

Reward Function Design: The reviewer may have misunderstood our reward function design. In our work, the reward function did not directly utilize mortality rates but rather included intermediate rewards (as shown in Appendices B.1 and B.2). For example, in the design for sepsis, intermediate rewards like the SOFA score were included. For different

diseases, we designed different reward functions based on previous literature. Additionally, we agree with the reviewer's comment that "complex reward design can facilitate the learning of strategies," but the design of reward functions is relatively challenging and requires the involvement of medical experts. We also recognize that the current reward function design may not account for hidden variables (potentially fatal). Therefore, we use a relatively simple reward function, incorporating a cost function with historical dependency, to take into account the changes in indicators like MAP and NEWS that were not considered in the reward function. This will guide the agent in learning safe and effective strategies.

To verify whether our penalty function can capture changes in NEWS and MAP, we conducted supplementary experiments as shown in Figure 1. When the NEWS score is too high, the penalty value increases accordingly; similarly, when MAP is outside the normal range, the penalty value also increases. This indicates that the penalty function can compensate for the shortcomings of the reward function design.

Evaluation Metrics: We acknowledge that the evaluation metric ω is not an ideal measure. Therefore, we also consider the relationship between the penalty value and medical metrics (section 5.1) and the probability of dangerous actions in the policy (section 5.2) as part of the evaluation criteria to compare the safety of different policies from multiple perspectives. In the experiments, all methods are based on the same reward function, so the design of the reward function is not a variable factor.

Behavior Policy Fitting Error: In our Offline RL, we did not fit the clinicians' policy using a neural network, so there is no fitting error in this part of the experiment. The reviewer has provided a valuable suggestion, highlighting that fitting the behavioral policy could indeed help eliminate some confounding factors. In our supplementary experiments, we utilized neural networks to fit the behavioral policy, corrected the model accordingly, and then used OPE evaluation to conduct the offline policy evaluation experiment.

Offline Policy Evaluation: We supplemented our work by referring to [1] for the offline policy evaluation as follows: Using the same behavioral fitting function and the same NEWS2 reward, we compared the results of the policy under different evaluation metrics, as shown in Table 1. The CDT+CT method performed better than other methods on the RMSE_IV, WIS, WIS_b, and WIS_bt evaluation metrics.

[1] Luo, Zhiyao, et al. "Position: Reinforcement Learning in Dynamic Treatment Regimes Needs Critical Reexamination." Forty-first International Conference on Machine Learning.

Model-Based Off-Policy Evaluation: Since we aim to evaluate whether there are instances of excessively high or sudden changes in medication dosage, our model's action space consists of the actual dosage values. However, DTR-Bench currently cannot perform online evaluations of medication dosages, as it only provides a discrete action space (i.e., "yes" or "no" for administering medication). In this case, there is no issue of overestimation, so this online testing method cannot assess whether the policy effectively avoids dangerous behaviors. Exploring a simulation environment based on continuous action spaces is a promising direction for future research.

Questions

Attention Mechanism: The paper highlights the importance of the causal attention mechanism in the Constraint Transformer. Since the two experimental datasets are both short-term time series datasets, is transformer really necessary? **Is it possible that an RNN can be better than a transformer?**

Compared to RNNs, the Transformer architecture has the following advantages:

1. **Computational Efficiency:** Due to the global computation allowed by the Transformer's self-attention mechanism, it can be highly parallelized during training, leading to faster training speeds. Transformers can significantly improve training efficiency. RNNs, on the other hand, require sequential processing of each time step, which makes parallelization difficult and results in slower training speeds.
2. **Capturing Complex Information:** Transformers utilize multi-head attention mechanisms to simultaneously focus on different parts of the sequence, allowing them to better capture complex relationships in medical data. In medical datasets, events might be recorded with non-uniform time intervals. The Transformer's self-attention mechanism does not rely on fixed time steps, making it more flexible in handling such situations.
3. **Interpretability:** The self-attention mechanism of Transformers makes it easier to understand which parts of the data sequence the model is focusing on, providing better interpretability, which is crucial in the medical field. RNNs, with their internal states and memory units, are more difficult to interpret, which may reduce the transparency of clinical decision-making.

Review3- Reviewer sHtP

Weaknesses:

- The proposed method depends heavily on the generated violating data, which defines the objective function in the constraint transformer. How sensitive is the estimated policy to the generative world model? Figure 12 shows that the action distributions in the expert dataset and the violating dataset are different. The VASO action seldom takes a large value in the violating dataset. Will this distribution difference cause any trouble in the learning of the constraint?

- (1) How sensitive is the estimated policy to the generative world model?

To explore the sensitivity of the estimated policy to the generative world model, we designed the following experiment. The quality of the data generated by the world model is determined by the target reward. As the target reward increases, the world model generates more aggressive data in order to obtain higher rewards. So we set the target reward to 1, 5, 10, 20, and 30, respectively, and observed the impact of the generated data on the policy, as shown in Table 2. As the target reward increases, the performance of the policy improves; however, there is an upper limit, and it will not increase indefinitely.

- (2) Will this distribution difference cause any trouble in the learning of the constraint?

We think that this impact exists, but it will not be significant. Firstly, the generative model used for creating the non-compliant dataset in this paper is a reinforcement learning (RL) model capable of generating data within legal boundaries (see Appendix B.1 in the paper). The actions, states, and rewards it generates must all

fall within these legal boundaries. Therefore, this generative model is less likely to produce extremely high drug dosages. However, previous work has confirmed that such models can indeed generate non-compliant data [1]. Consequently, this generative model still presents a "potential risk."

[1] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. arXiv preprint 507 arXiv:2302.07351, 2023.

- Could the authors provide some details on how the DIFF between the estimated policy and the physicians' policy is calculated through graphical analysis? It would also be helpful if the authors could explain how Figure 7 is plotted. Is it calculated based on the dosage differences at each timestamp? In addition, what are the implications of the three DIFF evaluation metrics in Table 2? Since both IV and VASO are part of the action space, is the ACTION DIFF alone sufficient to evaluate the estimated policy?
 - (1) In the real medical dataset, we know the patient's state and the drug dosage (a) under the doctor's policy. We use the estimated policy to provide the drug dosage (b) for the same patient state under the estimated policy. We then calculate the DIFF for each patient state, which is $b - a$, and we also compute the mortality rate and standard deviation (std) of patients under the doctor's policy for different DIFF values, thereby obtaining Figure 7.
 - (2) Yes, we calculate the dosage differences for each timestamp.
 - (3) **IV DIFF** represents the dosage difference of IV medication between the two policies at each timestamp. **VASO DIFF** represents the dosage difference of VASO medication between the two policies at each timestamp. **ACTION DIFF** represents the difference between the two policies for the vector composed of IV and VASO medications at each timestamp.
 - (4) **ACTION DIFF** can evaluate the policy individually. However, this might present a problem: even if we standardize the dosages of the two medications to the same dimension, there may still be cases where the policy performs better for only one specific action. Therefore, for a more comprehensive evaluation, we need to focus on the performance of actions in each dimension.

Review4- Reviewer cbZJ

1. We did indeed introduce our challenge in the introduction by raising several issues, where the challenges (Markov assumption and history dependence) are the key issue we aim to address. This writing style may have given the impression that the focus of the problem was not clear, so we will revise it to make it more straightforward. To aid your understanding, we will briefly re-explain it as follows:
 - (1) The current RL methods exhibit risky behavior, so we aim to address this with constrained reinforcement learning (CRL).
 - (2) The acquisition of constraints is crucial for the implementation of CRL. However, the custom constraint functions currently suffer from a lack of personalization. Therefore, ICRL emerges as a promising approach.
 - (3) However, the current ICRL faces challenges when applied to medical scenarios due

to the **Markov assumption** and **history dependence** issues (challenges), making it difficult to be effectively applied in healthcare.

Based on the above, we aim to achieve the goal of safe policy learning in RL by focusing on addressing these challenges. We hope this helps with your understanding.

2. We apologize for not explaining this clearly. By "too high," we meant a drug dosage that is dangerous for any patient state. We fully agree with the reviewer's comment that "unsafe behavior without conditioning on the patient state is inappropriate." Therefore, we have taken this into careful consideration in our design and have developed a personalized constraint function $C(\tau)$, which assigns a penalty to the current drug dosage based on the patient's historical treatment trajectory.
3. Undoubtedly, incorporating history into the state or redefining the state are the most direct approaches. However, there may be the following issues:
 - (1) Redundant Computation and Increased Complexity: Incorporating history into the state can lead to redundant calculations and increased complexity. For a patient's trajectory, the state at timestamp t will include the previous $t-1$ timestamps, which can result in excessive redundancy. The number of timestamps stored increased from $O(n)$ to $O(n^2)$.
 - (2) Inability to Represent Latent States: Latent states might be unobservable and unknown in medical contexts. As Reviewer 1 pointed out, the patient system cannot be modeled based on a Markov process due to the high-dimensional latent states that cannot be represented.
4. Organization & Flow
 - (1) Thank you for your suggestion. ICRL is part of previous work, which we have mentioned in the Introduction. However, we did not clarify this clearly in the Method section, which may have caused misunderstanding. We will provide additional explanation here.
 - (2) This is indeed a description of the experimental setup rather than the problem definition. We have corrected it.
5. Writing and notation
 - (1) Not necessarily; it could be either (s,a) or trajectory τ . Therefore, we do not impose any restrictions here.
 - (2) In L144, it is explained that $2N$ represents selecting $2N$ patients from the dataset, therefore N represents a constant, and $2N$ is less than or equal to the size of the dataset. And among these $2N$ patients, N patients died under the doctor's treatment, and N patients survived. We calculate the DIFF for the $2N$ patients (see L137 or next question) and rank them in ascending order of DIFF. The top N patients are then selected as the "top N ."
 - (3) We apologize for any confusion our previous statement may have caused. We will restate it to address your question:
 - i. Select $2N$ patients from the dataset. And among these $2N$ patients, N patients died under the doctor's treatment, and N patients survived.
 - ii. In the real dataset, we know the patient's state and the drug dosage (a) under the doctor's policy. We use the estimated policy to provide the drug dosage (b) for the same patient state.
 - iii. We then calculate the DIFF for each patient, defined as $\text{DIFF} = b - a$. For the $2N$

- patients, we can obtain $2N$ DIFF values.
- iv. Sort the $2N$ DIFF values in ascending order and observe the survival status of the top N patients recorded in the dataset (under the real doctor's policy).
 - v. The top N patients indicate that the difference between our policy and the doctor's policy is small. And if the survival rate of these top N patients is higher, it suggests that our policy is closer to the ideal optimal policy. If our policy is the ideal optimal policy, the survival rate of the top N patients after sorting will be the highest.
 - vi. In addition, we also need to consider the size of the DIFF values. For surviving patients, a smaller DIFF difference is preferable.
- (4) Eqn (3) what is $R(\tau)$, L157 is it ZM or ZMC, what is β
- i. In Eq (3), $R(\tau)$ is the reward of the trajectory τ .
 - ii. L157 is ZMC.
 - iii. $\beta \in [0, \infty)$ is a parameter describing how close the agent is to the optimal distribution (as $\beta \rightarrow \infty$ the agent becomes a perfect optimizer and as $\beta \rightarrow 0$ the agent simply takes random actions).
- (5) The cost function can be formulated as $C_\theta = 1 - \zeta_\theta$. $M^{\{\zeta_\theta\}}$ represents the MDP that results from adding the cost function C_θ to the original MDP M . The executing policy for this augmented MDP is denoted as $\pi_{M\zeta_\theta}$.
- (6) Thank you for the reminder; I have made the correction.
- (7) "Non-Markovians" refers to scenarios where the future state of a process depends not only on the current state but also on past states or actions. This concept is in contrast to "Markovian" processes, where the future state depends solely on the current state and is independent of past states or actions.
- (8) L207 What is "causality transformer"? In "generate the importance weights", importance weights has a specific meaning in RL. The authors probably meant the cost weights.
- i. "causality transformer" is the casual transformer, which refers to a variant of the Transformer model that incorporates causal (or temporal) relationships into its architecture.
 - ii. Yes, "the importance weights" is the cost weights, and I have made the correction.
- (9) Here's a revised version: We construct an expert dataset and a violating dataset to evaluate Equation 6 offline.
- (10) In L249, we add references [1] and [2], which experimentally demonstrate that excessively high rewards can incentivize agents to violate constraints.
- [1] Guiliang Liu, Yudong Luo, Ashish Gaurav, Kasra Rezaee, and Pascal Poupart. Benchmarking 431 constraint inference in inverse reinforcement learning. arXiv preprint arXiv:2206.09670, 2022.
- [2] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. arXiv preprint 507 arXiv:2302.07351, 2023.
- (11) We think that expert policy should have the highest possible survival rate to ensure the correctness of their policy. And it is likely that the deaths in the dataset are caused by issues with the doctors' policy, so we hope to exclude these disturbances.

6. Questions about experiments:
- (1) Firstly, mortality rate is not the reward in the experiment; the definition of the reward function is provided in Appendix B. The reward value is only related to the state and not directly related to mortality rate. Secondly, the experiment demonstrates the relationship between the cost function and mortality rate to observe whether the cost function can penalize actions that lead to higher mortality rates. Moreover, the cost function helps prevent policies from making unsafe decisions that might lead to higher mortality rates, making the model's decision-making process more interpretable. Clinicians need to understand not only the positive outcomes of actions but also their costs and risks. A separate reward function alone cannot achieve this.
 - (2) In our ablation study, as shown in Figure 5, when the attention layer is present, the cost value of CT is proportional to the mortality rate, indicating that the cost value can effectively penalize high mortality behaviors. However, when the attention layer is absent, there is no observed proportional relationship, making it difficult to determine whether the cost value can correctly penalize.
 - (3) We did not say that our method can evaluate unseen (s, a). What we want to express here is the reason for not using FQE. FQE is unable to evaluate unseen (s, a) pairs, which leads to inaccurate results. My evaluation method does not assess unseen (s, a) pairs; it only tests existing (s, a) pairs. Therefore, in this aspect, our evaluation method is accurate.
 - (4) We preprocess the data by normalizing it, so both training and evaluation processes operate on the same scale.
 - (5) In L264 "Tasks", we specify that the experiments are conducted in sepsis and mechanical ventilator environments, and we provide detailed locations of the related definitions for the two tasks.
 - (6) It's unclear what Fig 8 is showing. Table 3 not explained, unclear if lower or higher is better, what is $\max \Delta$?
 - i. Figures 8 and 7 use the same statistical methods. In Figure 8, in the mechanical ventilator environment, the relationship between different algorithm strategies (DDQN, CQL, CDT, and CDT+CT) and the action gap with the doctor's strategy (PEEP DIFF and FiO2 DIFF) and mortality rate is illustrated. The horizontal axis represents the action parameters' (PEEP and FiO2) DIFF, while the vertical axis represents the mortality rate.
 - ii. Lower is better. The lower the proportion of "too high" and "sudden change" the better.
 - iii. $\max \Delta$ represents the maximum change in medication dosage.

References:

- [1] Luo, Zhiyao, et al. "Position: Reinforcement Learning in Dynamic Treatment Regimes Needs Critical Reexamination." Forty-first International Conference on Machine Learning.