Offline Deep Reinforcement Learning and Off-Policy Evaluation for Personalized Basal Insulin Control in Type 1 Diabetes

Taiyu Zhu, *Member, IEEE*, Kezhi Li, *Member, IEEE*, and Pantelis Georgiou, *Senior Member, IEEE*

***Abstract*— Recent advancements in hybrid closed-loop systems, also known as the artificial pancreas (AP), have been shown to optimize glucose control and reduce the self-management burdens for people living with type 1 diabetes (T1D). AP systems can adjust the basal infusion rates of insulin pumps, facilitated by real-time communication with continuous glucose monitoring. Empowered by deep neural networks, deep reinforcement learning (DRL) has introduced new paradigms of basal insulin control algorithms. However, all the existing DRL-based AP controllers require a large number of random online interactions between the agent and environment. While this can be validated in T1D simulators, it becomes impractical in real-world clinical settings. To this end, we propose an offline DRL framework that can develop and validate models for basal insulin control entirely offline. It comprises a DRL model based on the twin delayed deep deterministic policy gradient and behavior cloning, as well as off-policy evaluation (OPE) using fitted Q evaluation. We evaluated the proposed framework on an *in silico* dataset containing 10 virtual adults and 10 virtual adolescents, generated by the UVA/Padova T1D simulator, and the OhioT1DM dataset, a clinical dataset with 12 real T1D subjects. The performance on the *in silico* dataset shows that the offline DRL algorithm significantly increased time in range while reducing time below range and time above range for both adult and adolescent groups. The high Spearman's rank correlation coefficients between actual and estimated policy values indicate the accurate estimation made by the OPE. Then, we used the OPE to estimate model performance on the clinical dataset, where a notable increase in policy values was observed for each subject. The results demonstrate that the proposed framework is a viable and safe method for improving personalized basal insulin control in T1D.**

***Index Terms*—Artificial pancreas, deep learning, diabetes, off-policy evaluation, offline reinforcement learning**

I. Introduction

TYPE 1 diabetes (T1D) is a group of metabolic disorders that affect millions of people worldwide and is the most common form of childhood diabetes [1]. Due to the insufficient insulin production by the pancreatic β-cell, people living with T1D require exogenous insulin administration in the long-term self-management of blood glucose (BG) concentration, aiming to maintain BG levels in a therapeutic target range (70-180 mg/dL) and minimize the presence of adverse glycemic events. Hyperglycemia, generally defined as BG above 180 mg/dL, is the main cause of long-term macrovascular and microvascular complications, such as retinopathy, neuropathy, nephropathy, and coronary heart disease [2]. Meanwhile, BG concentration below 70 mg/dL is usually classified as hypoglycemia, which is more dangerous than hyperglycemia and is often associated with intense complications, such as seizures, angina, and coma [3]. Severe hypoglycemia would lead to life-threatening events and even increase the risk of death [4].

The standard therapy of insulin replacement in T1D is known as basal-bolus insulin regimen, which mimics the insulin secretion by the pancreas and can be delivered through either multiple daily injections (MDIs) or continuous subcutaneous insulin infusion (CSII). Bolus insulin is administered at mealtime to compensate for the postprandial glucose increase, while basal insulin, referred to as background insulin, aims to regulate BG during fasting. Compared with conventional MDIs, CSII with insulin pump therapy has been demonstrated to effectively reduce HbA1c, glycemic variability, and the severity and frequency of hypoglycemia frequency in randomized controlled clinical trials [5]–[7].

The use of real-time continuous glucose monitoring (CGM) systems [8] has increased rapidly in recent years, and they have been shown to reduce severe hypoglycemia in people with T1D either using MDI therapy [9] or CSII therapy [10], [11]. Combining insulin pumps, CGM, and control algorithms, hybrid closed-loop glucose control systems, also known as the artificial pancreas (AP), are emerging in recent outpatient studies and have been shown to further improve glucose control while reducing the self-care burdens in T1D management [12], [13]. In this case, basal insulin is continuously delivered with small insulin doses 24 hours a day. The real-time communication between CGM and insulin pumps enable AP control algorithms to adjust basal rates (BRs), i.e., infusion rates of basal insulin, every few minutes according to the resolution of CGM. An ideal insulin regimen would ensure that BRs match individual's physiological needs and thus optimize glucose control. However, the key challenge lies in large inter- and intra-subject variability in clinical settings that affect insulin sensitivity [14], [15], due to a variety of internal

This work was supported by EPSRC EP/P00993X/1 and President's Ph.D. Scholarship at Imperial College London. *(Corresponding author: K. Li)*

T. Zhu, P. Georgiou are with Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom. (e-mail: {taiyu.zhu17, pantelis}@imperial.ac.uk).

K. Li is with Institute of Health Informatics, University College London, London, United Kingdom. (e-mail: ken.li@ucl.ac.uk).

and external factors, such as meal ingestion, physical activities, and recurrent illness. Therefore, intelligent algorithms for personalized basal insulin control are needed to fully exploit the benefits of AP, which can automatically adapt to real-time changes in physiological states.

Reinforcement learning (RL) is a category of machine learning technologies in artificial intelligence, which offers promise in various practical healthcare applications [16]. It has been increasingly integrated into glucose control over the past decades [17], [18]. Meanwhile, deep learning, another branch of machine learning, has been widely applied in T1D management [19] and achieved state-of-the-art performance in glucose prediction [20]–[23], by taking advantage of a large amount of CGM data. Integrating with deep learning technologies, deep reinforcement learning (DRL) employs deep neural networks to extract hidden representations from observations in the Markov decision process. DRL has achieved expert-level control in a number of complex tasks with a high-dimensional environment, such as robotics [24], the game of Go [25], drug design [26], and even nuclear fusion [27]. Pioneering work has demonstrated that DRL is an effective solution to optimize the policy of hormone administration in AP systems [28]–[34].

However, several notable limitations exist in these studies, which hamper the widespread adoption of DRL in actual T1D management. Existing approaches, including our previous work [31]–[33], have primarily relied on online learning in T1D simulators, involving long-term error and trial exploration, which is possible in virtual simulation but impractical and unsafe in clinical settings. Furthermore, none of these studies has evaluated algorithms on real clinical datasets to demonstrate the generalization of DRL models, due to the lack of methods for off-policy evaluation (OPE). Therefore, in this study, we propose an offline DRL-OPE framework to overcome these limitations, which addresses the challenges of policy learning and evaluation in basal insulin control. By utilizing both *in silico* and clinical datasets, we validate the effectiveness and generalization of the proposed algorithms. In particular, the OhioT1DM dataset is employed for evaluation, which consists of the data of 12 people with T1D on pump therapy. It should be noted that the length of the clinical dataset is eight weeks, which is shorter than that of the *in silico* dataset (nine months), mainly due to the high cost of clinical trials. The details of the datasets are presented in Section IV-A. The proposed framework exhibits promising performance and provides a viable method to develop and evaluate clinical decision support algorithms in T1D management.

II. Related Work

As mentioned earlier, a major obstacle to the use of DRL in the healthcare field is that in most cases, an agent needs to dynamically interact with the environment after being randomly initialized. In this regard, the UVA/Padova T1D simulators, including versions S2008 [35] and S2013 [36], provide a perfect platform that allows agents to freely explore different actions or exploit the learned policy. The version S2013 is upgraded from the version S2008 by incorporating a nonlinear hypoglycemia model and counter-regulation (i.e., glucagon administration), which was accepted by the US Food and Drug Administration for pre-clinical *in silico* trials and has been proved to match the observations in actual clinical trials [37]. Most existing studies on glucose control and DRL developed algorithms in the UVA/Padova T1D simulators and reported primary outcomes based on *in silico* population.

Employing the UVA/Padova T1D simulator (S2008), deep Q-networks (DQNs) [28] and proximal policy optimization [29] were proposed to control basal-bolus insulin with discrete insulin doses. Similarly, Fox *et al.* [30] proposed a soft actor-critic algorithm with continuous action space, which exhibited low glycemic risk on 2.1 million hours of simulated data. Meanwhile, Zhu *et al.* [31], [32] applied double DQNs based on dilated recurrent neural networks [22] to optimize single-hormone (basal insulin) and dual-hormone (basal insulin and glucagon) control in the UVA/Padova T1D simulator (S2013). Later, the authors proposed the deep deterministic policy gradient (DDPG) to recommend bolus insulin at mealtimes using the same simulated environment [33]. In [34], the authors used the Hovorka model [38] for *in silico* simulation and developed a trust-region policy optimization algorithm for basal insulin control.

Fortunately, recent breakthroughs in offline DRL [39]–[43] and OPE [44]–[46] respectively address the problem of policy learning and policy evaluation on fixed historical datasets instead of online interactions with environments. A major challenge for offline DRL is the trade-off between distributional shift and policy improvement, which is generally tackled by either using policy constraints [39]–[41] or value function regularization [42], [43]. Fitted Q evaluation (FQE) [45] is a promising OPE method that has been demonstrated to provide accurate estimates of policy values for several large benchmark datasets [47], [48], as well as for a healthcare application of sepsis treatment [49].

III. Methods

In this section, we describe the problem of basal insulin control using the framework of offline DRL and OPE, where offline DRL enables offline policy learning (i.e., model training), while OPE is used for offline policy evaluation (i.e., model validation and testing). As shown in Fig. 1, the framework includes a total of four steps. The first step is to train offline DRL agents with different hyperparameter settings to obtain multiple policies. Then we train a value function for each learned policy with the OPE method and the same training data in step two. Next, we use the trained OPE to estimate policy values on validation data and select the best policy with the highest value in step three. The final step is to evaluate the selected policy on testing data and/or in clinical trials to demonstrate unbiased performance. The optimal policy can provide the subject with the strategy of basal insulin control, and new transitions are generated to expand data collection. Table II in Appendix presents the notations used in this work.

A. Problem Formulation

The DRL environment of glucose control can be modelled as a Markov decision process, denoted by a tuple $\mathcal{M} =$

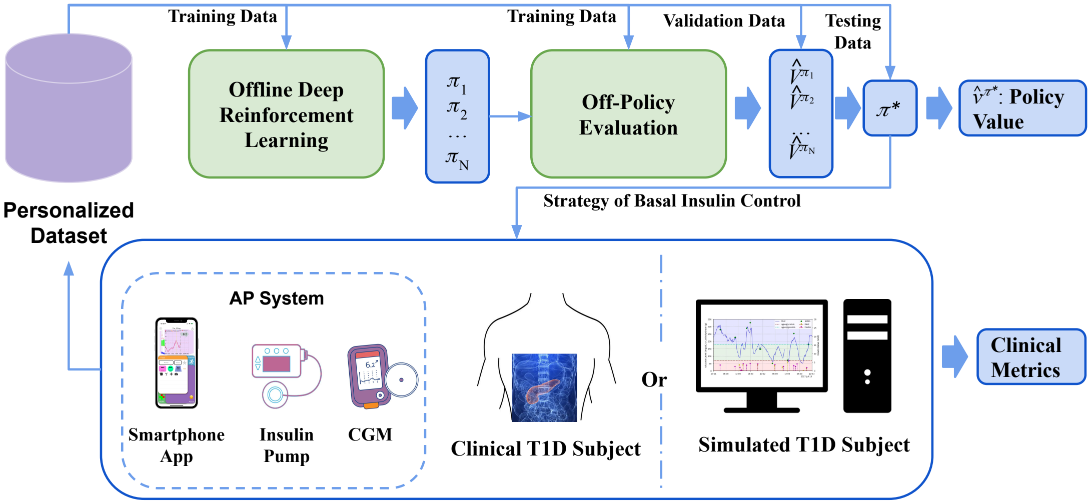


Fig. 1: System architecture of developing glucose control algorithms in T1D management using the proposed offline DRL and OPE framework. The thin and thick arrows indicate the input and output of each module, respectively. The model training, validation, and testing are performed in completely offline settings. The proposed algorithms can be applied to either a simulated T1D subject or a real T1D subject.

$(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \rho_0)$, where $\mathcal{S}$ denotes the state space; $\mathcal{A}$ is a set of actions; $\mathcal{T}(s_{t+1}|s_t, a_t)$ defines the transition distribution; $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function; $\gamma \in (0, 1]$ is a discount factor, and ρ_0 represents the distribution of initial states.

In general, the goal of DRL is to optimize the performance of a policy $\pi(a_t|s_t)$ in a given environment or on a historical dataset $\mathcal{D}$ in terms of offline settings. Typically, such an offline dataset is collected from one or multiple behavior policy $\pi_\beta(a_t|s_t)$, which may be different from $\pi(a_t|s_t)$, and contains a set of historical transitions $\mathcal{D} = \{s_t^i, a_t^i, s_{t+1}^i, r_t^i\}$, where $r_t = r(s_t, a_t)$. Denoting an episode (i.e., a trajectory) τ with a length of L, the dataset can also be defined as $\mathcal{D} = \{\tau^i\}$, where $\tau = (s_0, a_0, r_0, \ldots, s_L, a_L, r_L)$, and $s_0 \in \rho_0$. The state-value function is the excepted return when an agent starts from a state s and follows the policy π, which is defined as

$$V^\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s]. \tag{1}$$

To quantitatively estimate offline performance of DRL on $\mathcal{D}$, policy value v^π is a common metric [47]–[49] that computes the expected state value of initial states with the distribution ρ_0, which can be denoted as

$$v^\pi(\rho_0) = \mathbb{E}_{s_0 \sim \rho_0}[V^\pi(s_0)]. \tag{2}$$

In particular, we incorporate the glucose control into $\mathcal{M}$ as follows.

1) States: In an AP system, the BRs are frequently adjusted according to the real-time CGM measurements and the information of daily activities. Therefore, we model the states using the features extracted from a CGM sequence during the past hour and the external events recorded by insulin pumps or smartphone apps for diabetes management, which were selected according to the feature processing in our previous work and other studies in the literature [28], [29], [31]–[34]. We consider a total of seven features for glucose patterns, including the current BG level, the mean, maximum, and minimum values of the sequence, the maximum difference between adjacent measurements, and percentages of hypo- and hyperglycemia. The cyclical encoding of timestamps, time and amount of last carbohydrate ingestion and meal insulin bolus are also included in the states. In real-life scenarios, the information of carbohydrates and insulin bolus can be provided by built-in bolus calculators of insulin pumps and is unlikely to be missing, but it is common that dietary data are accompanied by variability and misestimation. In this case, personalized DRL algorithms are capable of adaptively learning this variability from personal habits or patterns and thus maintain good BG control. These features provide essential information to the DRL controller in three ways. First, glucose features represent glycemic variability, rates of glucose changes, and trends of glucose levels in the past hour, indicating current glucose dynamics. Second, exogenous events, including meal intake and insulin bolus delivery, have long-term impacts on glucose levels. Lastly, information on times of the day enables the model to learn time-varying basal rate profile.

During the feature selection process, we undertook model validation through exhaustive search. This enabled us to investigate various feature combinations and select the subset that achieved the best validation performance, i.e., the highest policy values, as shown in the third step of Fig 1. In this procedure, we excluded several factors from the full feature set: glucose sequences from the previous hour, low blood glucose index (LBGI), and high blood glucose index (HBGI).

2) Actions: The action space is continuous, defined by the amount of BRs. In offline settings, we do not need to explicitly specify the range for random exploration.

3) Rewards: We design a reward function based on the clinical targets of time in range (TIR), time above range

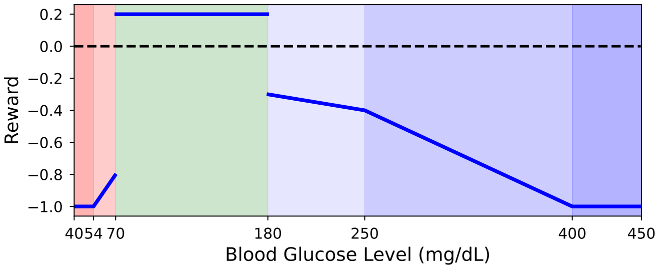


Fig. 2: Reward function based on the clinical metrics of TAR, TIR, and TBR in glucose control.

(TAR), and time below range (TBR), which are the standard metrics of glucose control recommended by the International Consensus [50]. In particular, TIR refers to the percentage of time that a T1D subject spends within the euglycemic range (70-180 mg/dL). TBR is the percentage of time spent in the hypoglycemic region and can be divided into level 1 (54-70 mg/dL) and level 2 (below 54 mg/dL). Similarly, TAR stands for the percentage of time spent in hyperglycemia and can be divided into level 1 (180-250 mg/dL) and level 2 (above 250 mg/dL). Hence, we use a piecewise function with multiple slopes to compute a reward r_t using the current BG level at s_t and the intervals of clinical targets, as depicted in Fig. 2, which penalizes the agent when BG levels move toward hyperglycemia or hypoglycemia. The design of this reward function is also based on the experiments of our previous study [31] that explored different reward functions for online DRL in basal insulin control. The step function enables significant differences between the rewards for euglycemia and hypo- and hyperglycemia and motivates the DRL agent to perform the actions that lead to high TIR. We denote this reward function as a TIR reward. Another TBR reward function is used in OPE to estimate hypoglycemia performance, which assigns -1 to a BG level below 70 mg/dL and 0 otherwise. We terminate episodes when a BG level is below 40 mg/dL or above 450 mg/dL, indicating that there is a medical emergency.

B. Offline Deep Reinforcement Learning

In our previous work, we demonstrated the effectiveness of Double DQNs and DDPG in improving glucose control through online learning [31]–[33]. However, in this study, we adopt a distinct approach for basal insulin control using the twin delayed deep deterministic policy gradient (TD3) algorithm [51] within the proposed offline learning framework. By using target networks, delayed policy updates, and target policy smoothing, TD3 outperformed DDPG in benchmark environments [51]. Furthermore, combining a behavior cloning regularization, TD3 has been demonstrated to be a minimalist method to achieve state-of-the-art performance in offline tasks [40], which has much lower complexity and computational cost when compared with other offline DRL algorithms [42], [43]. The behavior cloning regularization is a straightforward modification that incorporates a supervised learning loss to encourage the policy to prioritize actions present in the historical dataset. This technique is desirable in the context of glucose control, where the agent aims to achieve similarity to historical actions while also accommodating necessary adaptations to effectively handle diverse clinical scenarios. Therefore, we choose this variant of TD3 to learn personalized glucose control for T1D subjects using previously collected clinical data.

As an actor-critic approach, we first formulate the state-action value function (i.e., critic) Q^π updated by the Bellman equation as follows:

$$Q^\pi(s_t, a_t) = r_{t+1} + \gamma \mathbb{E}_{s_{t+1}\sim\mathcal{T}(s_{t+1}|s_t,a_t)}[V^\pi(s_{t+1})], \quad (3)$$

which is the expected return when an agent starts from a state s_t and takes action a_t. Parameterized by deep neural networks with parameters θ and ϕ, we denote the Q function and policy (i.e., actor) by Q_θ and π_ϕ, respectively. By doing this, we update Q_θ by temporal difference learning with the temporarily frozen target Q-network $Q_{\theta'}$ and actor network $\pi_{\phi'}$. To reduce overestimation bias, TD3 uses twin Q functions (i.e., Q_{θ_1} and Q_{θ_2}) to select a less biased value estimate in each updating step. In offline settings, we sample a mini-batch $\mathcal{B}$ with M transitions $d = (s_t, a_t, s_{t+1}, r_t)$ from $\mathcal{D}$ to calculate the loss $\mathcal{L}$ of temporal difference learning as follows

$$\begin{aligned}\mathcal{L}(\theta_i) = \mathbb{E}_{d\sim\mathcal{B}}[(r_{t+1} + \gamma \min_{j=1,2} Q_{\theta'_j}(s_{t+1},\pi_{\phi'}(s_{t+1}) + \epsilon) \\ -Q_{\theta_i}(s_t, a_t))^2],\end{aligned} \quad (4)$$

where $\epsilon \sim \texttt{clip}(\mathcal{N}(0, \sigma))$ is a clipped random noise to mitigate overfitting and smooth estimation in the deterministic policy. Furthermore, to mitigate the challenges of extrapolation error and distributional shift, we incorporated a behavior cloning term in the policy gradient of the actor update. This term, as proposed in [40], has been shown to effectively address these issues in offline DRL settings. The definition is as follows:

$$\mathcal{J}(\phi) = \mathbb{E}_{d\sim\mathcal{B}}[\lambda Q_{\theta_1}(s_t, \pi_\phi(s_t) - (a_t - \pi_\phi(s_t))^2], \quad (5)$$

where $\mathcal{J}$ is the loss function; $\lambda = \alpha(\mathbb{E}_{d\sim\mathcal{B}}|Q_{\theta_1}(s_t, a_t)|)^{-1}$ is a normalization factor calculated by the mean absolute values of Q_{θ_1}, and α is a weighting factor. The overall process of developing offline DRL is summarized in Algorithm 1.

Algorithm 1 Developing Offline DRL

```
Input: Randomly initialized $\theta_1$, $\theta_2$, and $\phi$, training data $\mathcal{D}_{\text{train}}$, interval to delay policy update $t_d$, averaging factor $\mu$
Output: Learned policy $\pi_\phi$
Set target networks: $\theta'_1 \leftarrow \theta_1$, $\theta'_2 \leftarrow \theta_2$, $\phi' \leftarrow \phi$
for steps $t \in 1, 2, \ldots, T_{\text{DRL}}$ do
    Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}_{\text{train}}$
    Update $Q_{\theta_i}$ for the critic using $\mathcal{L}(\theta_i)$ in Equation (4), for $i = 1, 2$
    if $t \mod t_d = 0$ then
        Update $\pi_\phi$ for the actor using $\mathcal{J}(\phi)$ in Equation (5)
        $\theta'_i \leftarrow (1 - \mu)\theta'_i + \mu\theta_i$, for $i = 1, 2$
        $\phi' \leftarrow (1 - \mu)\phi' + \mu\phi$
    end if
end for
```

C. Off-Policy Evaluation

The goal of OPE is to estimate the performance of DRL models using historical datasets and thus to rank and select polices, which enables us to assess the personalized glucose control algorithms without conducting actual clinical trials. Particularly, we choose FQE as the OPE method with the implementation by Paine *et al.* [47], considering it provided accurate and robust estimation in healthcare settings [49]. Given a evaluation policy π with parameters ϕ, FQE initializes a critic with parameters ψ and retrains it using bootstrapping targets of the Bellman equation and a supervised learning loss as follows:

$$\mathcal{L}(\psi) = \mathbb{E}_{d\sim\mathcal{B}}[(r_{t+1} + \gamma Q_{\psi'}(s_{t+1}, \pi_{\phi}(s_{t+1})) - Q_{\psi}(s_t, a_t))^2]. \tag{6}$$

The pseudo-code of OPE is presented in Algorithm 2, where the output is the estimated state-value function $\widehat{V}^{\pi}(s)$. Assuming a group of candidate policies $\{\pi_1, \pi_2, \ldots, \pi_N\}$ with different hyperparameters is obtained in offline DRL by Algorithm 1, we can apply OPE to estimate a set of state-value functions $\{\widehat{V}^{\pi_1}(s), \widehat{V}^{\pi_2}(s), \ldots, \widehat{V}^{\pi_N}(s)\}$. Using a validation set with the initial state distribution of ρ_0^{val}, we estimate policy values $\{\widehat{v}^{\pi_1}(\rho_0^{\text{val}}), \widehat{v}^{\pi_2}(\rho_0^{\text{val}}), \ldots, \widehat{v}^{\pi_N}(\rho_0^{\text{val}})\}$ by Equation (2), as scalar scores to rank the candidate policies. Various reward functions can be used as input r' to evaluate model performance in different ways. As a result, the optimal policy $\pi*$ with the best scores is selected. Then, the control performance can be measured by either policy value $\widehat{v}^{\pi^*}(\rho_0^{\text{test}})$ for testing data or clinical metrics obtained in clinical or *in silico* trials. Fig. 1 illustrates this process.

Algorithm 2 Developing OPE

Input: Randomly initialized ψ, training data $\mathcal{D}_{\text{train}}$, update interval t_u, policy to be evaluated π_ϕ, reward function r'
Output: Estimated $\widehat{V}^{\pi}(s) = Q_\psi(s, \pi(s))$

```
Set target networks: $\psi' \leftarrow \psi$
for steps $t \in 1, 2, \ldots, T_{\text{OPE}}$ do
    Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}_{\text{train}}$
    Update $Q_\psi$ for the critic using $\mathcal{L}(\psi)$, reward function $r'$ in Equation (6)
    if $t \mod t_u = 0$ then $\psi' \leftarrow \psi$ end if
end for
```

IV. Experiments

A. Offline Datasets

We conducted the experiments using two offline datasets to validate the clinical efficacy of the proposed algorithm, of which the details are summarized as follows. In both datasets, the BG levels were measured by CGM every five minutes.

1) In Silico Data: An *in silico* dataset was generated by the UVA/Padova T1D simulator (S2013) [36], which includes nine months of data of 10 virtual adults and 10 virtual adolescents. To emulate the variations of insulin sensitivity, we introduced a set of additional intra-subject variability [52] by adjusting meal intake protocols and the parameters of physiological models in the simulator. Specifically, the meal times of breakfast, lunch, and dinner followed normal distributions with mean values of 7:00, 14:00, and 21:00, respectively, and a standard deviation of 30 minutes. The corresponding carbohydrate amount for the three types of meals also followed normal distributions with mean values of 70, 110, and 90 grams, and a coefficient of variation of 10%. The misestimation of carbohydrate counting was assumed to follow a uniform distribution with an interval of [0.7, 1.1]. The variability of insulin sensitivity was simulated by adjusting parameters along a time-varying sinusoidal pattern with an amplitude of 30%. We assumed that the parameters of meal absorption and carbohydrate bioavailability were drawn from uniform distributions between ±30% and ±10%, respectively. These scenarios and parameters vary along the day or from meal to meal for each individual, which amounts to intra-subject variability.

During data generation, the virtual subjects used personalized BR profiles and the low glucose suspend (LGS) [53] to control basal insulin and employed a standard bolus calculator to compute meal insulin [54]. The LGS has been demonstrated to significantly reduce the exposure to hypoglycemia in clinical trials [55] and therefore is used as a baseline method to generate demonstration data for DRL. We divided the *in silico* dataset into a development set containing the first six-month data and an unseen testing set containing the remaining three months of data to provide an unbiased evaluation. The last two months of data in the development set were used as a hold-out validation set, while the training set included the first four-month data.

2) Clinical Data: The publicly available OhioT1DM dataset [56] is employed to the analyze the proposed framework. It contains data of 12 people with T1D over an eight-week clinical trial. Each participant wore a Medtronic Enlite CGM sensor to measure BG levels and a Medtronic 530G or 630G insulin pump to deliver basal and bolus insulin, where a personalized BR profile was used. Notably, the T1D participants frequently set temporary BRs during the self-management to manually adjust basal insulin delivery, including zero BR for suspension, to meet insulin needs for a specified period of time, such as physical activities.Thus, we use such personalized BR (PBR) control as a baseline method, in order to reflect the performance of real-world glucose control. Each subject in the OhioT1DM dataset is associated with two XML files. The first file contains approximately six weeks of data, which we employed as development data. We performed an 80/20 split to generate a training set and a hold-out validation set, with the latter consisting of the final 20% of the developmental data. The second file includes approximately two weeks of data, which we used as a hold-out testing set. Fig. 8 in Appendix visually illustrates this process.

B. Experimental Setup and Evaluation Metrics

The offline DRL algorithms developed by *in silico* data were evaluated by both simulation and OPE with the same meal scenarios and variability. Aiming to investigate the clinical performance, we first initialized the simulator using the initial state of the whole testing set for each subject and delivered basal insulin with the control strategy of the offline DRL algorithms through a three-month simulation. To evaluate the outcomes, we use a group of clinical metrics. Besides the aforementioned TBR (level 1 & 2), TIR, and TAR (level 1 & 2), we employed LBGI and HBGI to indicate the risk of adverse glycemic events. The mean of BG levels measured by CGM is also presented. We performed the paired t-test to indicate the statistical significance after using the Shapiro-Wilk test to confirm normality. Furthermore, various control algorithms have been investigated in basal insulin control for both MDI [57] and CSII therapy [30], [58]. Among these, proportional-integral-derivative (PID) control is a popular method, which we use as an additional baseline method in this work. We tuned the setpoint and three personalized parameters (i.e., gains) for proportional, integral, and derivative terms in online simulation with the same length of the development set (i.e., six months), where grid search [30] was performed. Then the personalized PID controller was tested on the same three-month simulation.

Secondly, we evaluated the offline DRL algorithms with OPE and investigated how well the OPE estimation matches actual policy values. In this case, we initialized the simulator using the initial state of each episode in a testing set (ρ_0^{test}) and obtained actual policy values by calculating rewards and state-values through *in silico* simulation for each episode. It is important to note that this is an additional step conducted purely for evaluation purposes and is not an integral part of the proposed offline DRL framework in Fig. 1.

For the real clinical data, the offline DRL algorithms were evaluated by the OPE only since the clinical trials on the same T1D subjects cannot be performed. All the deep learning algorithms were developed by Python 3.8, PyTorch 1.9, and NVIDIA GTX 1080 Ti GPU. The episodes and transitions of datasets were structured by the d3rlpy framework [59]. Table III in Appendix summarizes the values of hyperparameters.

C. Performance on In Silico Dataset

1) Clinical Metrics: Table I presents the results (Mean±STD) of glucose control for the adult group and adolescent group through the three-month simulation, which was evaluated by the clinical metrics. It is worth noting that, when compared with the LGS baseline method, the offline DRL algorithm significantly enhanced TIR for the two virtual cohorts and achieved smaller TAR and TBR. In particular, level 1 TBR, level 1 TAR, and level 2 TBR decreased in both groups, while level 2 TAR was maintained in the adult group. Meanwhile, we observed that the offline DRL algorithm reduced the HBGI and LBGI, indicating a lower risk of hypo- and hyperglycemia, and exhibited smaller mean CGM glucose, indicating more stable BG concentrations. This comprehensive analysis suggests that the offline DRL algorithm effectively improved glucose control for the subjects in the *in silico* dataset. Fig. 9 in Appendix shows the ensemble plots of the ambulatory glucose profile for an adult subject and an adolescent subject to indicate day to day variation of the simulation. It is observed that the performance on the adult group is better than that on the adolescent group. This is because glucose variability in adolescents is larger than that in adults, and adolescents usually have lower body weight in the simulation [36]. To improve the performance on the adolescents, a potential solution is to introduce dual-hormone control (i.e., glucagon delivery), as demonstrated by our previous study [31]. When compared with PID, significant improvements were also obtained by offline DRL for all the clinical metrics. In the experiments, we observed that PID is effective to minimize the distance between the process variable (i.e., mean CGM) and the setpoint (i.e., target glucose levels), but it is challenging for PID to handle scenarios with large variability, such as an overdose of meal insulin bolus, which limits its performance for reducing TAR and TBR.

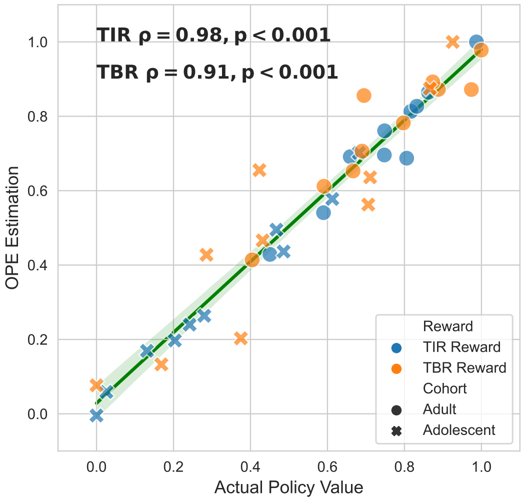


Fig. 3: Scatter plot of the comparison between OPE estimation and actual policy values for the two cohorts and the two reward functions.

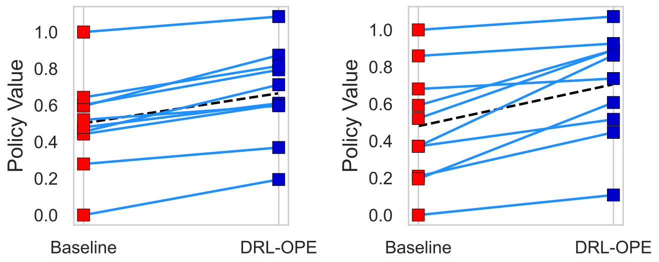
2) OPE Quality: To investigate whether OPE is a reliable method for policy evaluation, we analyze the quality of OPE by a direct measure of rank correlation, which has been widely adopted in existing studies [47]–[49]. Fig. 3 depicts the scatter plot of normalized OPE scores and actual policy values, where the TIR and TBR rewards were used for the adult and adolescent cohorts. We performed the Spearman correlation analysis and obtained rank coefficients ρ of 0.98 ($p < 0.001$) and 0.91 ($p < 0.001$) for TIR and TBR rewards, respectively, indicating high correlation and good ranking statistics. The solid green line stands for the results of the linear regression between the two variables, and the shaded area is a 95% confidence interval. A regression coefficient of 0.99 ($p < 0.001$) was achieved. These results demonstrate that the OPE method estimated accurate policy value and can be used to evaluate model performance in offline settings.

3) Policy Values: The ladder plots in Fig. 4 and 5 show the improvement of the normalized policy values achieved by the proposed offline DRL algorithm in the adult and adolescent groups, respectively, when compared with the LGS baseline method. Of note, the offline DRL algorithm enhanced the

TABLE I: Performance of glucose control on the virtual adult and adolescent cohort through a three-month simulation

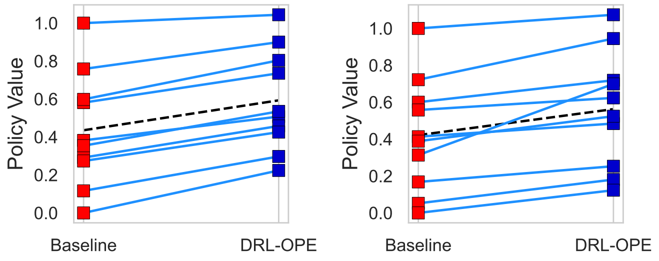
	Adults			Adolescents		
Metric	PID	LGS	Offline DRL	PID	LGS	Offline DRL
TIR (70 – 180 mg/dL) (%)	$72.9 \pm 5.2^{\dagger}$	$75.7 \pm 6.1^{\dagger}$	78.1 ± 6.7	$51.7 \pm 5.0^{\dagger}$	$57.4 \pm 12.7^{*}$	59.9 ± 8.6
TAR (> 180 mg/dL) (%)	23.0 ± 3.1	$21.3 \pm 6.4^{*}$	19.5 ± 6.8	$43.5 \pm 3.3^{*}$	38.4 ± 14.5	36.7 ± 9.8
Level 1 (181 – 250 mg/dL) (%)	20.1 ± 2.2	$19.5 \pm 5.1^{*}$	17.4 ± 5.4	$28.6 \pm 4.4^{*}$	$24.5 \pm 2.1^{*}$	23.8 ± 2.6
Level 2 (> 250 mg/dL) (%)	2.6 ± 1.4	2.1 ± 2.0	2.1 ± 1.9	$14.9 \pm 7.0^{*}$	14.6 ± 14.4	12.2 ± 9.1
TBR (< 70 mg/dL) (%)	$4.1 \pm 2.5^{*}$	$3.0 \pm 1.5^{*}$	2.4 ± 1.7	4.8 ± 2.5	$4.2 \pm 2.6^{\dagger}$	3.4 ± 2.2
Level 1 (54 – 69 mg/dL) (%)	2.3 ± 1.3	1.9 ± 0.9	1.6 ± 1.1	2.6 ± 1.4	$2.4 \pm 1.2^{\dagger}$	2.0 ± 0.9
Level 2 (< 54 mg/dL) (%)	$1.8 \pm 1.3^{*}$	1.1 ± 0.8	0.8 ± 0.5	2.2 ± 1.4	$1.8 \pm 1.8^{\dagger}$	1.4 ± 1.5
LBGI	1.20 ± 0.64	0.87 ± 0.41	0.76 ± 0.48	1.47 ± 0.65	$1.51 \pm 1.39^{\dagger}$	1.24 ± 1.10
HBGI	4.64 ± 0.51	4.24 ± 1.29	3.98 ± 1.67	$9.95 \pm 2.17^{*}$	9.82 ± 6.69	8.37 ± 3.27
Mean CGM (mg/dL)	146 ± 8	144 ± 10	141 ± 13	174 ± 13	$171 \pm 35^{*}$	165 ± 18

$^{*}p \leq 0.05$ $^{\dagger}p \leq 0.01$.



(a) TIR reward (b) TBR reward

Fig. 4: Comparison of policy values between the LGS baseline and the offline DRL algorithm for each T1D subject in the virtual adult group.



(a) TIR reward (b) TBR reward

Fig. 5: Comparison of policy values between the LGS baseline and the offline DRL algorithm for each T1D subject in the virtual adolescent group.

policy value for each T1D individual in the two groups. Specifically, as indicated by the black dashed lines, the mean policy values significantly increased by 32.6% ($p < 0.01$) and 47.0% ($p < 0.01$) for the TIR and TBR rewards, respectively, in the adult group, while the mean policy value was significantly improved by 36.2% ($p < 0.01$) for the TIR reward and 33.5% ($p < 0.01$) for the TBR reward in the adolescent group. These results indicate that the offline DRL algorithm is expected to improve glucose control by simultaneously increasing TIR and decreasing TBR for the virtual cohorts, which are consistent with the clinical results reported in Table I. Fig. 6 shows examples of BG trajectories and corresponding basal insulin delivered by the offline DRL and LGS algorithms. With the same initial state, the use of offline DRL algorithm successfully avoided severe reactive hypoglycemia, nocturnal hypoglycemia, and postprandial hyperglycemia. The offline DRL control promptly adjusted BRs when the measured BG levels tended to move outside of the target range, resulting in higher policy values for both reward functions in the same episodes. The insulin sensitivity of the virtual adult and adolescent population in the UVA/Padova T1D simulator varies from 30 to 90 mg/dL/U and has a median value of around 50 mg/dL/U [36]. Thus, a change of 0.6-1 U in basal insulin can result in an increase or decrease in BG levels by 30-50 mg/dL, as shown in Fig. 6.

D. Performance on Clinical Dataset

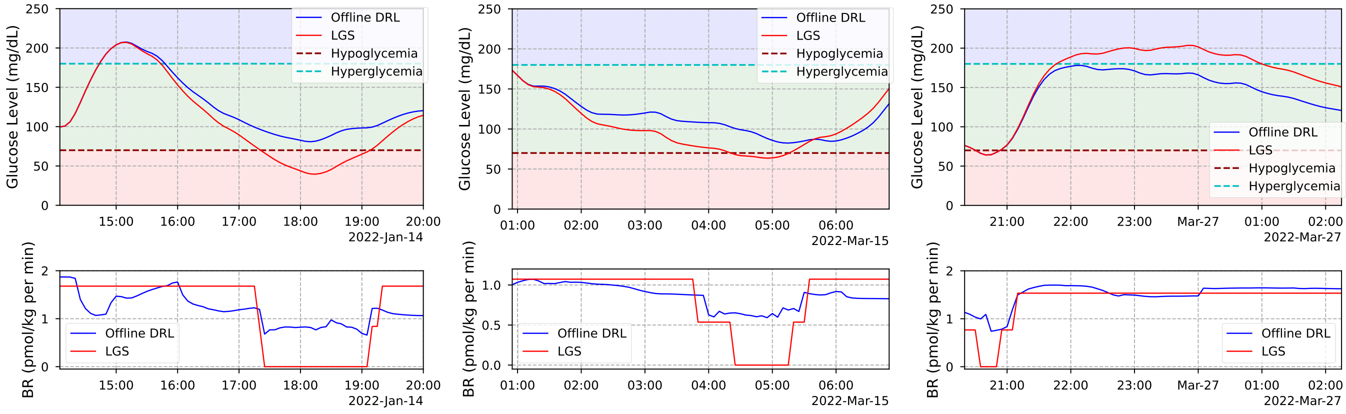
Fig. 7 depicts the normalized policy values of the PBR baseline method and offline DRL algorithm for real T1D subjects. A notable increase in policy values for each subject was observed in terms of the two reward functions. Compared with the PBR baseline, the offline DRL algorithm substantially improved the mean policy values by 45.3% ($p < 0.01$) for both the TIR and TBR rewards, as shown by the black dashed line. An upward trend in policy values indicates that the algorithm obtains larger returns when applied to the target T1D subjects, thereby yielding improved TIR and TBR performance. This is in line with the TIR and TBR reward functions that we carefully crafted in Section III-A.3. Therefore, as depicted by the logic sequence in Fig. 10 in Appendix, we anticipate that using the trained offline DRL models can outperform the temporary BR settings of the PBR control and achieve better glucose control for the OhioT1DM dataseset.

Table IV in Appendix shows the glycemic variability of the OhioT1DM dataset with PBR. It is noted that the TIR, TBR, TAR, LBGI, HBGI, and mean CGM of the real dataset are close to those of the virtual adult and adolescent groups (Table I), which indicates that the *in silico* simulation reflects the real-world scenarios well.

V. Discussion

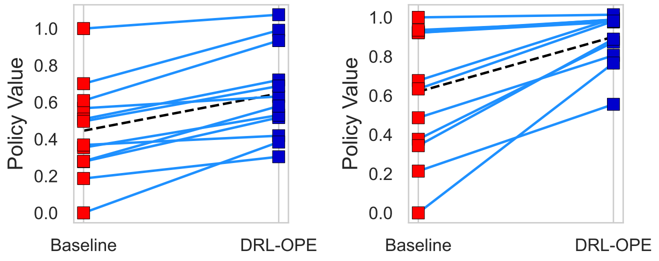
A. Comparison With Existing Studies

To the best of our knowledge, this is the first study that applies offline DRL or OPE to basal insulin control in T1D management, and this is also the first study that combines offline DRL and OPE to solve a healthcare problem. The



(a) Reactive hypoglycemia (b) Nocturnal hypoglycemia (c) Postprandial hyperglycemia

Fig. 6: Examples of BG trajectories and corresponding BRs controlled by the offline DRL algorithm and the LGS baseline method. The use of the offline DRL algorithm prevented the potential adverse glycemic events that occurred in the LGS control.



(a) TIR reward (b) TBR reward

Fig. 7: Comparison of policy values between the PBR baseline and the offline DRL algorithm for each T1D subject in the clinical dataset.

proposed framework has the potential to be applied to many other health problems, such as sepsis management [49], [60]. It should be noted that the novelty of this work also lies in the innovations in the aforementioned four-step framework, such as the design of a T1D-specific Markov decision process and a customized UVA/Padova simulator (S2013) with an OpenAI Gym interface. Due to different datasets and various experimental settings, it is difficult to draw a head-to-head comparison of the numeric results of clinical metrics or policy values. However, in our previous work [31], we identified that optimizing glucose control by adjusting a single hormone (i.e. basal insulin) is a challenging task. The reported TIR and TBR results on the *in silico* dataset (Table I) are comparable with those by the online DRL algorithm after training agents through thousands of simulated days. Although a two-step transfer learning framework was proposed in [31] to mitigate the high demand for personalized data, it still needs to fine-tune the policy with online interactions, where an undertrained agent may produce dangerous actions at the beginning of clinical trials. Moreover, the proposed DRL algorithm conservatively updated the policy by behavior cloning regularization with a weighting factor α (Equation 5), where a smaller α means the update is more inclined to imitation learning. As shown in Fig. 6, the BRs of the DRL algorithm were close to those of the LGS method, and notable differences were only observed when there was a risk of adverse glycemic events. Although this feature may lead to limited performance improvements, it prevents the DRL policy from deviating too far from behavior polices π_β, aiming to avoid out-of-distribution actions and provide better safety guarantees for healthcare applications.

Several studies [28]–[30] added meal insulin bolus into action space and obtained new control strategies for total insulin delivery through long-term random exploration (e.g., millions of simulated hours), which is not feasible for real-world scenarios. In the literature, the parameters of the bolus calculator, such as the insulin-to-carbohydrate ratio, insulin sensitivity (i.e., the correction factor), were optimized for meal insulin dosing in T1D simulators [52], [61], [62], including our previous work with online DRL [33]. However, these parameters are not available in current public T1D datasets. The number of transitions of meal insulin bolus is quite limited (i.e., three or four times per day), while a wide range of carbohydrate content in food ingestion requires a large action space. Therefore, these parameters have not been considered in developing personalized offline DRL control yet, due to insufficient historical data. If a large amount of real-world data with these parameters are available in the future, we will consider optimizing them by extending our previous work [33]. The scenarios considered in this study are the same as the real use case of most AP systems [12]. That is, users manually enter carbohydrate amounts into insulin pumps that calculate meal insulin bolus with bolus calculators, while BRs can be adjusted at the same resolution of CGM (i.e., every five minutes), using built-in personalized control algorithms. Such a real-time setting enables BR changes at different times of the day and thus optimizes basal profile for each individual.

Currently, the existing studies on DRL [28]–[34] validated their models by simulators rather than actual clinical datasets, and their therapeutic benefits for real T1D subjects were unproven. Expert assessment [63] and a Bayesian framework [64] with Markov-Chain Monte Carlo strategy to estimate physiological models trace by trace (i.e., episode by episode) [62] were used to evaluate supervised learning algorithms on glucose control, but these assessment methods are either costly or time-consuming, especially when the size of testing data is large. In this regard, OPE provides a convenient method to evaluate control algorithms on clinical data and can be adopted by many existing models without much difficulty.

B. Model Implementation

The model implementation is essential to the decision support systems for people living with T1D in daily self-management. Thus, aiming at on-device deep learning for model inference, we implemented the actor network of the offline DRL model on an iOS smartphone by converting the PyTorch model into the TensorFlow Lite format through the Open Neural Network Exchange framework. Then we analyzed the run-time and memory usage of the embedded model using a customized diabetes management app (Fig. 1) developed in our previous work [23]. The experiment was repeated 50 times. After the app received a BG measurement from Dexcom G6 CGM, it took an average of 5.2 ms and 2.1 MB of memory to compute a BR for basal insulin delivery.

Such end-to-end implementation of a personalized control algorithm would incentive the development of the do-it-yourself artificial pancreas (DIY AP) [65], [66], in which people with T1D are able to build AP systems by themselves. However, the existing controllers in DIY AP, such as OpenAPS, AndroidAPS, and Loop, adjust BRs based on fixed physiological parameters and simple formulas but lack personalized algorithms to meet real-world challenges of inter- and intra-subject variability [67]. By employing the proposed offline DRL framework, the users can train, update and evaluate personalized insulin control algorithms based on their own historical data collected during daily self-care.

C. Limitations and Future Work

A potential limitation of this study lies in matching the OPE metrics with actual clinical targets. In this work, we chose the policy value as a primary metric since it is a straightforward measure of the return of episodes and exhibited better accuracy in previous studies [47], when compared with other OPE metrics, such as soft off-policy classification [68]. Using the TIR reward function (Fig. 2), we can expect that good TIR performance would be achieved by the evaluated policy if the policy value is large. Nevertheless, we cannot guarantee that TBR would not rise in extreme cases. Improving TIR without the increase in TBR is a critical consideration of glucose control [50]. Therefore, we introduced a TBR reward function to compute an additional policy value in OPE and only selected policies that obtained both good TIR and TBR policy values in model validation. Another limitation is that the proposed offline DRL needs to be further improved to increase TIR before being tested in actual clinical trials. According to the recommendations from the International Consensus [50], the clinical targets of BG control for adults and adolescents are $> 70\%$ and $\sim 60\%$, respectively. Although we achieved these targets in the UVA/Padova T1D simulator (S2013) with additional intra-subject variability, it is much more challenging to control BG levels in real-life scenarios. Hence, the proposed framework will also be evaluated on large-scale clinical datasets with more diverse populations than that of the OhioT1DM dataset in future work. In addition, during the long-term use of the control algorithm, T1D subjects might experience changes in behavior or physiological conditions. In this case, transfer learning or meta-learning techniques can be integrated into the proposed DRL framework, which can fine-tune deep learning models using approximately two weeks of data [69]. If there are significant changes, we will retrain the DRL model from scratch using six months of data.

State-action values (i.e., Q values) have been used to compare RL policy with clinician policy in healthcare applications [60]. Q values are the same as the state values in Equation (1) if the policy is deterministic. Although policy values are highly correlated with the occurrence of adverse glycemic events, there is a lack of an explicit relationship between OPE metrics and specific BG trajectories. To this end, we will develop data-driven and physiological-based personalized simulators and integrate them with OPE methods in future work. Besides, more *in silico* experiments need to be conducted before the clinical use of the offline DRL algorithm, for which a new version of the UVA/Padova T1D simulator (S2017) [15] will be considered. It is noted that recent T1D simulators have incorporated physical activities and exercise [70], [71]. These data could be used as additional input features and integrated into the states of the Markov decision process as formulated in Section III-A.1. If high-quality exercise data are collected in real clinical trials, the offline DRL algorithm could be further applied. The models can also be directly implemented on insulin pumps by edge computing [23] and partnering with manufactures.

VI. Conclusion

Aiming to improve automated glucose control in T1D management, we propose an offline DRL-OPE framework to optimize BRs for basal insulin delivery. A TD3 model with behavior cloning regularization and an FQE-based OPE method were adopted to develop the control algorithms, which ensure that the model training, validation, and testing can be completed offline without performing actual clinical trials or requiring the assessment by human experts. The proposed algorithm was validated on *in silico* and clinical datasets, which significantly enhanced TIR while reducing TBR and TAR. This improvement also extended to other clinical targets in the three-month simulation. A promising increase in policy values was also noted in the OPE analysis for both datasets.

We envision that the proposed framework would benefit AP systems by using a safe offline process to build and implement personalized control models. It also has great potential to accelerate the development of RL-based algorithms for many other healthcare applications.

Appendix

A. Notations

Table II presents the notations used in this article.

TABLE II: Notation table

Variable	Definition
$\mathcal{M}, \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \rho_0$	Markov decision process, state space, action space, transition distribution, reward function, discount factor, distribution of initial states
π, a_t, s_t, r_t	Policy, the action, state, and reward at timestep t
$\mathcal{D}, \tau, L$	Dataset, trajectory, the length of trajectory
V, v, Q	State value, policy value, state-action value
$\theta, \theta', \phi, \phi'$	Weights of Q-network (critic), target Q-network (critic), actor network, target actor network in DRL
θ_1, θ_2	Weights of the first and the second Q-network of twin Q functions in DRL
$\mathcal{B}, M, d$	Mini-batch, batch size, transition
$\mathcal{L}, \epsilon, \mathcal{J}, \lambda, \mu$	Temporal difference (critic loss), clipped random noise, actor loss, normalization factor, Polyak averaging factor
ψ, ψ'	Weights of critic and target critic in FQE
$\widehat{V}, \widehat{v}$	Estimated state-value function, estimated policy value
t_d, t_u	Interval to delay policy update in DRL, interval to update target networks in FQE

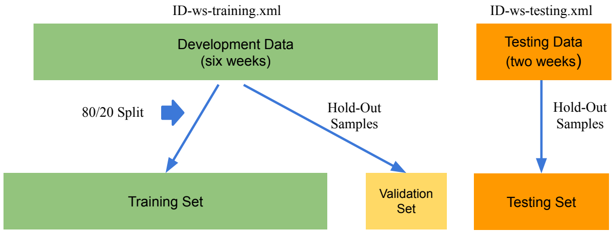


Fig. 8: Data split for the OhioT1DM dataset.

B. Clinical Data Split

Fig. 8 shows the process of dividing the OhioT1DM dataset into training, validation, testing sets.

C. Hyperparameter Tuning

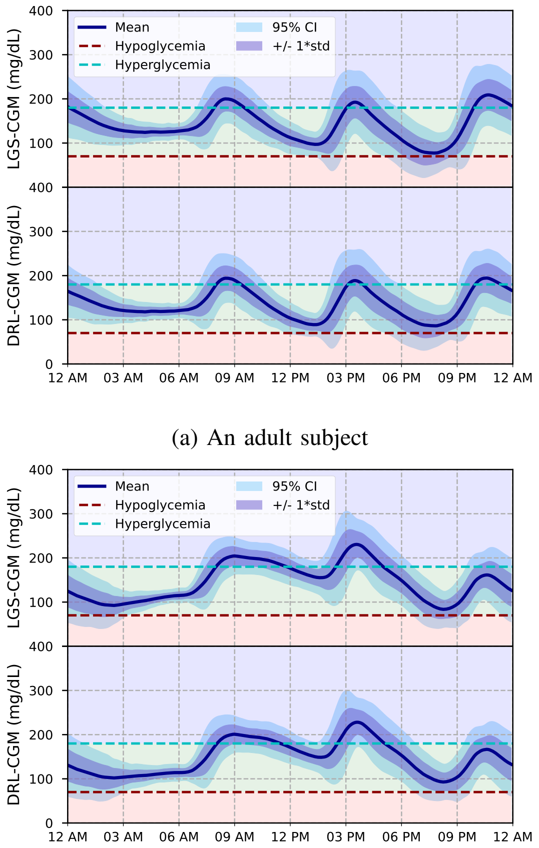
Table III summarizes the hyperparameters used in this work. The hyperparameters were selected by grid search in model validation using OPE, as described in Section III-C.

Parameter	Value
Actor learning rates	5×10^{-5}
Batch size M	64
Critic learning rates	1×10^{-4}
Discount factor γ	0.97
Episode length L	72 (6 hours)
Hidden units of network layers	[256, 256, 256]
Interval to delay policy update t_d	2
Interval to update target networkst_u	100
Normalization constant σ	1×10^{-3}
Number of DRL training steps T_{DRL}	3×10^4
Number of OPE training steps T_{OPE}	2×10^4
Proportion of soft update μ	0.01
Regularization factor α	2.5

TABLE III: List of hyperparameters

D. Ambulatory Glucose Profile

Fig. 9 shows the day-to-day ambulatory glucose profile for a virtual adult and a virtual adolescent using LGS and offline DRL control. It is noted that, compared with LGS, the proposed offline DRL algorithm simultaneously reduced TAR and TBR for the adult subject and notably improved hypoglycemia outcomes for the adolescent subject.



(a) An adult subject

(b) An adolescent subject

Fig. 9: Ambulatory glucose profile for a virtual adult and a virtual adolescent over the three-month testing period.

E. Evaluation Logic Sequence

Fig. 10 demonstrates how we evaluate the offline DRL algorithm on real data. The policy values achieved by the algorithm are strongly correlated with TIR and TBR, thereby providing a measure of glucose control performance.

F. Glycemic Variability of the Real Dataset

Table IV shows the glycemic variability of the OhioT1DM Dataset. The performance is obtained through PBR control.

Acknowledgment

The work is supported by EPSRC EP/P00993X/1 and the President's PhD Scholarship at Imperial College London (UK).

References

[1] A. Katsarou, S. Gudbjörnsdottir, A. Rawshani, D. Dabelea, E. Bonifacio, B. J. Anderson, L. M. Jacobsen, D. A. Schatz, and Å. Lernmark, "Type 1 diabetes mellitus," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–17, 2017.

[2] M. J. Fowler, "Microvascular and macrovascular complications of diabetes," *Clinical Diabetes*, vol. 26, no. 2, pp. 77–82, 2008.

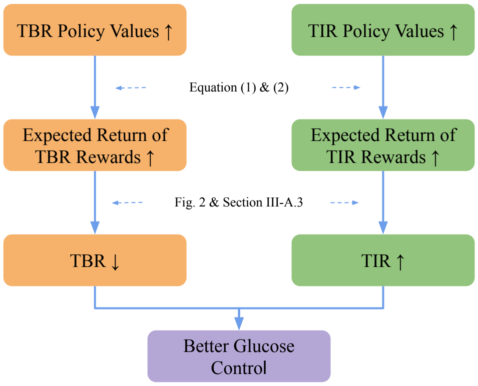


Fig. 10: Logic sequence for evaluating offline DRL algorithms on real-world datasets. The upward and downward arrows indicate increasing and decreasing trends, respectively.

TABLE IV: Glycemic variability of the OhioT1DM Dataset

Metrics	OhioT1DM
TIR ($70 - 180$ mg/dL) (%)	63.5 ± 9.7
TAR (> 180 mg/dL) (%)	33.2 ± 10.7
Level 1 ($181 - 250$ mg/dL) (%)	24.9 ± 6.2
Level 2 (> 250 mg/dL) (%)	8.3 ± 5.2
TBR (< 70 mg/dL) (%)	3.3 ± 2.3
Level 1 ($54 - 69$ mg/dL) (%)	2.6 ± 1.6
Level 2 (< 54 mg/dL) (%)	0.7 ± 0.7
LBGI	0.88 ± 0.48
HBGI	7.15 ± 2.45
Mean CGM (mg/dL)	159 ± 16

[3] G. Shafiee, M. Mohajeri-Tehrani, M. Pajouhi, and B. Larijani, “The importance of hypoglycemia in diabetic patients,” *Journal of Diabetes & Metabolic Disorders*, vol. 11, no. 1, pp. 1–7, 2012.

[4] S. Zoungas, A. Patel, J. Chalmers, B. E. De Galan, Q. Li, L. Billot, M. Woodward, T. Ninomiya, B. Neal, S. MacMahon *et al.*, “Severe hypoglycemia and risks of vascular events and death,” *New England Journal of Medicine*, vol. 363, no. 15, pp. 1410–1418, 2010.

[5] J. Pickup and A. Sutton, “Severe hypoglycaemia and glycaemic control in type 1 diabetes: meta-analysis of multiple daily insulin injections compared with continuous subcutaneous insulin infusion,” *Diabetic Medicine*, vol. 25, no. 7, pp. 765–774, 2008.

[6] S. S. Hussain and N. Oliver, *Insulin Pumps and Continuous Glucose Monitoring Made Easy*. Edinburgh, Scotland: Elsevier Health Sciences, 2015.

[7] I. Steineck, A. Ranjan, K. Nørgaard, and S. Schmidt, “Sensor-augmented insulin pumps and hypoglycemia prevention in type 1 diabetes,” *Journal of Diabetes Science and Technology*, vol. 11, no. 1, pp. 50–58, 2017.

[8] G. Cappon, M. Vettoretti, G. Sparacino, and A. Facchinetti, “Continuous glucose monitoring sensors for diabetes management: a review of technologies and applications,” *Diabetes & Metabolism Journal*, vol. 43, no. 4, pp. 383–397, 2019.

[9] L. Heinemann, G. Freckmann, D. Ehrmann, G. Faber-Heinemann, S. Guerra, D. Waldenmaier, and N. Hermanns, “Real-time continuous glucose monitoring in adults with type 1 diabetes and impaired hypoglycaemia awareness or severe hypoglycaemia treated with multiple daily insulin injections (HypoDE): a multicentre, randomised controlled trial,” *The Lancet*, vol. 391, no. 10128, pp. 1367–1377, 2018.

[10] T. T. Ly, J. A. Nicholas, A. Retterath, E. M. Lim, E. A. Davis, and T. W. Jones, “Effect of sensor-augmented insulin pump therapy and automated insulin suspension vs standard insulin pump therapy on hypoglycemia in patients with type 1 diabetes: a randomized clinical trial,” *JAMA*, vol. 310, no. 12, pp. 1240–1247, 2013.

[11] C. A. van Beers, J. H. DeVries, S. J. Kleijer, M. M. Smits, P. H. Geelhoed-Duijvestijn, M. H. Kramer, M. Diamant, F. J. Snoek, and E. H. Serné, “Continuous glucose monitoring for patients with type 1 diabetes and impaired awareness of hypoglycaemia (IN CONTROL): a randomised, open-label, crossover trial,” *The Lancet Diabetes & Endocrinology*, vol. 4, no. 11, pp. 893–902, 2016.

[12] H. Thabit and R. Hovorka, “Coming of age: the artificial pancreas for type 1 diabetes,” *Diabetologia*, vol. 59, no. 9, pp. 1795–1805, 2016.

[13] E. Bekiari, K. Kitsios, H. Thabit, M. Tauschmann, E. Athanasiadou, T. Karagiannis, A.-B. Haidich, R. Hovorka, and A. Tsapas, “Artificial pancreas treatment for outpatients with type 1 diabetes: systematic review and meta-analysis,” *BMJ*, vol. 361, 2018.

[14] Y. Ruan, M. E. Wilinska, H. Thabit, and R. Hovorka, “Modeling day-to-day variability of glucose–insulin regulation over 12-week home use of closed-loop insulin delivery,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 6, pp. 1412–1419, 2016.

[15] R. Visentin, C. Dalla Man, and C. Cobelli, “One-day bayesian cloning of type 1 diabetes subjects: toward a single-day UVA/Padova type 1 diabetes simulator,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 11, pp. 2416–2424, 2016.

[16] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, “Reinforcement learning for intelligent healthcare applications: A survey,” *Artificial Intelligence in Medicine*, vol. 109, p. 101964, 2020.

[17] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, “Reinforcement learning application in diabetes blood glucose control: A systematic review,” *Artificial Intelligence in Medicine*, vol. 104, p. 101836, 2020.

[18] I. Contreras and J. Vehi, “Artificial intelligence for diabetes management and decision support: literature review,” *Journal of Medical Internet Research*, vol. 20, no. 5, p. e10775, 2018.

[19] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Deep learning for diabetes: A systematic review,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2021.

[20] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, “A deep learning algorithm for personalized blood glucose prediction,” in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI 2018*, 2018, pp. 74–78.

[21] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, “GluNet: A deep learning framework for accurate glucose forecasting,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 414–423, 2020.

[22] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, “Dilated recurrent neural networks for glucose forecasting in type 1 diabetes,” *Journal of Healthcare Informatics Research*, vol. 4, no. 3, pp. 308–324, 2020.

[23] T. Zhu, L. Kuang, J. Daniels, P. Herrero, K. Li, and P. Georgiou, “IoMT-enabled real-time blood glucose prediction with deep learning and edge computing,” *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3706–3719, 2023.

[24] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[25] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, “Mastering the game of go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[26] M. Popova, O. Isayev, and A. Tropsha, “Deep reinforcement learning for de novo drug design,” *Science Advances*, vol. 4, no. 7, p. eaap7885, 2018.

[27] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de las Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J.-M. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis, and M. Riedmiller, “Magnetic control of tokamak plasmas through deep reinforcement learning,” *Nature*, vol. 602, no. 7897, pp. 414–419, 2022.

[28] I. Fox and J. Wiens, “Reinforcement learning for blood glucose control: Challenges and opportunities,” in *Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning (ICML)*, 2019.

[29] S. Lee, J. Kim, S. W. Park, S.-M. Jin, and S.-M. Park, “Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 536–546, 2020.

[30] I. Fox, J. Lee, R. Pop-Busui, and J. Wiens, “Deep reinforcement learning for closed-loop blood glucose control,” in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 508–536.

[31] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1223–1232, 2021.

[32] T. Zhu, K. Li, and P. Georgiou, “Personalized dual-hormone control for type 1 diabetes using deep reinforcement learning,” in *International Workshop on Health Intelligence (W3PHIAI-20) in the 34th AAAI Conference on Artificial Intelligence*, 2020.

[33] T. Zhu, K. Li, L. Kuang, P. Herrero, and P. Georgiou, "An insulin bolus advisor for type 1 diabetes using deep reinforcement learning," *Sensors*, vol. 20, no. 18, p. 5058, 2020.

[34] J. Nordhaug Myhre, M. Tejedor, I. Kalervo Launonen, A. El Fathi, and F. Godtliebsen, "In-silico evaluation of glucose regulation using policy gradient reinforcement learning for patients with type 1 diabetes mellitus," *Applied Sciences*, vol. 10, no. 18, p. 6350, 2020.

[35] B. P. Kovatchev, M. Breton, C. Dalla Man, and C. Cobelli, "In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes," *Journal of Diabetes Science and Technology*, vol. 3, no. 1, pp. 44–55, 2009.

[36] C. Dalla Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 diabetes simulator: new features," *Journal of Diabetes Science and Technology*, vol. 8, no. 1, pp. 26–34, 2014.

[37] R. Visentin, C. Dalla Man, B. Kovatchev, and C. Cobelli, "The university of Virginia/Padova type 1 diabetes simulator matches the glucose traces of a clinical trial," *Diabetes Technology & Therapeutics*, vol. 16, no. 7, pp. 428–434, 2014.

[38] R. Hovorka, V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. O. Federici, T. R. Pieber, H. C. Schaller, L. Schaupp, T. Vering *et al.*, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiological Measurement*, vol. 25, no. 4, p. 905, 2004.

[39] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2052–2062.

[40] S. Fujimoto and S. S. Gu, "A minimalist approach to offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[41] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, "Stabilizing off-policy q-learning via bootstrapping error reduction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[42] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.

[43] I. Kostrikov, R. Fergus, J. Tompson, and O. Nachum, "Offline reinforcement learning with fisher divergence critic regularization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5774–5783.

[44] Y. Xie, B. Liu, Q. Liu, Z. Wang, Y. Zhou, and J. Peng, "Off-policy evaluation and learning from logged bandit feedback: Error reduction via surrogate policy," in *International Conference on Learning Representations*, 2018.

[45] H. Le, C. Voloshin, and Y. Yue, "Batch policy learning under constraints," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3703–3712.

[46] M. R. Zhang, T. Paine, O. Nachum, C. Paduraru, G. Tucker, M. Norouzi *et al.*, "Autoregressive dynamics models for offline policy evaluation and optimization," in *International Conference on Learning Representations*, 2021.

[47] T. L. Paine, C. Paduraru, A. Michi, C. Gulcehre, K. Zolna, A. Novikov, Z. Wang, and N. de Freitas, "Hyperparameter selection for offline reinforcement learning," in *The Offline Reinforcement Learning Workshop, NeurIPS 2020*, 2020.

[48] J. Fu, M. Norouzi, O. Nachum, G. Tucker, A. Novikov, M. Yang, M. R. Zhang, Y. Chen, A. Kumar, C. Paduraru *et al.*, "Benchmarks for deep off-policy evaluation," in *International Conference on Learning Representations*, 2021.

[49] S. Tang and J. Wiens, "Model selection for offline reinforcement learning: Practical considerations for healthcare settings," in *Machine Learning for Healthcare Conference*. PMLR, 2021, pp. 2–35.

[50] T. Battelino, T. Danne, R. M. Bergenstal, S. A. Amiel, R. Beck, T. Biester, E. Bosi, B. A. Buckingham, W. T. Cefalu, K. L. Close *et al.*, "Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range," *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019.

[51] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1587–1596.

[52] P. Herrero, J. Bondia, O. Adewuyi, P. Pesl, M. El-Sharkawy, M. Reddy, C. Toumazou, N. Oliver, and P. Georgiou, "Enhancing automatic closed-loop glucose control in type 1 diabetes with an adaptive meal bolus calculator–in silico evaluation under intra-day variability," *Computer Methods and Programs in Biomedicine*, vol. 146, pp. 125–131, 2017.

[53] C. Liu, P. Avari, Y. Leal, M. Wos, K. Sivasithamparam, P. Georgiou, M. Reddy, J. M. Fernández-Real, C. Martin, M. Fernández-Balsells *et al.*, "A modular safety system for an insulin dose recommender: a feasibility study," *Journal of Diabetes Science and Technology*, vol. 14, no. 1, pp. 87–96, 2020.

[54] S. Schmidt and K. Nørgaard, "Bolus calculators," *Journal of Diabetes Science and Technology*, vol. 8, no. 5, pp. 1035–1041, 2014.

[55] T. Danne, O. Kordonouri, M. Holder, H. Haberland, S. Golembowski, K. Remus, S. Bläsig, T. Wadien, S. Zierow, R. Hartmann *et al.*, "Prevention of hypoglycemia by using low glucose suspend function in sensor-augmented pump therapy," *Diabetes Technology & Therapeutics*, vol. 13, no. 11, pp. 1129–1134, 2011.

[56] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *The 5th KDH workshop, ECAI 2020*, 2020, pp. 71–74.

[57] A. El Fathi, C. Fabris, and M. D. Breton, "Titration of long-acting insulin using continuous glucose monitoring and smart insulin pens in type 1 diabetes: A model-based carbohydrate-free approach," *Frontiers in Endocrinology*, vol. 12, p. 1787, 2022.

[58] T. K. Ritschel, A. T. Reenberg, and J. B. Jørgensen, "Large-scale virtual clinical trials of closed-loop treatments for people with type 1 diabetes," *CoRR*, vol. abs/2205.01332, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.01332

[59] M. I. Takuma Seno, "d3rlpy: An offline deep reinforcement library," in *The Offline Reinforcement Learning Workshop, NeurIPS 2021*, 2021.

[60] 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, no. 11, pp. 1716–1720, 2018.

[61] G. Cappon, M. Vettoretti, F. Marturano, A. Facchinetti, and G. Sparacino, "A neural-network-based approach to personalize insulin bolus calculation using continuous glucose monitoring," *Journal of Diabetes Science and Technology*, vol. 12, no. 2, pp. 265–272, 2018.

[62] G. Noaro, G. Cappon, M. Vettoretti, G. Sparacino, S. Del Favero, and A. Facchinetti, "Machine-learning based model to improve insulin bolus calculation in type 1 diabetes therapy," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 1, pp. 247–255, 2020.

[63] N. S. Tyler, C. M. Mosquera-Lopez, L. M. Wilson, R. H. Dodier, D. L. Branigan, V. B. Gabo, F. H. Guillot, W. W. Hilts, J. El Youssef, J. R. Castle *et al.*, "An artificial intelligence decision support system for the management of type 1 diabetes," *Nature Metabolism*, vol. 2, no. 7, pp. 612–619, 2020.

[64] G. Cappon, A. Facchinetti, G. Sparacino, and S. Del Favero, "A bayesian framework to identify type 1 diabetes physiological models using easily accessible patient data," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 6914–6917.

[65] N. Oliver, M. Reddy, C. Marriott, T. Walker, and L. Heinemann, "Open source automated insulin delivery: addressing the challenge," *NPJ Digital Medicine*, vol. 2, no. 1, pp. 1–5, 2019.

[66] J. Kesavadev, S. Srinivasan, B. Saboo, M. Krishna B, G. Krishnan *et al.*, "The do-it-yourself artificial pancreas: a comprehensive review," *Diabetes Therapy*, vol. 11, no. 6, pp. 1217–1235, 2020.

[67] R. Armiger, M. Reddy, N. S. Oliver, P. Georgiou, and P. Herrero, "An in silico head-to-head comparison of the do-it-yourself artificial pancreas loop and bio-inspired artificial pancreas control algorithms," *Journal of Diabetes Science and Technology*, vol. 16, no. 1, pp. 29–39, 2022.

[68] A. Irpan, K. Rao, K. Bousmalis, C. Harris, J. Ibarz, and S. Levine, "Off-policy evaluation via off-policy classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[69] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 193–204, 2023.

[70] N. Resalat, J. El Youssef, N. Tyler, J. Castle, and P. G. Jacobs, "A statistical virtual patient population for the glucoregulatory system in type 1 diabetes with integrated exercise model," *PloS one*, vol. 14, no. 7, p. e0217301, 2019.

[71] C. Fabris, B. Ozaslan, and M. D. Breton, "Continuous glucose monitors and activity trackers to inform insulin dosing in type 1 diabetes: the University of Virginia contribution," *Sensors*, vol. 19, no. 24, p. 5386, 2019.