InfCTI-ImpCTI: Inferring and Implementing Clinicians' Treatment Intentions

Jinsheng Shi, Yuyu Yuan*, Yuang Cai, Rui Han, Zhenyu Zhao, Zijie Shi
School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications, Beijing, China
Key Laboratory of Trustworthy Distributed Computing and Service (BUPT),
Ministry of Education, Beijing, China
{jinsheng,yuanyuyu,cyang,hanr,zhaozhenyu,shizijie}@bupt.edu.cn
* Corresponding author

Abstract

In the field of medical decision-making, understanding the treatment intentions of clinicians is crucial for effective treatment strategies. However, these intentions are often implicit and challenging to quantify. In this paper, we propose a novel two-module model to infer and implement clinicians' treatment intentions through treatment records. We construct the InfCTI module, which infers intentions and quantifies them numerically, and the ImpCTI module, which generates treatment strategies based on inferred intentions. Our experiments demonstrate that the treatment strategies obtained by ImpCTI reflect clinicians' intentions and the intention values obtained by InfCTI are reasonable. This model has the potential to improve the quality of care provided to patients.

Keywords: intention inference; computational cognitive modeling; inverse reinforcement learning; imitation learning

Introduction

In the field of medical decision-making, clinicians often face the challenge of making complex treatment decisions that rely on a combination of medical evidence, personal experience, and treatment goals. However, the intentions behind these treatment decisions are usually implicit and difficult to quantify. Additionally, real-time monitoring of patients can be resource-intensive. To address these challenges, researchers have turned to AI techniques, such as imitation learning and reinforcement learning, to assist in decision-making therapy.

Imitation learning(Hussein, Gaber, Elyan, & Jayne, 2017) has shown great promise in developing medical treatment models by training the model to imitate expert clinicians(Wang, Tang, He, & He, 2022). However, relying solely on imitation learning may limit the model's ability to handle unexpected situations that deviate from the training dataset, and it may struggle with generalization to different patients or populations. To improve the model's performance, it is necessary to incorporate the intent of the clinician into the learning process.

On the other hand, reinforcement learning(Sutton, Barto, et al., 1999) provides an approach for an agent to learn sequential decision-making through trial and error. However, training an agent in a real-world medical setting can be time-consuming, costly, and potentially risky. While simulation environments attempt to capture the complexities and dynamics of medical situations, it is difficult to achieve a perfect representation of the real-world.

To address these challenges, Offline Reinforcement Learning(Levine, Kumar, Tucker, & Fu, 2020) has been proposed as a solution to avoid the need for real-time interaction with the environment. This approach allows the agent to learn from pre-collected trajectories, but it still requires experts to annotate the dataset with reward signals to guide the agent toward learning good policies. The challenge here is that experts need to assign appropriate rewards to each state or action based on their domain knowledge and expertise, which is a difficult task for experts to quantify the reward value that should be marked.

Regardless of whether using imitation learning, reinforcement learning, or offline reinforcement learning, understanding the clinicians' intentions is of utmost importance. By understanding clinicians' treatment intentions, we can improve the efficacy and quality of medical treatments. In this paper, we propose a novel two-module model to infer clinicians' treatment intentions through treatment records and gain a better treatment policy through the intentions of the clinicians that are obtained. Our contributions are as follows:

- We constructed the InfCTI(Inferring Clinicians' Treatment Intentions) module, which is capable of inferring clinicians' treatment intentions and quantifying them numerically which can help improve the quality of care provided to patients.
- We constructed the ImpCTI(Implementing Clinicians' Treatment Intentions) module to obtain a treatment strategy. This module combines the intentions inferred by InfCTI to obtain a treatment policy.
- We prove through experiments that the treatment strategies obtained by ImpCTI can reflect the clinicians' treatment intentions, and analyzed that the intention value obtained by InfCTI is reasonable.

Background and Related Work

In the field of cognitive science, inferring intentions(Catmur, 2015) refers to the process of understanding and predicting the goals or motivations behind an individual's actions or behaviors. This area of research is crucial for understanding human social interactions, communication, and decision-making processes.

The inferring intention from behavior(Royka, Török, & Jara-Ettinger, 2023) is a complex and multi-disciplinary research area that has attracted the attention of scholars from

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

various fields such as psychology, philosophy, neuroscience, artificial intelligence, and robotics. The main focus of this research is to understand how humans make inferences about the intentions of others based on their behavior and how this knowledge can be applied to various domains.

ToM(Theory of Mind)(Den Ouden, Frith, Frith, & Blakemore, 2005) is a theory that suggests people can infer the mental states of themselves and others, such as beliefs, desires, and intentions. This ability allows individuals to predict and understand others' behavior, comprehend their motivations and intentions, and engage in complex social interactions. In essence, ToM(Berke & Jara-Ettinger, 2022) forms the foundation for inferring intentions. By observing how someone behaves, we can make educated guesses about their thoughts and emotions. For example, a smile and laughter during a conversation(Lv, Li, Wang, & Zeng, 2022) may indicate happiness or amusement, while a frown and avoidance of eye contact may suggest discomfort or unhappiness. By using behavior as a cue, we gain insight into others' intentions and adjust our own behavior accordingly.

Cognitive neuroscience(Gazzaniga, 2004) investigates the neural mechanisms underlying cognitive processes, including intention inference. Neuroimaging techniques, such as functional magnetic resonance imaging(fMRI) and electroencephalography(EEG), have been employed to study the brain regions and networks involved in inferring intentions. These studies aim to uncover the neural basis of intention understanding. However, this research is focused on understanding the fundamental mechanisms of intention inference and does not directly relate to practical applications. In contrast, our research focuses on using computational cognitive modeling to infer intentions from behavior.

In modeling higher-level cognitive processes, researchers have also used computational cognitive modeling(Farrell & Lewandowsky, 2018) to capture the underlying mechanisms and algorithms that govern various aspects of human cognition, such as perception, attention, memory, learning, decision-making, and problem-solving. Computational cognitive modeling is an interdisciplinary field that combines principles from cognitive science, psychology, neuroscience, and computer science to create computational models that simulate and explain how humans infer intentions. It formalizes cognitive theories and hypotheses into mathematical or computational frameworks.(Wu, Sridhar, & Gerstenberg, 2023) They translate psychological theories and empirical findings into a set of rules, algorithms, or equations that can be implemented and simulated on a computer. The models take inputs, process information, and produce outputs that mimic human cognitive behavior. Common computational modeling paradigms include deep neural networks, reinforcement learning, Bayesian modeling, and probabilistic graphical models.

To enable intelligent robots to effectively collaborate with humans, Inagaki, Sugie, Aisu, Ono, and Unemi (1995)proposed a method for inferring intentions from human behavior, which consists of three levels: perception, recognition, and intention inference. The intention inference level utilizes groups of fuzzy rules that match qualitative expressions to specific situations related to the cooperative task. By reasoning with these fuzzy rules, the system infers the human's intention and determines when to exit from the task. However, qualitative inference of intention is only applicable to limited states and cannot express intention in more detail.

The methods introduced in the work of Marken (2013), known as the "Test for the Controlled Variable" (TCV), are based on the premise that intentional behavior is equivalent to the process of control. By applying the TCV, one can objectively infer the intentions underlying behavior by examining the perceptual variables that are under control and the goal states associated with those variables. This approach provides psychologists with an empirical ToM framework grounded in active experimentation rather than passive observation. It is important to note that the applicability of this method is contingent upon the ability to conduct active experimentation. However, in clinical treatment settings, conducting active experiments with patients is often impractical or unfeasible. Instead, in these scenarios, it becomes necessary to infer intentions based on past treatment records and observations rather than direct experimentation.

It is relatively easy for humans to make knowledge inferences by observing another person's actions to complete a goal. Rafferty, LaMar, and Griffiths (2015) developed a general framework for automating such inferences based on observed actions, which allows us to gain insights into student knowledge by observing their behavior. However, only the knowledge is inferred, and the strategy is not combined with the knowledge. For example, decision-making in healthcare to assist clinicians.

In multi-agent reinforcement learning, Raileanu, Denton, Szlam, and Fergus (2018) proposed Self Other-Modeling (SOM) approach, where an agent uses its own policy to predict the actions of the other agents and updates its belief of the hidden state in an online manner. It focuses solely on predicting the actions of other agents and does not involve inferring their intentions from historical behavior sequences.

However, many methods in cognitive science for inferring intentions rely on active online interactions or implicit reflections in the prediction or decision-making process. These approaches often lack the ability to express intentions at a finer granularity.

In response to these challenges, we propose the InfCTI-ImpCTI model, which allows for the inference of clinicians' treatment intentions from offline data and quantifies them into a reward function. By utilizing this reward function, which represents the treatment intentions of clinicians, we can train an agent to develop effective treatment strategies.

InfCTI-ImpCTI Model

Our proposed InfCTI-ImpCTI is a two-module model, which aims to infer clinicians' treatment intentions and obtain treat-

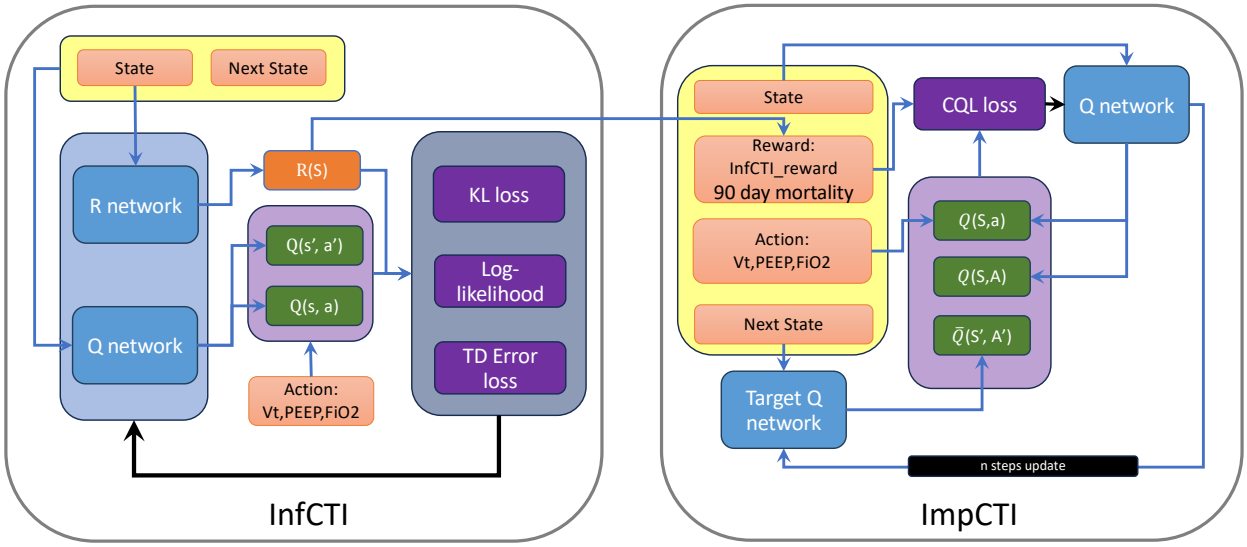


Figure 1: InfCTI-ImpCTI: The left module, InfCTI, infer clinicians' treatment intentions. The right module, ImpCTI, leverages the inferred treatment intentions obtained from InfCTI to learn a treatment policy.

ment strategies based on these intentions. As shown in Figure 1, the left module is InfCTI (Inferring Clinicians' Treatment Intentions), which uses the idea of inverse reinforcement learning(Chan & van der Schaar, 2021). It uses the patient's treatment records to infer clinicians' intentions and quantifies it into a reward function. The right module is ImpCTI (Implementing Clinicians' Treatment Intentions), which generates treatment strategies based on inferred clinician's intentions and treatment records through offline reinforcement learning to assist clinicians.

Inferring Clinicians' Treatment Intentions

We utilize two neural networks: the Q-network Q_θ, which computes the value of state-action pairs, i.e., Q(s, a), and the R-network r_ϕ, which estimates the reward distribution that reflects the intentions of clinicians. These two networks adopt the same structure, using 2 hidden layers of 64 units and exponential linear unit(ELU) activation functions. The R-network generates a Gaussian distribution, and then we use its mean as the reward value. We jointly train these two networks using three constraints.

The first constraint is the log-likelihood between the reward distribution and the TD error generated by the Q-values Q(s, a) and Q(s', a') for current and next state-action pairs. This constraint aims to optimize the reward function using the TD error as a supervisory signal. The second constraint is based on the Kullback-Leibler Divergence between the reward distribution and the standard normal distribution. This constraint is designed to normalize the reward distribution. The third constraint considers the log-likelihood between Q-values and selected actions in the dataset. This constraint allows the Q-network to imitate the behavior of clinicians, thereby correcting the Q-value. By incorporating these three constraints, we jointly train the Q-network and R-network in a unified manner. This approach enables us to effectively capture reward function based on the intentions of clinicians, ultimately leading to improved learning performance.

$$\begin{aligned}\mathcal{L}(\phi,\theta,\mathcal{D}) = \sum_{(s,a,s',a')\in\mathcal{D}} & log\frac{exp(Q_\theta(s,a))}{\sum_{b\in\mathcal{A}} exp(Q_\theta(s,b))} \\ & - D_{KL}(r_\phi(R(s))||p(R(s))) \\ & + log r_\phi(Q_\theta(s,a) - \gamma Q_\theta(s',a'))\end{aligned} \tag{1}$$

Equation(1) is the loss function formed by the combination of three constraints. We optimize ϕ and θ simultaneously to maximize Equation(1). In this way, a reward function can be obtained that contains clinicians' treatment intentions. Next, we use this reward function for the implementation of the treatment strategy.

Implement Clinicians' Treatment Intentions

The ImpCTI module combines offline reinforcement learning with clinicians' intentions obtained using InfCTI to generate a treatment strategy that aligns with the clinicians' intentions. The module architecture consists of two hidden layers, each containing 256 units and utilizing the ReLU activation function.

Since the module can only learn from pre-collected data without interacting with the environment, the CQL(Kumar, Zhou, Tucker, & Levine, 2020) objective is introduced to address the problem of overestimation outside the data distribution. To estimate a lower-bound value function for the policy's actual performance value, ImpCTI employs a two-objective approach in training the Q-function. The first objective serves as a regularizer that minimizes the Q-values on unseen actions with overestimated values while also maximizing the expected Q-value on the dataset. The second ob-

jective is the standard TD(temporal difference) error, which helps accurately estimate the Q-values.

$$\hat{Q}^{\pi}_{CQL} \leftarrow \arg\min_{Q} \max_{\mu(a|s)} \underbrace{(E_{s\sim data, a\sim\mu} Q[s,a] - E_{s,a\sim data} Q[s,a])}_{\text{CQL regularizer}} + \frac{1}{2\alpha} \underbrace{E_{s,a,s'\sim data}[(r(s,a) + \gamma E_{\pi}[\bar{Q}(s',a')] - Q(s,a))^2]}_{\text{standard TD error}} \quad (2)$$

In this formulation, μ represents the treatment strategy that is continuously optimized, and α is a tunable weight. When α is smaller, the impact of the CQL regularization term is relatively minor, leading to a learned policy that is more inclined towards utilizing clinicians' treatment intentions.

Experiment

This section aimed to evaluate the similarity of the ImpCTI and DeepVent(Kondrup et al., 2023) with the treatment actions of clinicians, demonstrate the effectiveness of ImpCTI in approximating clinician intentions through case analyses, and verify the rationality of the reward function by calculating the covariance between medical indicators and reward values.

We constructed a dataset following Kondrup et al. (2023)'s method. This dataset is treatment trajectories of patients undergoing mechanical ventilation. InfCTI-ImpCTI uses this dataset to infer clinicians' treatment intentions and learns treatment strategies based on those intentions. DeepVent is a treatment model based on offline reinforcement learning. It does not consider clinicians' treatment intentions but customizes a reward function for policy learning.

Preliminary

The dataset employed in this study was obtained from the MIMIC-III database and underwent identical preprocessing procedures to those detailed in (Kondrup et al., 2023). More specifically, we extracted data within the initial 72 hours following intubation and constructed trajectories with a time window of 4 hours. Consequently, each episode had a maximum length of 18, and every time-step state included 34 measurement indicators.

State Space The state space S consists of 34 measurement indicators, as shown in Table 1.

Action Space The action space consists of 3 types of actions, each type has 7 action ranges, so there are 343 action combinations, as shown in Table 2.

Reward Function The reward function is R(s) which learned by InfCTI. Given a state s, it can generate a reward value that reflects the intentions of clinicians.

Comparative Analysis

If a model acquires the intentions of clinicians, its treatment strategy should resemble that of clinicians. This comparative analysis aimed to assess the similarity between the behavior of ImpCTI and the treatment decisions made by clinicians.

Table 1: State Space

Demographics: Age, Gender, Weight, Readmission to the ICU, Elixhauser score
Vital Signs: SOFA (Sequential Organ Failure Assessment), SIRS (Systemic Inflammatory Response Syndrome), GCS (Glasgow Coma Scale), Heart Rate, SysBP (Systolic Blood Pressure), DiaBP (Diastolic Blood Pressure), MeanBP (Mean Blood Pressure), Shock Index, Respiration Rate, Temperature, SpO2 (Peripheral Capillary Oxygen Saturation)
Lab Values: Potassium, Sodium, Chloride, Glucose, BUN (Blood Urea Nitrogen), Creatinine, Magnesium, Carbon Dioxide, Hemoglobin, WBC (White Blood Cell Count), Platelet, PTT (Partial Thromboplastin Time), PT (Prothrombin Time), INR (International Normalized Ratio), pH, PaCO2 (Partial Pressure of Carbon Dioxide in Arterial Blood), Base Excess, Bicarbonate

Table 2: Action Space

Ideal Weight Adjusted Tidal Volume (Vt)		Positive End Expiratory Pressure (PEEP)		Fraction of Inspired Oxygen (FiO2)	
Index	Actions	Index	Actions	Index	Actions
0	0-2.5	0	0-5	0	25-30
1	2.5-5	1	5-7	1	30-35
2	5-7.5	2	7-9	2	35-40
3	7.5-10	3	9-11	3	40-45
4	10-12.5	4	11-13	4	45-50
5	12.5-15	5	13-15	5	50-55
6	>15	6	>15	6	>55

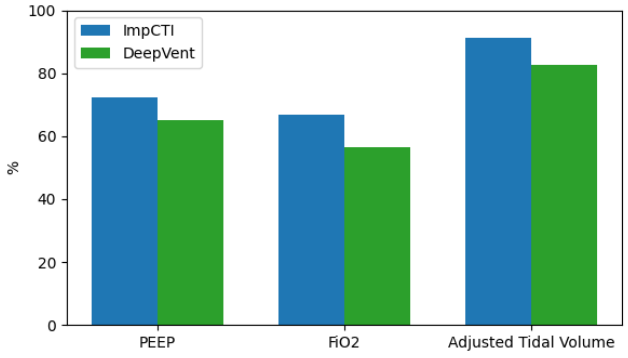
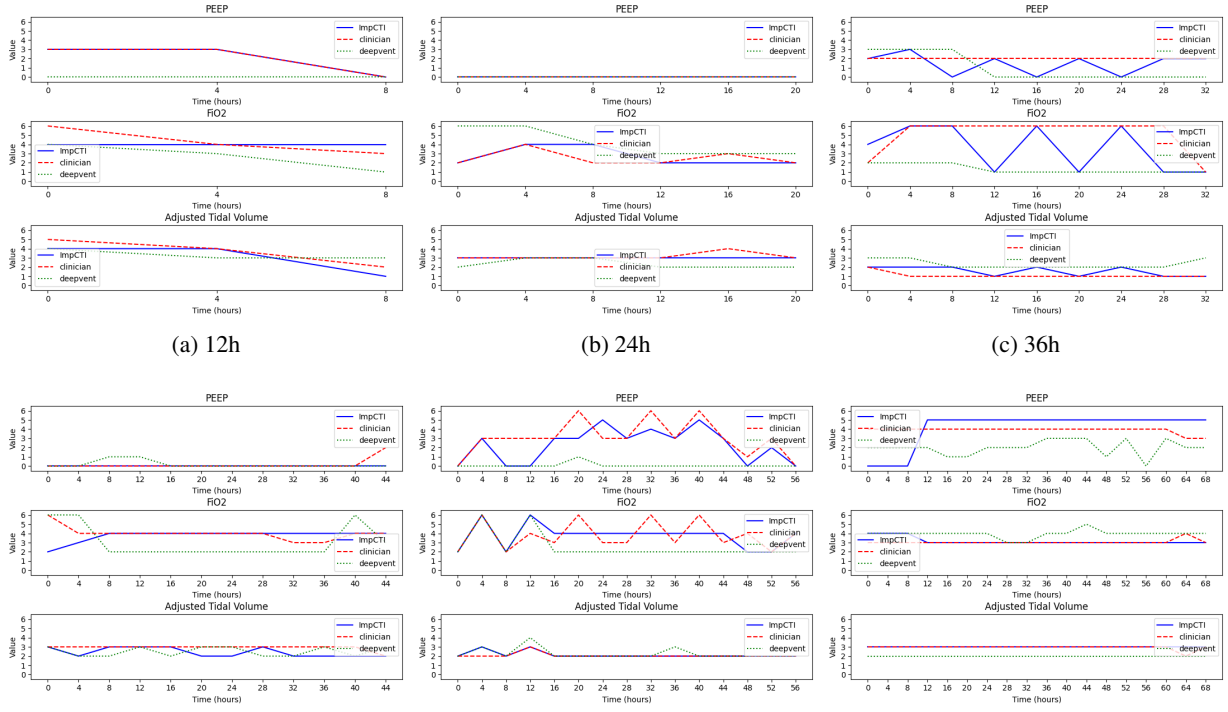


Figure 2: Similarity to Clinicians. The left blue column is the percentage of actions similarity between ImpCTI and clinicians. The right green column is the percentage of actions similarity between DeepVent and clinicians.

The validation method employed in this study involved comparing the proportion of consistent actions between the



(a) 12h (b) 24h (c) 36h

(d) 48h (e) 60h (f) 72h

Figure 3: Case Presentation. The subplots a, b, c, d, e, and f are the treatment action records of patients who have been treated for 12, 24, 36, 48, 60, and 72 hours respectively. The blue solid line is the treatment action of ImpCTI, the red broken line is the treatment action of clinicians, and the green dotted line is the treatment action of DeepVent.

model and clinicians. We presented the percentages of similarity between ImpCTI and the actions performed by clinicians, as well as between DeepVent and clinician actions.

Based on the experimental results provided, the similarity between ImpCTI and clinicians in terms of PEEP actions is 73%, and the similarity between DeepVent and clinicians is 65%. For FiO2 actions, the similarity between ImpCTI and clinicians is 67%, while the similarity between DeepVent and clinicians is 56%. In the case of Adjusted Tidal Volume action, ImpCTI demonstrates an 91% similarity to clinicians, while DeepVent shows an 82% similarity to clinicians.

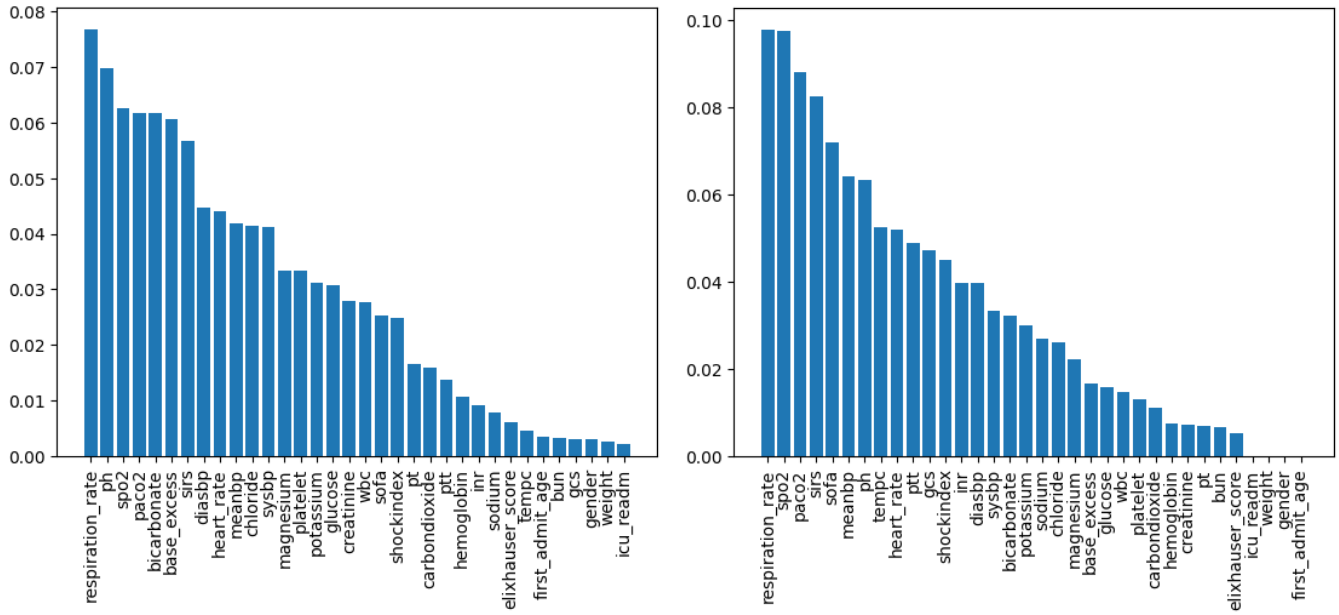
The results revealed that ImpCTI exhibited a higher degree of similarity with the actions of the clinicians compared to DeepVent. This indicates successful inference and implementation of the clinical intentions by our model, highlighting its efficacy in capturing and replicating the decision-making process of clinicians.

Case analysis

In order to demonstrate the effectiveness of the ImpCTI in approximating the intentions of clinicians, we conducted several case analyses. We divided the dataset into six categories based on the duration of treatment: 12, 24, 36, 48, 60, and 72 hours. From each category, we randomly selected a patient's treatment trajectory. We then compared the treatment strategies of ImpCTI and DeepVent with those of the attending clinicians. Our findings revealed that ImpCTI's treatment strategies aligned more closely with the clinicians' treatment intentions.

As shown in Figure 3, from a global perspective, ImpCTI's treatment actions are more consistent with clinicians, DeepVent's treatment strategies generally favored lower action values compared to those chosen by clinicians. This tendency was particularly evident in the PEEP graph (Figure 3a), the FIO2 graph (Figure 3c), the FiO2 graph (Figure 3d), and the PEEP graph (Figure 3e). Looking at the temporal dimension, in the short-term and long-term treatment scenarios depicted in Figure 3a, 3b, and 3f, ImpCTI and the attending clinicians exhibited similar and relatively stable actions. However, in the PEEP subplot of Figure 3f, DeepVent's actions displayed greater instability and deviated significantly from those of the clinicians. In terms of flexibility (Figure 3e), ImpCTI was able to adjust its actions in accordance with the attending clinicians or remain stable within the clinicians' adjustment range. On the other hand, DeepVent exhibited minimal changes and had substantial differences from the clinicians' actions. As for volatility, as observed in the FiO2 graph (Figure 3c), when the attending clinicians selected the highest-risk action, ImpCTI, prioritizing safety, intermittently chose smaller, safer action values, resulting in larger fluctuations.

Overall, these results from the second part of our experimental study support the notion that ImpCTI successfully approximates the treatment intentions of clinicians, while



(a) Covariance between indicators and rewards

(b) Covariance of changes between indicators and rewards

Figure 4: Correlation analysis

DeepVent shows some discrepancies in its treatment strategies compared to those of attending clinicians.

Correlation analysis

During the treatment process, clinicians make decisions based on medical indicators they are concerned about, which implies that the reward value reflecting their intentions should be correlated with those indicators. In this section, we used covariance to reflect the correlation between the indicators and the reward, thereby demonstrating the rationality of the reward function. Figure 4a presents the covariance between the indicators and the rewards. Figure 4b shows the covariance of changes between indicators and rewards. Both calculations involve taking the absolute value of the covariance, normalizing it, and ranking it. Figure 4a demonstrates a strong correlation between respiratory rate, SpO2, PaCO2, blood pressure, pH values, and the reward value. Similarly, Figure 4b reveals noteworthy correlations between the changes in respiratory rate, SpO2, PaCO2, blood pressure, pH values, and the associated changes in the reward value. This aligns with the focus of clinicians on blood gas analysis (Gattinoni, Pesenti, & Matthay, 2018)during treatment. Moreover, It is clearly observed that age, gender, and weight have very low correlations with reward value, which is consistent with intuition and objective facts.

This shows that InfCTI infers clinicians' intentions, and its reward function can reasonably and effectively reflect the intention through numerical values, consistent with the treatment goals of maintaining adequate respiratory support and improving patient prognosis.

Conclusion

In this paper, we have addressed the challenge of inferring clinicians' treatment intentions in medical decision-making. By leveraging treatment records, we proposed the InfCTI-ImpCTI model to infer intentions and generate treatment strategies. Our results indicate that the treatment strategies obtained align with clinicians' intentions, highlighting the effectiveness of our model. By incorporating clinicians' intent into the learning process, we can enhance the quality and efficacy of medical treatments. The finding also underscores the potential of our model as a reliable tool for assisting healthcare professionals in their decision-making processes.

By demonstrating a greater similarity with clinician actions, our model offers valuable insights into the effectiveness of its inference and implementation capabilities. This signifies a significant step forward in the development of intelligent systems that can support and enhance clinical decision-making. The alignment between our model and clinician actions also establishes a foundation for further research and development in the field of medical AI, paving the way for the integration of machine learning algorithms into clinical practice.

Overall, our experiments provide compelling evidence of the capacity of our model to infer the intentions of clinicians and implement the treatment strategy of clinicians, thereby contributing to the advancement of AI-assisted healthcare.

References

Berke, M., & Jara-Ettinger, J. (2022). Integrating experience into bayesian theory of mind. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Catmur, C. (2015). Understanding intentions from actions: Direct perception, inference, and the roles of mirror and mentalizing systems. *Consciousness and cognition*, *36*, 426–433.

Chan, A. J., & van der Schaar, M. (2021). Scalable bayesian inverse reinforcement learning. *arXiv preprint arXiv:2102.06483*.

Den Ouden, H. E., Frith, U., Frith, C., & Blakemore, S.-J. (2005). Thinking about intentions. *Neuroimage*, *28*(4), 787–796.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Forehand, M. (2000). Extending overjustification: the effect of perceived reward-giver intention on response to rewards. *The Journal of applied psychology*, *85 6*, 919-31. Retrieved from https://api.semanticscholar.org/CorpusID:44566229

Gattinoni, L., Pesenti, A., & Matthay, M. (2018). Understanding blood gas analysis. *Intensive care medicine*, *44*, 91–93.

Gazzaniga, M. S. (2004). *The cognitive neurosciences*. MIT press.

Hussein, A., Gaber, M. M., Elyan, E., & Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, *50*(2), 1–35.

Inagaki, Y., Sugie, H., Aisu, H., Ono, S., & Unemi, T. (1995). Behavior-based intention inference for intelligent robots cooperating with human. In *Proceedings of 1995 ieee international conference on fuzzy systems.* (Vol. 3, p. 1695-1700 vol.3). doi: 10.1109/FUZZY.1995.409904

Kondrup, F., Jiralerspong, T., Lau, E., de Lara, N., Shkrob, J., Tran, M. D., ... Basu, S. (2023). Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 15696–15702).

Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, *33*, 1179–1191.

Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Lv, G., Li, J., Wang, X., & Zeng, Z. (2022). Inferem: Inferring the speaker's intention for empathetic dialogue generation. *arXiv preprint arXiv:2212.06373*.

Marken, R. S. (2013). Making inferences about intention: Perceptual control theory as a "theory of mind" for psychologists. *Psychological Reports*, *113*(1), 257–274.

Rafferty, A. N., LaMar, M. M., & Griffiths, T. L. (2015). Inferring learners' knowledge from their actions. *Cognitive Science*, *39*(3), 584–618.

Raileanu, R., Denton, E. L., Szlam, A., & Fergus, R. (2018). Modeling others using oneself in multi-agent reinforcement learning. In *International conference on machine learning.* Retrieved from https://api.semanticscholar.org/CorpusID:3622509

Royka, A. L., Török, G., & Jara-Ettinger, J. (2023). Guiding inference: Signaling intentions using efficient action. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Sutton, R. S., Barto, A. G., et al. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*, *11*(1), 126–134.

Wang, L., Tang, R., He, X., & He, X. (2022). Hierarchical imitation learning via subgoal representation learning for dynamic treatment recommendation. In *Proceedings of the fifteenth acm international conference on web search and data mining* (pp. 1081–1089).

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences.

Zhang, H., Parkes, D. C., & Chen, Y. (2009). Policy teaching through reward function learning. In *Acm conference on economics and computation.* Retrieved from https://api.semanticscholar.org/CorpusID:11788078