Chinese Medicine

RESEARCH

Open Access

PrescDRL: deep reinforcement learning for herbal prescription planning in treatment of chronic diseases

Kuo Yang[1], Zecong Yu[1†], Xin Su[1], Fengjin Zhang[2], Xiong He[1], Ning Wang[1], Qiguang Zheng[1], Feidie Yu[1], Tiancai Wen[3] and Xuezhong Zhou[1*]

Abstract

Treatment planning for chronic diseases is a critical task in medical artificial intelligence, particularly in traditional Chinese medicine (TCM). However, generating optimized sequential treatment strategies for patients with chronic diseases in different clinical encounters remains a challenging issue that requires further exploration. In this study, we proposed a TCM herbal prescription planning framework based on deep reinforcement learning for chronic disease treatment (PrescDRL). PrescDRL is a sequential herbal prescription optimization model that focuses on long-term effectiveness rather than achieving maximum reward at every step, thereby ensuring better patient outcomes. We constructed a high-quality benchmark dataset for sequential diagnosis and treatment of diabetes and evaluated PrescDRL against this benchmark. Our results showed that PrescDRL achieved a higher curative effect, with the single-step reward improving by 117% and 153% compared to doctors. Furthermore, PrescDRL outperformed the benchmark in prescription prediction, with precision improving by 40.5% and recall improving by 63%. Overall, our study demonstrates the potential of using artificial intelligence to improve clinical intelligent diagnosis and treatment in TCM.

Keywords Deep reinforcement learning, Traditional Chinese medicine, Herbal prescription planning, Chronic disease, Artificial intelligence

Introduction

Intelligent diagnosis and automatic drug recommendation have become important topics in medical artificial intelligence[1]. The optimization problem of dynamic diagnosis and treatment scheme (DDTS) considers a patient's treatment as a sequential decision-making process[2], aiming to identify the best sequential treatment schema[3]. In the field of Traditional Chinese Medicine (TCM)[4], DDTS optimization typically requires consideration of a patient's status (e.g., symptoms and signs) at each stage, and generates an herbal prescription treatment plan (HPTP). TCM doctors obtain a patient's symptom descriptions and corresponding syndromes through the "Four Examinations" method of "watching, listening, asking, and feeling" [5]. Unlike prescription recommendation, which predicts the appropriate prescription based on a patient's current situation, DDTS optimization focuses on providing the best treatment at each stage to maximize the treatment effect of the entire sequence and

†Kuo Yang and Zecong Yu are contributed equally to this work

*Correspondence:
Xuezhong Zhou
xzzhou@bjtu.edu.cn
[1] Department of Artificial Intelligence, Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer Science & Technology, Beijing Jiaotong University, Beijing 100044, China
[2] Department of Nephrology, the Third Hospital of Hebei Medical University, Shijiazhuang 050051, China
[3] Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

identify the best sequential decision-making path. DDTS optimization prioritizes the outcome of a sequential treatment process rather than the outcome of a particular treatment.

With the explosive development of deep learning technology, it has gradually been applied to a variety of biomedical problems, such as disease gene prediction[6, 7], drug target prediction[8], drug repositioning[9]. Since the appearance of AlphaGO[10] in 2015, deep reinforcement learning (DRL) has emerged as a research hotspot in medical artificial intelligence, combining the depth perception of deep learning[11] with the decision-making of reinforcement learning (RL) to achieve optimal decision-making control[12]. Many excellent diagnosis and treatment planning models based on RL have been proposed by researchers in recent years[13–21]. For example, Shamim et al. proposed a circular decision-making framework based on RL, which provides personalized dose schemes for patients[13]. Liu et al. constructed an RL model for the prevention and treatment of graft-versus-host disease in leukemia patients[22]. Wang et al. proposed a supervised RL model based on cyclic neural networks to recommend dynamic diagnosis and treatment schemes[23]. In the field of Traditional Chinese Medicine (TCM), Feng proposed the use of a partially observable Markov decision process (POMDP) model to mine the optimal DDTS[24]. Hu proposed a deep RL algorithm framework for optimizing the sequential diagnosis and treatment scheme of TCM[25].

With the intricate mechanisms of herb combinations in prescriptions, combined diseases in patients, and individual differences among patients, designing an appropriate DDTS optimization model remains a challenge[26]. Current RL-based DDTS optimization algorithms, on one hand, do not effectively learn from the medication rules of experienced TCM doctors and fail to achieve satisfactory results. On the other hand, they do not fully represent the patient's state space and action space. Consequently, there is a pressing need for more accurate and dependable models to enhance the practicality of auxiliary diagnosis and treatment and to recommend more reliable HPTP for patients.

With the availability of large-scale real-world clinical data[27] and advancements in artificial intelligence[28], it is now possible to construct robust computational models for recommending appropriate prescriptions[29]. Three main categories of prescription recommendation methods have emerged, including traditional machine learning-based[30, 31], topic model-based[32, 33], and deep learning-based methods[34–36].For instance, Li et al.[36] proposed an improved seq2seq model to generate herbal prescriptions, while Yu et al.[37] developed a model based on CNN and topic model to predict TCM prescriptions. Liao et al.[38] proposed a CNN-based model that extracts facial image features and maps the relationship between facial features and drugs to predict herbal prescriptions. Zhou et al.[39] proposed an effective formula recommendation framework called FordNet, which integrates macro and micro information using deep neural networks. Dong et al.[40] proposed a sub-network-based symptom term mapping method (SSTM), and constructed a SSTM-based TCM prescription recommendation method TCMPR. Recently, Dong et al. [41] proposed a novel herbal prescription recommendation algorithm for real-world patients with integration of syndrome differentiation and treatment planning, which effectively integrated the embedding vectors of the knowledge graph for progressive recommendation tasks. Wang et al.[42] proposed the feature fusion and bipartite decision networks to leverage external knowledge and improve medication recommendation accuracy and drug-drug interaction rate. Tan et al.[43] proposed a logically-pretrained and model-agnostic medical ontology encoders for medication recommendation that addressed data sparsity problem with medical ontologies. Mi et al. [44] proposed an attention-guided collaborative decision network for medication recommendation, which effectively captured patient health conditions and medication records, utilizing the similarity between medication records and medicine representation to facilitate the recommendation process. Zheng et al.[45] proposed a novel end-to-end drug package generation framework, which developed a new generative model for drug package recommendation that considered the interaction effects between drugs that are affected by patient conditions. Despite the growing number of studies on herbal prescription recommendation, a significant challenge remains in bridging the gap between treatment planning based on reinforcement learning and recommending specific prescriptions.

In our study, we present PrescDRL, a novel model for optimizing diagnosis and treatment schemes using deep reinforcement learning (Fig. 1A-1E). Initially, we constructed a high-quality benchmark dataset for sequential diagnosis and treatment of diabetes, and subsequently designed the PrescDRL framework for herbal prescription treatment planning. Unlike traditional reward-driven approaches, PrescDRL focuses on long-term effectiveness to ensure better outcomes for patients. We formulated the optimization of Diagnosis and Treatment Treatment Scheme (DDTS) as a reinforcement learning task, with patient symptom observations as inputs and High-Performance Treatment Plan (HPTP) as the optimization goal. We then employed a multi-layer neural network to predict TCM prescriptions using patient symptoms and recommended HPTP as inputs. Finally,

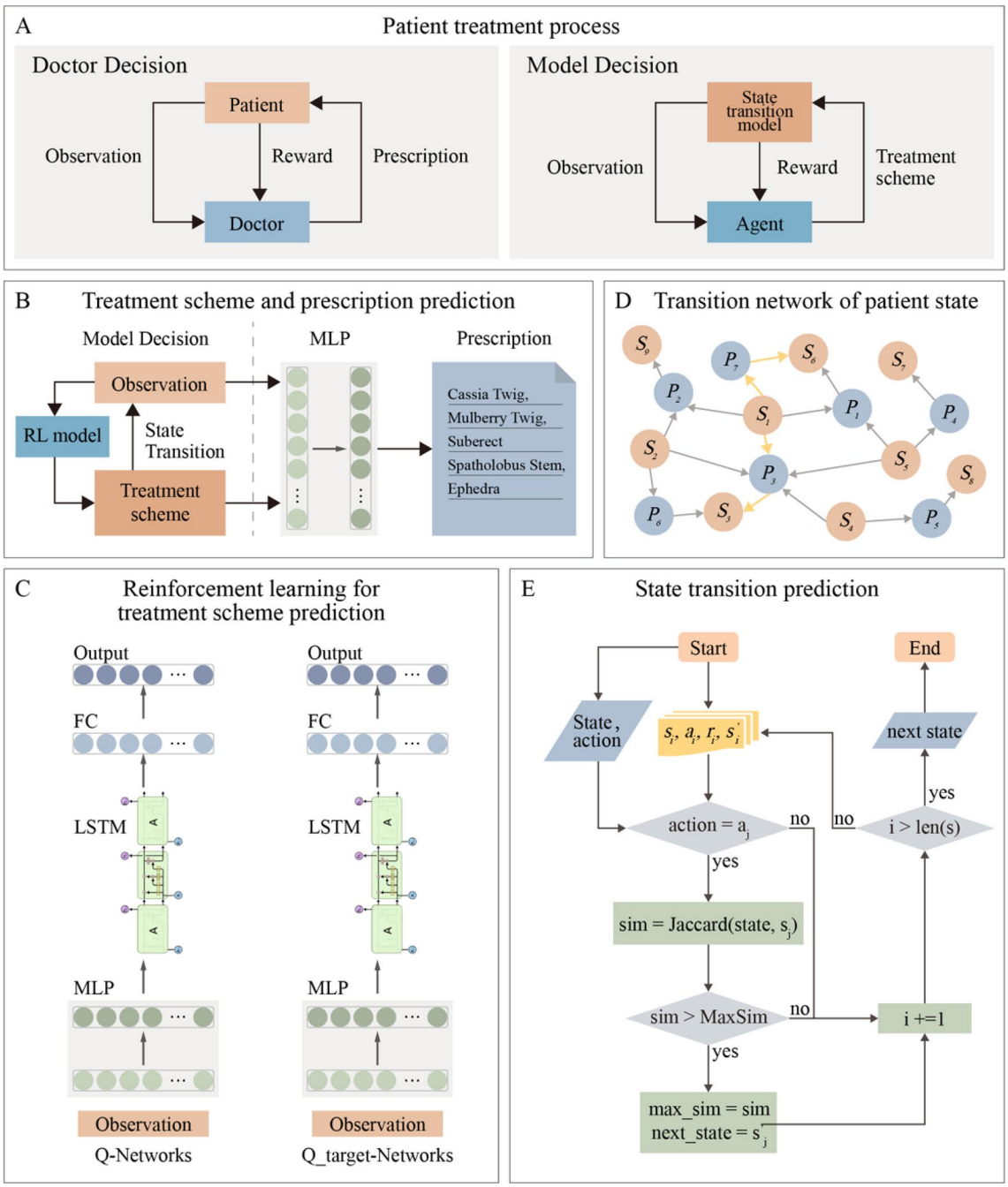

Fig. 1 Overall framework of PrescDRL. **A** Diagnose and treatment process of doctors and intelligence decision model. **B** The macro framework of PrescDRL. In this sub-figure, MLP represents multi-layer perception. **C** Reinforcement learning based prediction module of diagnose and treatment scheme. In this sub-figure, FC denotes a fully connect neural layer, and LSTM denotes a neural network of long short-term memory. **D** Transition network of patient states. In this sub-figure, *S* and *P* denotes the states and prescriptions of patient, respectively. **E** The prediction module of state transition. In this sub-figure, *i* and *j* denote $i-th$ and $j-th$ iterations of the deep reinforcement learning model. *s*, *a* and *r* denote the state of patient, the action and reward of the model, respectively

the recommended HPTP and herbal prescriptions are proposed to patients as a treatment scheme. Moreover, PrescDRL includes a prediction module for TCM prescription based on patient symptoms and HPTP. Our comprehensive experiments demonstrate that PrescDRL outperforms doctors in providing HPTP with better

expected effectiveness and has a higher prediction performance for TCM prescription.

Materials and methods

Clinical sequential data of diabetes

In this section, we present a benchmark dataset of clinical sequential diagnosis and treatment for diabetes, which serves as an example to train the optimization method of DDTS. (Ethics approval of this study has been obtained from ethics committee of institute of Clinical Basic Medicine of Traditional Chinese Medicine (NO. 2016NO.11-01)). In this dataset, the symptom observations of patients are selected as the states, and herbal prescription prescribed by doctors as actions in reinforcement learning (RL) model.

To construct a standard dataset for sequential decision-making in TCM, we first extracted 10,666 medical records of 2,895 diabetic patients from Guang'anmen Hospital. As depicted in Fig. 2A, 49.6% of the patients had only one medical record and each patient had an average of 3.68 medical records. For each medical record, we extracted the patient's symptoms and an herbal prescription consisting of multiple herbs for treatment. With the exception of 334 medical records with over 40 symptoms, the number of symptoms per patient was normally distributed (Fig. 2B), with an average of 10.386 symptoms per medical record. Similarly, the number of herbs per prescription was normally distributed (Fig. 2C), with an average of 10.059 herbs per prescription. We screened 1459 patients with more than one medical visit and obtained 5,638 medical records, which were arranged into diagnosis and treatment sequences based on clinic time.

A deep reinforcement learning framework to optimize herbal prescription treatment planning

The optimization of DDTS is essentially a Markov Decision Process (MDP, [46]). To tackle this problem, we propose an optimization model for herbal prescription treatment planning based on two high-performance deep RL models, namely, DRN [47] and DRQN [48]. DQN is a combination of Q-learning and convolutional neural network that can perform RL tasks. On the other hand, DRQN first extracts features using two fully connected layers, followed by a LSTM layer, and then predicts the action value using a final fully connected layer (Fig. 1C).

In the RL framework, the agent acts as a virtual intelligent doctor, with the patient's state serving as the environment and prescribing herbal medication to the patient as the agent's action. The key components of RL models are defined as follows: 1) The state space is denoted by S, where a state $s \in S$ represents the observation of a patient's symptoms; 2) The action space is denoted by A, where an action $a \in A$ represents the herbal medication prescribed to the patient; 3) The reward function is denoted by $R(s, a)$, which returns a reward after the

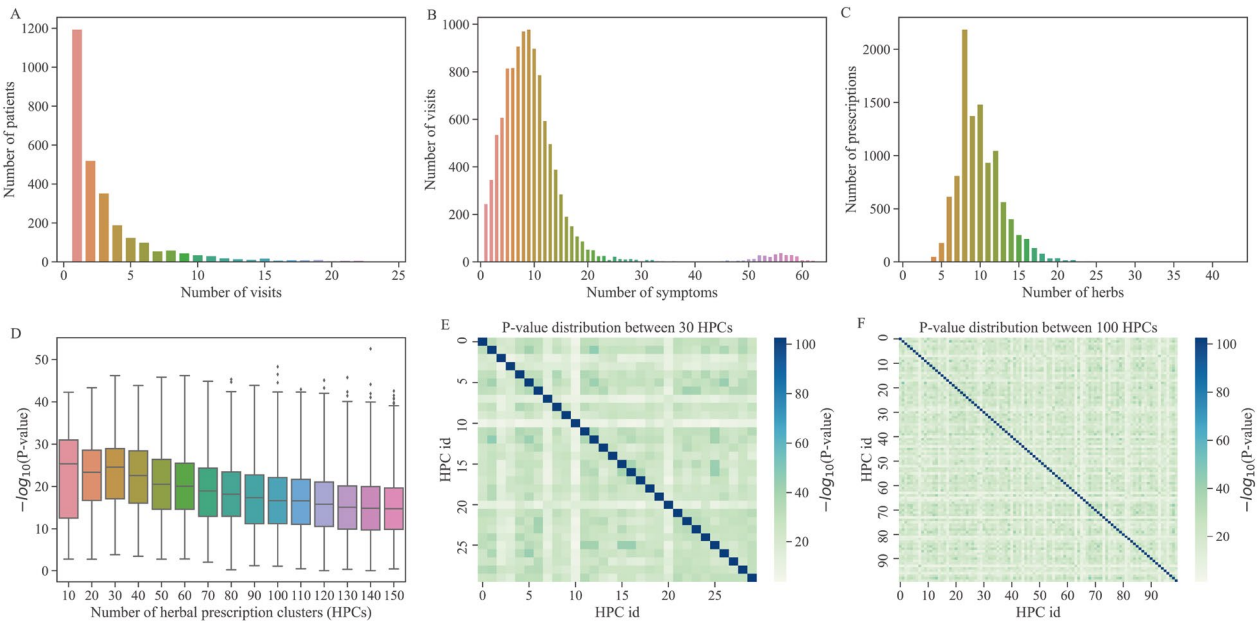

Fig. 2 Distribution of sequential diagnosis and treatment data. **A** Distribution of the number of patient visits. **B** Distribution of the number of patients' symptoms. **C** Distribution of the number of the herbs in prescriptions. **D** P-values distribution of different number of herbal prescriptions clusters. **E** P-value distribution of 30 prescription clusters. **F** P-value distribution of 100 prescription clusters

agent takes action in state s; 4) The virtual environment is denoted by E, which is an offline virtual environment based on sequential clinical data; 5) The state transition is denoted by T, where each transition is obtained using a prediction strategy.

The state observations of patients

In the DDTS optimization problem, the patient's state is a key component of the reinforcement learning model. In TCM clinics, doctors obtain symptom descriptions of patients through "seeing, hearing, asking, and cutting" "辨证论治", summarize the syndrome type, and prescribe appropriate treatments. However, since the true state of the patient is not available, even experienced doctors cannot fully determine the specific conditions of patients. Therefore, the patient's symptoms observed by the doctors are used to approximate the patient's state.

In the diabetes dataset, the distribution of patient symptoms (Fig. 2B) shows that the number of symptoms varies among patients (average of 10 symptoms per patient). The core symptoms for each disease typically differ, and different symptoms may have varying importance. However, it is challenging to obtain a precise symptom grading for diabetes, and thus different symptoms are typically considered to have equal weight. As a result, a patient's state is represented by a symptom vector, where the symptoms present in the patient are marked as 1 and those that are not are marked as 0.

The action spaces of virtual doctor

In TCM diagnosis and treatment, doctors prescribe herbal prescriptions based on the patient's symptoms. From all the medical records, we obtained 9,695 distinct herbal prescriptions. Considering all these prescriptions as actions of the RL model would greatly increase the difficulty of training and convergence of RL algorithms due to the large number of actions. Therefore, it is necessary to reduce the number of actions by converting prescription numbers into a suitable discrete space. This will reduce the model complexity and improve the convergence speed.

To reduce the number of prescriptions, we employed the K-means clustering algorithm[46] to cluster these prescriptions and used prescription's herb information as the feature. We performed a parameter tuning experiment to obtain a proper number of herbal prescription clusters (HPC) which is considered a hyperparameter. We tested different values of HPC ranging from 30 to 150 with increments of every 10 categories. A good HPC result is expected to have different categories with significantly different herbs. To achieve this, we used the Chi-square test[49] to calculate the statistical difference between any two clusters based on the composition of herbs prescribed in different clusters. The resulting HPC is used as the action of the RL models.

Design of reward function

The aim of RL-based DDTS optimization is to use a vast number of medical records to predict the optimal sequence of herbal prescriptions for patients. The objective is not only to maximize the treatment effect but also to ensure that the predicted prescriptions are reasonable. This means that the efficacy of the predicted prescriptions should be within a reasonable range, and they should not have side effects on patients or contradict drug indications.

Due to the absence of curative effect evaluation data in the diabetes dataset, we utilized the change in symptom scores between two consecutive visits before and after treatment as the immediate reward value for the current patient action. The symptom score is used to evaluate the severity of the patient's disease state and is supposed to be the weighted sum of all the patient's symptoms (the weight indicates the importance of the symptoms). However, the weight of symptoms is difficult to define, so we set the weight of all symptoms to 1, then the symptom score is simply defined as the number of symptoms of the patient. For example, the patient has 5 symptoms, then the symptom score is 5. Additionally, we calculated the Jaccard coefficient to measure the similarity between the predicted action and the actual prescription provided by the doctor. A higher reward value was assigned to actions that had a higher similarity to the doctor's prescription. Therefore, the reward function was formulated as follows:

$$\mathcal{R}(s,a) = \gamma \sum_{i=1}^{n} \alpha_i \left(s_i - s_i'\right) + \beta Jac\left(a, a'\right) \tag{1}$$

$$\mathrm{Jac}\left(a, a'\right) = \frac{\left|a \cap a'\right|}{\left|a \cup a'\right|} \tag{2}$$

where α_i represents the weight of patient's $i-th$ symptom, s_i represents the $i-th$ symptom of the patient at the current visit, and s_i' represents the $i-th$ symptom of the patient at the next visit. γ denotes the weight of therapeutic effect of patients, β denotes the weight of risk, a denotes the prescription given by the doctor, and a' denotes the prescriptions predicted by the model.

Virtual environment construction

To overcome the impossibility of training the proposed DDTS optimization model in the real diagnosis and treatment process, we developed an off-line virtual environment based on the available medical records of patients.

We constructed a tetrad, represented as (s_1, a, r, s_2), using the symptom observation and prescriptions of each patient in the current and next diagnosis and treatment. In this tetrad, s_1 denotes the current symptom observation of the patient, denotes the action based on s_1, r denotes the reward received after performing the action a, and s_2 denotes the new symptom observation of the patient after the action a. We obtained 4,179 tetrads from the medical records, which served as a virtual environment to train the deep RL model.

State transition prediction and termination

In the optimization of DDTS with the deep RL model, one of the main challenges is obtaining the next symptom observations after conducting an action based on the current symptoms due to the lack of tetrads constructed in the training stage. To address this issue, we utilized the state transition network, which includes states and actions (Fig. 1D), to predict the patient's symptoms after treatment. Specifically, we developed a prediction strategy that involves screening out all tetrads (s_1, a, r, s_2) with the same predicted action in the training set, calculating the Jaccard similarity between s_1 the symptom observations in each tetrad and the current symptoms, and selecting the s_1 tetrad with the highest similarity to the current symptoms. Finally, s_2 in the same tetrad as s_1 is selected as the state of the patient after treatment.

The distribution of symptoms in patients (Fig. 2B) indicates that 94.7% of patients have between 1 and 20 symptoms. Based on TCM expert recommendations and the symptom distribution, we define the first sequence termination condition (STC) as a patient's symptom score $\leq$3. According to the evaluation criteria of diabetes treatment effect, a 30% reduction in symptom score is considered effective, while a 70% reduction is considered markedly effective. Therefore, the second STC is defined as a 60% reduction in the patient's symptom score. The distribution of consultations (Fig. 2A) shows that 93% of patients have between 1 and 10 consultations (average number is 3.7). The last STC is number of iterations bigger than 15.

A multi-layer neural network for herbal prescription recommendation

In clinical practice of TCM, the ultimate goal of intelligent decision-making for diagnosis and treatment is to recommend effective herbal prescriptions to patients. By utilizing the trained deep RL models, we can obtain the sequential HPC for patients. In order to predict appropriate prescriptions, we model the prescription recommendation as a task of multi-label prediction. To achieve this, we constructed a multi-layer neural network (i.e., multi-layer perception), which takes the patient's symptoms and the HPC predicted by the RL models as input features, and outputs the predicted herbal prescription (Fig. 1B).

Experimental design

Parameter setting

In the DDTS optimization experiment, we used a total of 1,495 patient samples, of which 80% (1,203 samples) were used for training, and the remaining 20% (i.e., 292 samples) were used for testing. Similarly, there are also 80% samples for training and 20% for testing in the experiment of prescription recommendation.

In our proposed PrescDRL, the DQN network framework consists of three fully connected (FC) layers with 400, 300, and 30 neurons, respectively. For the DRQN network, the first two layers are FC layers with 300 and 512 neurons. The middle layer is an LSTM layer with 512 neurons, and the final layer is a FC layer with 30 neurons. Since there are 30 well-tuned HPCs, which correspond to 30 actions in modeling RL models, the DQN and DRQN layers have 30 neurons. During the training of these two models, the learning rate is 0.01, the discount coefficient of the reward value is 0.9, the random exploration probability is 0.1, and the batch size is 32. The parameters are copied to the Q-target network every 100 training batches.

Evaluation metrics

In clinical practice, evaluating the effectiveness of TCM treatment for chronic diseases, such as diabetes, can be challenging due to the long duration of treatment and the unsuitability of western medicine's clinical mortality as an evaluation metric[50]. In this study, we evaluated the performance of DDTS optimization results based on the improvement of symptom score, which is represented as the return values of RL models. To assess the effectiveness of the optimization models, we considered three commonly used metrics: single-step return (SSR), single-step cumulative return (SCR), and multi-step cumulative return (MCR). For SSR, the optimization models are trained based on the symptom observations of the first visit of each patient, and the differential between the symptom observations before and after the models provide an HPC is defined as SSR. In contrast, SCR considers all visits of each patient, and the average of all returns is computed. MCR is a more comprehensive metric, where the models predict an HPC based on the first visit of each patient and then use the state transition function to generate follow-up

symptom observations until a set stopping condition is reached.

In addition, predicting prescriptions is considered a multi-label classification problem, for which precision, recall, F1 score, and IoU are used as evaluation metrics.

Results

Clustering and validation of diagnosis and treatment plan of diabetes

Based on the PMET information of the prescriptions used by patients, we utilized the K-means clustering algorithm [46] to obtain prescription clusters. To determine the optimal number of HPCs, we conducted comprehensive experiments for parameter tuning and calculated the distribution of statistical difference (i.e., the negative logarithm of P-value) for each clustering result. A higher difference between categories implies better HPC results and indicates that the HPC obtained by clustering is more personalized. Analytical results showed that as the number of categories increases, the statistical difference of clustering results decreases (Fig. 2D). We used a heatmap to illustrate the $-\log_{10} P$ distributions of the results of 10 and 150 categories, respectively (Fig. 2E and Fig. 2F). The statistical results indicated that the result of 30 categories had the highest statistical difference $\left(-\log_{10} P = 23.27 \pm 8.27\right)$. We ultimately chose to cluster herbal prescriptions into 30 clusters as the action space of the deep RL model.

Prescription treatment planning optimization of PrescDRL

In this study, we proposed a deep RL-based method for predicting DDTS. To evaluate the performance of our proposed PrescDRL in predicting sequential HPCs, we compared the multiplex return values of the HPCs predicted by our model with those given by clinical doctors. In this experiment, we considered the results of 30 HPCs as the action space of PrescDRL. The comparison results (Fig. 3A-3C and Table 1) reveal that PrescDRL obtains higher return values than clinical doctors on three evaluation metrics.

In terms of the single-step return (SSR), the clinical doctors achieved a score of 1.39, while $PrescDRL_{DQN}$ and $PrescDRL_{DRQN}$ improved by 117% and 153% compared to doctors, respectively. From the term of single-step cumulative return (SCR), compared to doctors, $PrescDRL_{DQN}$ and $PrescDRL_{DRQN}$ improved by 269% and 292% respectively. And for the multi-step cumulative return (MCR), $PrescDRL_{DQN}$ and $PrescDRL_{DRQN}$ improved by 387% and 402% than doctors respectively. Meanwhile, the results also showed $PrescDRL_{DRQN}$ obtain higher rewards than than $PrescDRL_{DQN}$, improved by 16.2% for SSR, 6.48% for SCR and 3.16% for MCR.

We compared the length of diagnosis and treatment sequences of the PrescDRL model with that of doctors (Fig. 3D). The results showed that doctors had the shortest sequence length, while both the $PrescDRL_{DQN}$ and $PrescDRL_{DRQN}$ models had longer sequence length than doctors. This indicates that although PrescDRL has high diagnostic performance, it increases the number of diagnosis and treatment. Additionally, we also compared the performance differences between $PrescDRL_{DQN}$ and $PrescDRL_{DRQN}$ models with different numbers of HPCs (Fig. 3E and 3F). The results showed that the number of HPCs has little influence on the prediction performance of PrescDRL. However, with the increase of HPC number, the $PrescDRL_{DRQN}$ model showed better stability than the $PrescDRL_{DQN}$ model.

The above results showed that our PrescDRL outperformed the doctors in terms of rewards, indicating better curative effects. Additionally, the sequential HPCs predicted by PrescDRL have shorter curative periods compared to those given by doctors. These results indicated that PrescDRL, based on RL, has significant advantages in DDTS optimization.

Prescription prediction of PrescDRL

In clinical practice, a prescription recommendation system needs to provide specific recommended prescriptions for each patient. However, PrescDRL provides an herbal prescription cluster rather than an actual herbal prescription. Therefore, we need to combine the HPC given by PrescDRL with the symptom observations of patients to make prescription predictions. The previous experimental results showed that the HPC provided by PrescDRL obtains higher rewards than those provided by clinicians. Therefore, the model that combines the HPC given by PrescDRL with symptom observations should recommend better prescriptions than doctors. Thus, it is not possible to evaluate the performance of PrescDRL using doctors' prescriptions as a benchmark. To address this, we conducted a degradation experiment by combining the original HPC with symptom observations and used prescriptions given by doctors as a benchmark. If the prediction performance of this method is better than that using only symptom observation, it can be concluded that the prescription recommendation based on symptom observations and the HPC is better than that of doctors. Therefore, PrescDRL should have higher predictive performance.

In the experiment, we constructed a three-layer fully connected neural network to compare prescription recommendation models. The first model considered both the patient's symptom information and the HPC given by PrescDRL, while the second model only considered the patient's symptom information and served as the

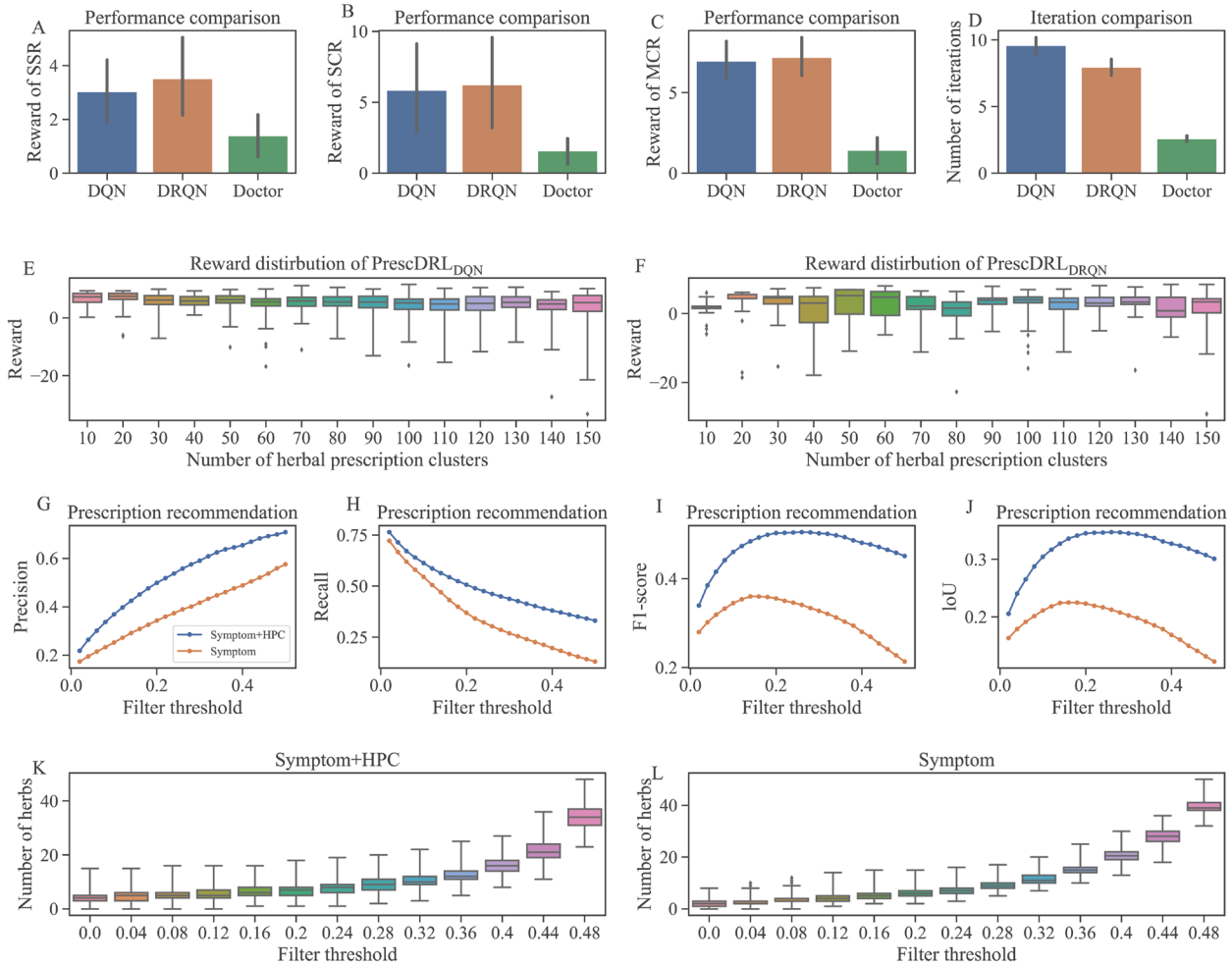

Fig. 3 Experimental results of PrescDRL. **A** Performance comparison of single-step return. **B** Performance comparison of single cumulative return. **C** Performance comparison of multi-step cumulative return. **D** Comparison of iterations, i.e., the sequence length of diagnosis and treatment plan. **E** Reward distribution of *PrescDRL*$_{DQN}$ with different HPCs. **F** Reward distribution of *PrescDRL*$_{DQN}$ with different HPCs. **G** Precision comparison of prescription recommendation. **H** Recall comparison of prescription recommendation. **I** F1-score comparison of prescription recommendation. **J** LoU comparison of prescription recommendation. **K** Number distribution of the predicted herbs given by PrescDRL that considers symptom and scheme. **L** Number distribution of the predicted herbs given by PrescDRL that only considers symptom

Table 1 Reward comparison of DDTS optimization

Models	Single-step return	Single-step cumulative return	Multi-step cumulative return
Doctor	1.39	1.59	1.42
PrescDRL$_{DQN}$	3.03 (117%)	5.86 (269%)	6.96 (387%)
PrescDRL$_{DRQN}$	3.52 (153%)	6.24 (292%)	7.18 (402%)

The rate of improvement of our method compared to the doctor is shown in parentheses

benchmark. We evaluated the performance of these models under different filter thresholds and evaluation metrics (Fig. 3G-3J and Table 2). It's important to note that the filter threshold is a hyperparameter given to the neural network. The size of this parameter is inversely proportional to the strictness of the screening, meaning that smaller values correspond to stricter screening and larger values correspond to looser screening.

Our experimental results showed that the recommendation performance of our model was significantly higher than that of the benchmark model. For instance, with a filter threshold of 0.2, our model achieved higher precision (improved by 40.5%), recall (improved

Table 2 Performance comparison of prescription recommendation

Inputs of model	Filter threshold	No.of predicted herbs	Precision	Recall	F1-score	IoU
Symptom [1]	0.1	3.77±1.34	0.49	0.20	0.28	0.17
	0.2	6.07±1.48	0.42	0.27	0.33	0.20
	0.3	10.12±1.77	0.34	0.37	0.36	0.22
	0.4	20.52±2.63	0.25	0.55	0.35	0.21
Symptom + *scheme* [2]	0.1	5.44±2.39	**0.65**	0.38	0.48	0.33
	0.2	7.04±2.60	0.59	0.44	**0.50**	**0.34**
	0.3	9.72±2.80	0.50	0.51	**0.50**	**0.34**
	0.4	16.00±3.03	0.37	**0.61**	0.46	0.30

[1] "*Symptom*" denotes that the input of the recommendation algorithm is the patient's symptoms

[2] "*Symptom* + *scheme*" denotes that the input of the recommendation algorithm includes the patient's symptoms and the diagnosis and treatment plan provided by PrescDRL

The bold numbers represent the highest performance

by 63%), and F1-score (improved by 51.5%) than the benchmark model. Additionally, we compared the number of herbs recommended by the two models under different filter thresholds (Fig. 3K-3L). The results showed that, as the filter threshold increased, the number of herbs recommended by both models increased. However, the number of herbs recommended by our model was lower than that of the benchmark model. When the filter threshold was set to 0.3, our model predicted an average of 9.72±2.80 herbs, while the benchmark model predicted an average of 10.12±1.77 herbs. Notably, in the real prescription data set, the mean and standard deviation of drugs in each prescription were 9.84±2.96. Thus, our model's predictions were closer to the true number of herbs in prescriptions when the appropriate filter threshold was selected.

Case study of the diagnosis and treatment sequence of PrescDRL

In the case study section, we first presented the distribution of herb occurrence frequency predicted by PrescDRL, with a filter threshold of 0.22, and compared the frequency of these herbs with those in the original prescriptions (Fig. 4). The results showed that the predicted herbs had a similar frequency distribution to the original herbs, which indirectly confirmed the reliability of the predictive results generated by PrescDRL.

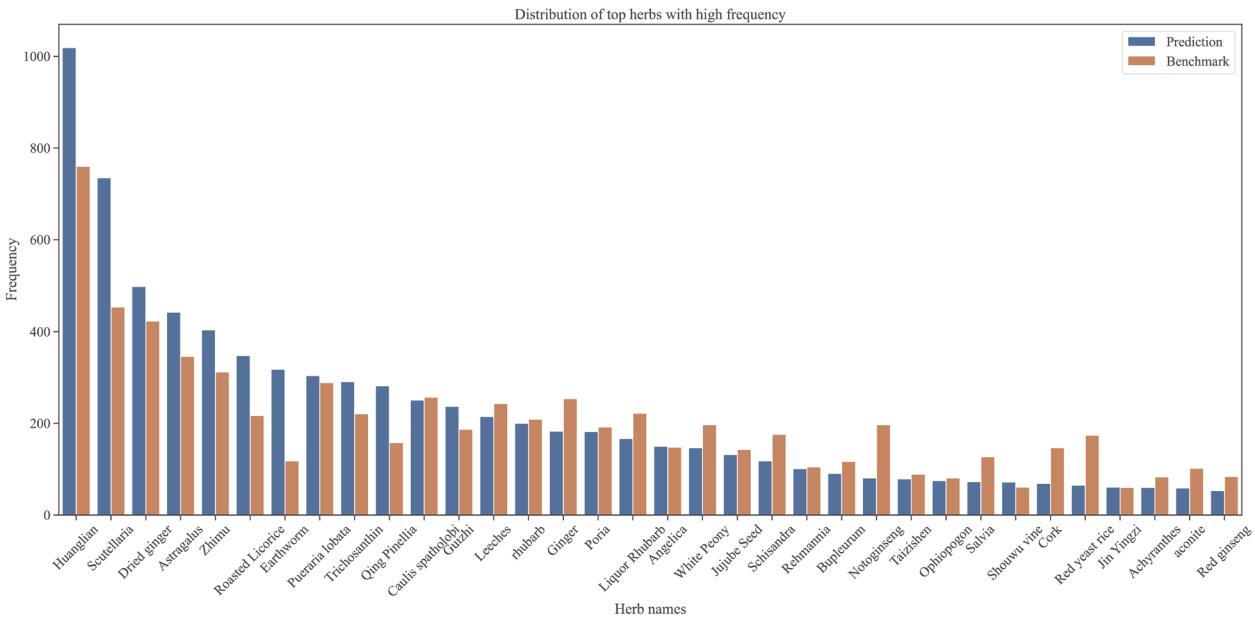

Fig. 4 Frequency distribution of top herbs predicted by PrescDRL

Table 3 Diagnosis and treatment sequence of doctors

Treatment number	Symptom of patients	Treatment plan	Prescription	Symptom score
1	Tongue or coating with blood stasis, dark tongue or coating, rapid pulse, irregular stool, ...	67	Coptis chinensis, donkey-hide gelatin beads, chicken yellow, white peony root, skullcap, jujube seed, ...	13
2	Thick and greasy tongue or coating, stasis of tongue or coating, dark tongue or coating, aversion to cold, ...	37	Dried ginger, Coptis chinensis, Scutellaria baicalensis, American ginseng, Poria, Panax notoginseng	10
3	The whole body is heavy, the tongue or fur is stagnated, the tongue or fur is dark, the whole body is weak, ...	67	Angelica, Astragalus, Coptidis, Cinnamon, Anemarrhena, Golden Cherry, Ginger, ...	8
4	Loose stools, stagnant tongue or coating, fluttering tongue or coating, thin tongue or coating, ...	36	wine rhubarb, aconite, Alisma, fenugreek, Gorgon, yam, ...	11

Table 4 Diagnosis and treatment sequence given by PrescDRL

Treatment number	Symptom of patients	Treatment plan	Prescription	Symptom score
1	Tongue or coating with blood stasis, dark tongue or coating, rapid pulse, irregular stool, ...	81	Pueraria Root, Helichrysum, Trichosanthes, Ginger, Leeches, ...	13
2	Thick and greasy tongue or fur, stasis of tongue or fur, aversion to cold,dry or secretive stool, ...	81	Astragalus, Salvia, Zhigancao, Leech, Sophora Radix, Rhizoma Coptidis, ...	8
3	Swelling of lower extremities, thick and greasy tongue or coating, stasis of tongue or coating, rapid pulse, ...	12	Treats, Astragalus, Suanzaoren, Dilong, Sophora japonica, ...	11
4	Tinnitus, dark tongue or coating, sweating, yellow tongue or coating, urination and nocturia, ...	71	Pueraria, Monascus, Dilong, Trichosanthes, Dried Ginger, Achyranthes, ...	7
5	tinnitus, general fatigue	–	–	2

To illustrate the predictive performance of PrescDRL, we showed a real diagnosis and treatment sequence (Table 3 and 4) and a predicted sequence given by PrescDRL based on the first diagnosis of the real sequence. It is important to note that the two sequences are based on the results of 100 HPCs. Based on the doctor's sequence, the patient had a total of four visits with an initial symptom score of 13 and a symptom score of 11 for the last visit, indicating that the treatment effect was not optimal and there were recurrent conditions. This could be due to variations in patients' physical quality and medication, as well as differences in doctors' experience. Therefore, even if the same patient is at the same stage of disease, each doctor may prescribe different prescriptions according to their own experience.

After analyzing the diagnosis and treatment sequence provided by the PrescDRL algorithm, it was found that the patient had undergone a total of five visits and his symptom score had gradually decreased throughout the treatment, indicating a gradual improvement in the patient's condition. At the end of the sequence, the patient's symptom score was 2, which suggests that the patient's disease had significantly improved through the entire diagnosis and treatment process.

Through a comparison of the diagnosis and treatment sequences provided by the doctor and the PrescDRL algorithm, we observed that patient symptoms tend to improve in repeated fluctuations rather than directly. In the doctor's sequence, the symptom score gradually decreases in the first three visits, but then increases in the fourth visit, indicating that the fourth prescription may not have been appropriate. In contrast, the PrescDRL model aims to maximize long-term effectiveness by selecting the best medicine based on the

patient's current symptoms, without necessarily expecting maximum reward at every step. As demonstrated in the sequence provided by PrescDRL, the symptom score does increase in the third visit, but ultimately drops to 2 in the fifth visit, suggesting that the diagnosis and treatment plan generated by the PrescDRL model based on reinforcement learning is more reasonable and effective.

In addition to the above example, we also provide three additional examples of treatment process comparison given by doctors and PrescDRL algorithm in Section 1 of supplemental file 1. In the first example, the algorithm performed better than the doctor. In the second example, the doctor performed better than the algorithm, and in the third example, both the doctor and the algorithm performed better. These comparison examples show that compared with doctors, most of the treatment plans given by PrescDRL can reduce the symptom score of patients, thereby improving the disease status and treating the disease of patients.

Discussion

With the advancement of real-world clinical medicine, a significant amount of diagnosis and treatment data from famous and experienced TCM doctors have been accumulated. As a result, how to extract medication rules from this data and develop an effective model for recommending reasonable prescriptions has become a research hotspot in TCM intelligence. In light of this, we propose a RL-based prediction model for optimizing diagnosis and treatment schemes. This model is a sequential optimization approach that prioritizes long-term effectiveness to provide rational TCM prescriptions. We designed the comprehensive experiments following the algorithm evaluation guidelines in the network pharmacology [51], and the experimental results indicated that HPTP given by PrescDRL have better curative effect than doctors and higher performance on prescription prediction.

Although our proposed PresDRL algorithm have achieved the excellent performance, prescription optimization and recommendation algorithm still faces some challenges in clinical application. For example, first, incompleteness, inconsistency, or errors in clinical data may affect the accuracy and reliability of drug recommendation algorithms. Second, algorithms need to be able to process data across different patient populations and disease states to ensure effective recommendations in multiple clinical settings. Third, treatment recommendation algorithms should aid clinical decision-making, not replace the professional judgment of doctors, and it is necessary to ensure the transparency and interpretability of algorithms to gain the trust of medical professionals.

There are still some areas that require further exploration in the future. First, the diagnosis and treatment sequences generated by PrescDRL are longer compared to those provided by doctors. In the future, it would be necessary to investigate a diagnosis and treatment optimization model that can provide shorter sequences while maintaining high efficacy. Second, all the experimental results presented in this study are based on simulation experiments. To validate the effectiveness of our proposed PrescDRL model, it is crucial to apply it in real-world clinical diagnosis and treatment systems to evaluate the specific effects. Finally, with the explosive development of large language model (LLM) technology [52], LLMs has been applied to the medical field [53]. Due to the advantages of strong model generalization ability and knowledge reasoning of LLMs, relevant scholars have focused on the field of TCM and proposed several LLMs of TCM, such as ShenNong-TCM [54] and HuaTuo [55]. However, there are currently no LLMs optimized specifically for TCM treatment planning. Therefore, in the future, we will collect more data of clinical sequential diagnosis and treatment in order to train LLM for treatment planning optimization, improving the efficiency of clinical diagnosis and treatment.

Conclusions

In this study, we proposed PrescDRL, a deep RL-based optimization model for herbal prescription treatment planning, that prioritizes long-term effectiveness to provide reasonable TCM prescriptions. The experimental results demonstrated that PrescDRL generated herbal prescription treatment plans with better curative effects than doctors and achieved higher performance in herbal prescription recommendation. Overall, PrescDRL provides an exemplary approach to employ RL to learn the patient's optimal treatment path, which can help minimize medication errors, reduce the patient's treatment cost, and improve treatment effectiveness.

Abbreviations

TCM	Traditional Chinese medicine
DDTS	Dynamic diagnosis and treatment scheme
PrescDRL	Herbal prescription planning based on deep reinforcement learning
RL	Reinforcement learning
MDP	Markov decision process
STC	Sequence termination condition
SSR	Single-step return
SCR	Single-step cumulative return
MCR	Multi-step cumulative return
HPC	Herbal prescription cluster
HPTP	Herbal prescription treatment planning

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13020-024-01005-w.

Supplementary File 1.

Acknowledgements
This work is partially supported by the Fundamental Research Funds for the Central Universities (No. 2022RC022), the National Key Research and Development Program (Nos. 2023YFC3502604 and 2021YFC1712901), the National Natural Science Foundation of China (Nos. 82174533, 82374302, 82204941, and U23B2062), the Natural Science Foundation of Beijing (No. L232033), and Key R& D Program Project of Ningxia Hui Autonomous Region(2022BEG02036).

Author contributions
K.Y., X.Z. were involved in the conception and design of the work. Z.Y., X.H. and N.W. were involved in data collection and model construction. K.Y., Z.Y., X.H., Q.Z., F.Y. and X.S. were involved in data analysis and interpretation. K.Y., F.Z., T.W. and X.Z. were involved in the drafting and revision of the article. K.Y. and X.Z. approved the final version to be published.

Data Availability
The data of case study presented in the study are included in the Supplementary Material. Further inquiries about data can be directed to the corresponding author.

Received: 6 April 2024 Accepted: 14 September 2024
Published online: 16 October 2024

References

1. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, Abu-Hanna A. The coming of age of artificial intelligence in medicine. Artif Intell Med. 2009;46(1):5–17.
2. Alagoz O, Hsu H, Schaefer AJ, Roberts MS. Markov decision processes: a tool for sequential decision making under uncertainty. Med Decis Making. 2010;30(4):474–83.
3. Deng N, Zhang Q. The application of dynamic uncertain causality graph based diagnosis and treatment unification model in the intelligent diagnosis and treatment of hepatitis b. Symmetry. 2021;13(7):1185.
4. Li S. Mapping ancient remedies: applying a network approach to traditional Chinese medicine. Science. 2015;350(6262):S72–4.
5. Cui J. Diagnosis and treatment technologies of traditional Chinese medicine: application and prospect in context of artificial intelligence. Acad J Second Mil Univ. 2018;15:846–51.
6. Yang K, Zheng Y, Lu K, Chang K, Wang N, Shu Z, Yu J, Liu B, Gao Z, Zhou X. PDGNet: predicting disease genes using a deep neural network with multi-view features. IEEE/ACM Trans Comput Biol Bioinf. 2020;19(1):575–84.
7. Yang K, Wang N, Liu G, Wang R, Yu J, Zhang R, Chen J, Zhou X. Heterogeneous network embedding for identifying symptom candidate genes. J Am Med Inform Assoc. 2018;25(11):1452–9.
8. Zhang S, Yang K, Liu Z, Lai X, Yang Z, Zeng J, Li S. DrugAI: a multi-view deep learning model for predicting drug-target activating/inhibiting mechanisms. Brief Bioinform. 2023;24(1):526.
9. Yang K, Yang Y, Fan S, Xia J, Zheng Q, Dong X, Liu J, Liu Q, Lei L, Zhang Y, et al. DRONet: effectiveness-driven drug repositioning framework using network embedding and ranking learning. Brief Bioinf. 2023;24(1):bbac518.
10. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. Mastering the game of go with deep neural networks and tree search. Nature. 2016;529(7587):484–9.
11. Lei Y, Li S, Liu Z, Wan F, Tian T, Li S, Zhao D, Zeng J. A deep-learning framework for multi-level peptide-protein interaction prediction. Nat Commun. 2021;12(1):5465.
12. Li H, Kumar N, Chen R, Georgiou P (2018) A deep reinforcement learning framework for identifying funny scenes in movies," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3116–3120
13. Nemati S, Ghassemi M M, Clifford G D. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach[C]//2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2016: 2978-2981.
14. Padmanabhan R, Meskin N, Haddad WM. Optimal adaptive control of drug dosing using integral reinforcement learning. Math Biosci. 2019;309:131–42.
15. Ghassemi MM, Alhanai T, Westover MB, Mark RG, Nemati S, Personalized medication dosing using volatile data streams, in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
16. Lin R, Stanley MD, Ghassemi MM, Nemati S, A deep deterministic policy gradient approach to medication dosing and surveillance in the icu, in,. 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE. 2018; 2018:4927–31.
17. Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M, Deep reinforcement learning for sepsis treatment, *arXiv preprint*arXiv:1711.09602, 2017.
18. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M, Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach, in Machine Learning for Healthcare Conference. PMLR, 2017; 147–163.
19. Raghu A, Komorowski M, Singh S, Model-based reinforcement learning for sepsis treatment, arXiv preprint arXiv:1811.09602, 2018.
20. Futoma J, Lin A, Sendak M, Bedoya A, Clement M, O'Brien C, Heller K, Learning to treat sepsis with multi-output gaussian process deep recurrent q-networks, 2018.
21. Lopez-Martinez D, Eschenfeldt P, Ostvar S, Ingram M, Hur C, Picard R, Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep q networks, in,. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE. 2019; 2019:3960–3.
22. Liu Y, Logan B, Liu N, Xu Z, Tang J, Wang Y, Deep reinforcement learning for dynamic treatment regimes on medical registry data, in,. IEEE international conference on healthcare informatics (ICHI). IEEE. 2017;2017: 380–5.
23. Wang L, Zhang W, He X, Zha H, Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation, in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018; 2447–2456.
24. Feng Q, Combining mortality and longitudinal measures in clinical trials. Ph.D. dissertation, Beijing Jiaotong University, 2011.
25. Hu X, "Research on optimization method of traditional chinese medicine sequential diagnosis and treatment scheme based on deep reinforcement learning." Master's thesis, Beijing Jiaotong University, 2019.
26. Gijsen R, Hoeymans N, Schellevis FG, Ruwaard D, Satariano WA, van den Bos GA. Causes and consequences of comorbidity: a review. J Clin Epidemiol. 2001;54(7):661–74.
27. Zhang X, Zhou X, Zhang R, Liu B, Xie Q, Real-world clinical data mining on tcm clinical diagnosis and treatment: a survey, in *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2012;88–93.
28. Russell S, Norvig P, Artificial Intelligence: A Modern Approach, (2016)[J]. doi, 10: 363.
29. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559–67.
30. Wang Z, Poon J, Poon S, TCM Translator: A sequence generation approach for prescribing herbal medicines, in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019; 2474–2480.
31. Wu Y, Pei C, Ruan C, Wang R, Yang Y, Zhang Y. Bayesian networks and chained classifiers based on svm for traditional Chinese medical prescription generation. World Wide Web. 2022;25(3):1447–68.
32. Yao L, Zhang Y, Wei B, Zhang W, Jin Z. A topic modeling approach for traditional Chinese medicine prescriptions. IEEE Trans Knowl Data Eng. 2018;30(6):1007–21.
33. Zhang X, Zhou X, Huang H, Feng Q, Chen S, Liu B. Topic model for Chinese medicine diagnosis and prescription regularities analysis: case on diabetes. Chin J Integr Med. 2011;17(4):307–13.
34. Jin Y, Ji W, Zhang W, He X, Wang X, Wang X, A KG-enhanced multi-graph neural network for attentive herb recommendation, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021.
35. Li S, Wang W, He J, KGAPG: Knowledge-aware neural group representation learning for attentive prescription generation of traditional Chinese

medicine, in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021;450–455.
36. Li W, Yang Z, Exploration on generating traditional chinese medicine prescriptions from symptoms with an end-to-end approach, in CCF International Conference on Natural Language Processing and Chinese Computing. Springer, 2019; 486–498.
37. Hu Y, Wen G, Liao H, Wang C, Dai D, Yu Z. Automatic construction of chinese herbal prescriptions from tongue images using cnns and auxiliary latent therapy topics. IEEE Trans Cybernet. 2019;51(2):708–21.
38. Liao H, Wen G, Hu Y, Wang C, Convolutional herbal prescription building method from multi-scale facial features, Multimedia Tools and Applications, vol. 78, no. 24, pp. 35 665–35 688, 2019.
39. Zhou W, Yang K, Zeng J, Lai X, Wang X, Ji C, Li Y, Zhang P, Li S. FordNet: recommending traditional Chinese medicine formula via deep neural network integrating phenotype and molecule. Pharmacol Res. 2021;173: 105752.
40. Dong X, Zheng Y, Shu Z, Chang K, Yan D, Xia J, Zhu Q, Zhong K, Wang X, Yang K, *et al.*, TCMPR: TCM prescription recommendation based on subnetwork term mapping and deep learning, in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021;3776–3783.
41. Dong X, Zhao C, Song X, Zhang L, Liu Y, Wu J, Xu Y, Xu N, Liu J, Yu H, et al. PresRecST: a novel herbal prescription recommendation algorithm for real-world patients with integration of syndrome differentiation and treatment planning. J Am Med Inform Assoc. 2024;31(6):1268–79.
42. Wang Z, Liang Y, Liu Z, FFBDNet: Feature fusion and bipartite decision networks for recommending medication combination, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases.Springer, 2022;419–436.
43. Tan W, Wang W, Zhou X, Buntine W, Bingham G, Yin H. OntoMedRec: Logically-pretrained model-agnostic ontology encoders for medication recommendation. World Wide Web. 2024;27(3):28.
44. Mi J, Zu Y, Wang Z, He J. ACDNet: Attention-guided Collaborative Decision Network for effective medication recommendation. J Biomed Inform. 2024;149: 104570.
45. Zheng Z, Wang C, Xu T, Shen D, Qin P, Zhao X, Huai B, Wu X, Chen E. Interaction-aware drug package recommendation via policy gradient. ACM Trans Inf Syst. 2023;41(1):1–32.
46. Bellman R. A Markovian decision process. J Math Mech. 1957;20:679–84.
47. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M, Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602, 2013.
48. Hausknecht M, Stone P, Deep Recurrent Q-Learning for Partially Observable MDPs, in 2015 aaai fall symposium series, 2015.
49. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinburgh Dublin Philos Mag J Sci. 1900;50(302):157–75.
50. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Stat Med. 1999;18(11):1341–54.
51. Li S, et al. Network pharmacology evaluation method guidance-draft. World J Tradit Chin Med. 2021;7(1):146.
52. Wu T, He S, Liu J, Sun S, Liu K, Han Q-L, Tang Y. A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA J Auto Sinica. 2023;10(5):1122–36.
53. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D *et al.*, Towards expert-level medical question answering with large language models, arXiv preprint arXiv:2305.09617, 2023.
54. Zhu W, Wang X, Wang L, Chatmed: A Chinese medical large language model, Retrieved September, vol. 18, p. 2023, 2023.
55. Wang H, Liu C, Xi N, Qiang Z, Zhao S, Qin B, Liu T, HuaTuo: Tuning LLaMA model with Chinese medical knowledge, arXiv preprint arXiv:2304.06975, 2023.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.