Artificial Intelligence In Medicine 104 (2020) 101836

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Reinforcement learning application in diabetes blood glucose control: A systematic review

Miguel Tejedor[a,*], Ashenafi Zebene Woldaregay[a], Fred Godtliebsen[b]

[a] Department of Computer Science, University of Tromsø-The Arctic University of Norway, Norway
[b] Department of Mathematics and Statistics, University of Tromsø-The Arctic University of Norway, Norway

ARTICLE INFO

Keywords:
Reinforcement learning
Blood glucose control
Artificial pancreas
Closed-loop
Insulin infusion

ABSTRACT

Background: Reinforcement learning (RL) is a computational approach to understanding and automating goal-directed learning and decision-making. It is designed for problems which include a learning agent interacting with its environment to achieve a goal. For example, blood glucose (BG) control in diabetes mellitus (DM), where the learning agent and its environment are the controller and the body of the patient respectively. RL algorithms could be used to design a fully closed-loop controller, providing a truly personalized insulin dosage regimen based exclusively on the patient's own data.

Objective: In this review we aim to evaluate state-of-the-art RL approaches to designing BG control algorithms in DM patients, reporting successfully implemented RL algorithms in closed-loop, insulin infusion, decision support and personalized feedback in the context of DM.

Methods: An exhaustive literature search was performed using different online databases, analyzing the literature from 1990 to 2019. In a first stage, a set of selection criteria were established in order to select the most relevant papers according to the title, keywords and abstract. Research questions were established and answered in a second stage, using the information extracted from the articles selected during the preliminary selection.

Results: The initial search using title, keywords, and abstracts resulted in a total of 404 articles. After removal of duplicates from the record, 347 articles remained. An independent analysis and screening of the records against our inclusion and exclusion criteria defined in Methods section resulted in removal of 296 articles, leaving 51 relevant articles. A full-text assessment was conducted on the remaining relevant articles, which resulted in 29 relevant articles that were critically analyzed. The inter-rater agreement was measured using Cohen Kappa test, and disagreements were resolved through discussion.

Conclusions: The advances in health technologies and mobile devices have facilitated the implementation of RL algorithms for optimal glycemic regulation in diabetes. However, there exists few articles in the literature focused on the application of these algorithms to the BG regulation problem. Moreover, such algorithms are designed for control tasks as BG adjustment and their use have increased recently in the diabetes research area, therefore we foresee RL algorithms will be used more frequently for BG control in the coming years. Furthermore, in the literature there is a lack of focus on aspects that influence BG level such as meal intakes and physical activity (PA), which should be included in the control problem. Finally, there exists a need to perform clinical validation of the algorithms.

Abbreviations: AC, actor-critic; ADP, model-free approximate/adaptive dynamic programming; AP, artificial pancreas; AR, average-reward; BAL, Bayesian active learning; BG, blood glucose; CALA, continuous action-set learning automata; CGM, continuous glucose monitoring; CHO, carbohydrate; CONT, continuous; CVGA, control variability grid analysis; DISC, discrete; DM, diabetes mellitus; DP, dynamic programming; DQN, deep Q-network; GP, Gaussian process; GPRL, Gaussian processes reinforcement learning; HBGI, high blood glucose index; IHD, infinite-horizon discounted; LBGI, low blood glucose index; LSMDP, linearly-solvable Markov decision process; MAGE, mean amplitude of glucose excursion; ML, machine learning; PA, physical activity; RL, reinforcement learning; RLFF, reinforcement learning with feedforward; RLOC, reinforcement-learning optimal control; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus; TDI, total daily insulin; TG, truncated Gaussian; TIR, time in range; UN, uniform

* Corresponding author.
E-mail address: miguel.tejedor@uit.no (M. Tejedor).

https://doi.org/10.1016/j.artmed.2020.101836
Received 30 July 2018; Received in revised form 3 August 2019; Accepted 19 February 2020

1. Introduction

Diabetes Mellitus (DM) is characterized by chronic high blood glucose (BG) level as a consequence of a metabolic disorder that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces, leading to long-term damage, dysfunction and failure of various organs [1]. According to the International Diabetes Federation approximately 1 in 11 adults has diabetes, which means 425 million adults worldwide suffered from these conditions in 2017. This represents 9.1 % of the adult population, while trends suggest the rate would continue to rise. Furthermore, DM at least doubles a person's risk of early death, resulting in approximately 1.5–5.0 million deaths each year, while 12 % of global health expenditure is spent on diabetes ($727 billion) [2]. Because of the high incidence and prevalence of diabetes, the share of research devoted to the disease is continuously increasing [3].

There exist three main types of diabetes: Type 1 Diabetes Mellitus (T1DM), in which the patient presents a deficient insulin production and requires daily administration of insulin, Type 2 Diabetes Mellitus (T2DM), characterized by an ineffective use of insulin in the body, and gestational diabetes, produced by a high BG levels during pregnancy. All of them require continuous management from patients and physicians in order to avoid complications [1].

Recent technological advances in medical wearable devices and sensor technologies, as well as the increase of processing power in mobile phones, have made an extensive acceleration of research activities possible in all aspects of diabetes. This new scenario has led to the application of machine learning (ML) and data mining techniques in the DM research field [4], with BG prediction appearing to be the most popular focus [5], indicating that artificial intelligence is increasingly common in DM solutions [6]. Among DM management tasks, the development of BG control strategies has been one of the most important issues during the last years [7]. For this reason, the design of control algorithms for DM is a very active research area approached from many different angles by a large number of scientists in different fields. Furthermore, there is a great need for more data-driven control strategies in this problem and the disadvantages of traditional algorithms suggest the use of data-driven ML algorithms [8]. Among these, reinforcement learning (RL) algorithms provide a highly promising approach that has been increasingly adopted in the area of control algorithms. Indeed, over the last few decades, RL has offered an appealing framework for the treatment and long-term management of chronic diseases. In this review, the goal is to analyze and assess existing RL algorithms for a closed-loop controller in DM.

2. Diabetes and blood glucose control using reinforcement learning

DM is often self-managed by the patient through multiple glucose level measurements throughout the day and administration of insulin via injection or a pump, which become a really challenging task for the patients, who have to deal with many complications during their daily life. Even with a due amount of vigilance, many patients may still suffer significant diabetes-associated complications. This traditional and manual BG control framework is shown in Fig. 1.

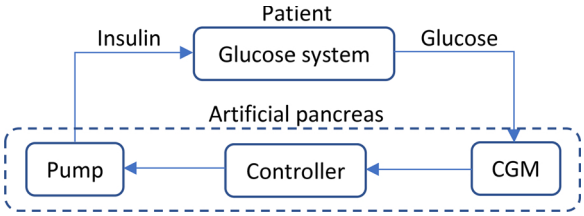

Fig. 2. Blood glucose management based on artificial pancreas.

The artificial pancreas (AP) offers an efficacious and safe approach for treating DM [9], therefore it has become the holy grail of diabetes research [10]. The successful development of an AP consists of three primary components: a continuous glucose monitoring (CGM) system to continuously measure BG every five minutes or monitor glucose readings over a period of time, an insulin pump that can deliver precise amounts of insulin, and a control algorithm that translates data streaming from CGM into instructions for insulin pump. While the first two components have seen rapid technological gains in recent years, state-of-the-art controllers still require regular patient or caregiver intervention, operating in open-loop control with the user. Fig. 2 shows a flowchart of the artificial pancreas BG control framework. This is a closed-loop model [11], where BG levels are measured by the CGM and, based on glucose concentrations, the controller determines the proper amount of insulin needed. This insulin dosage is applied by the insulin pump, affecting glucose system and changing BG level. Based on the changes produced in BG concentration, a new insulin dosage is calculated and applied. This process implies that only information measured from the patient is used to make decisions by the controller, without knowledge of external data [12].

This framework can be extended to a broader scope using mobile communication and wearables devices for health services, information, and data collection, obtaining a complete mHealth system [13]. The system would be able to monitor the patient physiological status while supervising the healthcare plan, allowing to include additional relevant information for diabetes care, such as food intake, physical activity (PA), infections and stress level.

The principle of RL is based on the interaction between a decision-making agent and its environment [14]. In RL, the goal is to train an agent to take actions that result in preferable states. At each decision time point, the agent chooses an action for some given current state of the system. The environment reacts to this action and transitions to a new state. For the previous action taken, the agent now receives a positive or negative reinforcement from the environment. The mapping of state to action is called the policy. The goal of RL is to learn an optimal policy that maximizes the amount of rewards it receives over time. Fig. 3 shows this RL framework, where the agent is the decision maker and learner while the environment is the thing the agent interacts with, encompassing everything outside the agent [14].

Furthermore, in this framework there are additional sub-elements: the policy defining the behavior of the agent, the reward function defining the goal of the problem and the value function specifying the long-term desirability of states. Concretely, the value function indicates

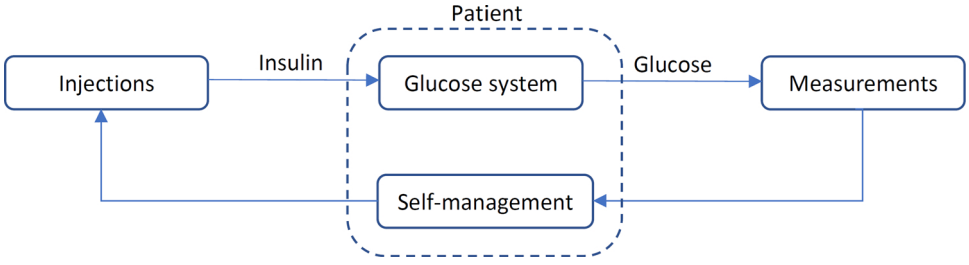

Fig. 1. Self-managed blood glucose control.

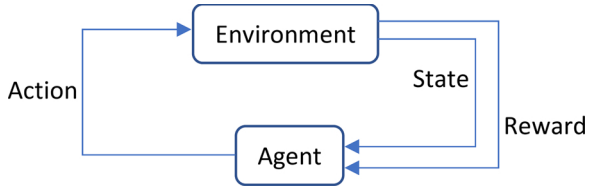

Fig. 3. Reinforcement learning framework.

the total amount of reward expected by an agent when it starts from a given state and follows a given policy thereafter. Similarly, the action-value function indicates the total amount of reward expected by an agent when it starts from a given state, takes a given action and follows a given policy thereafter. Finally, some problems have the model of the environment, a sub-element predicting future states and rewards [14].

Several approaches have been used in the literature in order to reach the RL goal: learn the optimal policy, which is the policy that is better than or equal to all other policies based on the values of the states. This have originated many RL methods such as temporal-difference learning, which learn by bootstrapping and perform updates from the current estimate of the value function, or actor-critic (AC) learning, which are algorithms formed by two different parts: an *actor* following a policy to select actions and a *critic* used to estimate the value function and criticizes the actions taken by the *actor*. Therefore these algorithms are characterized by a separate memory structure to explicitly represent the policy independent of the value function [14].

In the DM reinforcement learning task, the interstitial glucose curve is taken to be the state variable, as measured by the CGM. The action space consists of insulin dosage amounts. The agent is the controller. The environment is the patient's glucose system. Finally, the reward function should measure the discrepancy between ideal and actual glucose levels.

RL is particularly suited to situations where decisions are made sequentially along a timeline, actions depend on the observed state, effects manifest at later points in time than the actions that induced them (time delay), and there is some notion of preferred state(s). These features are certainly present in the DM controller challenge.

Another advantage is that modeling the glucose-insulin dynamics can be entirely bypassed in RL. Furthermore, labeled training data is not required as in supervised learning strategies, but instead the agent can learn optimal policies without the necessity of first being trained on examples of "correct" actions to take.

RL algorithms are uniquely suited to problems with inherent time delays. This presents a strong advantage in the diabetes application due to the time lags in both continuous glucose monitors (which actually measures subcutaneous glucose measurements) and insulin effect. RL naturally accommodates for these time delays because actions are allowed to have delayed effects and rewards are given for good behavior in the long run.

Finally, this algorithm continuously adapts and evolves with the user, which leads to a truly personalized analysis. In contrast, traditional statistics and ML often operate by borrowing strength across subjects. Additional convincing arguments for the use of RL in the DM scenario are given in [8].

3. Methods

The purpose of the review is to identify, assess and analyze the state-of-the-art RL algorithms and strategies focusing on its applications towards BG control in people with diabetes. As a result, a comprehensive literature search was conducted from 5th June 2019 to 3rd August 2019. The search was performed using different online databases such as ACM digital library, DBLP Computer Science Bibliography, Google Scholar, IEEE Xplore, Journal of American Medical Informatics Association (JAMIA), PubMed and ScienceDirect. Relevant papers were further extracted from the reference lists of the selected articles. The search process covers a specified timeframe from 1990 to 2018 and considered peer reviewed journal articles and conference proceedings. The search was conducted using different combination of strings along with "reinforcement learning" including "artificial pancreas", "blood glucose control", "closed-loop in diabetes", "decision making in diabetes", "decision support in diabetes", "insulin infusion", "insulin pump" and "personalized feedback in diabetes". For the purpose of effective searching strategy, the search strings were combined using Boolean function such as "And" and "Or". During the search, relevant articles were identified by reviewing the title, keywords, and abstracts for a preliminary filter based on the inclusion and exclusion criteria. A full-text assessment was done on only articles that seemed relevant according to our inclusion and exclusion criteria. Information extraction were also done based on some structured predefined categories that is in line with our inclusion and exclusion criteria, which were defined based on discussions and brainstorming among the authors.

3.1. Inclusion and exclusion criteria

To be considered in this review, the study should develop and test RL algorithms and strategies based on people with diabetes and in addition fulfil the following conditions: focus on BG control and be published between 1990 and 2019.

As a result, studies outside of the stated scope were excluded from the review including all studies presented in other languages than English.

3.2. Data categorization and data collection

Extraction of information from the selected studies was conducted using some predefined and structured categories, which were defined based on discussions and brainstorming among the authors. The categories were defined to fully assess and evaluate the state-of-the-art of RL algorithms and strategies developed and tested on BG control for people with diabetes.

3.2.1. Subjects

This category defines the nature and characteristics of the subject used in algorithm development and testing, which includes age, gender, type of DM and nature of the subjects; in silico and real subjects.

3.2.2. Data sources

This category defines different kind of data sources the studies have used to develop and test the RL algorithms, which include data sources like CGM devices, insulin pumps, different BG dynamics simulators and others.

3.2.3. Preprocessing

This category defines the kind of preprocessing performed on the raw data and the various approaches employed in the processes, including glycemic ranges, sparsification (detecting novel information) and others.

3.2.4. RL approach

This category defines the reinforcement algorithm approach used to develop the control algorithm, including tabular solution methods and approximate solution methods.

3.2.5. Class of RL

This category defines the class of RL algorithms used to develop and test the control algorithm, which includes AC learning, Q-learning, Sarsa and others.

3.2.6. Exploitation versus exploration

This category encompasses the exploitation-exploration dilemma in

RL algorithms, which involves making the best decision given the current information or gathering more information with sacrifices for a long-term benefit. In this regard, it pinpoints the approached favored by the studies to solve the dilemma.

3.2.7. State space

This category encompasses the definition of the state space, its nature and defining parameters used in the control algorithms, that is the actual situation of the environment in which the agent finds itself. The nature of the state space is either continuous or discrete. The defining parameters include key diabetes parameters such as BG, insulin, diet, PA and others.

3.2.8. Action space

This category encompasses the definition of the action space, its nature and defining parameters, which is a set of all possible actions the agent is entitled to choose. The nature of the action space is either continuous or discrete. The defining parameters include different actions such as insulin dose, food intake, PA and others.

3.2.9. Planning

This category encompasses the planning techniques used in the reinforcement algorithms. It includes either a model-based or model-free approach.

3.2.10. Generalization approaches

This category determines the approaches to address the problem of learning in large spaces. Among these techniques we can find policy gradient method, Gaussian process (GP) regression and others.

3.2.11. Performance metrics or evaluation criteria

This category defines performance metrics the studies have used to evaluate the developed BG control algorithms. It includes different approaches such as predefined target ranges, control variability grid analysis (CVGA), comparison with reference value and others.

3.2.12. Model of optimal behavior

This category considers the different models of optimality, where there are three main models in this area: the finite-horizon model, the infinite-horizon discounted (IHD) model and the average-reward (AR) model.

3.2.13. Reward function

This category defines the kind of reward function used to develop the control algorithms, which measures the success or failure of an agent according to a set of chosen actions. A reward is defined based on the objective of the task at hand and the expert knowledge. As a result, various kinds of reward functions have been defined in the literature and this category pinpoint widely adopted reward functions.

3.3. Literature evaluation

Papers were evaluated based on the above predefined categories to evaluate the state-of-the-art approaches and strategies used in RL algorithms for BG control in people with diabetes. The first evaluation and analysis was done based on data characteristics including data sources, subjects and preprocessing approach. The second evaluation and analysis were conducted based on RL strategies including class of RL algorithms and its approaches. The third analysis was carried out based on exploitation versus exploration, to reveal the state-of-the-art approaches in solving the dilemma involved. The fourth evaluation and analysis was conducted based on state and action space including their respective nature and defining parameters. The fifth evaluation and analysis was carried out based on planning approaches employed during development. The sixth evaluation and analysis was conducted based on reward function used to learn the agent. Note that the number of features extracted might exceed the number of reviewed articles since many features are reported in the literature. Therefore, the number of findings in each category might vary from the number of total studies included in the review, since more than one approach can be considered in the same article.

4. Results

4.1. Relevant literatures

RL is a quickly growing field, and its application to diabetes BG control is growing even more rapidly, as found in the literature publication dates, with only 2 publications before 2012 while 27 publications between 2012 and 2019. From those articles, 8 were published in just the last year.

The initial search using title, keywords, and abstracts resulted in a total of 404 articles. After removal of duplicates from the record, 347 articles remained for further analysis. An independent analysis and screening of the records against our inclusion and exclusion criteria resulted in removal of 296 articles, leaving 51 relevant articles. A full-text assessment was conducted on the remaining relevant articles, which resulted in 29 relevant articles that were critically analyzed as shown in Fig. 4 below. The inter-rater agreement was measured using Cohen Kappa test [15], and any differences were resolved through discussion among the authors.

4.2. Evaluation of literature

The reviewed articles are evaluated, as described earlier, based on the above predefined categories. The results obtained are showed below in Table 1.

4.3. Data characteristics

4.3.1. Subjects

The reviewed articles are mainly based on real and in silico (simulated) subjects for T1DM and/or T2DM, as shown in Table 1 above. Almost all studies developed and tested algorithms for T1DM (82.75 %, 24/29), while only 2 studies (6.9 %) are based on T2DM, 2 other studies (6.9 %) consider both types of diabetes, and 1 study (3.45 %) does not specify the type of diabetes. Moreover, most of the studies (76.67 %, 23/30) have relied on in silico subjects and only 20 % of the studies (6/

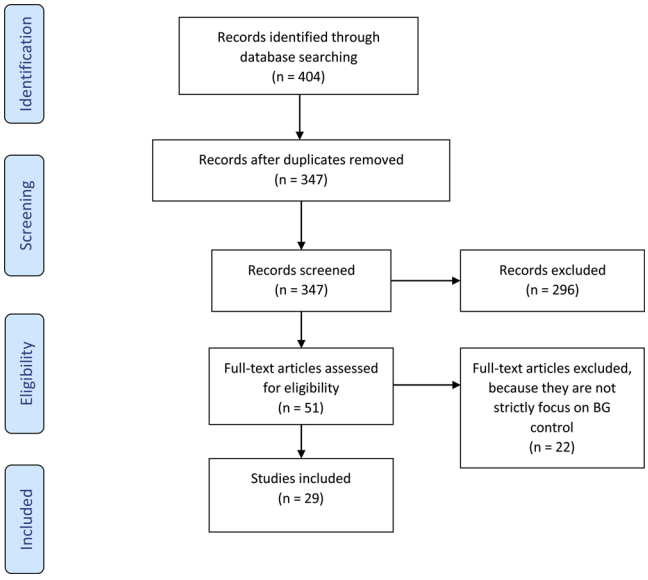

Fig. 4. Flow diagram of the process.

Table 1
Features extracted from the papers.

Ref.	Subjects	Type of DM	Data source	Preprocessing	Class of RL	Exploitation vs. exploration	State space	Action space	Planning	Generalization approaches	Performance metrics	Model of optimal behavior
[16]	1 in silico patient	T1DM	AIDA model [17]	BAL and sparsification	GPRL	BAL	CONT. BG and insulin	CONT. insulin	Model-based	Nonparametric regression (GP)	Target ranges	IHD model
[18]	52 real patients	T2DM	Clinical data. Clinical study.	No	Learning automaton	Gaussian distribution	CONT. BG	CONT. insulin	Model-free	CALA	Target ranges	N/A
[19]	28 in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	N/A	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	CVGA	IHD model
[21]	28 in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	N/A	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	CVGA and LBGI	N/A
[22]	70 real patients	N/A	Clinical data. Public data set.	Glycemic features	Q-learning	N/A	DISC. BG	DISC. insulin	Model-free	Tabular method (Q-learning)	Target ranges	IHD model
[23]	1 in silico patient	T1DM	AIDA model [17]	BAL	GPRL	BAL	CONT. BG and insulin	CONT. insulin	Model-based	Nonparametric regression (GP)	Target ranges	IHD model
[24]	3 in silico patients	T1DM	Bergman's minimal model [25]	Glycemic features	Q-learning	ε-greedy policy	DISC. BG	DISC. insulin	Model-free	Tabular method (Q-learning)	Meal disturbance rejection and overcoming variability	IHD model
[26]	2 real patients	T1DM	Clinical data. Private data set.	Glycemic features	Q-learning	ε-greedy policy	DISC. BG, weight and PA	DISC. insulin	Model-free	Tabular method (Q-learning)	N/A	IHD model
[27]	3 in silico patients	Both	Palumbo model [28]	No	Sarsa	ε-greedy policy	DISC. BG and insulin	DISC. insulin	Model-free	Tabular method (Sarsa)	Reference value	IHD model
[29]	10 in silico patients	T1DM	UVA/PADOVA Simulator [20]	No	AC	N/A	CONT. BG and insulin	CONT. insulin infusion rate	Model-free	Policy Gradient Method	CVGA	IHD model
[30]	28 in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	N/A	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	CVGA	IHD model
[31]	1 in silico patient	T1DM	AIDA model [17]	Bayesian surprise and sparsification	GPRL	BAL	CONT. BG and insulin	CONT. insulin	Model-free	Nonparametric regression (GP)	On-line behavior monitoring	IHD model
[32]	1 in silico patient	T1DM	AIDA model [17]	Bayesian surprise	LSMDP	Greedy policy	CONT. BG and insulin	CONT. insulin	Model-free	Nonparametric regression (GP)	On-line behavior monitoring	IHD model
[33]	1 in silico patient	T1DM	Bergman's minimal model [25]	BAL	GPDP	BAL	CONT. BG and insulin	CONT. insulin	Model-based	Nonparametric regression (GP)	Target ranges	IHD model
[34]	100 FDA accepted in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	Gaussian distribution	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	Target ranges	IHD model
[35]	"Various" real patients	T2DM	Clinical data. Public data set.	Glycemic features	DP	Greedy policy	DISC. BG, glucose absorption rate, measurement times, CHO and PA	DISC. insulins	Model-based	Tabular method (DP)	Optimal insulin treatment policy	IHD model
[36]	20 in silico patients	T1DM	UVA/PADOVA Simulator [20]	No	Sarsa and AC	UN and TG. Gaussian stochastic policy.	DISC and CONT. BG and CHO	DISC and CONT. insulin	Model-free	Tile-coding and Policy Gradient Method	Target ranges	IHD model
[37]	100 in silico and 31 real patients	T1DM	Simulated data generated by themselves and clinical data. Public data set.	No	V-learning	Randomized decision rule	DISC. BG, PA and CHO	DISC. insulin, PA and CHO	Model-based	Tabular method	Target ranges	IHD model
[38]	3 in silico patients	T1DM	Bergman's minimal model [25]	No	ADP	Exploration noise	CONT. BG	CONT. insulin	Model-free	Function approximation	Reference value	N/A
[39]	1 in silico patient	T1DM	Hovorka model [40]	Glycemic features	AC	Gaussian distribution	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	Target ranges	AR model
[41]	5565 real patients	Both	Clinical data. Public data set.	No and Sparse autoencoder	DP	Greedy policy	DISC. Patient variables, BG, vital signs and laboratory values	DISC. BG	Model-based	Tabular method (DP)	Comparison with reference value	IHD model

(*continued on next page*)

Table 1 (*continued*)

Ref.	Subjects	Type of DM	Data source	Preprocessing	Class of RL	Exploitation vs. exploration	State space	Action space	Planning	Generalization approaches	Performance metrics	Model of optimal behavior
[42]	1 in silico patient with Real meal data.	T1DM	Combination of minimal and Hovorka models with actual meal data [25,40]. Private meal data.	No	RLOC	Least square algorithm and exploration noise	CONT. BG and interstitial insulin activity	CONT. insulin	Model-free	Mix from Policy gradient and function approximation	Comparison between algorithms with a reference value	IHD model
[43]	1 in silico patient	T1DM	Bergman's minimal model [25] and Hovorka model [40]	No	Fitted Q-iteration	UN distribution	CONT. BG and insulin	DISC. insulin	Model-free	Nonparametric regression (kernel and random forest)	Target ranges	IHD model
[44]	1 in silico patient	T1DM	Combination of minimal and Hovorka models [25,40].	No	RLFF	Least square algorithm and exploration noise	CONT. BG and interstitial insulin activity	CONT. insulin	Model-free	Function approximation	Comparison between algorithms with a reference value	IHD model
[45]	30 in silico patients	T1DM	UVA/PADOVA Simulator [20]	No	DQN	ε-greedy policy	CONT. BG and insulin	DISC. insulin	Model-free	Function approximation	Risk function [46]	IHD model
[47]	100 in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	Gaussian distribution	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	TIR, LBGI, HBGI, MAGE and TDI	IHD model
[48]	10 in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	Gaussian distribution	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	TIR, LBGI and HBGI	IHD model
[49]	11 in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	Gaussian distribution	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	TIR	IHD model
[50]	30 in silico patients	T1DM	UVA/PADOVA Simulator [20]	Glycemic features	AC	Gaussian distribution	CONT. BG	CONT. insulin	Model-free	Policy Gradient Method	TIR and TDI	IHD model

Table 2
Data sources used by the studies.

Data sources	Count	Percentages
UVA/PADOVA simulator	11	35.48 %
Bergman's minimal model	4	12.90 %
AIDA model	4	12.90 %
Public available data set (Real data)	4	12.90 %
Hovorka model	2	6.45 %
Combination models	2	6.45 %
Palumbo model	1	3.23 %
Private data set (Real data)	1	3.23 %
Clinical study (Real data)	1	3.23 %
Simulated data generated by themselves	1	3.23 %

30) have tried to test the algorithm on real subject data sets, while the remaining group (3.33 %, 1/30) relies on mixed data sets such as using simulated BG and insulin along with real meal data sets.

4.3.2. Data sources

The reviewed articles have used various kinds of data sources for the development of the control algorithm using RL, as shown in Table 2 below. Accordingly, the most used data source is the UVA/PADOVA simulator [20] (35.48 %, 11/31) followed by the Bergman's minimal model [25] (12.90 %, 4/31), AIDA model [17] (12.90 %, 4/31) and public available real datasets (12.90 %, 4/31). The third most used are the Hovorka model [40] (6.45 %, 2/31) and combination of the minimal model with part of the Hovorka model, one of them using actual meal data (6.45 %, 2/31). The fourth most used data sources includes, private datasets (3.23 %, 1/31), Palumbo model [28] (3.23 %, 1/31), real datasets from a clinical study (3.23 %, 1/31) and simulated data generated by researchers (3.23 %, 1/31). The real datasets are mainly from CGM (3.23 %, 1/31), insulin pump (9.68 %, 3/31), accelerometer (3.23 %, 1/31), automatic electronic recording device (3.23 %, 1/31), paper records (3.23 %, 1/31), multiple daily injections (3.23 %, 1/31), and actual meal data records (3.23 %, 1/31).

4.3.3. Preprocessing

Preprocessing is a crucial component in RL strategies. In this regard, extracting a range of glycemic features ranked as most used (40.63 %, 13/32) followed by the absence of a preprocessing stage (34.38 %, 11/32), as shown in the Table 3 below. Bayesian active learning (BAL) (9.37 %, 3/32) and sparsification (9.37 %, 3/32) are the third most used techniques followed by Bayesian surprise (6.25 %, 2/32).

4.4. Reinforcement learning strategies

4.4.1. Class of reinforcement learning algorithms

There are various classes of RL algorithms such as AC learning, Q-learning, Sarsa to mention a few. In this regard, the most popular RL algorithms is found to be the AC learning (36.67 %, 11/30) followed by Q-learning (10 %, 3/30) and Gaussian processes reinforcement learning (GPRL) (10 %, 3/30), as shown in the Table 4 below. Sarsa (6.68 %, 2/30) and dynamic programming (DP) (6.68 %, 2/30) are ranked as the third most popular reinforcement learning algorithms followed by Gaussian process dynamic programming (GPDP) (3.33 %, 1/30), learning automaton (3.33 %, 1/30), V-learning (3.33 %, 1/30), model-

Table 3
Preprocessing techniques used in the reviewed literature.

Preprocessing	Count	Percentages
Extracting a range of glycemic features	13	40.63 %
No preprocessing	11	34.38 %
BAL	3	9.37 %
Sparsification	3	9.37 %
Bayesian surprise	2	6.25 %

Table 4
Class of reinforcement learning algorithms.

Class of reinforcement learning algorithms	Count	Percentages
AC learning	11	36.67 %
Q-learning	3	10 %
GPRL	3	10 %
Sarsa	2	6.68 %
DP	2	6.68 %
GPDP	1	3.33 %
Learning automaton	1	3.33 %
V-learning	1	3.33 %
ADP	1	3.33 %
RLOC	1	3.33 %
LSMDP	1	3.33 %
Fitted Q-iteration	1	3.33 %
RLFF	1	3.33 %
DQN	1	3.33 %

free approximate/adaptive dynamic programming (ADP) algorithm (3.33 %, 1/30), reinforcement-learning optimal control algorithm (RLOC) (3.33 %, 1/30), linearly-solvable Markov decision process (LSMDP) (3.33 %, 1/30), fitted Q-iteration (3.33 %, 1/30), reinforcement learning with feedforward (RLFF) (3.33 %, 1/30), and deep Q-network (DQN) (3.33 %, 1/30).

4.4.2. Reinforcement learning approaches

The approaches in RL in the reviewed literature could be roughly categorized as tabular solution methods and approximate solution methods. In this regard, as shown in Table 5 below, approximate solution methods (73.33 %) are more popular than the tabular solution methods (26.67 %).

4.5. Exploitation-exploration dilemma

In RL algorithm applications, exploitation-exploration dilemma is one of the most important constituents of the design choices. In this regard, Gaussian distribution function (24.25 %, 8/33) is the most popular choice, as shown in Table 6 below. BAL (12.12 %, 4/33) and ε-greedy policy (12.12 %, 4/33) are the second most important choices followed by greedy policy (9.09 %, 3/33) and exploration noise (9.09 %, 3/33). Least squares algorithm (6.06 %, 2/33) and uniform distribution (UN) (6.06 %, 2/33), are the fourth most popular choices followed by truncated gaussian (TG) (3.03 %, 1/33) and randomized decision rule (3.03 %, 1/33). However, surprisingly (15.15 %, 5/33) of the studies either did not report their choices or did not consider it at all.

Table 5
Approaches to reinforcement learning for blood glucose control in diabetes patient.

RL solution	Count	Percentages
Approximate Solution Methods	22	73.33 %
Tabular Solution Methods	8	26.67 %

Table 6
Various design choices towards exploitation-exploration dilemma.

Exploitation-exploration dilemma	Count	Percentages
Gaussian distribution	8	24.25 %
BAL	4	12.12 %
ε-greedy	4	12.12 %
Greedy policy	3	9.09 %
Exploration noise	3	9.09 %
Least squares algorithm	2	6.06 %
UN	2	6.06 %
TG	1	3.03 %
Randomized decision rule	1	3.03 %
Unspecified	5	15.15 %

Table 7
Nature of the state space.

State space nature	Count	Percentage
Continuous	22	73.33 %
Discrete	8	26.67 %

4.6. State and action spaces

The other most important constituents design choices of RL applications is defining the nature and parameters of the agent state and action spaces. In this section, we will present the nature of the state and action spaces along with their defining parameters.

4.6.1. State space

4.6.1.1. Nature of the state space. Based on the reviewed studies, the nature of the state space could be grouped in two; continuous and discrete state space. In this regard, most of the studies have relied on continuous state space (73.33 %), as shown in Table 7 below.

4.6.1.2. State space defining parameters. Various key diabetes parameters have been used to define the state spaces, as shown in Table 8 below. Based on the reviewed studies, the most popular parameter is BG level (43.34 %, 13/30) followed by BG level and insulin dose (30 %, 9/30). BG level and carbohydrate (CHO) intake (6.67 %, 2/30), and BG level and the interstitial insulin activity (6.67 %, 2/30) are the third most used parameters. The fourth most used parameters include the following combinations:

- BG level, glucose absorption rate, measurement times during the day, CHO intake and PA (3.33 %, 1/30).
- BG level, weight and PA (3.33 %, 1/30).
- BG level, PA and CHO intake (3.33 %, 1/30).
- Patient level variables, BG related variables, periodic vital signs and laboratory values (3.33 %, 1/30).

4.6.2. Action space

4.6.2.1. Nature of the action space. As for the state spaces, the nature of the action space is inline and could be grouped into continuous or discrete as shown in Table 9 below. Accordingly, most of the studies have relied on continuous action spaces (66.67 %, 20/30), while only 33.33 % of the studies have relied on a discrete space.

4.6.2.2. Action space defining parameters. Various action parameters taken by the diabetes patients to manage his/her BG are considered in the reviewed studies, as show in Table 10 below. In this regard, insulin dose is the most popular action parameter used in the studies

Table 8
State space defining parameters.

State space defining parameters	Count	Percentages
BG level	13	43.34 %
BG level and insulin dose	9	30 %
BG level and CHO intake	2	6.67 %
BG level and interstitial insulin activity	2	6.67 %
BG level, glucose absorption rate, measurement times during the day, CHO intake and PA	1	3.33 %
BG level, weight and PA	1	3.33 %
BG level, PA and CHO intake	1	3.33 %
Patient level variables, BG related variables, periodic vital signs and laboratory values	1	3.33 %

Table 9
Nature of the action spaces.

Action Space Nature	Count	Percentage
Continuous	20	66.67 %
Discrete	10	33.33 %

Table 10
Action space defining parameters.

Action Space Parameters	Count	Percentage
Insulin dose	29	93.54%
Insulin dose, PA and food intake	1	3.23 %
Targeted BG level	1	3.23 %

followed by insulin dose, PA and food intake (3.23 %, 1/31) and targeted BG level (3.23 %, 1/31).

4.7. Planning

Planning is another important constituent of the design choices in the RL applications. Accordingly, based on the studied articles planning approaches could be roughly categorized as model-based or model-free approaches. In this regard, a model-free approach (79.31 %, 23/29) is the most widely exploited approach in diabetes BG control algorithms, as shown in the Table 11 below.

4.8. Generalization approaches

Generalization is a straight forward approach for high dimensional and continuous state and action spaces in real world control tasks, where a discrete representation is intractable. In this regard, the reviewed literatures have exploited various generalization approaches as shown in Table 12 below. The most used generalization approach is policy gradient method (11/24, 45.83 %) followed by nonparametric regression (7/24, 29.16 %). Function approximation (3/24, 12.5 %) is the third most used generalization approach. The fourth most used generalization approaches include continuous action-set learning automata (CALA) (1/24, 4.17 %), tile-coding (1/24, 4.17 %), and mix from policy gradient and function approximation (1/24, 4.17 %).

4.9. Performance metrics or evaluation criteria

Various kinds of evaluation criteria have been used to measure the performance of the algorithm towards the specified goal as shown in Table 13 below. In this regard, the most used approach is predefined target ranges (14/36, 38.90 %) followed by comparison with reference value (5/36, 13.89 %) and CVGA (4/36, 11.11 %). Low blood glucose index (LBGI) (3/36, 8.33 %) is the fourth most used approaches followed by on-line behavior monitoring (2/36, 5.55 %), high blood glucose index (HBGI) (2/36, 5.55 %), and total daily insulin (TDI) (2/36, 5.55 %). The sixth most used performance metrics are risk function (1/36, 2.78 %), mean amplitude of glucose excursion (MAGE) (1/36, 2.78 %), optimal insulin treatment policy (1/36, 2.78 %) and ability to reject the effect of meal disturbance and to overcome the variability in the glucose-insulin dynamics from patient to patient (1/36, 2.78 %).

Table 11
Planning approaches.

Planning	Count	Percentage
Model-free	23	79.31 %
Model-based	6	20.69%

Table 12
Generalization Approaches.

Generalization issues	Count	Percentages
Policy Gradient Method	11	45.83 %
Nonparametric regression	7	29.16 %
Function approximation	3	12.5 %
CALA	1	4.17 %
Tile-coding	1	4.17 %
Mix from Policy gradient and function approximation	1	4.17 %

Table 13
Performance metrics or evaluation approaches.

Performance metrics or evaluation criteria	Count	Percentages
Predefined target ranges	14	38.90 %
Comparison with reference value	5	13.89 %
CVGA	4	11.11 %
LBGI	3	8.33 %
HBGI	2	5.55 %
TDI	2	5.55 %
On-line behavior monitoring	2	5.55 %
Risk function	1	2.78 %
MAGE	1	2.78 %
Meal disturbance rejection and overcoming variability	1	2.78 %
Optimal insulin treatment policy	1	2.78 %

Table 14
Model of optimal behavior.

Model of optimal behavior	Count	Percentages
IHD model	25	86.20 %
AR model	1	3.45 %
Unspecified	3	10.35 %

4.10. Model of optimal behavior

Another important constituent of reinforcement algorithm design choices includes the description of model of optimal behavior, as shown in Table 14 below. In this aspect, the reviewed papers mainly exploited the IHD model (25/29, 86.20 %) and only (1/29, 3.45 %) used the AR model. Surprisingly, (3/29, 10.35 %) have not stated anything related to the optimal behavior model.

4.11. Reward function

The reward function is also among the crucial constituents of design choices for a successful RL design. In this regard, choosing the reward function relies on the expert designing and developing the algorithms. As a result, the expert is free to choose the reward function based on the specific task and objective he/she is in need of achieving. With the same token, the reviewed studies have reported various types of reward functions based on their nature and defining parameters of the state and action spaces as shown in Table 15 below.

5. Discussion

Over the last decade, there has been an increase in the use of ML techniques for diabetes management, which has meant important advances in this research area. Concretely, RL algorithms have arisen as a competitive solution for BG control in diabetes patients during recent years, especially in T1DM where its use is more extended. These algorithms were applied on in-silico subjects in most cases. Clinical data is usually hard to obtain because the patients have to collect carefully their data and in addition, there are ethical issues related to the use of such data. However, although the current situation could be marked by the difficulties of obtaining real data from diabetic patients, there exists

Table 15
Reward functions.

Reference	Reward/Cost function	Comments
[16]	$r(G(t)) = -1 + e^{-\frac{(G(t)-G_X)^2}{2a^2}};\ r \in [-1,\ 0]$	Gaussian reward function where: $G(t)$ - Instantaneous reading from the glucose sensor G_X - Reference value of the glucose concentration a - Width of the desired glucose band for normoglycemia
[18]	$\beta_k = \frac{\lvert G_A(k) - \overline{G}_N \rvert}{G_A(k)}$	$G_A(k)$ - Actual BG level $\overline{G}_N$ - BG average normal value
[19]	$c(x_k) = a_h F_1^k + a_l F_2^k$	F_1^k and F_2^k - Features describing the glycemic profile a_h and a_l - weights for scaling the hypo and hyperglycemia components
[21]	N/A	N/A
[22]	Reward +1 if next BGL measurement is within a predefined range Penalty -1 if next BGL measurement is out of a predefined range	N/A
[23]	$r(G(t)) = -1 + e^{-\frac{(G(t)-G_X)^2}{2a^2}};\ r \in [-1,\ 0]$	Gaussian reward function where: $G(t)$ - Instantaneous reading from the glucose sensor G_X - Reference value of the glucose concentration a - Width of the desired glucose band for normoglycemia
[24]	$r_t(x, a) = -\lvert(G - 80)\rvert$	The reward is set equal to the difference of the glucose concentration from its target value of 80 mg/dl. This value has been considered as a reference set point in normoglycemic range of BG.
[26]	N/A	Function of the difference of the A1C from its target value 7.
[27]	$r_t(s, a) = -\lvert G(t) - G_{ref}(t)\rvert$	The reward is set equal to the difference of the plasma glycaemia signal from a reference signal.
[29]	$R(t) = \begin{cases} a_h \bullet E(t)\ if\ G(t) \geq G_H \\ a_l \bullet E(t)\ if\ G(t) < G_L \\ 0\ otherwise \end{cases}$ Where $E(t) = \lvert G(t) - G_{ref}\rvert$	a_h - Hyperglycemia penalty a_l - Hypoglycemia penalty G_H - Hyperglycemia bound G_L - Hypoglycemia bound $E(\bullet)$ - Current error between the measured and the desired glucose concentration value G_{ref} - reference glucose concentration value
[30]	N/A	The state is used by the algorithm for the estimation of the long-term expected costs
[31]	$l^p(\mathbf{x},\ u^p) = q(\mathbf{x}) + KL(p(\hat{\mathbf{x}}\lvert\mathbf{x},\ u^p)\lvert\lvert h(\hat{\mathbf{x}}\lvert\mathbf{x}))$	$q(\mathbf{x})$ - State cost $KL(\bullet\lvert\lvert\bullet)$ - Kullback–Leibler distance $p(\hat{\mathbf{x}}\lvert\mathbf{x},\ u^p)$ - Optimal actions under uncertainty $h(\hat{\mathbf{x}}\lvert\mathbf{x})$ - Passive system dynamics $\mathbf{x}$ - Actual state $\hat{\mathbf{x}}$ - Next state u^p - Control action
[32]	$l(\mathbf{x},\ u) = hq(\mathbf{x}) + KL(p^u(\mathbf{x}_{k+1}\lvert\mathbf{x}_k)\lvert\lvert p^0(\mathbf{x}_{k+1}\lvert\mathbf{x}_k))$	$q(\mathbf{x})$ - State cost $KL(\bullet\lvert\lvert\bullet)$ - Kullback–Leibler distance $p^u(\mathbf{x}_{k+1}\lvert\mathbf{x}_k)$ - Controlled diffusion process $p^0(\mathbf{x}_{k+1}\lvert\mathbf{x}_k)$ - Passive dynamics $\mathbf{x}_k$ - State at time k $\mathbf{x}_{k+1}$ - State at time k + 1 u - Control action
[33]	$g(G_t) = -1 + e^{-\frac{(G_t-\overline{G})^2}{2a^2}};\ g \in [-1,\ 0]$	Gaussian reward function where: G_t - Instantaneous reading from the glucose sensor $\overline{G}$ - Reference value of the glucose concentration a - Width of the desired glucose band for normoglycemia
[34]	$c(x_k) = a_h x_k^1 + a_l x_k^2$	x_k^1 and x_k^2 - Features describing the glycemic profile a_h and a_l - weights for scaling the hypo and hyperglycemia components
[35]	$r(g) = r^l(g) - r^n(g)$ $r^l(g) = 1 - \frac{\lvert gl - gl_h\rvert}{gl_1 - gl_L}$ $r^n(g) = \mathbf{I}_{gl < gl_L{}^c} \bullet \left[1 - \frac{gl - gl_L{}^c}{gl_L{}^c}\right]$	Heuristically defined. Positive rewards are obtained for the healthiest states and negative rewards are obtained at undesired BG levels. gl - BGL-state gl_h - Most healthy BGL $\mathbf{I}$ - Standard indicator function
[36]	Mean-reward (Sarsa): $\widetilde{R}_{t+1} = \frac{\int_{\tau_t}^{\tau_t+1} score(BG_\tau)\, d\tau}{\tau_{t+1} - \tau_t}$ Cumulative-reward (Actor-Critic): $R_{t+1}^+ = \int_{\tau_t}^{\tau_t+1} score(BG_\tau)\, d\tau$	They define a score function that matched their objectives. This function penalizes when glucose level is out of the ideal range (4−8 mmol/L).
[37]	Weighted sum of glycaemic events (hypo- and hyperglycaemic episodes) over the 60 minutes preceding and following time t.	Weights are: −3 when glucose ≤ 70 (hypoglycemic) −2 when glucose > 150 (hyperglycemic) −1 when 70 < glucose ≤ 80 or 120 < glucose ≤ 150 (borderline hypo- and hyperglycemic) 0 when 80 < glucose ≤ 120 (normal glycaemia)
[38]	$J = \int_0^\infty (\alpha G_\Delta^2 + \beta u_\Delta^2) d\tau$	G - BG concentration u - Infusion rate of the insulin pump $\alpha > 0$ and $\beta > 0$ - Weighting constants

(*continued on next page*)

Table 15 (*continued*)

Reference	Reward/Cost function	Comments
[39]	$c(\boldsymbol{x}_t) = a_h x_t^1 + a_l x_t^2$	x_t^1 and x_t^2 - Features describing the glycemic profile a_h and a_l - weights for scaling the hypo and hyperglycemia components
[41]	90-day mortality status: +100 for patients who survived 90 days after their admission −100 for those who were deceased before 90 days after their admission	N/A
[42]	$r^k = \mathbf{x}_k^T \mathbf{Q} \mathbf{x}_k + u_k^T R u_k$	$\mathbf{x}$ - State of the model formed by BG level and interstitial insulin activity u - Insulin dose $\mathbf{Q}$ and R - Weighting factors
[43]	$r_i^{'} = \lvert g_{i+1} - 90 \rvert$	g_i - Plasma glucose value 90 mg/dl = 5 mmol/L is taken as the optimal blood glucose level
[44]	$r_{k+1} = \mathbf{x}_k^T \mathbf{Q} \mathbf{x}_k + u_k^T R u_k$	$\mathbf{x}$ - State of the model formed by BG level and interstitial insulin activity u - Insulin dose $\mathbf{Q}$ and R - Weighting factors
[45]	$R = risk(\mathrm{b}_{t+1}) - risk(\mathrm{b}_t)$	Where *risk* is the asymmetric blood glucose risk function defined as: $risk(b) = 10 * (1.509 * \log(b)^{1.084} - 5.381)$ b_t - Blood glucose value
[47]	$c_k = a_{hyper} F_{k\_hyper} + a_{hypo} F_{k\_hypo}$	F_{k_hyper} and F_{k_hypo} - Features describing the glycemic profile a_{hyper} and a_{hypo} - weights for scaling the hypo and hyperglycemia components
[48]	$c_k = a_{hyper} F_{k\_hyper} + a_{hypo} F_{k\_hypo}$	F_{k_hyper} and F_{k_hypo} - Features describing the glycemic profile a_{hyper} and a_{hypo} - weights for scaling the hypo and hyperglycemia components
[49]	$c_k = a_{hyper} F_{k\_hyper} + a_{hypo} F_{k\_hypo}$	F_{k_hyper} and F_{k_hypo} - Features describing the glycemic profile a_{hyper} and a_{hypo} - weights for scaling the hypo and hyperglycemia components
[50]	$c_k = a_{hyper} F_{k\_hyper} + a_{hypo} F_{k\_hypo}$	F_{k_hyper} and F_{k_hypo} - Features describing the glycemic profile a_{hyper} and a_{hypo} - weights for scaling the hypo and hyperglycemia components

a need to move the studies from simulated data to clinical data in order to facilitate the validation of the algorithms. Regarding the source of these data, most of the studies relies on the in-silico patient cohort provided by UVA/PADOVA simulator [20] to evaluate the algorithms. The main reason for this is presumably that it is the only in-silico diabetes model accepted by the FDA as a substitute for pre-clinical animal testing of new treatment strategies for T1DM, which is the prelude to the clinical studies on humans. This simulator is followed by AIDA and Bergman's minimal models [17,25], which are the second most common option, probably because these are simple models and for this reason it is easier to work with them. For example, Bergman's minimal model does not present any delay in insulin action, which fits better with the RL framework. Once again, the lack of real data sets is evident and so is the validation of the problem since we only found one clinical study in the literature [18]. After obtaining the data, preprocessing is performed to extract a range of glycemic features, which in some of the studies is used to establish different glycemic ranges in order to discretize the state space [22,24,26]. However, studies using raw BG levels also occurs frequently. Other techniques such as BAL, which samples only relevant data, and sparsification, which determines whether arriving data provide valuable information are interesting options in future research [16].

Moving the discussion to the RL framework, we can find two different solutions: tabular methods and approximate methods, the latter being most used for this BG control problem. Tabular methods are used to face problems with small state and action spaces, while approximate methods are well-suited to problems with large state and action spaces. Since current BG control research is focused on developing the AP, which includes the use of a CGM and insulin pump that generate continuous blood glucose measurements and continuous insulin infusion, we found that we are facing continuous spaces and therefore approximate solutions fit well given the nature of the problem. Moreover, in scenarios with continuous or large discrete state and action spaces we need to use generalization techniques to learn information and transfer knowledge between similar states and actions, since in a large and smooth state space we generally expect similar states to have similar values and similar optimal actions [51]. In this regard, we found in the literature that the most used generalization technique is policy gradient method, characterized by learning a parametrized policy that does not use the value function to select actions [14]. Another much used generalization technique is GP regression, which is an interpolation method with the interpolated values modeled by a GP governed by prior covariances. Further information about GP in ML can be found in [52].

Among the RL algorithms analyzed during this review, AC methods are most used. These algorithms produce an approximate solution based on policy gradient methods that learn a parametrized policy instead of learning which action is better in each state. Therefore, action-value functions are not directly used by these methods to select actions [14]. Regarding tabular methods Q-learning is the most used approach, which is an off-policy temporal-difference control algorithm in which the learned action-value function directly approximates the optimal action-value function [14]. During a temporal-difference learning process, previous predictions are used as a targets for next predictions in order to solve the prediction problem [36]. Furthermore, most of those algorithms found in the literature are on-policy methods that evaluate the same policy that is used to make decisions. Otherwise, off-policy methods evaluate a policy which is different than the policy used to obtain the data. Moreover, although in most of the literature learning method information is not included, we found more cases based on on-line learning, in which learning is performed as the data is coming in, than on off-line learning where there is a static dataset. It is worth mentioning articles in which a policy is learning off-line in a first stage using stored data, and then this policy is adapted on-line for the patient [16,29,33]. Finally, most of the articles in the literature use the IHD model to decide how the future is considered in the actions made by the agent about how to behave in the current time step. These are typical situations in mHealth applications, in which we usually have an on-line estimation of optimal treatment strategies as data continuously accumulate, as well as no definite time horizon taking into account the long-run reward of the agent [37]. This scenario is reflected in the BG control task, where a CGM yields a continuous flux of BG measurements.

Further comparison between different RL algorithms is performed in [36], were policy gradient and tabular methods are compared. In this paper, AC algorithm shows better performance than sarsa. This is because sarsa starts completely from scratch, while AC starts from a reasonable policy from which knows its structure. Furthermore, we are trying to face a continuous action task and sarsa is designed for discrete

action space, while AC is designed for continuous action space [36]. This paper also compares traditional supervised learning with RL methods. In this regard, RL does not require any knowledge on the parameters of the policy, but supervised learning needs this information. Moreover, supervised learning needs shorter training period than RL because of the generalization ability of the former. However, RL algorithms continuously learn from new data, while supervised learning does not adjust to the patient after the training period, loosing this extra information. Therefore, glucose pattern in diabetes keeps changing and RL methods can adapt to this change, but supervised learning algorithms cannot [53].

Proportional-integral-derivative control algorithm and self-managed control by the patient are compared with RL methods in [45]. From this study, RL algorithms were able to outperform traditional approaches under certain circumstances, although they do not outperform the proportional-integral-derivative controller across all settings [45]. This kind of control algorithms are considered one of the most used techniques in the AP framework [54]. Moreover, the impact of errors in CHO estimation is analyzed in [49]. This paper tests the performance of proportional-integral-derivative controller, bolus calculator [55] and RL algorithm under different CHO estimation error levels. In this work, RL algorithm outperforms traditional approaches, achieving stable blood glucose control performance under all different conditions. Furthermore, categorical CHO announcement using three different levels (small, medium, and large) has low or no impact on the blood glucose control when errors in CHO estimation are lower than ± 25 %, indicating that the algorithms do not need accurate meal announcements [49].

The trade-off between exploration and exploitation is one of the unique characteristics that differentiate the RL algorithms from others ML approaches. Therefore, how to perform it is one of the choices we must make when we are going to implement a RL algorithm. However, we extracted from the results that in many cases this issue is not defined. This is because in most of that cases AC algorithms and therefore policy gradient methods are used, and for these algorithms we only generally require that the policy never becomes deterministic in order to ensure the exploration [14]. Therefore, in practice it is enough to choose a stochastic policy to solve the exploration-exploitation dilemma, and in some of these studies those policies are not specified. Moreover, we found that Gaussian distribution functions are very frequently used to deal with this issue. It is worth mentioning the use of ε-greedy exploration, since it is a really simple method in which instead of taking in each state always the action with greatest value, we choose from time to time a random action with small probability ε in order to ensure the exploration.

Another of the most important choices we must take during RL algorithm implementation is the definition of the state and action spaces. First of all, we found that most of these spaces are defined as continuous. As we mentioned above, this is because of the nature of the problem, in which we expected to have continuous BG measurements and continuous insulin infusion rate. Accordingly, in the BG control problem we will always have at least two information sources: BG level and insulin doses. Therefore, it is natural in the RL framework to relate that information with the states and actions respectively. There are various definitions of the state space in the reviewed literatures, all of them somehow related to the BG level. Concretely, most of the authors define the state space based only on the BG level, followed by these studies in which the states take into account not only the BG level, but also the insulin doses. Regarding the action space, there is only one study in which the actions are not based on the insulin doses [41]. In this paper, the authors take the actions choosing the best glycemic target under different circumstances, leaving the choice of agents and doses to achieve that target to the clinicians. It is worth to mention two articles in which not only the quantity of insulin is used as an action, but also the kind of insulin used [22], such us short-acting, intermediate-acting or long-acting, and even a combination of those different insulins [35]. However, several additional factors affect the BG level such us CHO intake, PA, stress level, infections, etc [56]. This means that the use of this information is useful in order to face the BG control problem, so we expected to find this data as part of the state and action spaces. However, there are few papers in which for example CHO intakes and PA are included in the state space, although this information is really relevant for the algorithm and facilitates its operation. Furthermore, there is a lack of automatic CHO recording since in those cases this task relies on manual recording. In order to reduce the burden on the patient, as well as increase the objectivity during the control task, the combination of RL algorithms with meal detection algorithms such as [57,58] could be part of future perspectives in order to work in a fully closed-loop system. Concerning the action space, we found that despite the importance of the PA and CHO intakes, there is only one paper in the literature in which this value information is indeed taking into account as part of the actions [37]. This action space is formed by a hypothetical mHealth intervention where insulin injections are administered using an insulin pump while suggestions for food intake and PA are administered using a mobile app, considering all possible combinations of insulin injection, food intake, and PA.

The model of the environment is another element of model-based RL systems. The models are used for planning or predicting the next state and the next reward. In this stage we have to decide if we want to use a model-based method or a model-free method in which the learner behavior is based on trial and error. What we found here is that most of the authors based their algorithms on model-free methods. It can be explained by the fact that it is difficult to obtain realistic metabolic models for a real person. Furthermore, it is expected that RL algorithms becomes a personalized solution learning from the real patient, and each person presents different characteristics due to the inter- and intra-subject variability of insulin absorption and insulin action [59].

Finally, the choice of a good reward function is crucial for the correct performance of the algorithm. This is the way we have to communicate to the agent what we want to achieve, thereby defining the goal in the RL problem [14]. Therefore, in our BG control problem, the reward function should reflect our desire to stay inside the normal glycemic range. In general, these may be stochastic functions of the state of the environment and the actions taken. Since the reward function is freely defined by the authors, in this category we found very varied reward functions as we can see in Table 15. In general terms, we found that most of reward functions are related with the BG level in some way and consequently with the state of the environment. There is only one case that does not take into account the BG level [41]. This is because the study is focused on severely ill septic patients and in this situation the survival of patients is the main objective of clinicians for critical care. It is also common to find some reference values related to normal, hyper and hypoglycemia ranges in order to establish good rewards and penalties. However, we found that only five papers include the actions taken in the reward function [31,32,38,42,44]. We think it could be interesting to also consider the insulin doses in the reward function, which for example can lead to take less aggressive actions for the patients. The success of a RL application strongly depends on how well the reward function frames the goal of the application's designer and how well the function assesses progress in reaching that goal [14].

In order to measure the performance of these algorithms, the authors usually predefine target ranges since in the BG control problem we aim to spend as much time as possible in normal glycemia, which is between 70 and 130 mg/dl with a mean normal value of 100 mg/dl. This means that in this task it is quite easy to establish desired ranges and reference values. Another quite common technique to evaluate the efficacy of the glucose regulation algorithms is the CVGA, which shows the glucose excursions caused by a control algorithm in a group of patients, providing a summary of the quality of glycemic regulation for a population of subjects [60]. This method is complementary to the low blood glucose indices (LBGI) measurement, which characterize a single glucose trajectory for a single patient and is used to estimate the risk of hypoglycemia [61].

6. Conclusion

Recent research in diabetes area has produced new advances and technologies such as sensors, new insulins, monitoring devices, etc. On the one hand these discoveries facilitate the adoption of new techniques such as ML methods and the idea of the AP, but on the other hand the problem becomes more complex. At this point, RL algorithms emerge as a smart, personalized and optimal solution to calculate insulin delivery. In this regard, it is worth to mention this recent patent related to estimate insulin dose based on RL [62], and this patent that uses RL combined with neural network to optimize patient treatment recommendations [63], in which diabetes is used as a practical example of application. However, RL is still a recent approach in the diabetes area and there are few papers which explicitly use this class of algorithms in the BG control problem. For such purpose, we expected to find a model-free RL algorithm based on an approximate solution method, using continuous state an action spaces, learning on-line and following the IHD model, some of them being typical characteristics of mHealth systems. This is because of the nature of the problem, in which we continuously expect to receive BG measurements from a CGM indefinitely and learning according to the data is obtained, while at the end stage we are not able to know the model of the patient. Those expected features perfectly match with the trends we found in the literature during this systematic review.

Moreover, despite several factors, such as CHO intakes, PA, infections, or stress level, influence the BG, there are few papers in the reviewed literature which include these factors in the state and action spaces. This is, in particular, the case if we talk about the action space where there is only one study that considers PA and food intakes as part of the possible actions [37]. Therefore, we consider inclusion of some of these factors in the BG control problem to be a very important future research direction. For example, it would be possible to use meal detection [57,58] or CHO counting algorithms [64] to include the food intake information as a part of the state and action spaces. Another option could be a sensor mounted on a tooth transmitting information on glucose intake [65]. Moreover, nowadays the use of mobile devices and other wearables is quite common, therefore the inclusion of the PA in the state and action spaces would be really easy. This would allow the creation of a mHealth system for self-management diabetes controlled by a mobile app [66], in which BG level, insulin doses, food intake and PA are combined to deal with the BG control problem. However, although the inclusion of that additional information would be easy, the difficulties come with how such information can be correctly used by the RL algorithm, which in our opinion is the next challenge developers have to overcome to obtain a fully closed-loop AP system. In addition to the integration of additional systems for the estimation of the accurate CHO intake during meals as well as PA, an early warning system in order to forecast and predict hyper/hypoglycemic events would be extremely valuable [67].

Furthermore, to perform evaluation experiments on diabetic patients may be neither possible, appropriate, convenient nor desirable, since some of these experiments cannot be done at all or are too difficult, dangerous and not ethical [68]. Moreover, different countries have different execution procedures and regulatory conditions. For this reason, simulators are really necessary in order to deal with the diabetes framework, because these allow us to design, evaluate and verify the effectiveness of the BG controller before clinical tests. This is particularly important in the case of RL, where a continuous interaction with the patient is needed in order to learn the correct amount of insulin for each situation. However, there exist few papers in the literature using real data, therefore it is necessary to obtain and use more clinical data in order to clinically validate the algorithms.

Finally, traditional RL algorithms requires carefully chosen feature representations. Therefore, it would be interesting to test other RL approaches such us deep reinforcement learning [45], in which deep learning is used for learning feature representations, that in the traditional framework are usually hand-engineered [69]. Another possibility would be to combine supervised learning with RL, since the latter requires an extensive amount of training data in order to converge to a meaningful solution, restricting its usage for complex input spaces [70]. In such scenarios, it would be possible to learn from the past historical records of the subject BG level before start to learn directly from the patient, accelerating convergence and reducing the amount of time needed by the controller to stay in normoglycemic range, thereby facilitating clinical trials. Other approaches have been used in the literature for that purpose, for example [21,30,34,47] and [49] use transfer entropy to automatically initialize the control algorithm in a personalized fashion, providing faster learning rate. This method is a measurement of the information transfer between insulin and glucose signals, with promising application in biomedical signal analysis [71].

Declaration of Competing Interest

None.

Acknowledgements

This work was supported by the Tromsø Research Foundation. We are grateful for funding from the University of Tromsø - The Arctic University of Norway. We would also like to thank Susan Wei, PhD for comments that greatly improved the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.artmed.2020.101836.

References

[1] Diabetes. WHO; 2017 [cited 2018 25 June 2018]; Available from:. http://www.webcitation.org/719KGYXpa.
[2] International Diabetes Federation. IDF diabetes atlas. 8th edn Brussels, Belgium: International Diabetes Federation; 2017.
[3] ADA. American diabetes association research programs 2018-01-17 [cited 2018 24 July 2018]; Available from:. 2018http://www.webcitation.org/719Iz6gfm.
[4] Kavakiotis I, et al. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104–16.
[5] Oviedo S, et al. A review of personalized blood glucose prediction strategies for T1DM patients. Int J Numer Method Biomed Eng 2017;33(6).
[6] Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018;20(5):e10775.
[7] Lunze K, et al. Blood glucose control algorithms for type 1 diabetic patients: a methodological review. Biomed Signal Process Control 2013;8(2):107–19.
[8] Bothe MK, et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Rev Med Devices 2013;10(5):661–73.
[9] Bekiari E, et al. Artificial pancreas treatment for outpatients with type 1 diabetes: systematic review and meta-analysis. BMJ 2018;361:k1310.
[10] Hovorka R. Closed-loop insulin delivery: from bench to clinical practice. Nat Rev Endocrinol 2011;7(7):385–95.
[11] Kumareswaran K, Evans ML, Hovorka R. Closed-loop insulin delivery: towards improved diabetes care. Discov Med 2012;13(69):159–70.
[12] Farmer Jr TG, Edgar TF, Peppas NA. The future of open- and closed-loop insulin delivery systems. J Pharm Pharmacol 2008;60(1):1–13.
[13] Adibi SE. Mobile health: a technology Road map. Springer series in bio-/Neuroinformatics. 1 ed. Springer International Publishing; 2015.
[14] Sutton RS, Barto AG. Reinforcement learning: an introduction. 1998.
[15] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012:276–82.
[16] De Paula M, Acosta GG, Martínez EC. On-line policy learning and adaptation for real-time personalization of an artificial pancreas. Expert Syst Appl 2015;42(4):2234–55.
[17] Lehmann ED, Deutsch T. A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. J Biomed Eng 1992;14(3):235–42.
[18] Akbari Torkestani J, Ghanaat Pisheh E. A learning automata-based blood glucose regulation mechanism in type 2 diabetes. Control Eng Pract 2014;26:151–9.
[19] Daskalaki E, Diem P, Mougiakakou SG. An Actor-Critic based controller for glucose regulation in type 1 diabetes. Comput Methods Programs Biomed 2013;109(2):116–25.
[20] Man CD, et al. The UVA/PADOVA type 1 diabetes simulator: new features. J Diabetes Sci Technol 2014;8(1):26–34.

[21] Daskalaki E, Diem P, Mougiakakou SG. Personalized tuning of a reinforcement learning control algorithm for glucose regulation. Conf Proc IEEE Eng Med Biol Soc 2013;2013:3487–90.
[22] Patil P, Kulkarni P, Shirsath R. Padma Suresh L, Sekhar Dash S, Panigrahi BK, editors. Sequential decision making using Q learning algorithm for diabetic patients. New Delhi: Springer; 2014. p. 313–21.
[23] De Paula M, Ávila LO, Martínez EC. Controlling blood glucose variability under uncertainty using reinforcement learning and Gaussian processes. Appl Soft Comput 2015;35:310–32.
[24] Yasini S, Naghibi-Sistani MB, Karimpour A. Agent-based simulation for blood glucose control in diabetic patients. Int J Appl Sci Eng Technol 2009;5:40–7.
[25] Bergman RN. Minimal model: perspective from 2005. Horm Res 2005;64(Suppl 3):8–15.
[26] Javad MOM, Zeid I, Kamarthi S. Reinforcement learning algorithm for blood glucose control in diabetic patients. ASME 2015 international mechanical engineering congress and exposition. Texas, USA: Houston; 2015. p. 9.
[27] Noori A, Sadrnia MA, Sistani MbN. Glucose level control using temporal difference methods. Iranian Conference on Electrical Engineering (ICEE). 2017.
[28] Palumbo P, Panunzi S, Gaetano A. Qualitative behavior of a family of delay-differential models of the Glucose-Insulin system. Discret Contin Dyn Syst - Ser B 2006;7(2):399–424.
[29] Daskalaki E, et al. Preliminary results of a novel approach for glucose regulation using an actor-critic learning based controller. UKACC International Conference on Control. 2010.
[30] Daskalaki E, Diem P, Mougiakakou S. Adaptive algorithms for personalized diabetes treatment. Data-driven modeling for diabetes. 2014. p. 91–116.
[31] Avila L, Martínez E. Behavior monitoring under uncertainty using Bayesian surprise and optimal action selection. Expert Syst Appl 2014;41(14):6327–45.
[32] Avila L, Martínez E. An active inference approach to on-line agent monitoring in safety–critical systems. Adv Eng Inform 2015;29(4):1083–95.
[33] De Paula M, Martínez EC. Probabilistic optimal control of blood glucose under uncertainty. 22nd European Symposium on Computer Aided Process Engineering 2012:1400.
[34] Daskalaki E, Diem P, Mougiakakou SG. Model-free machine learning in biomedicine: feasibility study in type 1 diabetes. PLoS One 2016;11(7):e0158722.
[35] Shifrin M, Siegelmann H. Insulin Regimen ML-based control for T2DM patients. 2017. p. 11.
[36] Bastani M. Model-free intelligent diabetes management using machine learning. 2013. p. 161.
[37] Luckett DJ, et al. Estimating dynamic treatment regimes in mobile health using V-learning. 2017. p. 26.
[38] Jiang Y, Jiang Z-P. Computational adaptive optimal control with an application to blood glucose regulation in type 1 diabetics. Control Conference (CCC), 2012 31st Chinese. 2012. p. 6.
[39] Mösching A. Reinforcement learning methods for glucose regulation in type 1 diabetes, in mathematical engineering statistics and applied probability. Ecole Polytechnique Federale de Lausanne; 2016. p. 100.
[40] Hovorka R, et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. Physiol Meas 2004;25(4):905–20.
[41] Weng W-H, et al. Representation and reinforcement learning for personalized glycemic control in septic patients. 31st Annual Conference on Neural Information Processing Systems (NIPS 2017) Workshop on Machine Learning for Health (ML4H) 2017. p. 5.
[42] D. Ngo P, et al. Reinforcement-learning optimal control for type-1 diabetes. IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). 2018. p. 4.
[43] Myhre JN, et al. Controlling blood glucose levels in patients with type 1 diabetes using fitted Q-iterations and functional features. 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP) 2018:1–6.
[44] Ngo PD, et al. Control of blood glucose for Type-1 diabetes by using reinforcement learning with feedforward algorithm. Comput Math Methods Med 2018;2018:4091497.
[45] Fox I, Wiens J. Reinforcement learning for blood glucose control: challenges and opportunities. International Conference on Machine Learning (ICML). 2019.
[46] Clarke W, Kovatchev B. Statistical tools to analyze continuous glucose monitor data. Diabetes Technol Ther 2009;11(Suppl 1):S45–54.
[47] Sun Q, et al. A dual mode adaptive basal-bolus advisor based on reinforcement learning. IEEE J Biomed Health Inform 2018.
[48] Sun Q, Jankovic MV, Mougiakakou SG. Reinforcement learning-based adaptive insulin advisor for individuals with type 1 diabetes patients under multiple daily injections therapy. Engineering in Medicine and Biology Conference. 2019.
[49] Sun Q, Jankovic MV, Mougiakakou SG. Impact of errors in carbohydrate estimation on control of blood glucose in type 1 diabetes. 2018 14th Symposium on Neural Networks and Applications (NEUREL) 2018:1–5.
[50] Sun Q, et al. Personalised adaptive basal-bolus algorithm using SMBG/CGM data. 11th International Conference on Advanced Technologies & Treatments for Diabetes (ATTD2018). 2018.
[51] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res 1996;4:237–85.
[52] Rasmussen CE. Gaussian processes in machine learning. Advanced lectures on machine learning. Berlin, Heidelberg: Springer; 2004. p. 63–71.
[53] Gao F, Jia W. Perspectives on continuous glucose monitoring technology. Continuous Glucose Monitoring 2018:207–15.
[54] Steil GM. Algorithms for a closed-loop artificial pancreas: the case for proportional-integral-derivative control. J Diabetes Sci Technol 2013;7(6):1621–31.
[55] Schmidt S, Norgaard K. Bolus calculators. J Diabetes Sci Technol 2014;8(5):1035–41.
[56] Factors affecting blood glucose. ADA; 2015 [cited 2018 28 June 2018]; Available from:. http://www.webcitation.org/719KXfVyv.
[57] Wang Y, et al. Automatic bolus and adaptive basal algorithm for the artificial pancreatic beta-cell. Diabetes Technol Ther 2010;12(11):879–87.
[58] Hughes CS, et al. Anticipating the next meal using meal behavioral profiles: a hybrid model-based stochastic predictive control algorithm for T1DM. Comput Methods Programs Biomed 2011;102(2):138–48.
[59] Heinemann L. Variability of insulin absorption and insulin action. Diabetes Technol Ther 2002;4(5):673–82.
[60] Magni L, et al. Evaluating the efficacy of closed-loop glucose regulation via control-variability grid analysis. J Diabetes Sci Technol 2008;2(4):630–5.
[61] Kovatchev BP, et al. Assessment of risk for severe hypoglycemia among adults with IDDM: validation of the low blood glucose index. Diabetes Care 1998;21(11):1870–5.
[62] Mougiakakou S, Daskalaki E, Diem P. Estimation of insulin based on reinforcement learning. United States.; 2019.
[63] Mei J, et al. Optimizing patient treatment recommendations using reinforcement learning combined with recurrent neural network patient state simulation. United States; 2019.
[64] Bally L, et al. Carbohydrate estimation supported by the GoCARB system in individuals with type 1 diabetes: a randomized prospective pilot study. Diabetes Care 2017;40(2):e6–7.
[65] Tseng P, et al. Functional, RF-Trilayer sensors for tooth-mounted, wireless monitoring of the oral cavity and food consumption. Adv Mater 2018;30(18):e1703257.
[66] Jia G, et al. A framework design for the mHealth system for self-management promotion. Biomed Mater Eng 2015;26(Suppl 1):S1731–40.
[67] Waidyanatha N. Towards a typology of integrated functional early warning systems. Int J Crit Infrastruct 2010;6(1).
[68] Dalla Man C, Rizza RA, Cobelli C. Meal simulation model of the glucose-insulin system. IEEE Trans Biomed Eng 2007;54(10):1740–9.
[69] Duan Y, et al. Benchmarking deep reinforcement learning for continuous control. Proceedings of the 33rd International Conference on Machine Learning. 2016.
[70] Kangin D, Pugeault N. Combination of supervised and reinforcement learning for vision-based Autonomous control. International Conference on Learning Representations. 2018.
[71] Lee J, et al. Transfer entropy estimation and directional coupling change detection in biomedical time series. Biomed Eng Online 2012;11:19.