A Deep Reinforcement Learning Approach for Type 2 Diabetes Mellitus Treatment

Zhuo Liu
Ping An Healthcare Technology
Beijing, China
liuzhuo046@pingan.com.cn

Linong Ji
Department of Endocrinology and Metabolism
Peking University People's Hospital
Beijing, China
jiln@bjmu.edu.cn

Xuehan Jiang
Ping An Healthcare Technology
Beijing, China
jiangxuehan073@pingan.com.cn

Wei Zhao
Ping An Healthcare Technology
Beijing, China
zhaowei798@pingan.com.cn

Xiyang Liao
Ping An Healthcare Technology
Beijing, China
liaoxiyang852@pingan.com.cn

Tingting Zhao
Ping An Healthcare Technology
Beijing, China
zhaotingting757@pingan.com.cn

Siqi Liu
NUS Graduate School for Integrative Sciences and Engineering, Saw Swee Hock School of Public Health
National University of Singapore
Singapore, Singapore
e0272316@u.nus.edu

Xingzhi Sun
Ping An He Healthcare alth Technology
Beijing, China
sunxingzhi820@pingan.com.cn

Gang Hu
Ping An Healthcare Technology
Beijing, China
hugang898@pingan.com.cn

Mengling Feng
Saw Swee Hock School of Public Health
National University of Singapore
Singapore, Singapore
ephfm@nus.edu.sg

Guotong Xie
Ping An Healthcare Technology
Beijing, China
xieguotong@pingan.com.cn

2020 IEEE International Conference on Healthcare Informatics (ICHI) | 978-1-7281-5382-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHI48887.2020.9374313

***Abstract*—Type 2 diabetes mellitus (T2DM) is a chronic disease that requires continuous treatments. T2DM treatments aim to achieve not only short-term but, more importantly, long-term control of the patient's glucose to a normal level. We believe Reinforcement Learning (RL) can be an effective approach to learn and further recommend the ideal sequence of treatments that optimize the patient's long-term outcome. In this paper, we implement RL with the deep Q network. Our RL model learns from a T2DM patient registry dataset that consists of the clinical data of newly diagnosed T2DM patients over a twelve-month period including four follow-ups since their first visit. The RL model is trained to recommend the number of oral antidiabetic drugs and the number of insulins. Our experiments have shown that our RL model can improve the long-term hemoglobin A1c goal achieving rate by up to 15%. This confirms that the model learns good medication patterns that favor the long-term glucose control.**

Keywords—Type 2 diabetes mellitus, treatment recommendation, reinforcement learning, deep Q network

I. INTRODUCTION

Type 2 diabetes mellitus (T2DM) is a world-wide chronic disease featured by higher-than-optimal blood glucose that can lead to multiple complications and increase the overall risk of mortality. According to the global report on diabetes of World Health Organization, 3.7 million people died of diabetes and higher-than-optimal blood glucose in 2012 and the prevalence is increasing in the past three decades [1]. The treatment of T2DM is often a life-long process. Thus, T2DM can lead to substantial economic burdens to both the families and society. The core of T2DM treatment lies in keeping control of blood glucose level. In traditional clinical practices, physicians make sequential decisions of prescriptions based mainly on clinical guidelines, which are not patient-specific and heavily depend on the expertise of physicians [2-4]. Nowadays, thanks to the widely adoption of electronic health record (EHR) systems, automatic recommendation models can now be trained on the EHR data to support the decision-making process in T2DM treatment [5-10].

A study by Bertsimas et al. [11] proposed a personalized treatment algorithm based on the EHR data from 48,140 patients. They built a k-nearest neighbors (kNN) model with regression function to estimate the resulting hemoglobin A1c (HbA1c) from a series of treatment options for each subset of patients who have similar disease conditions. The treatment leading to the best HbA1c was recommended. Chen et al. [12] applied similar kNN-based predictive models on the EHR data

978-1-7281-5382-7/20/$31.00 ©2020 IEEE

of 12,016 patients who were admitted to the Boston Medical Center. Chen et al. showed that their method led to a larger reduction of the HbA1c than the standard of care. Wang et al. [13] considered both efficacy of reducing HbA1c and safety issues in deriving an optimal treatment regime. They applied regression-based learning framework on a randomized trial of 2,091 T2DM patients to guide optimal personalized utilization of insulin while controlling the risks of hypoglycemia events in the regression model.

The mentioned prior works on treating T2DM are based on predictive models using traditional machine learning approaches, such as kNN and regression, which predominantly optimize short-term patient outcome according to a snapshot of symptoms of patients. However, the management of T2DM is a longitudinal and sequential decision-making process where choices of treatments would consider both short-term and long-term impacts. In complementary to the traditional machine learning models, reinforcement learning (RL), has the distinctive advantage in the situations where impacts of long-term outcomes are non-trivial comparable to or even more important than those of short-term outcomes. In our case, medications should be prescribed to a T2DM patient to reduce their blood glucose level (i.e., the short-term goal), and at the same time improve the glycemic control in a patient's future disease trajectory (i.e., the long-term goal). Achieving both the short-term and long-term goals are crucial for improving patients' overall outcomes.

RL has been proved to achieve the human-level capacity for learning complex sequential decisions in many domains, such as playing games [14], robotics control [15], autonomous driving [16], etc. And it has also been applied in recommending treatments in both critical care [17-22] and chronic diseases contexts [23-26].

For critical clinical care, RL was applied to support treatment decision for physicians in various applications, such as recommending the optimal type and dosing of a medication [17-20], or choosing the best timing of an intervention [21, 22]. In one recent study, Komorowski et al. [20] managed to use the RL to recommend personalized optimal dosage for intravenous fluids and vasopressors to improve septic patient's mortality. Their result showed that constantly following the recommendations from the RL model would result in the lowest risk of 90-day mortality. In terms of chronic diseases, RL has been used for treating type 1 diabetes mellitus patients by administrating a precise insulin dose using a continuous glucose monitoring system and a closed loop controller [23, 24]. Besides, there are also applications of using RL to recommend treatment (e.g., chemotherapy [25] and radiotherapy [26]) for cancer patients. Their results showed that RL could learn meaningful knowledge from the observational EHR data and have the potential to improve patients' long-term outcome.

Inspired by RL's success in critical care, we developed RL framework for recommending treatments for T2DM patients. The patient's clinical condition at each follow-up visit is defined as a *state*: consists of the demographics, lifestyle, laboratory measurements and clinical events. At each time step, there are predefined candidate actions that a physician can prescribe to a patient. Here the *action* refers to a medication pattern such as a monotherapy of oral antidiabetic drug (OAD) or a duel therapy of an insulin and an OAD. For each patient, we obtained a sequence of <*state*, *action*, *next state*> triples to represent the disease trajectory of all the follow-up visits for the patient. A 'reward' function is customized to quantify the effectiveness of the treatment (action) at each visit. It is determined by the current state, the action is taken, and the next state. The ultimate goal of RL is to learn a policy, by which for any given state, the RL model can select the action that maximizes the cumulative future rewards. As mentioned earlier, the RL approach has a few intrinsic advantages. Firstly, by considering the accumulative rewards as the optimization goal, the long-term effect of current treatment decision is taken into account. Secondly, the design of RL leverages all samples in the model training, by reinforcing the action leads to good clinical outcome and punishing the one which resulted in worse outcome. Thirdly, new data collected from the clinical practice could be applied to refine the existing model incrementally rather than re-training the model.

Our key contributions are: 1) We formulate the task of optimizing medication prescription for individual T2DM patient as an RL problem. 2) With real T2DM patient data, we show that RL is capable of extracting and summarizing the knowledge from the observational data for treatment recommendation. 3) Our experiments demonstrate that when physicians constantly follow our RL model's recommendation, both the patients' short-term and long-term clinical outcomes would be improved. We are confident that our model has great potential in assisting physicians to make more informed decisions and to better manage individual patients' conditions.

II. Methods

A. Data

We obtain data from the prospective study of Evaluation of Effectiveness of Treatment Paradigm for Newly Diagnosed Type 2 Diabetes Patients in China (NEW2D) [27] from June 2012 to February 2014. Patients which included in current study satisfy the following important criteria: at least 20 years old; newly diagnosed with T2DM within 6 months; and neither pregnant (or planning to be within one year) nor lactating women. These patients participated in the investigation through case report forms. After a baseline visit, patients are expected to be followed up every 3 months for 1 year. The data consists of 5 visits, a baseline and 4 follow-ups at the 3rd, 6th, 9th and 12th month.

The data set we use includes 5,193 T2DM patients from around 80 hospitals. Although the number of patients participating the visits decreases with time, there are still 4,097 (78.9%) patients participating at the 12th month. The main reason is the loss of follow-up visits which due to loss of connection, moving out of their local area, leaving their original hospitals, withdrawal of the informed consent, other illness, or death. About 50 features are collected at each patient visit in this study, including demographic information, life styles, physical examination, drug usage, medical history, lab test, prescription and medical events.

In current study, we stratify patients' states based on their HbA1c levels. HbA1c is a result of a medical test detecting blood density of glycated hemoglobin, representing the average blood glucose level over the past 8 to 12 weeks [28]. We divide HbA1c into 3 levels, defined as follow: *Low* for HbA1c<7%, *Med* for $7\% \leqslant \text{HbA1c} \leqslant 9\%$, and *High* for HbA1c>9%.

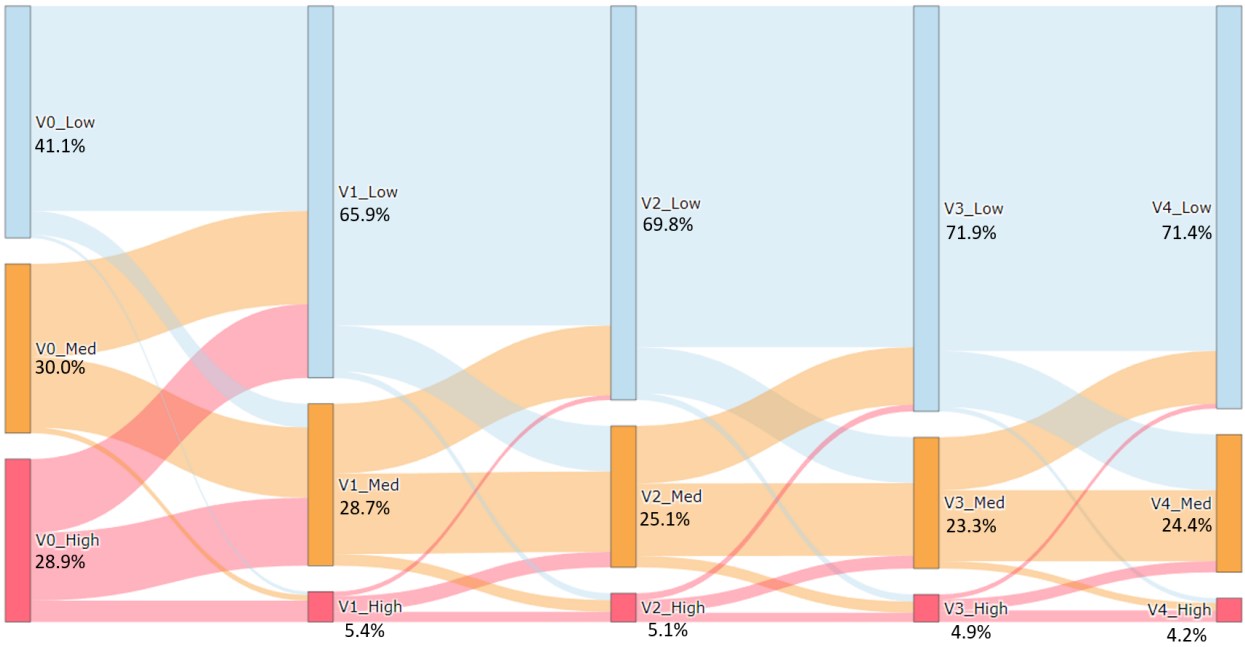

Fig. 1 Visualization of the change of HbA1c level during 5 visits. From left to right, the 5 sets of vertical bars represent 5 visits arranged in chronological order (denoted as V0 to V4, respectively). There are 3 bars in each vertical set corresponding to 3 HbA1c levels, namely *Low*, *Med* and *High*. The curved bands linking two consecutive bars represent the changing number of patients in adjacent visits. The height of the bars and the width of the bands are proportional to the number of patients.

To assess the quality of our data, we analysis the change of HbA1c level as the number of visit increase. The result shows in Fig. 1, from left to right, more and more patients have lower HbA1c level as the visit progressing. The proportion of low-HbA1c patients increased from the baseline 41.1% to 71.4% when data collection finished, which indicates that the physicians' treatment generally achieved a good clinical outcome. Yet, some patients remain *Med* or *High* HbA1c level till the final visit. Learning from this data set, our reinforcement learning model aims to capture effective treatment policies and avoid inappropriate ones.

B. Data Preprocessing

To incorporate with the RL framework of T2DM treatment, we future select samples from NEW2D according to the considerations list below. Firstly, we choose 4,243 patients that have no less than two visits. Secondly, by filtering out visits with HbA1c missing, there are 3,893 patients and 16,562 visits left. We then generate $<s_t, a_t, s_{t+1}>$ triples based on the 16,562 visits, where s_t represents the patient state in the current visit, s_{t+1} represents that in the next follow-up visit, and a_t represents the action of medication policy in the current visit. For each patient, a visit cannot be used to generate a triple when the next step of the visit is missing. Therefore we get 12,530 $<s_t, a_t, s_{t+1}>$ triples for modeling at last.

Aberrant values in our dataset are deleted, which may due to mis-recording or a preconceived unit. Those aberrant records are deleted. As HbA1c is an indispensable feature that defines the reward, visits with HbA1c absent are also removed. For other features, missing values are imputed using interpolation method, namely by averaging the value recorded at last and next visits or one side only if the other side is also absent.

C. States

To define the state of a patient, we select features that closely related with physician's prescription process as listed in Table I. Those features are all reviewed by physicians. A

TABLE I. FEATURES SELECTED FROM THE DATASET TO DEFINE STATE.

Class of feature	Feature	Class of feature	Feature
Demographic information	Gender	Life styles	Physical activity quantity
	Age		Smoking history
Physical examination	Body mass index	Previous drug usage	Oral antidiabetic drugs
	Systolic blood pressure		Insulins
	Diastolic blood pressure	Lab test result	Hemoglobin A1c
Medical history	Duration of T2DM		High-density lipoprotein cholesterol
	Whether have complications		Serum creatinine
	Whether have nephropathy		Fasting blood glucose
	Whether have hypertension		Low-density lipoprotein cholesterol
	Whether have dyslipidemia		Total cholesterol
	Whether have hypoglycemia		Triglyceride

state is represented with a vector of selected variables. The continuous variables are normalized to change values to common scales, which benefits the training of neural network models. The categorical variables are converted to one-hot encoding which as the same as binary variables donated by 1 and 0. Finally, a state is embedded with a 35-dimension vector.

D. Actions

In routine treatment of blood glucose control, physicians mainly consider OADs and insulins. Thus, we define the action as a combination of the number of OADs and the number of insulins in the prescription. Compared with directly using drug names as actions, our simplification reduces the complexity of action space comparable to the limited data size. Here both OADs and insulins are counted at the drug class level. Table II shows the drug classes of 5 OADs and 3 insulins in the data set. For example, a prescription with metformin, sulfonylureas, and basal insulin is represented as the action (2,1) that means 2 OADs and 1 insulin. Note that a special action (0,0) means that no drug is prescribed. The maximum number of OADs and insulins in the data are 4 and 2, respectively. Since the data set of newly diagnosed patients does not contain overly strong prescriptions of (3,2) and (4,2), we define the action space as {0,1,2,3,4} × {0,1,2} except (3,2) and (4,2), leading to 13 actions in total.

Also, defining the action by the number of OADs and insulins is reasonable in clinical practice. As specified in the many clinical guidelines, such as Guidelines for the Prevention and Treatment of Type 2 Diabetes in China (2017 edition) [29] , American Diabetes Association (ADA) guidelines [30], and Ministry of Health (MOH) clinical practice guidelines for diabetes mellitus [31], the initial treatment often starts with mono-therapy of OAD. If the patient's glucose level cannot be well-controlled, dual-therapy of OAD is recommended. Basal insulin or premix insulin is the option with better glucose control effect than OAD. Finally, basal insulin together with prandial insulin is used as the intensive treatment.

To further prove the rationality of our action definition, we partition patient visits according to the number of OADs and insulins used in the prescription. For each action-specified group of visits, the corresponding sample size, mean value, and median value of HbA1c at the time of visit are listed in Table III. It shows that the number of OADs in the prescriptions increases with the increment of HbA1c mean and median value. At the same time, the number of insulins increases with the increment of HbA1c mean and median value when the OAD number is fixed. It indicates that the number of OADs and the number of insulins correlates to the treatment of patients with different clinical conditions.

E. Reward Function

Reward plays a key role in RL as it guides the model to choose the best action for a given state. Considering that the primary goal for T2DM is glucose control, our reward function is defined based on change of HbA1c value. According to the Chinese guideline of T2DM [29], the control target of HbA1c for diabetes patients is less than 7%. Thus, an action should be rewarded if the HbA1c in the next state is less than 7%. Also, since the decrease of HbA1c in the next state is beneficial (even if not as low as 7%), it's reasonable that the current action gets a reward as well. Otherwise, the action should get a penalty. Importantly, to avoid the hypoglycemia (i.e., a severe adverse event in the T2DM treatment), we penalize an action if hypoglycemia is reported between visit t and t+1. Formally, we use $r(s_t, s_{t+1})$ to denote the reward of taking an action at state s_t and getting to the next state s_{t+1}. Equation (1) gives the reward function,

$$\begin{aligned} r(s_t, s_{t+1}) &= \alpha\,(\tanh(7- s_{t+1}^{\mathrm{HbA1c}})) + \alpha\,(\tanh(s_t^{\mathrm{HbA1c}} - s_{t+1}^{\mathrm{HbA1c}})) \\ &\quad - (1-\alpha)\, s_{t+1}^{\mathrm{HQ}} \quad \text{, if } t\text{+1 is not the final visit} \\ &= \beta \cdot \mathrm{sgn}(7-s_{t+1}^{\mathrm{HbA1c}}) \quad \text{, otherwise} \end{aligned} \tag{1}$$

where s_t^{HbA1c} is the value of HbA1c (unit: %) of state s_t, s_{t+1}^{HQ} is the binary value representing whether hypoglycemia occurred between visit t and t+1, and α, β are the parameters with positive value. Firstly, for the intermediate visits (i.e., t+1 is not the last step), the reward is the weighted sum of three items: tanh (tangent hyperbolic function) of 7 (upper limit of normal HbA1c range) minus s_{t+1}^{HbA1c}, tanh of s_t^{HbA1c} minus s_{t+1}^{HbA1c}, and hypoglycemia variable s_{t+1}^{HQ}. We apply tanh to make sure that the intermediate rewards are not too large. The factor α gives the relative weight of importance to the first two

TABLE II. DRUG CLASSES OF ORAL ANTIDIABETIC DRUGS AND INSULINS IN THE DATA SET.

Type	Drug class
Oral antidiabetic drug	Metformin
	Sulfonylureas or glinides
	Dipeptidyl peptidase-4
	Alpha-glucosidase inhibitors
	Thiazolidinediones
Insulin	Basal insulin
	Prandial insulin
	Premix insulin

TABLE III. HBA1C DISTRIBUTION OF VISITS USING A DIFFERENT NUMBER OF ORAL ANTIDIABETIC DRUGS AND INSULINS

Insulin number	Oral antidiabetic drug number	Sample percentage	Hemoglobin A1c (%)	
			Mean value	*Median value*
0	0	11.74%	6.74	6.50
0	1	39.40%	6.85	6.54
0	2	21.79%	7.25	6.80
0	> 3	2.73%	7.50	7.00
1	0	10.96%	7.83	7.20
1	1	7.60%	8.00	7.27
1	2	2.58%	8.29	7.50
1	> 3	0.28%	9.11	8.75
2	0	2.10%	8.90	8.10
2	1	0.70%	9.69	9.50
2	2	0.12%	9.75	9.70

items, and 1−α to the third one. Secondly, for the final visits (i.e., t+1 is the last step) whose HbA1c are considered as the final outcomes, a large reward β is given if s_{t+1}^{HbA1c}−7, as well as a penalty of −β if otherwise.

To learn a model that mainly benefits the long-term outcome, we tune parameters α, β so that the samples with good final outcomes get positive rewards. For a sense of scale, the reward of each intermediate visit is limited in the interval (−1, 1). Since there are at most 3 intermediate visits (t=0,1,2) before the last visit, we choose β = 3.0 to ensure that the accumulation of the rewards or penalties of intermediate visits do not exceed that of the final visit. We further optimize the factor α in the range of [0, 1] to obtain the long-term outcome that ensures the convergence of the loss function. Finally, we choose α = 0.8.

F. Model architecture and training

In this study, we tackle the problem of personalized treatment for T2DM patients with the RL approach. As introduced earlier, RL has been applied in various types of applications in healthcare. Following [17, 20] of treating Sepsis in the ICUs, we apply similar value-based RL model, Double Deep-Q-Network (Double DQN) [32], in our work to optimize antidiabetic medication treatment for individual T2DM patients. Double DQN is an extension of simple value based RL model to address several shortcomings in the networks, such as the overestimation problem.

Fig. 2 shows the architecture and training process of our Double DQN model. The model has two architecture-identical neural networks, namely evaluation network and target network. The evaluation network is the main network that will be used to obtain optimal medication after training. The target network is used to estimate the target output value to calculate the loss function. Each network has two hidden layers of 128 units (neurons) with batch normalization and Leaky-ReLU activation.

The neural network gets a state vector as input and a Q vector of action space as output. The state vector is a 35-dimension numeric representation of a patient's clinical status during a certain visit, including demographic information, life style, previous drug usage, etc. The Q vector is the evaluation of all 13 medication actions' effects on the patient's status. In other words, the neural network in our DQN model is used to calculate $Q(s, a)$ of the given state s for all actions a.

When training the Double DQN model, the two networks get different ways of updating their parameters. The evaluation network is trained by optimizing its parameters to minimize the loss function L, while the target network updates its parameters by adding parameters of the evaluation network by a proportion τ periodically. Given states s_t, s_{t+1} and action a_t, the function of L is defined in (2).

$$L = (Q_{\mathrm{true}}(s_t, s_{t+1}) - Q_{\mathrm{eval}}(s_t, a_t))^2 + \lambda \max(|Q_{\mathrm{eval}}(s_t, a_t)| - r_{\mathrm{reg}}, 0) \quad (2)$$

where

$$Q_{\mathrm{true}}(s_t, s_{t+1}) = r(s_t, s_{t+1}) + \gamma \cdot Q_{\mathrm{target}}(s_{t+1}, \mathrm{argmax}_a(Q_{\mathrm{eval}}(s_t, a))) \quad (3)$$

, r_{reg} = 3.0 is the maximum reward of all $r(s_t, s_{t+1})$ which is used in the regularization term to penalize excessively large Q value, and Q_{eval}, Q_{target} are the Q value output of the evaluation and target networks. To speed up the training approach, we use Prioritized Experience Replay method [33]: instead of uniformly sampling, the input data of each training batch is sampled according to the samples' importance, which measured by their temporal-difference error.

We attempt to apply multiple combination of neural network training hyper parameters on our model. Finally, we select the target network update parameter τ equals to 0.01, the learning rate equals to 0.001, and training batch size equals to 128. We train our DQN model for maximum 4,000 iterations using the Adam Optimizer [34]. Given the state s of a patient visit, the evaluation network gives Q outputs as expected outcome for all 13 actions. Then, the action with the largest Q, i.e., $\mathrm{argmax}_a(Q_{\mathrm{eval}}(s_t, a))$, is recommended as the optimal medication policy for the patient visit.

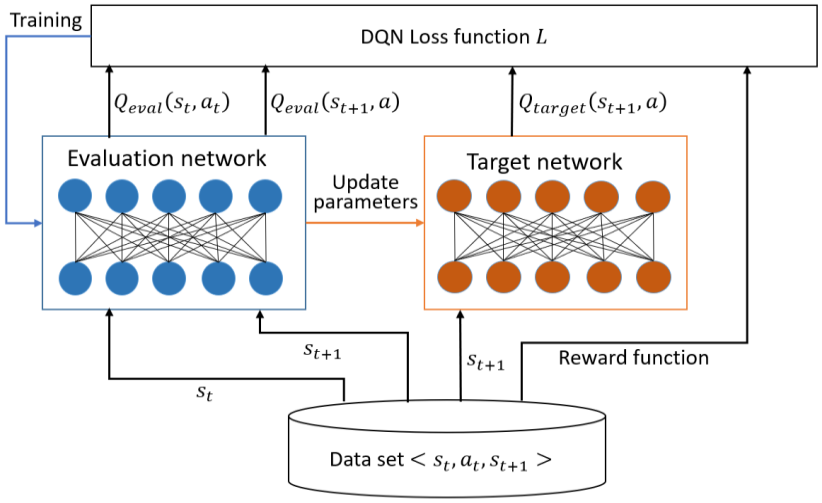

Fig. 2 The architecture and training process of the Double DQN model. The blue network in block represents the evaluation network, and the brown one represents the target network. The arrows between blocks represent the flows of variables.

III. Results

To evaluate our DQN model on both short-term and long-term clinical outcomes, we choose 2,847 patients that have their final visits at the 12th month as test set, corresponding to 10,741 triples. The demographics, life styles and medical history of the 2,847 patients are described in Table IV.

After applying our DQN model on the test set, 53.66% of the samples' actual prescriptions are concordant with DQN recommendations according to our definition of actions. It is worth mention that the reinforcement learning method does not simply mimic the physician's prescription (i.e., not targeting on 100% concordant). Instead, it targets to learn the prescription patterns that lead to good clinical outcome. As shown in Fig. 3, we compare physician's policy with DQN policy. All test samples are grouped based on the current HbA1c level defined previously in Methods, namely *Low*, *Med*, and *High*, and colored differently. For each group, the upper two-dimension heatmap represent policy of physicians and the lower one represents DQN policy. The x-axis in the two-dimension heatmap represents the number of OADs and y-axis represents the number of insulins. The darkness of the color is in proportion to the sample size of corresponding action. As shown in Fig. 3, physicians' policy with good clinical outcomes are consistent with clinical knowledge, as the majority of patients in *Low* group are mainly prescribed with a single OAD, while for patients in *Med* and *High* group, dual-OADs, insulin, and insulin plus OADs are increasingly used. Moreover, the DQN policy is similar to the physician's policy at the first glance, which indicates the DQN model somehow learns the good prescription pattern from the physicians.

Next, we evaluate the effectiveness of our DQN model for both short-term and long-term clinical outcomes (Table V). We partition all the data samples into three groups (i.e., *Low*, *Med*, and *High*) based on HbA1c levels defined previously in

TABLE IV. Demographics, Life Style, and Medical History of Patients in the Data Set (*N*=2,847).

Feature	Value[a] or percentage
Age (years)	57.3 (48.9-65.7)
Female, *N* (%)	1,452 (51.0)
Smoker or ex-smoker, *N* (%)	850 (29.9)
Physical activity quantity, *N* (%)	
No physical activity	598 (21.0)
Less than 5 times a week	1,207 (42.4)
At least 5 times a week	1,042 (36.6)
Days since first diagnosis in EHR	65.7 (7.3-138.7)
Hypertension, *N* (%)	1,185 (41.6)
Dyslipidemia, *N* (%)	1,359 (47.7)
Kidney disease, *N* (%)	116 (4.1)
With at least 1 complication, *N* (%)	324 (11.4)
HbA1c (%)	7.3 (6.5-9.2)
Fasting blood glucose (mmol/L)	7.8 (6.5-10.2)
Body mass index (kg/m^2)	24.77 (22.86-26.95)
Systolic blood pressure (mmHg)	129 (120-137)
LDL-c (mmol/L)	2.78 (2.21-3.35)
HDL-c (mmol/L)	1.20 (1.01-1.41)
Serum Creatinine (μmol/L)	67.7 (56.0-82.5)

[a] Continues variables are in the form of "median (interquartile range)".

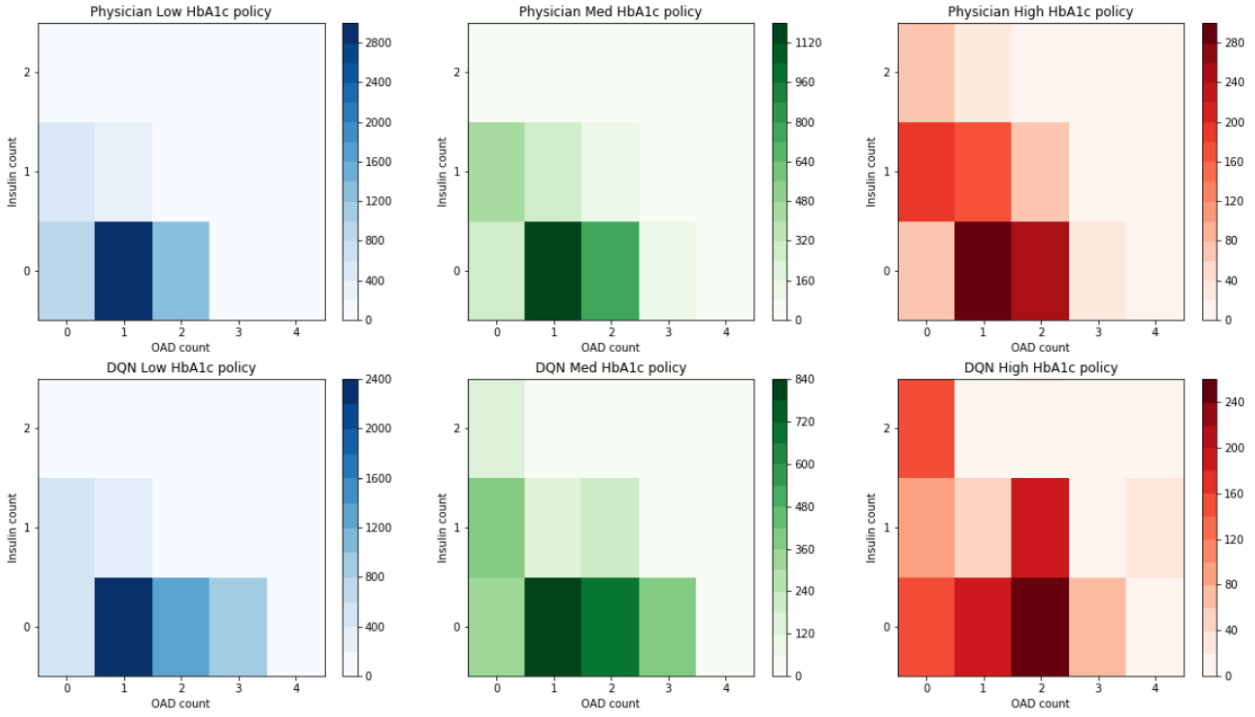

Fig. 3. Two-dimension heatmaps of the treatment policies made by the physicians (the upper panel) and by the DQN agent (the lower panel). The blue, green, and red heatmaps represent HbA1c level of *Low*, *Med,* and *High* respectively.

TABLE V. FINAL AND NEXT HBA1C-GOAL ACHIEVING RATES OF DIFFERENT PATIENT GROUPS AND DQN-CONCORDANCE

Patient visit groups			Final HbA1c-goal achieving (long-term clinical outcome)				Next HbA1c-goal achieving (short-term clinical outcome)			
HbA1c[a] level	DQN[b]-concordant	Number of visits	Visit count	Rate	Difference of rates	P-value	Visit count	Rate	Difference of rates	P-value
Low	Yes	3,785	3,246	85.76%	8.14%	<0.001	3,291	86.95%	4.26%	<0.001
	No	2,507	1,946	77.62%			2,073	82.69%		
Med	Yes	1,609	863	53.64%	12.53%	<0.001	752	46.74%	10.84%	<0.001
	No	1,652	679	41.10%			593	35.90%		
High	Yes	370	213	57.57%	18.40%	<0.001	154	41.62%	14.45%	<0.001
	No	817	320	39.17%			222	27.17%		
All visits	Yes	5,764	4,322	74.98%	15.79%	<0.001	4,197	72.81%	14.76%	<0.001
	No	4,977	2,946	59.19%			2,889	58.05%		

[a] HbA1c is the abbreviation of hemoglobin A1c. [b] DQN is the abbreviation of deep Q network.

Methods section. For each group, we further partition them into DQN-concordant and non-concordant subset, and compare their clinical outcomes in terms of the HbA1c-goal achieving rate, which is computed as the number of the samples with HbA1c<7% in the future visit over the total number of samples. We perform such comparison based on the HbA1c in the next visit (i.e., short-term clinical outcome) and in the final visit (i.e., long-term clinical outcome), respectively. As shown in Table V, for all settings, the clinical outcomes of DQN-concordant samples are significantly better than those of non-concordant ones (p-value of Chi-square test is less than 0.001). For all patient visits, the difference of final HbA1c-goal achieving rates is 15.79%, and the difference of next HbA1c-goal achieving rate is 14.67%. Besides, compared with the short-term clinical outcome, the DQN-model achieves better performance on the long-term clinical outcome (referring to the difference of HbA1c-goal achieving rate between DQN-concordant group and non-concordant group). This also confirms that the model learned by the reinforcement learning approach favors optimizing the long-term clinical outcome.

Further, we perform the long-term clinical outcome evaluation at patient level. For chronic diseases like T2DM, a patient's final state could be related to all the prescriptions from his previous visits. Here, we introduce a new concept called model-concordant proportion to represent the relation between a sequence of DQN recommendations and a list of actual prescriptions from the patient visits. model-concordant proportion of a given patient is defined as the number of the patient visits with the actual prescription in concordant with model recommendation, over the total number of visits of this patient. From all 2,847 patients in the test set, there are 2,499 patients who have a baseline and all 4 follow-ups. Given a patient, we compute the model-concordant proportion according to his/her first 4 visits and use the final HbA1c-goal achieving as the outcome.

To demonstrate the effectiveness of our DQN model at patient level, we compare the treatment recommendation of our model with two supervised learning methods: kNN and Random forest. Rather than mimicking physician's prescriptions by using kNN or Random forest, our DQN model learns from good practices from clinicians and penalizes those actions that worsen the clinical outcomes. Thus, the overall concordant rate of our DQN model (53.66%) is smaller than those of kNN and Random forest, which are 74.38% and 86.67% respectively.

Fig. 4 shows the relationship between model-concordant proportion and the finial HbA1c-goal achieving rate for the DQN, kNN and Random Forest models. The overall trends of the three models are gradually rising, which indicates that the higher model-concordant proportion rate is, the greater chance the patient achieves her/his HbA1c-goal at the end. More importantly, compared with kNN and Random forest, our DQN model achieves much better performance at patient level (the red line in Fig. 4), gaining 40% increase of final HbA1c goal-achieving rate from 0% model-concordant proportion to 100%.

In addition, the treatments recommended by our model lead to less hypoglycemia events. We obtain from the data

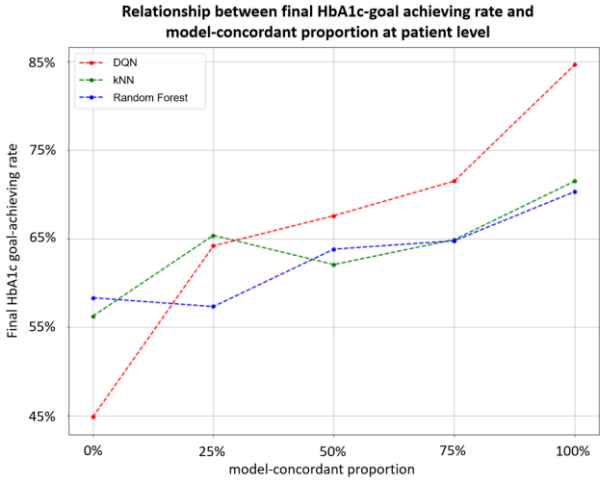

Fig. 4. Relationship between final HbA1c goal-achieving rate and model-concordant proportion. Each patient gets a model-concordant proportion, computed as the ratio between the number of the visits with model-concordant medications and the total number of the visits (i.e., 4), where 0% represents none is model-concordant and 100% represents all are model-concordant. The upward trend of the line chart shows that our model is effective for the long-term treatment.

TABLE VI. FINAL AND NEXT HBA1C-GOAL ACHIEVING RATES OF DIFFERENT PATIENT GROUPS AND DQN-CONCORDANCE

DQN[a]-concordant	Number of visits	Visit count	Hypoglycemia event occurrence rate	Difference of rates	P-value
Yes	5,764	183	3.17%	-1.29%	<0.001
No	4,977	222	4.46%		

[a] DQN is the abbreviation of deep Q network.

whether a patient had any hypoglycemia events within three months after the current visit, including sudden sweating, rapid hunger, dizziness, unconsciousness, awkward, trembling activity, etc. Table VI shows that the hypoglycemia event occurrence rate of DQN-concordant visits is significantly smaller than that of non-concordant visits (difference of rates: -1.29%, Chi-square test p-value < 0.001).

IV. DISCUSSION

During T2DM treatment, the intensive therapy often leads to good glucose control but more hypoglycemia events [33]. Therefore, it is challenging to control blood glucose to a normal range and lower hypoglycemia occurrence rate at the same time. The comprehensive evaluation results show that model-concordant treatments are associated with better glycemic control in all HbA1c-stratified subgroups, and meanwhile causes less hypoglycemia events.

To the best of our knowledge, our study is the first to tackle the treatment of T2DM using the RL approach. The strengths of this study are twofold. Firstly, the T2DM patient registry data used for model building and evaluation is of good quality. They are from a prospective study encompassing a large patient population with regular follow-up visits. Secondly, to demonstrate the effective of the RL-based model, we evaluate multiple types of outcomes, namely the glycemic control in the short-term and the long-term, as well as short-term hypoglycemia events.

We nonetheless acknowledge limitations. First of all, our study adopts a uniform HbA1c control target, (i.e., 7%) as recommended by the Chinese T2DM control guideline [30]. In future work, we may extend this uniform HbA1c control target to a personalized HbA1c target for individual patients. For instance, the HbA1c target may be less stringent for the elderly and those with recurrent hypoglycemia events. Secondly, the data used in our model are equal time intervals. In order for our method to be widely adopted in other clinical setting, there is a need to consider unequal time intervals since patient hospital visits are on demand and irregular rather than fixed duration. Irregular visit data could still be added on to our method by weighting the time impact coefficient to the reward design corresponding to various time intervals.

Our work can be further extended to different scenarios. First, our current RL model is used to learn the optimal treatment strategy for blood glucose control, while it can also be applied for complication control of T2DM patients (e.g., stroke and diabetic kidney disease) by setting the occurrence of complications as the long-term outcome. Secondly, our methods can be integrated with medical knowledge as constraints to enhance guideline adherence. To be more specific, the medication recommendations from our model could be filtered by knowledge extracted from clinical guidelines so as to pick best candidate action in the RL model. Last but not least, our RL method can be generalized to other chronic diseases with customization of state, action, reward and other related parameters.

V. CONCLUSION

In this paper, we developed a reinforcement learning method with the DQN model for T2DM treatment recommendation. The effectiveness of our approach is validated with the NEW2D data set from a prospective study. Our experiments have demonstrated that treatments that are in agreement with our RL model's recommendation not only achieved better clinical outcomes in terms of HbA1c control, but they also lead to better long-term glucose control after one year. Our experimental results also suggest that our RL method is capable of learning the good medication pattern from physicians that leads to effective long-term glycemic control.

ACKNOWLEDGMENT

This work was funded by the China National Key R&D Program (Grant No. 2018YFC0910700).

Siqi Liu was funded by the NUS Graduate School for Integrative Sciences and Engineering Scholarship (NGSS). This research was partially supported by the National Research Foundation Singapore under its AI Singapore Programme [Award No. AISG-GC-2019-002] and the NMRC Health Service Research Grant [HSRG-OC17nov004].

REFERENCES

[1] World Health Organisation, "Global report on diabetes," ISBN 978-92-4-156525-7, 2016, Available from: http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf

[2] Z. Chen, K. Marple, E. Slazar, and G. Gupta, "A physician advisory system for chronic heart failure management based on knowledge patterns," Theory and Practice of Logic Programming, vol. 16(5-6), pp. 604-618, 2016, DOI: 10.1017/S1471068416000429

[3] K. Hannes, M. Leys, E. Vermeire, B. Aertgeerts, F. Buntinx, and A-M. Depoorter, "Implementing evidence-based medicine in general practice: a focus group based study," BMC family practice, vol. 6(1), p. 37, 2005, PMID: 16153300

[4] A. Hutchinson, and R. Baker, "Making use of guidelines in clinical practice," Radcliffe Publishing, 1999, PMID: 10521225

[5] J.C. Weiss, S. Natarajan, P. L. Peissig, C. A. McCarty, and D. Page, "Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records," AI Magazine, vol. 33(4), pp. 33-45, 2012.

[6] R. Iniesta, D. Stahl, P. McGuffin, "Machine learning, statistical learning and the future of biological research in psychiatry," Psychological medicine, vol. 46(12), pp. 2455-2465, 2016

[7] R. C. Kessler, H. M. van Loo, K. J. Wardenaar, R. M. Bossarte, L. A. Brenner, *et al.*, "Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports," Molecular psychiatry, vol. 21, pp. 1366-1371, 2016.

[8] P. Lambin, R. T.H. Leijenaar, T. M. Deist, J. Peerlings, E. E.C. de Jong, *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," Nature Reviews Clinical Oncology, vol. 14, pp. 749-762, 2017.

[9] K. Donsa, S. Spat, P. Beck, T. R. Pieber, and A. Holzinger, "Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges," Smart Health. Springer, Cham, pp. 237-260, 2015.

[10] M. J. Patel, C. Andreescu, J. C. Price, K. L. Edelman, C. F. Reynolds III, and H. J. Aizenstein, "Machine learning approaches for integrating clinical and imaging features in late‐life depression classification and response prediction," International journal of geriatric psychiatry, vol. 30(10), pp. 1056-1067, 2015.

[11] D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo, "Personalized diabetes management using electronic medical records," Diabetes care, vol. 40(2), pp. 210-217, 2017.

[12] R. Chen, and I. Paschalidis, "Learning Optimal Personalized Treatment Rules Using Robust Regression Informed K-NN," arXiv preprint, 2018, arXiv:1811.06083

[13] Y. Wang, H. Fu, and D. Zeng, "Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies," Journal of the American Statistical Association, vol. 113(521), pp. 1-13, 2018.

[14] F-Y. Wang, J. J. Zhang, X. Zheng, X. Wang, Y. Yuan, *et al.*, "Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond," IEEE/CAA Journal of Automatica Sinica 3.2, 2016, pp. 113-120.

[15] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," The International Journal of Robotics Research, vol. 32(11), pp. 1238-1274, 2013.

[16] A. E.L. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," Electronic Imaging, Autonomous Vehicles and Machines, 2017, pp. 70-76.

[17] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi, "Deep reinforcement learning for sepsis treatment," arXiv preprint, 2017, arXiv:1711.09602.

[18] R. Padmanabhan, N. Meskin, and W. M. Haddad, "Optimal adaptive control of drug dosing using integral reinforcement learning," Mathematical biosciences, vol. 309, pp. 131-142, 2019

[19] M. M. Ghassemi, T. AlHanai, M. B. Westover, R. G. Mark, and S. Nemati, "Personalized medication dosing using volatile data streams," The Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018. Available from: https://www.aaai.org/ocs/index.php/WS/AAAIW18/paper/viewPaper/17234

[20] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care," Nature medicine, vol, 24(11), pp. 1716-1720, 2018

[21] N. Prasad, L-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," arXiv preprint, 2017, arXiv:1704.06300

[22] L-F. Cheng, N. Prasad, and B.E. Engelhardt, "An Optimal Policy for Patient Laboratory Tests in Intensive Care Units," PSB, World Scientific, 2019.

[23] P. D. Ngo, S. Wei, A. Holubová, J. Muzik, and F. Godtliebsen, "Reinforcement-learning optimal control for type-1 diabetes," IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2018, pp. 333-336.

[24] M. O.M. Javad, S. Agboola, K. Jethwani, I. Zeid, and S. Kamarthi, "Reinforcement learning algorithm for blood glucose control in diabetic patients," ASME 2015 International Mechanical Engineering Congress and Exposition, 2015.

[25] Y. Zhao, M. R. Kosorok, and D. Zeng, "Reinforcement learning design for cancer clinical trials," Statistics in Medicine, vol. 28, no. 26, pp.3294–3315, 2009.

[26] A. Jalalimanesh, H. S. Haghighi, A. Ahmadi, and M. Soltani, "Simulation-based optimization of radiotherapy: Agent-based modelingand reinforcement learning," Mathematics and Computers in Simulation, vol. 133, pp. 235–248, 2017

[27] F. Lv, X. Cai, D. Hu, C. Pan, D. Zhang, *et al.*, "Characteristics of Newly Diagnosed Type 2 Diabetes in Chinese Older Adults: A National Prospective Cohort Study," Journal of Diabetes Research, 2019.

[28] D. M. Nathan, H. Turgeon, and S. Regan, "Relationship between glycated hemoglobin levels and mean glucose levels over time," Diabetologia, vol. 50(11), pp. 2239-2244, 2007.

[29] Chinese diabetes society, "Chinese guideline for prevention and treatment of type 2 diabetes mellitus (version 2017)," Chinese Journal of Diabetes Mellitus, vol. 10(1), pp. 4-67, 2018.

[30] American Diabetes Association, "Pharmacologic approaches to glycemic treatment: Standards of Medical Care in Diabetes-2018," Diabetes Care, vol. 41(Suppl 1), pp. S73-S85, 2018.

[31] S. Y. Goh, S. B. Ang, Y. M. Bee, R. Y.T. Chen, D. Gardner, *et al.*, "Ministry of Health clinical practice guidelines: diabetes mellitus," Singapore Medical Journal, vol. 55(6), pp. 334-347, 2014.

[32] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," Thirtieth AAAI Conference on Artificial Intelligence Proc, pp. 2094-2100, 2016.

[33] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," arXiv preprint, 2016, arXiv:1511.05952

[34] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint, 2014, arXiv:1412.6980

[35] C. V. Desouza, G. B. Bolli, and V. Fonseca, "Hypoglycemia, Diabetes, and Cardiovascular Events," Diabetes Care, vol. 33(6), pp. 1389-1394, 2010.