Simulation Modeling of Clinical Decision Making for Personalized Policy Identification

Ashish T. S. Ireddy[0000−0003−2964−4371] and Sergey V. Kovalchuk[0000−0001−8828−4615]

ITMO University, Saint Petersburg, Russia
{ireddy,kovalchuk}@itmo.ru

Abstract. With Human – Artificial Intelligence (AI) collaboration booming in all fields, the pace of task-based cooperation is ever-expanding. Yet, in most applications, AI induction is sidelined to test beds and is perceived skeptically as a competitor rather than a collaborator. The healthcare domain is one field where AI support is viewed as theoretical and far from practical. While most focus is directed towards developing and training AI models, the human expert and their interactions with the AI model are often overlooked. We present an experiment that incorporates, the personalization of human experts into the AI's model training, aiming to improve collaboration and overall outcome. Using a simulation-based approach, we optimize the AI learning policy of a domain expert's behaviour when evaluating decision support data of a patient's risk of acquiring type 2 diabetes mellitus (T2DM). With Linear and Maximum Entropy inverse reinforcement learning (IRL) algorithms, we analyze various learning strategies by including context, rewards and sampling rates to show personalized expert characteristics with optimal policies and effective reward functions respectively. Our results provide insights into experts' personalized evaluation policy and the AI model's learning behaviour in various environmental scenarios and further the implicit difference in the evaluation of domain experts.

Keywords: Inverse Reinforcement Learning · Decision Support Systems · Behaviour Optimization · Policy Simulation · Diabetes Mellitus.

1 Introduction

Artificial intelligence - the world of recent times partially revolves around it. From minor hints to complex solutions, the use of AI assistance in the general domain workflow is ever-expanding. The healthcare domain is one of many fields that has seen a rise of clinical decision support systems (CDSS) incorporated into a quasi-practical state where recommendations provide valuable insight to the decision maker therefore aiding in achieving a better outcome. While most decision support systems are based on offline training using historical observations, few approaches account for the remaining aspects of decision support, i.e. interaction with experts, environment variables, information context, perceptional

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49

states and the human expert [7]. Context is one aspect that is overlooked during training, the inclusion of which can improve the outcome with perceptional states [4]. However, in instances where we have only an expert's demonstration of performing a task, IRL can be an effective approach to recover the behaviour and implicit decision-making policy of the human user [1]. Together with the inclusion of context, background and perceptional states, we can improve recommendations from CDSS to be more aligned with the domain expert's solution and scenario. Further, the type of training for the AI model (i.e. offline or online) has been given little attention when working with expert demonstrations. Our work aims to simulate the AI model's learning curve and investigate the difference in policy behaviour of doctors from various specializations via IRL in a clinical decision-making scenario. We present a simulation-based approach to model the domain expert's (medical professional) personalized policy when evaluating recommendation data of a prediction model, to assess the patient's risk of acquiring type 2 diabetes mellitus (T2DM). Medical professional's evaluation is acquired through subjective metrics (understandability, agreement and usability). We perform personalization, using Maximum entropy (MaxEnt) and Linear Inverse Reinforcement Learning (IRL) to extract the underlying reward functions and the real optimal policies. We simulate the personalization of policies on three groups of data, comprising individual doctors, by specialization and a global dataset, to show the sensitivity of respective reward functions via an entropy measure. Using context, strategy of evaluation and information levels, we scrutinize our observations to reveal behavioural patterns of medical professionals on real-world data. Our results provide a collective insight into identifying personalized policies based on expert behaviour. Further, the paper is structured as follows: Section 2 introduces the methodology of modelling CDSS for personalized policies. Section 3 shows the interpretation of policies from IRL, Section 4 investigates results after simulation. Section 5 is the conclusion.

2 Modeling Personalization in Clinical Decision Making

This section introduces our approach to simulating and modelling expert personalization using IRL algorithms. We define notations that will be used throughout the paper. A set of n (finite) **expert trajectories** $E_T = \{\tau_1; \tau_2; ....\tau_n\}$ constitute to a combination of **states** $S = \{s_1, s_2, ...s_n\}$ and **actions** $A = \{a_1, a_2, ...a_n\}$ that an agent can take in E_T where $T_{PA}(.)$ is the state **transition probabilities** of moving to state s' from s upon taking action a (i.e. $T(s, a, s')$. A **discount factor** $\gamma \in [0, 1)$ dictates the weightage for long-term-short-term reward strategy. $L1 \in [0, 1)$ is the **regularization factor**. π is the **policy** function that defines the action to be taken in each state i.e. $(\pi : S \rightarrow A)$, π^* is the **optimal policy** that defines the optimal actions to take in each state such that the generated reward is maximum. $\tau = \{(s_0, a_1, s_1); (s_1, a_2, s_2); ....(s_{n-1}, a_n, s_n)\}$ is a **trajectory** describing one complete iteration of the agent in the MDP. $R(s_n, a_n) \in R_f$: is the **reward** received for reaching state s_n by taking action a_n where R_f is the collective reward function for all policies π in trajectories in E_T. To identify

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49

personalization, we have used a combination of linear and maximum entropy algorithms to generate individual reward functions, optimal policies and analyze the impact across doctors and specializations extending our prior experiment [3]. Our linear IRL [6] is based on the approach of using RL inside IRL to iterate across all policies and identify the maximal reward using the assumed optimal policy π^* over trajectories E_T. This procedure provides us with a complete overview of all possible rewards in the state space. The maximum entropy IRL is implemented as in [8] i.e. maximizing the reward function relative to their weights $\theta^* = \underset{\theta}{argmax} \sum_{E_T} log\ P(E_T|\theta, E_T)$. Further, using the maximum entropy algorithm we recover the real optimal policy π^* therefore describing true expert behaviour. Figure 1 showcases our approach. We use the setup of the IRL algorithm by [2] and customized Python scripts to perform simulations.

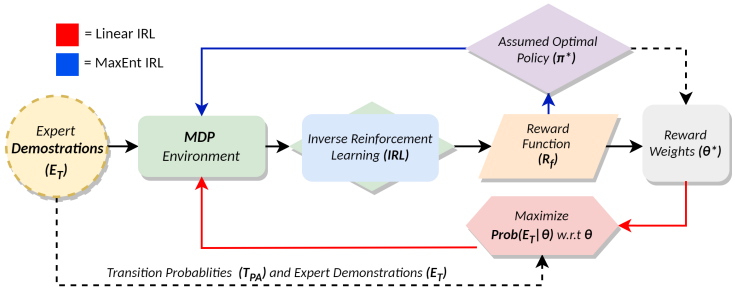

Fig. 1. Linear and MaxEnt IRL in our experiment. An MDP is setup using E_T. A tuple of trajectories is fed to IRL, generating reward function R_F. Linear IRL iterates through all policies π, while MaxEnt maximizes reward weights θ^* relative to T_{PA} and E_T

CDSS Data: The dataset used in our experiment is from an experimental survey [5] conducted at a medical research center where the authors analyze the effect of having decision makers (doctors) supported by information from a prediction model, a FINDRISK measure and case-explanation to assess the perceptional state through subjective metrics of patients suffering from T2DM. Physicians were provided with the patients' basic information (age, BMI etc), one of three prediction information and asked to assess the data via three subjective perception measures on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). The measure being *Understandability* denoting interpretation, *Agreement* of model's prediction and *Usability* of prediction data in diagnosis. A total of 541 cases of patient assessment data were found to be usable for our experiment. We include internal context (as doctors' experience, specialization), external context (as patient risk) to generate inferences.

MDP Setup and Policy Iteration: In [3], we used CDSS data to model the internal perceptional state of the medical experts using IRL. We extend the MDP of $S : 5$ **states** $= \{End,\ Understandability,\ Agreement,\ Usability,\ Completion\}$, $A : 2$ **actions** $= \{Continue,\ Terminate\}$ with T_P: Transition probabilities of moving from state s to s' extracted from E_T using $T_P(s, a, s') = \frac{\#\ of\ times\ (s \rightarrow s'\ )\ occurs\ in\ T_E}{Total\ \#\ of\ occurences\ in\ T_E}$. Where the agent is initialized at understandability state and takes actions based on the scored evaluation from the physician. A

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49

deterministic function compares the doctor's assessment to a metric threshold M_T (i.e. reflecting lenient, strict evaluation) and is used to decide the action of the agent. We thresholds of $M_T = [2, 3, 4]$ relating to intensity of assessment.

3 Analysis of Reward Functions and Maximum Entropy Policies

To derive doctor's personalized evaluation policy, we first assume that all doctors have universal behaviour and run IRL on all 541 cases of CDSS data (i.e. global demonstrations). We select a metric threshold $M_T = 2$ to ensure the inclusion of a broad range of evaluations and create a subset of CDSS data trajectories. We feed the linear IRL a Linear IRL algorithm with a tuple (S, A, T_P, $[\pi]$, R_{max}, γ, $L1$), where $[\pi]$ is a set of all policies extracted from the trajectories, $L1$ and γ were set to (0.9, 0.9) aimed at long term rewards. We obtain the reward function as shown in table 1 representing reward/penalties for reaching each state respective to the policy, here the assumed optimal policy π^* is [0,1,1,1,0] as per reward-penalty distribution. Here we encounter optimal ambiguity where the same rewards are obtained for multiple policies. Therefore, we introduce entropy to reflect the collective behaviour of individual policies. It is defined as the probability of reward-penalty received during the resolution of trajectories $H(R_f) = \sum_{R(s_n,a_n)\in R_f} P(R(s_n, a_n))\ log(P(R(s_n, a_n)))$. Next, we use maximum entropy IRL to recover the real optimal policy using the same trajectory subsets.

Table 1. The reward function of global demonstrations generated over policies π and with reward terms and respective entropy $H(R_F)$. For all other π the reward and entropy were 0

Policy	*End*	*Understand*	*Agree*	*Use*	*Complete*	*Entropy*
[0,0,1,1,0]	0.0	0	0	0	**10.00**	0.217
[0,1,1,1,0]	0.0	0	0	**10.00**	**10.00**	0.293
[0,1,1,0,0]	0.0	0	0	**10.00**	0	0.218
[0,1,0,1,0]	0.0	0	0	0	**10.00**	0.218

Assessment by Specializations: On running linear IRL and MaxEnt for the specialization subset of CDSS data, we identify collective reward functions and optimal policies shown in 2. We analyse the variance of policy-reward space and compare it with MaxEnt policies. To ensure an unbiased analysis, we selected specializations where more than 3 individual doctors' evaluations were available i.e. endocrinologists, cardiologists and general medicine [E, C, GM] while sparsely involving other doctors since they had less than 2 samples or their behaviour was erroneous. Table 3 shows the reward function for the three specializations [E, C, GM] at $M_T = [2, 3]$. On analysing the MaxEnt policies and respective R_F, we observe the π^* for [E, C, GM] doctors to be relatively the same across M_T with the entropy of endocrinologists being consistent compared to all specializations. The average score of evaluation metrics for [E, C, GM] were (4.30, 3.72, 3.75), (4.9, 4.2, 4.7) and (4.27, 4.30, 4.0) respectively. This

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49

shows the volatility in the optimal policy followed and given the high metric scores, the IRL assumes that the amount of information does not impact the evaluation thus moderately saturating the reward function. The infectionsists and behaviour follow closely with endocrinologists. Gynaecologists have a peculiar deviation of policy at $M_T = 2$ due to a higher number of evaluations at low M_T. For neurologists and rheumatologists, the R_F remained the same throughout all M_T since they had only trajectory. Overall, at a higher M_T, the optimal policy for all specializations follows similar behaviour, and at MT = 2,3, the policy reveals individual assessment tendencies (i.e. strict/lenient) which is reflected in the average evaluation scores. At $M_T = 4$ the behaviour converges as in table 1 with rewards awarded for reaching usability states across specializations.

Table 2. The maximum entropy optimal policies π^* for all doctors per specialization representing their personalized decision policy at various levels of assessment M_T

Specialization	***No of Docs***	**MT = 2**	**MT = 3**	**MT = 4**
Endocrinologist	5	[0, 0, 0, 0, 0]	[0, 0, 0, 1, 0]	[0, 0, 1, 1, 0]
Cardiologist	3	[0, 0, 0, 0, 0]		[0, 0, 1, 1, 0]
General Medicine	4	[0, 0, 0, 0, 0]	[0, 0, 0, 1, 0]	[0, 0, 1, 1, 0]
Gynaecologist	2	[0, 1, 1, 0, 0]	[0, 0, 0, 1, 0]	[0, 0, 1, 1, 0]
Ophthalmologist	1	[0, 0, 0, 0, 0]	[0, 0, 0, 1, 0]	[0, 1, 1, 1, 0]
Infection Specialist	2	[0, 1, 0, 0, 0]	[0, 1, 1, 0, 0]	[0, 1, 1, 1, 0]
Rheumatologist	1		[0, 1, 1, 1, 0]	
Neurologist	1		[0, 1, 1, 1, 0]	

4 Simulating Personalized Learning via Policy Iteration

The simulation setup for personalized learning behaviour uses the same MDP and IRL as in section 3. To simulate behaviour, we extend the IRL cycle to randomly resample the trajectories from the subsets of specialization and individual doctors to train the model over multiple steps. To investigate optimal policies we select subsets of three specializations as per CDSS data, i.e. 5 endocrinologists, 4 general medicine specialists and 3 cardiologists since their sample trajectories are large enough for simulating individual and groupwise behaviour. For uniformity in experiments, we selected our IRL parameters to be focused on having a long-term reward strategy of $(L1 : 0.9, \gamma : 0.9)$ and $M_T = 2$ to cover all possible scenarios. Figure 2 gives an overview of our simulation in three cases.

Case 1: Simulation using Individual Demonstrations The linear IRL algorithm is initialized with a doctor's full trajectory set and trained. On concluding one iteration of training, we randomly select a single sample from the original set of doctors' trajectories and append it to the training set. The process is continued for N Sample additions and the entropy scores for all sample addition steps up to are averaged over M cycles. To make the computing process more efficient, we selected two combinations of sample addition and randomization (M, N) i.e. (25, 600) and (100, 100) since at smaller intervals, the behaviour was incomplete and larger intervals tended to have constant variance while also taking immense computing power and time. We evaluate the training by assessing the reward entropy $H(R_f)$ across N and M randomization cycles.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49

Table 3. The reward functions and entropy of Endocrinologists (E), Cardiologists (C) and General medicine therapists (GM) at $M_T = [2, 3]$ when running IRL on subsets of specialization. The rewards are constant for states of ***End***, ***Understandability*** (Und) and ***Agreement*** (Ag) while varying for ***Usability*** (Use) and ***Complete***

Policy	*End*	*Und*	*Agree*	*Use*		*Complete*		Entropy	
Specializations	[E, C, GM]			[E]	[C, GM]	[E]	[C, GM]	[E]	[C, GM]
[0,0,0,0,0]	0	0	-10	0	-10	0	-10	0.217	0.292
[0,0,0,1,0]	0	0	-10	0	-10	0	10	0.217	0.458
[0,0,1,1,0]	0	0	-10	0	10	10	10	0.412	0.458
[0,1,1,1,0]	0	0	10	10	10	10	10	0.292	0.292
[0,1,1,0,0]	0	0	10	10	10	0	0	0.292	0.292
[0,1,0,0,0]	0	0	10	0	0	0	0	0.217	0.217
[0,0,1,0,0]	0	0	-10	0	10	0	0	0.217	0.412
[0,1,0,1,0]	0	0	10	0	0	10	10	0.292	0.218

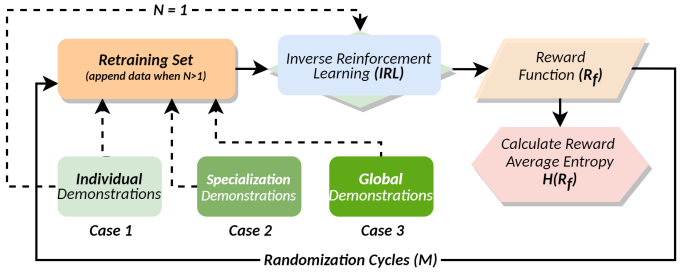

Fig. 2. Our simulation setup to learn personalized policies of doctors in various training scenarios. Given a single doctors demonstrations at $N = 1$ we run IRL on the current set of trajectories to obtain R_F and its entropy. The initial set of trajectories is appended with one sample from; Case 1: doctor undergoing personalization; Case 2: doctor's specialization; Case 3: global demonstrations; This cycle is continued for M steps.

Across all specializations, we observe spikes in the entropy with the addition of new trajectories reflecting new unlearned behaviour. After cycles of random addition, the entropy gradually saturates. Figure 3A shows the learning behaviour of cardiologists (A,B,C) trained with 600 iterations and 25 randomization cycles. On comparing with respective MaxEnt policies and R_F, we observe that the physicians have a strict policy of evaluation at lower M_T with penalties. Whereas for general medicine therapists trained with 100 iterations of sampling and 100 randomization cycles, (Figure 3B), the reward entropy does not stabilize for specialists (A,C) until 100 samples. We attribute this behaviour to higher metric scores. Specialist (D) has a constant entropy due to erroneous evaluation scores. We observe that an increase in sample additions (N) elongates the learning behaviour with sharp deviations in R_F, while an increase in randomization (M) results in smooth learning behaviour with fewer steps to adopt a policy.

Case 2: Resampling using Specialization Trajectories: Here, we initialize the MDP and IRL algorithm as in the previous case, however, our resampling data is sourced from the data of specializations (i.e. grouped data of doctors from the same specialization except the doctor being assessed). We consider the case of endocrinologists, where we feed specialization data to all doctors

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49

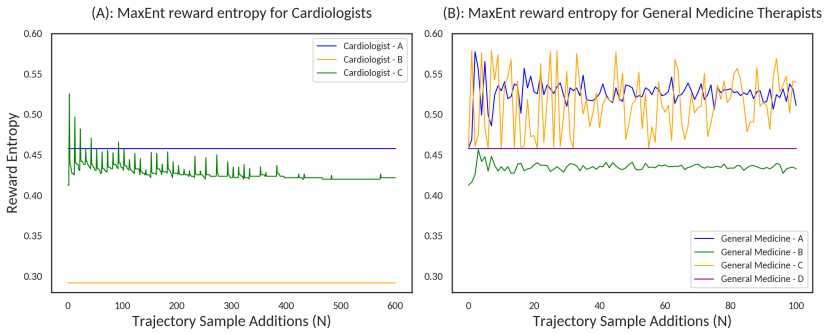

Fig. 3. The personalized learning curve of reward entropy at Maximum Entropy policy [0, 0, 1, 1, 0] for; (A): cardiologists after training with (25,600) randomized -sampling iterations; (B): Personalized learning curve for general medicine specialists after training with (100, 100) randomized-sampling iterations

individually. Figure 4A shows the results of the endocrinologists. Unlike in case 1, the reward entropy takes comparatively longer to converge and stabilize while there are sections of constant trajectory in between. This behaviour is the impact of deviating policies given the breadth of added samples (i.e. trajectories of four endocrinologists were added during training) therefore requiring more cycles of randomization to stabilize.

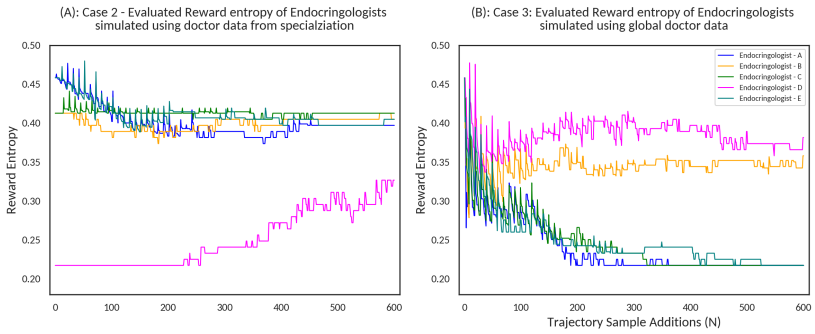

Fig. 4. Learning curve of endocrinologist's reward entropy of MaxEnt policy [0,0,1,1,0] when simulated at (25,600) cycles of randomization; (A): Sampling data from endocrinologists specialization; (B): Sampling data from global doctors

Case 3: Resampling using Global Trajectories: The final case is aimed at providing a broad perspective of the learning behaviour as the trajectories are added from global demonstrations. The setup of MDP and IRL follows case 1. We experiment with endocrinologists again as they provide a wider base for analysis. At 25 randomization and 600 iterations, we observe the behaviour as in Figure 4B. Compared to case 1 and case 2, we notice the magnitude of reward entropy is much lower than the latter, when analyzed, the reward functions are observed to vary frequently with changes in penalty and reward terms across all policies. The variance in behaviour does not subside despite 600 additions

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49

and 25 randomization cycles, we believe this is attributed to the nature of the dataset since it consists of 541 samples consisting of doctor trajectories of all specializations, the ideal number of iterations required to stabilize the reward function for collective policies should be much larger. Hence, the number of trajectories required to reach a constant policy is much larger.

5 Conclusion and Future work

Overall, the results of our experiment provide insights into modelling and simulating clinical decision-making data using inverse reinforcement learning to identify personalized policies of doctors when evaluating prediction data of patients' risk of type 2 diabetes mellitus. We perform simulations to evaluate the personalized learning of doctors' individual policies in three cases of trajectory sampling. From our investigation, we observed an increase in evaluated entropy when trained by experts of different specializations whereas when trained using one's own data, the identification of policy is faster (i.e. less than 600 steps of sampling). We see these results as a crucial step towards the development of personalization in DSSs, the inclusion of which can improve the AI model's alignment with domain experts enabling a more efficient collaboration. In future works, we plan to extend AI model simulation with online and offline training constructed around theory of mind and feedback learning to improve human-AI collaboration in universal domains.

Acknowledgments. The research was supported by The Russian Science Foundation, agreement №24-11-00272, https://rscf.ru/project/24-11-00272/.

References

1. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the 21st international conference on Machine learning. p. 1 (2004)
2. Alger, M.: Inverse reinforcement learning (2017). `https://doi.org/10.5281/zenodo.555999`, `https://doi.org/10.5281/zenodo.555999`
3. Ireddy, A.T., Kovalchuk, S.V.: Modelling information perceiving within clinical decision support using inverse reinforcement learning. In: International Conference on Computational Science. pp. 210–223. Springer (2024)
4. Kovalchuk, S., Ireddy, A.T.S.: Prediction of users perceptional state for human-centric decision support systems in complex domains through implicit cognitive state modeling. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 46 (2024)
5. Kovalchuk, S.V., Kopanitsa, G.D., Derevitskii, I.V., Matveev, G.A., Savitskaya, D.A.: Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. Journal of Biomedical Informatics **127**, 104013 (2022)
6. Ng, A.Y., Russell, S., et al.: Algorithms for inverse reinforcement learning. In: Icml. vol. 1, p. 2 (2000)
7. Schmidt, P., Biessmann, F.: Calibrating human-ai collaboration: impact of risk, ambiguity and transparency on algorithmic bias. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 431–449. Springer (2020)
8. Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K., et al.: Maximum entropy inverse reinforcement learning. In: Aaai. vol. 8, pp. 1433–1438. Chicago, IL, USA (2008)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_49