Patient-specific deep offline artificial pancreas for blood glucose regulation in type 1 diabetes

Yixiang Deng\textsuperscript{1,2,3}, Kevin Arao\textsuperscript{4}, Christos S. Mantzoros\textsuperscript{4,5}, and George Em Karniadakis\textsuperscript{1,6†}

\textsuperscript{1}School of Engineering, Brown University, Providence, RI 02912, USA
\textsuperscript{2}Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139, USA
\textsuperscript{3}Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
\textsuperscript{4}Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02215, USA
\textsuperscript{5}VA Boston Healthcare System, Harvard Medical School, Boston, MA 02215, USA
\textsuperscript{6}Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
\textsuperscript{†}e-mail: george_karniadakis@brown.edu

Abstract

Due to insufficient insulin secretion, patients with type 1 diabetes mellitus (T1DM) are prone to blood glucose fluctuations ranging from hypoglycemia to hyperglycemia. While dangerous hypoglycemia may lead to coma immediately, chronic hyperglycemia increases patients’ risks for cardiorenal and vascular diseases in the long run. In principle, an artificial pancreas – a closed-loop insulin delivery system requiring patients manually input insulin dosage according to the upcoming meals – could supply exogenous insulin to control the glucose levels and hence reduce the risks from hyperglycemia. However, insulin overdosing in some type 1 diabetic patients, who are physically active, can lead to unexpected hypoglycemia beyond the control of common artificial pancreas. Therefore, it is important to take into account the glucose decrease due to physical exercise when designing the next-generation artificial pancreas. In this work, we develop a deep reinforcement learning algorithm using a T1DM dataset, containing data from wearable devices, to automate insulin dosing for patients with T1DM. In particular, we build patient-specific computational models using systems biology informed neural networks (SBINN), to mimic the glucose-insulin dynamics for a few patients from the dataset, by simultaneously considering patient-specific carbohydrate intake and physical exercise intensity.

Keywords: artificial pancreas, blood glucose regulation, type 1 diabetes, physical exercise, wearable devices, offline reinforcement learning, digital twin, precision medicine

1 INTRODUCTION

Diabetes mellitus is a growing epidemic and its prevalence has been increasing in nearly all countries (1). Global estimates show that the prevalence of adults aged 20-79 years is 8.8% in 2015, and it is predicted to rise to 10.8% in 2040 (2). In the US, according to the Centers for Disease Control and Prevention (CDC), a total of 34.2 million people have diabetes or 10.5% of the US population in 2018 (3). The most common forms of diabetes are type 1 and type 2 diabetes mellitus. In 2016, US data showed type 1 and type 2 diabetes accounted for approximately 6% and 91% of all cases of diagnosed diabetes, respectively (4). Type 1 diabetes (T1DM) is due to autoimmune B-cell destruction, which usually leads to absolute insulin deficiency. On the other hand, type 2 diabetes (T2DM) is due to progressive loss of adequate B-cell insulin secretion with associated insulin resistance (5, 6). Glucose is integral in energy consumption as it serves as a primary metabolic fuel. Under normal physiology, in a fasting state, there is a basal insulin secretion to help match hepatic gluconeogenesis to maintain a blood glucose target between 70 and
130 mg/dl. After a meal, there is a rise in blood glucose levels resulting in concomitant increase secretion of insulin from pancreas (7). The major effects of insulin on glucose metabolism are the following: (a) increases glucose transport across the cell membrane in adipose tissue and muscle, (b) increases glycolysis in muscle and adipose tissue, (c) stimulates glycogenesis and inhibits glycogenolysis in muscle, and liver, and (d) inhibits gluconeogenesis in the liver (8). A few hours after the meal, as the blood glucose concentration falls, glucagon is secreted to release glucose back to the blood which decreases glucose fluctuations. The main goal of treatment, especially for T1DM patients, is to mimic these physiologic insulin secretion by providing appropriate basal and prandial insulin doses. Patients with T1DM are frequently suffering from complications associated with unstable glucose levels, when the blood glucose (BG) regulation is not well controlled. According to the Diabetes Control and Complications Trial, better BG control leads to lower HbA1c levels, which finally leads to better outcomes in terms of both microvascular and macrovascular complications (9). Hence, achieving stable glucose levels has been a well established goal in diabetes management.

Unlike a fingerstick test providing one-time reading of blood glucose from a blood sample (10), wearable minimally-invasive continuous glucose monitoring (CGM) sensors can provide real-time blood glucose concentration measurements for days, by measuring from interstitial fluid (11). When integrated with CGM, another implantable or wearable device, namely “insulin pump”, provides continuous insulin administrations through subcutaneous infusion instead of needle injection, greatly improve the quality of life for patients with diabetes (12, 13, 14). Combining a glucose sensor, an insulin infusion device, and a control algorithm, artificial pancreas (AP) is a closed-loop system designed for patients with T1DM to improve their BG regulation, and consequently decrease the risk of diabetic complications (15, 16, 17, 18). While a few control algorithms in the AP design (19, 20), show sufficient performance at the general population level, very few of them address the inter- and intra-patient variability. More importantly, although patients benefit from regular physical activity in terms of the diabetes management (21, 22, 23), the exercise-induced glucose uptakes also leads to negative effect of the pre-programed insulin usage, leading to increased risk of hypoglycemia during physical exercises (24, 25). Unfortunately, most existing AP algorithms consider only limited external factors like meal intakes which increase plasma glucose levels and insulin infusion which lower the glucose levels (26, 27, 28, 29, 30), and some also require meal announcement (31).

Therefore, the next-generation AP in precision medicine requires a holistic design considering features such as the patient-specificity, physical exercises, and less human intervention. Fortunately, rapid growth of artificial intelligence (32, 33), especially deep neural networks based data-driven machine learning algorithms, provides us with opportunities to characterize the complex dynamics in patients’ metabolic environment, and further enhance the development of control algorithms like reinforcement learning. However, regular reinforcement learning algorithms demands large training datasets in the learning process and constant interaction between the learned agent and environment, i.e., patient’s body, in healthcare related tasks. Despite clinical trials providing abundant medical data from the wearable or implantable devices on patients (34, 35), the specific requirement of agent-environment interaction in reinforcement learning undoubtedly is a huge barrier considering the safety of patients during data collection. Fortunately, utilizing only previously collected offline data without additional online interaction with the environment, offline reinforcement learning algorithms may provide a good opportunity to mitigate the potentially high expense and danger of data collection in healthcare related problems (36).

Motivated by the features highlighted for next-generation AP design, we propose a novel framework to design a patient-specific artificial pancreas integrating wearable device data, systems biology informed neural networks and offline reinforcement learning algorithms (Fig. 1). By simultaneously considering patient-specificity, meal intakes, insulin infusion, and most importantly physical exercises, we first build a patient-specific simulator (digital twin) using a system of ordinary differential equations (ODE) developed by Roy and Parker (37), systems biology informed neural networks (SBiNN) (38) and wearable sensor data from the OhioT1DM dataset (39, 40). After obtaining the dynamics of hidden states and hidden parameters for the simulator, we implement offline reinforcement learning algorithms on the patient-specific simulator to improve the plasma glucose dynamics with a safer and more efficient insulin dosage planning. In particular, our offline agents learn the insulin dosage with only a short sequence of past glucose levels without any meal or exercise announcements, which significantly further the step towards an authentic closed-loop system for artificial pancreas design. Our results suggest that the proposed framework provides in-depth diagnosis on patients’ diabetes management and leads to an insightful prognosis on medical
improvements at a patient-specific level. We are confident that this framework could benefit the design of in silico clinical trials, where some patient-specific data and clinically verified physiological (parametrized) models are available.

**Figure 1: Motivation of digitize blood glucose management in type 1 diabetes.** With interaction to the real world, T1D patients experience fluctuating blood glucose from hypoglycemia (BG levels lower than 70 mg/dl, 80 mg/dl if measured by CGM), normoglycemia (BG levels between 80 and 180 mg/dl), and hyperglycemia (BG levels greater than 180 mg/dl). In the digital world, a digital twin mimicking the glucose-insulin dynamics can be created and optimized using computational algorithm with real-world inputs, i.e., insulin pumps provided exogenous insulin along with carbohydrate intakes, fitness bands provided the exercise intensity, and CGM provided real-time glucose levels. BG, blood glucose. CGM, continuous glucose monitoring. Part of the image is created using Biorender.

### 2 METHODS

#### 2.1 Framework of the study

In this work, we developed a computational framework to design a patient-specific automated insulin delivery system for six patients with type 1 diabetes using patient-specific data from the OhioT1DM dataset. The OhioT1DM dataset contains eight-week continuous glucose monitoring, insulin, physiological sensor, and self-reported life-event data for 12 patients with type 1 diabetes, among which 6 patients participated in the 2018 cohort and the other 6 patients in the 2020 cohort (40). A workflow of this framework is shown in Fig. 2A. We first implemented systems biology informed neural networks (SBINN) on the OhioT1DM dataset for parameter inference of the patient-specific Roy-Parker model (37), based on historical records of total exogenous insulin (bolus insulin and basal insulin), carbohydrate intakes, heart rate, and CGM measured glucose level. We then trained a deep offline reinforcement learning neural network to build a patient-specific automated insulin delivery system for two representative patients in the OhioT1DM dataset. The final optimized agent, represented by a deep neural networks, can serve as the patient-specific artificial pancreas, leading to a better insulin dosage scheme for the patient.

#### 2.2 OhioT1DM dataset

**Dataset overview.** All patients in the OhioT1DM were on insulin pump therapy with continuous glucose monitoring (CGM), throughout the 8-week data collection period. During these 8 weeks of monitoring, the patients also reported life-event data via a custom smart-phone app and provided physiological data from a fitness band. Based on the form of the ODE model, i.e., the Roy-Parker model, studied in this work, we selected from the OhioT1DM dataset the following historical measurements: 1) the CGM blood glucose level, 2) insulin doses, both bolus insulin...
Figure 2: Overview of patient-specific deep offline reinforcement learning for blood glucose regulation. (A) Schematic of the framework proposed. Data collected from wearable or implantable devices, such as plasma glucose levels, carbohydrate intakes, insulin infusion and heart rates, were used in two ways. Firstly, the data is applied to reconstruct patient’s metabolic environment, represented by an ODE model and systems biology informed neural networks (SBINN). Meanwhile, the data is also used to train an offline agent composed of neural networks with reinforcement learning algorithms. The reconstructed environment is used to validate the trained offline agent. Part of the image is created using Biorender. (B) The compartment diagram of an ODE model, namely modified Roy-Parker model, captures the intrinsic relationships among the observed state $G$ and other hidden state variables, considering external inputs like carbohydrate intakes and physical exercises intensity. (C) Schematic of the time-dependent systems biology informed neural networks (SBINN) with self-adaptive weights proposed to learning the dynamics of hidden states and the hidden model parameters. (D) Schematic of artificial pancreas design using deep offline reinforcement learning, specifically, batch-constrained Q learning (BCQ) algorithm. A buffer derived from the OhioT1DM dataset was used to train a variational autoencoder (VAE), which partially mimics the patient’s metabolic environment in an offline fashion. An agent represented by a deep neural networks was then trained using merely $n$ plasma glucose levels before time $t$ ($t$ included) to provide the insulin dosage at time $t + 1$. By iteratively marching forward, a better insulin dosing plan is generated.
and basal insulin, 3) self-reported meal times and the amounts of carbohydrate intakes, and 4) the heart rate. We note that only the data of those 6 patients in the 2018 cohort is used in our analysis, due to the lack of heart rate monitoring in the 2020 cohort.

**Data preprocessing.** The exogenous insulin is calculated as the sum of basal insulin and bolus insulin at each moment. According to the user manual of the insulin pumps, i.e., Medtronic 530G and 630G, while the basal insulin is given at a rate and is provided explicitly, bolus insulin is a one-time dose and can be released into the blood stream using different modes, i.e., “normal”, “normal dual”, “square” and “square dual”. Given limited information for the exact releasing process of these different modes in the OhioT1DM dataset, we assume the conversion formula based on literature (41) and the user guides of the corresponding commercial insulin pumps (42, 43). For “normal” type, the single x mU dose bolus insulin is converted to a constant pseudo basal insulin at a rate of x/10 mU lasting for 10 minutes. For “square” type, the single x mU dose bolus insulin is converted to a constant pseudo basal insulin at a rate of x/t mU lasting for t minutes, where t denotes the time elapse between two adjacent boluses. For “normal dual” and “square dual”, the single dose is divided evenly into two identical half doses. These two half doses are released sequentially by “normal” and “square” type in “normal dual”, while by “square” and “normal” type in “square dual”, respectively. Additionally, we also consider “temp basal” insulin, which overrides the basal insulin set previously. The glucose consumption rate due to meal intakes is computed by the amount of meal carbohydrate using an exponential decay function (38) as follows,

\[ u_2(t) = 0.0083 \sum_{j=1}^{N} m_j \exp(0.0083(t_f - t)), \]  

where \( m_j \) gram of carbohydrate intake is recorded at \( t_j \), \( N \) is the total number of meals. The percentage of \( \text{VO}_2^{\text{max}} \) (\( \text{PVO}_2^{\text{max}} \)) denoting a percentage of the maximum rate of oxygen consumption due to exercise is approximated by the heart rate (HR) as follows (44),

\[ \text{PVO}_2^{\text{max}}(t) = 0.888HR - 71.91, \]
\[ u_3(t) = \text{PVO}_2^{\text{max}}(t) - 8, \text{ if } u_3(t) > 0 \text{ else } 0 \]  

where \( 8 \) denotes the average \( \text{PVO}_2^{\text{max}} \) for a person at the basal state (37). After converting all data into the same resolution, we further smoothed the data by using a rolling window and generate a formatted dataset, where the sampling interval is 1 hour between neighbouring time points. Fig. S2 shows the processed insulin infusion \( u_1(t) \), carbohydrate intake \( u_2(t) \) and exercise intensity \( u_3(t) \) for 6 patients in the OhioT1DM dataset.

### 2.3 Systems biology informed neural networks (SBINN) with the Roy-Parker model

**Roy-Parker model.** With the aim of developing a robust closed-loop insulin delivery system under changing physiological conditions, Roy and Parker developed a model that can predict blood glucose levels at rest and during physical exercises (37). Since the patients participated in OhioT1DM only performed sporadic and light physical activities, we modified the ODE system in Roy and Parker (37) by omitting the \( G_{gl}(t) \) term representing the decline of the glycegenolysis rate during prolonged exercise due to the depletion of liver glycogen stores (Fig. 2B). The resulting ODE system for the 6 state variables \([I, X, G, G_{prod}, G_{up}, I_c]\) is shown in Eqs. (3)-(8),
\[
\frac{dI(t)}{dt} = -nI(t) + p_4u_1(t) - I_e(t); \quad I(0) = I_0 = \frac{p_4}{n}u_1b
\]
\[
\frac{dX(t)}{dt} = -p_2X(t) + p_3[I(t) - I_b]; \quad X(0) = 0
\]
\[
\frac{dG(t)}{dt} = -p_1[G(t) - G_b] - X(t)G(t) + \frac{W}{V_{olG}}G_{prod}(t) - \frac{W}{V_{olG}}G_{up}(t) + u_2(t); \quad G(0) = G_b
\]
\[
\frac{dG_{prod}(t)}{dt} = a_1PV_{G_2}^{max}(t) - a_2G_{prod}(t); \quad G_{prod}(0) = 0
\]
\[
\frac{dG_{up}(t)}{dt} = a_3PV_{G_2}^{max}(t) - a_4G_{up}(t); \quad G_{up}(0) = 0
\]
\[
\frac{dI_e(t)}{dt} = a_5PV_{G_2}^{max}(t) - a_6I_e(t); \quad I_e(0) = 0
\]

It captures the exercise-induced dynamics of plasma insulin concentration \(I(t)\), remote insulin concentration \(X(t)\), the plasma glucose level \(G(t)\), exercise-induced hepatic glucose production \(G_{prod}(t)\), exercise-induced glucose uptake \(G_{up}(t)\), exercise-induced insulin removal from the circulatory system \(I_e(t)\), exogenous infusion \(u_1(t)\), and external glucose uptake \(u_2(t)\). The instant parameters \(I_0\) and \(G_b\) represent the basal plasma insulin and glucose concentrations, respectively. Given the multi-scale nature of the glucose levels in a month-long observation, we found that our algorithm learns better the glucose dynamics when we allow the parameters in the ODE to vary over time. Table S1 shows the nomenclature, physiological meaning and reference values of patient-specific parameters to be inferred in the ODE. The ranges of the parameters were set to \([0.2 \cdot x_{ref}, 1.8 \cdot x_{ref}]\), where \(x_{ref}\) represents the corresponding reference value of the variable in Table S1.

**Systems biology informed neural networks (SBINN).** Yazdani et al. developed a novel systems-biology-informed deep learning framework, namely systems biology informed neural networks (SBINN), which successfully incorporates the system of ordinary differential equations (ODE) into the neural networks (38). Inspired by the physics-informed neural networks, SBINN, shown in Fig. 2C, is sequentially composed of an input-scaling layer to allow input normalization for the robust performance of the neural networks, a feature layer marking different patterns of state variables in the ODE system and the output-scaling layer to convert normalized state variables back to physical units. By effectively adding constraints derived from the ODE system to the optimization procedure, SBINN is able to simultaneously infer the dynamics of unobserved species, external forcing, and the unknown model parameters.

Given the measurements of \(y_1, y_2, \ldots, y_M\) at times \(t_1, t_2, \ldots, t_{N_{data}}\), SBINN enforces the network to satisfy the ODE system at the time point \(\tau_1, \tau_2, \ldots, \tau_{N_{ode}}\). The total loss is defined as a function of both the parameters of the neural networks, denoted by \(\theta\) and parameters of the ODE system, denoted by \(p\).

\[
\mathcal{L}(\theta, p) = \mathcal{L}^{data}(\theta) + \mathcal{L}^{ode}(\theta, p) + \mathcal{L}^{aux}(\theta),
\]

where \(\mathcal{L}^{data}\) is associated with the \(M\) sets of observations of the state variables \(y\) in the ODE system; \(\mathcal{L}^{ode}\) enforces the structure imposed by the system of ODE; \(\mathcal{L}^{aux}\) is introduced as an additional source of information for the system identification. We recommend Ref. (38) that provides further details of the loss functions. The final step of SBINN is to infer the neural network parameters \(\theta\) as well as the unknown ODE parameters \(p\) simultaneously by minimizing the aforementioned loss function via gradient-based optimizers.

In this work, the known observed state variable \(y\) is the CGM measured glucose record in the OhioT1DM dataset, i.e., \(G(t)\), which is used for minimizing the data loss, \(\mathcal{L}^{data}\). We used \(\mathcal{L}^{ode}\) to minimize the residual terms in the ODE, shown in Eqs. (3)-(8). Following Yazdani et al. (38), we imposed the initial condition as the auxiliary loss \(\mathcal{L}^{aux}\). To improve the training of SBINN and speed up the convergence, we implemented self-adaptive weights over each iteration on the weights of each loss terms (45). The self-adaptive weighted loss \(\mathcal{L}'\) is as follows,

\[
\mathcal{L}'(\theta, p, \lambda_0, \lambda_d) = \mathcal{L}^{data}(\theta) + \lambda_0\mathcal{L}^{ode}(\theta, p) + \lambda_d\mathcal{L}^{aux}(\theta),
\]
where $\lambda_o$ and $\lambda_a$ are trainable, non-negative self-adaptation weights associated to the ODE loss term and auxiliary loss term, respectively. Hence, the objective of the training of neural networks is updated to

$$
\min_{\theta} \max_{\lambda_o, \lambda_a} L'(\theta, p, \lambda_o, \lambda_a).
$$

(11)

The update rules for the self-adaptive weights are given by,

$$
\lambda_o^{k+1} = \lambda_o^k + \rho_o^k \nabla_{\lambda_o} L'(\theta, p, \lambda_o, \lambda_a),
$$

(12)

$$
\lambda_a^{k+1} = \lambda_a^k + \rho_a^k \nabla_{\lambda_a} L'(\theta, p, \lambda_o, \lambda_a),
$$

(13)

where $\rho_o^k$ and $\rho_a^k$ denotes the learning rates for the corresponding weights, which was set to be $\rho_o^k = \rho_a^k$ in this work.

### 2.4 Offline reinforcement learning

Unlike supervised learning where a direct input and output mapping is given explicitly, reinforcement learning algorithms define the control problem as a Markov Decision process, and train an agent, which collects a step-by-step reward through interactions with the environment of interest. The goal of the agent is to respond to the environment changes such that the total reward of the series of responses, namely actions, is maximized by the end of training. While (online) reinforcement learning requires a demanding setting where the agent is trained through trial-and-error interactions with a dynamic environment, offline reinforcement learning algorithms utilize previously collected data, with no extra interaction with the environment (36, 46). However, the benefits of offline reinforcement learning also come with some disadvantages, for example the distributional shift, i.e., while the function approximations might be trained under one distribution, it will be evaluated on a different distribution. Efforts to address this challenge can be generally grouped into two categories: (1) policy constraint methods, which constrain the learned policy to lie close to the behavior policy, which was used to collect the offline dataset; and (2) uncertainty based methods, which attempt to estimate the epistemic uncertainty of value functions, and then utilize this uncertainty to detect distributional shift. A comprehensive review of offline reinforcement learning algorithms could be found in (36). Despite of some disadvantages, offline reinforcement learning is a favorable choice for tasks demanding a lot of time and cost when interacting with the environment, for example, healthcare related tasks. As shown in Fig. 2D, in offline reinforcement learning, a buffer dataset is collected by some behavior policy $\pi_B$, before the training starts. Afterwards, the agent represented by a deep neural networks is trained with an offline reinforcement learning algorithm, specifically, batch-constrained Q learning (47). Finally, a policy outperforming the behavioral policy $\pi_B$ is deployed in the patient-specific artificial pancreas. In this work, we generated a buffer by sampling from the OhioT1DM dataset and generate the states, i.e., glucose levels, carbon intakes and physical exercises, along with the action denoted by total exogenous insulin at a specific time point, i.e., a sum of the bolus insulin and basal insulin, and the corresponding returns depending on the resulting glucose levels.

**Batch-constrained Q learning (BCQ).** Herein, we implemented the batch constrained Q-learning (BCQ) algorithm, which is one of the most popular offline reinforcement learning algorithms (47). By restricting the action space in order to force the agent towards behaving close to a subset of the given data, BCQ is able to learn successfully without interacting with the environment by considering extrapolation error. The details of the BCQ algorithm are shown in Algorithm 1. Following Fujimoto et al. (47), we implemented a fully-connected feed forward neural networks for the Q-networks and the variational autoencoders (VAE). The hyper-parameters used in this work can be found in Table S2.

### 3 RESULTS

To build a surrogate environment for our agent to interact with, we first performed patient-specific parameter inference using SBINN with a system of ODE developed by Roy and Parker (37). The primary step of parameter
inference for a system of ODE is to examine its identifiability. After obtaining the patient-specific parameters of the ODE, which is essential to reconstruct the dynamics of the state variables, we developed a patient-specific offline reinforcement learning algorithm to learn an optimal planning for the external insulin infusion for two representative patients, which helped them decrease the risk of hypoglycemia and hyperglycemia, respectively.

**ODE identifiability of the Roy-Parker model**

We first perform structural identification on the ODE for the set of parameters

\[ \{n, \text{Vol}_G, p_1, p_2, p_3, p_4, a_1, a_2, a_3, a_4, a_5, a_6, W, u_{1b}\} \]

appeared in Eqs. (3)-(8). Although most of the ODE parameters are not readily available in the OhioT1DM dataset, there are several parameters that are practically available for patients. For example, the patient’s body weight \( W \) can be measured practically but is not available in OhioT1DM. While \( u_{1b} \), the exogenous insulin infusion rate to maintain basal plasma insulin, is not available in the OhioT1DM, but it could be inferred from the mode of the exogenous insulin profile of the specific patient. We considered the identifiability of the aforementioned parameters under different scenarios depending on whether \( W \) and \( u_{1b} \) are known for the given patient. Table 1 suggests that when \( W \) and \( u_{1b} \) are both known, other parameters are either globally identifiable or locally identifiable. We adopted this scenario in the following analysis by assuming \( W = 60 \) kg and patient-specific \( u_{1b} \) to construct the patient-specific model by inferring a subset of the aforementioned parameters \( \{n(t), \text{Vol}_G(t), p_1(t), p_2(t), p_3(t), p_4(t), a_1(t), a_2(t), a_3(t), a_4(t), a_5(t), a_6(t)\} \) for one-month period.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( W ) and ( I_b ) unknown</th>
<th>( W ) known, ( I_b ) unknown</th>
<th>( W ) unknown, ( I_b ) known</th>
<th>( W ) and ( I_b ) known</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
</tr>
<tr>
<td>( \text{Vol}_G )</td>
<td>globally</td>
<td>globally</td>
<td>globally</td>
<td>globally</td>
</tr>
<tr>
<td>( p_1 )</td>
<td>nonidentifiable</td>
<td>globally</td>
<td>globally</td>
<td>globally</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>nonidentifiable</td>
<td>nonidentifiable</td>
<td>locally</td>
<td>locally</td>
</tr>
<tr>
<td>( p_4 )</td>
<td>nonidentifiable</td>
<td>nonidentifiable</td>
<td>locally</td>
<td>locally</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>nonidentifiable</td>
<td>locally</td>
<td>nonidentifiable</td>
<td>locally</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
</tr>
<tr>
<td>( a_3 )</td>
<td>nonidentifiable</td>
<td>locally</td>
<td>nonidentifiable</td>
<td>locally</td>
</tr>
<tr>
<td>( a_4 )</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
<td>locally</td>
</tr>
<tr>
<td>( a_5 )</td>
<td>nonidentifiable</td>
<td>nonidentifiable</td>
<td>locally</td>
<td>locally</td>
</tr>
<tr>
<td>( a_6 )</td>
<td>globally</td>
<td>globally</td>
<td>globally</td>
<td>globally</td>
</tr>
<tr>
<td>( I_b )</td>
<td>nonidentifiable</td>
<td>-</td>
<td>nonidentifiable</td>
<td>-</td>
</tr>
<tr>
<td>( W )</td>
<td>nonidentifiable</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: **Parameter identifiability in the Roy-Parker model under four different scenarios.** \( W \) denotes the bodyweight of the patient, \( I_b \) denotes the basal plasma insulin level. Details of other parameters can be found in (37).

**Patient-specific parameter inference and kinetics reconstruction using OhioT1DM**

In the OhioT1DM dataset, we assumed the patient’s weight is 60 kg, i.e., \( W = 60 \) kg. We also assume that the initial condition of the ODE system is given by \[ [I, X, G, G_{prod}, G_{up}, I_e, PV/O_{max}]_{t=0} = [I_b, 0, G_b, 0, 0, 0, 0], \] where \( I_b \) is estimated from the basal insulin rate and \( G_b \) is the initial blood glucose level, both of which are patient-specific estimation from the OhioT1DM dataset. We imposed the initial condition of the state variables as the auxiliary loss. We also imposed smoothing using a moving window of 30 data points on the model inputs to speed up the convergence of parameter inference. Fig. 3 shows the inference of model parameters and hidden kinetics on patient
ID 588, who experienced repeated hyperglycemic events during the data collection period. Fig. 4 shows the same inference on another patient ID 591, who experienced a short period of hypoglycemia (glucose level below 80 mg/dl, we adjusted the threshold of hypoglycemia due to an overestimation of glucose levels by CGM (48)) around December 22, 2021. The corresponding results for the other four patients (IDs: 559, 563, 570, 575) can be found in Fig. S3-S6 in the Supplementary Material. Note that the time stamps in OhioT1DM dataset are pre-processed to avoid privacy leakage, hence they do not represent the real collection time. Inspired by observed fluctuation of metabolic reaction rates in oral glucose tolerance test (OGTT) (49), we adopted a time-varying parameters setting for SBINN to improve the flexibility of our model. Interestingly, while patient 588 and patient 591 followed similar insulin infusion and carb intakes, the only lifestyle difference in the exercise intensity seemingly changed the outcomes of their glucose management (Fig. S2).

We found frequently elevated plasma insulin and remote insulin in patient 591 compared to patient 588 (Fig. 3A and Fig. 4A), suggesting that patient 591 may have higher risk in developing hyperinsulinemia than patient 588. We also observed that some of the hidden parameters of patient 588 do not fluctuate as significantly over time as those of patient 591 (Fig. 3B and Fig. 4B). These parameters are $p_4$ denoting the rate of insulin addition into the plasma from exogenous insulin, $a_1$ denoting the rate of exercise-induced hepatic glucose production, $a_3$ denoting the rate of exercise-induced glucose uptake, $a_5$ denoting the rate of exercise-induced plasma insulin depletion during recovery period. Especially, we observed a sudden increase in terms of $a_5$ for patient 591 right after the occurrence of exercise-induced hypoglycemia right before Jan 01. We also found some parameters are significantly different between patient 591 and patient 588 on average. These parameters are $n$ denoting the rate of plasma insulin clearance, $p_3$ denoting the rate of insulin addition in the remote insulin compartment and $p_4$ denoting the rate of insulin addition into the plasma from exogenous insulin. Interestingly, all these parameters point to the balance of remote insulin and plasma insulin with $n$ and $p_3$ being almost doubled in 591 while $p_4$ being frequently higher in 588. These findings together imply that the glucose level fluctuation is a complex process involving multiple organs and tissues, and exercise contributes to the occurrence of hypoglycemia in a more complicated way rather than merely lowering the plasma glucose level.

Our results suggest that time-depedent SBINN successfully infer the fluctuating hidden kinetics as well as the parameters for these 6 patients with high accuracy, albeit the inter-patient variability due to different daily routine of physical activities and insulin injection. More importantly, we also observed that time-dependent SBINN were able to perform robustly under 5 different random seeds with the uncertainty band confined to an acceptable range. We specifically emphasized the accuracy of our parameter inference, which helped us reconstruct the dynamics by solving a forward problem with high accuracy, indicated by the good match between pink curves (forward solver) and blue curves (real data collection).
Figure 3: Time-dependent SBINN accurately infers the dynamics of state variables and hidden parameters in the Roy-Parker model for a patient in OhioT1DM with recurrent hyperglycemia. (A) The trajectory of state variables, plasma insulin level $I(t)$, remote insulin level $X(t)$, plasma glucose level $G(t)$, exercise-induced hepatic glucose production $G_{prod}(t)$, exercise-induced glucose uptake $G_{up}(t)$, exercise-induced insulin removal from the circulatory system $I_e(t)$. Patient ID 588, shown here, experienced repeated hyperglycemic period, suggested by the plasma glucose levels $G$ repeatedly exceeding 180 mg/dl. The blue solid dots denote plasma glucose levels collected by CGM. The pink line denotes mean of the forward solution of the ODE from 5 different runs using inferred parameters, with the pink shade denoting the corresponding standard deviation. The green line denotes mean of denotes the predicted values by SBINN from 5 different runs using inferred parameters, with the green shade denoting the corresponding standard deviation. (B) The trajectory of patient-specific time-dependent hidden parameters in the Roy-Parker model for patient ID 588. The black dots denote the mean of each hidden parameters from 5 different runs and the gray shade denotes the corresponding standard deviation.
Figure 4: Time-dependent SBINN accurately infers the dynamics of state variables and hidden parameters in the Roy-Parker model for a patient in OhioT1DM with occasional hypoglycemia. (A) The trajectory of state variables, plasma insulin level \( I(t) \), remote insulin level \( X(t) \), plasma glucose level \( G(t) \), exercise-induced hepatic glucose production \( G_{prod}(t) \), exercise-induced glucose uptake \( G_{up}(t) \), exercise-induced insulin removal from the circulatory system \( I_{e}(t) \). Patient ID 591, shown here, experienced a short hypoglycemic period, suggested by the plasma glucose levels went below 80 mg/dl for a short period. The blue solid dots denotes plasma glucose levels collected by CGM. The pink line denotes mean of the forward solution of the ODE from 5 different runs using inferred parameters, with the pink shade denoting the corresponding standard deviation. The green line denotes mean of denotes the predicted values by SBINN from 5 different runs using inferred parameters, with the green shade denoting the corresponding standard deviation. (B) The trajectory of patient-specific time-dependent hidden parameters in Roy-Parker model for patient ID 591. The black dots denote the mean of each hidden parameters from 5 different runs and the gray shade denotes the corresponding standard deviation.
Offline reinforcement learning

We formulated the reinforcement learning problem for glucose regulation provided with the glucose infusion rate $u_2$ and exercise intensity $u_3$ from the OhioT1DM dataset, and we trained the agent to provide the optimal exogenous insulin infusion rate $u_1^{opt}$ for a month. The return function depending on the resulting glucose levels at time point $t$ is defined as follows,

$$r_t(G_t) = \begin{cases} -(G_t - G_{target})^2, & G_{min} \leq G \leq G_{max}; \\ -10, & \text{otherwise.} \end{cases}$$  \hspace{1cm} (14)$$

where we set the target glucose level to be 120 mg/dl, $G_{min} = 90$, $G_{max} = 150$. The performance of the best agent is given by the solid blue curve in Fig. 5 (A) for patient 588 and (C) for patient 591. We note here that the glucose trajectories for patient 588 and patient 591 are distinctly reflecting patients suffering from repeated hyperglycemia and hypoglycemia, respectively. Interestingly, we found that extending the permitted range of the insulin dosage for the agent by doubling the maximum action documented in the offline dataset for patient 588, who experienced repeated hyperglycemia, yielded a better glucose trajectory. When we trained the offline reinforcement learning agent, we allowed the maximum action (insulin dosage) to reach 80 $\mu$U/mL, given the maximum insulin for patient 588 is around 40 $\mu$U/mL. However, for patient 591 who did not have frequent hyperglycemic events, we adopted the suggestion by Fujimoto et al. by restricting the maximum insulin dosage within the maximum insulin, i.e., 35 $\mu$U/mL, recorded in the offline dataset. Figs. 5 (C) and (D) show that the offline RL agent could provide significantly better insulin dosage plans for both patient 588 and patient 591, suggested by the blue curves staying in the safe glucose region (glucose level higher than 80 mg/dl and lower than 180 mg/dl) denoted by the green shade.
Figure 5: Deep offline reinforcement learning stabilizes plasma glucose levels variation and increases the time plasma glucose levels staying in normoglycemia for both patients with ID 588 and ID 591, compared to the archived data in OhioT1DM. (A) The solid blue curve denotes the insulin infusion rate over time provided by the offline RL agent, and the solid black curve denotes the patient’s self-operated insulin infusion rate recorded in the OhioT1DM dataset for patient ID 588. (B) The solid blue curve denotes the glucose trajectory using the insulin infusion plan provided by the offline RL agent for patient ID 588. The solid black curve denotes the glucose trajectory recorded in the OhioT1DM dataset for patient ID 588. Red shade denotes hypoglycemia region (BG lower than 80 mg/dl), green shade denotes normoglycemia region (BG within 80 mg/dl and 180 mg/dl), and yellow shade denotes hyperglycemia region (BG higher than 180 mg/dl). (C) Scatter plots of insulin infusion-glucose level pairs in OhioT1DM dataset (yellow) and the optimized offline RL (green) for patient ID 588, suggesting offline RL significantly decreases the occurrence of hyperglycemia. (D) The solid blue curve denotes the insulin infusion rate over time provided by the offline RL agent, and the solid black curve denotes the patient’s self-operated insulin infusion rate recorded in the OhioT1DM dataset for patient ID 591. (E) The solid blue curve denotes the glucose trajectory using the insulin infusion plan provided by the offline RL agent for patient ID 591. The solid black curve denotes the glucose trajectory recorded in the OhioT1DM dataset for patient ID 591. Red shade denotes hypoglycemia region (BG lower than 80 mg/dl), green shade denotes normoglycemia region (BG within 80 mg/dl and 180 mg/dl), and yellow shade denotes hyperglycemia region (BG higher than 180 mg/dl). (F) Scatter plots of insulin infusion-glucose level pairs in OhioT1DM dataset (yellow) and the optimized offline RL (green) for patient ID 588, suggesting offline RL prevents the dangerous hypoglycemia and decreases the time of hyperglycemia.

4 CONCLUSION

Insulin is the mainstay of treatment for patients with type 1 diabetes mellitus and long-standing type 2 diabetes mellitus to achieve good glycemic control (50). Overestimation of the necessary insulin dosage can be extremely dangerous and may lead to fatally low blood glucose levels below 80 mg/dl when measured by CGM, namely
hypoglycemia, while underestimated insulin leaving blood glucose above 180 mg/dl may result in hyperglycemia, which is believed to be responsible for micro- and macrovascular diseases in the long run. In modern medicine, the use of insulin pumps along with continuous glucose monitors has made it easier but requires significant resources and patient education. Fortunately, a closed-loop control system, also called artificial pancreas, which automates insulin infusion so as to maintain a consistently stable blood glucose level, undoubtedly relieves the burden of both patients and doctors and saves medical costs.

In this work, we attempted to address a few of the challenges with our framework, which effectively combines three key components to build a patient-specific artificial pancreas simultaneously considering data collected from wearable devices, i.e., meal intake, insulin infusion and physical exercises. These important components are: 1) a real-world historical medical dataset, namely OhioT1DM dataset, containing patient-specific glucose, insulin, meal intake and exercise intensity; 2) a flexible ODE model defining glucose-insulin dynamics by systematically prioritizing two significant external factors, i.e., meal intake and physical exercise intensities; and 3) an offline reinforcement learning algorithm, namely batch constrained Q-learning, without directly interacting with the patient’s metabolic environment. With time-dependent systems biology informed neural networks (SBINN) along with patient-specific data, our model not only correctly predicts the hidden states that cannot be measured with current diabetes technology, but also accurately infers patient-specific parameters governing the patient-specific Roy-Parker model by reconstructing all states of interest. More importantly, the uncertainty quantified hidden parameters provide patient-specific clinical interpretations on how patients’ behavioral pattern shape their corresponding glucose-insulin dynamics and can contribute to the prognostic suggestion on future patient-specific behaviors.

Furthermore, we were able to train an agent to automate insulin infusion and optimize the performance of the agent on the patient-specific ODE model with same glucose infusion and exercise intensity over time. Our results suggest that the best trained agent has a better performance in terms of maintaining blood glucose levels within the safe range, compared to the self-operated insulin infusion by patients themselves. This suggests that an agent trained by offline reinforcement learning is able to learn a better insulin dosage depending solely on past glucose level sequences without meal or exercise announcements, implying the advantageous potential of offline reinforcement learning in translational healthcare.

In spite of the improved glycemic control provided by our offline agent and minimal human intervention demanded by our framework, we can still identify possible improvements to the proposed framework, considering ODE model development, disease characterization and data processing. As is believed that there are both a time delay in the effect of insulin on the glucose production and that on the glucose utilization (27), it may be helpful to modify the ODE model to address the sluggish effects. Additionally, given that insulins are categorized into fast-acting, intermediate-acting and long-acting depending on the timing of their action in body, we could extend the data preprocessing of external insulin as well as update the ODE model to allow variations in insulin types (51, 52). Another avenue to be considered in improving closed-loop systems is by adding amylin or glucagon, a hormone also produced by the pancreas, resulting to a multi-hormone closed-loop system (53). With glucagon being the additional action option, the reinforcement learning agent will be able to explore insulin action space with a higher degree of freedom.

Despite of the aforementioned possible improvements, our framework shows inspiring translational potentials in precision medicine by enriching the limited options of digital models utilizing medical data collected from wearable devices. For example, with a few slight modifications on the ODE model to address the insulin resistance of tissue, we could model and optimize the insulin usage in patients with type 2 diabetes who develop insulin resistance (54). Interestingly, novel wearable devices are being designed to probe insulin and other metabolites that are impossible to quantify with existing techniques (55, 56), which will clearly benefit the calibration and the extension of our framework. Furthermore, the rapid development in non-invasive wearable devices tremendously increases patients’ compliance to the use of wearable devices (57, 58), and hence provides data with more frequent readouts, higher resolution and diverse modalities, which will obviously enhance the performance of our framework beyond diabetes related research. In summary, by seamlessly integrating real-world data from wearable sensors, systems biology informed numerical models and offline reinforcement learning algorithms, we believe our novel framework could shed light on the translational precision medicine where adequate biological numerical models are established and enough amount of medical data from wearable devices are available.
5 DATA AVAILABILITY

The OhioT1DM dataset used in the current study is publicly available on http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html.

6 CODE AVAILABILITY

The code will be released upon acceptance of publication.

ACKNOWLEDGEMENTS. G.E.K. would like to acknowledge support by NIH U01HL142518 and R01HL154150. C.S.M. would like to acknowledge support by NIH DK081913.


COMPETING INTERESTS. No potential conflicts of interest relevant to this article were reported.

References


[39] C. Marling, R. C. Bunescu, The ohiot1dm dataset for blood glucose level prediction, in: KHD@ IJCAI.

[40] C. Marling, R. Bunescu, The ohiot1dm dataset for blood glucose level prediction: Update 2020, KHD@ IJCAI (2020).


[48] C. M. Farrell, A. D. McNeilly, S. M. Hapca, R. J. McCrimmon, Real-time continuous glucose monitoring during a hyperinsulinemic-hypoglycemic clamp significantly underestimates the degree of hypoglycemia, Diabetes Care 43 (2020) e142–e143.


7 Supplementary Material

<table>
<thead>
<tr>
<th>Parameter</th>
<th>physiological meaning</th>
<th>reference value</th>
<th>time-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>rate of plasma insulin clearance</td>
<td>$0.142 \text{ min}^{-1}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$\text{Vol}_G$</td>
<td>glucose distribution space</td>
<td>$117 \text{ dl}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$p_1$</td>
<td>rate of insulin-independent glucose removal from the plasma space</td>
<td>$0.035 \text{ min}^{-1}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$p_2$</td>
<td>rate of insulin removal from the remote insulin compartment</td>
<td>$0.05 \text{ min}^{-1}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$p_3$</td>
<td>rate of insulin addition in the remote insulin compartment</td>
<td>$2.8e-5 \text{ ml/µU} \cdot \text{min}^2$</td>
<td>Yes</td>
</tr>
<tr>
<td>$p_4$</td>
<td>*rate of insulin addition into the plasma from exogenous insulin</td>
<td>$0.098 \text{ ml/min}^2$</td>
<td>Yes</td>
</tr>
<tr>
<td>$a_1$</td>
<td>*rate of exercise-induced hepatic glucose production</td>
<td>$1.58e-3 \text{ mg/kg-min}^2$</td>
<td>Yes</td>
</tr>
<tr>
<td>$a_2$</td>
<td>*rate of exercise-induced hepatic glucose consumption</td>
<td>$0.056 \text{ min}^{-1}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$a_3$</td>
<td>*rate of exercise-induced glucose uptake</td>
<td>$1.95e-3 \text{ mg/kg-min}^2$</td>
<td>Yes</td>
</tr>
<tr>
<td>$a_4$</td>
<td>*rate of exercise-induced glucose accumulation</td>
<td>$0.0485 \text{ min}^{-1}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$a_5$</td>
<td>rate of exercise-induced plasma insulin depletion during recovery period</td>
<td>$1.25e-3 \text{ µU/ml} \cdot \text{min}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$a_6$</td>
<td>rate of exercise-induced plasma insulin repletion during recovery period</td>
<td>$0.0108 \text{ min}^{-1}$</td>
<td>Yes</td>
</tr>
<tr>
<td>$u_{1b}$</td>
<td>exogenous insulin infusion rate to maintain $I_b$</td>
<td>patient-specific</td>
<td>No</td>
</tr>
<tr>
<td>$W$</td>
<td>body weight</td>
<td>$60 \text{ kg}$</td>
<td>No</td>
</tr>
</tbody>
</table>

Table S1: **Nomenclature and physiological meanings of patient-specific parameters in the Roy-Parker model.** Physiological meanings marked with * were not explained in the original publication, hence standing for the interpretation given the ODE form.
Patient record example

```xml
<patient id="563" insulin_type="Humalog 200">
    <glucose_level>
        <event ts="2021-10-28 00:01:00" value="231"/>
        ...
        <event ts="2021-11-08 08:14:00" value="145"/>
    </glucose_level>
    <basal>
        <event ts="2021-10-28 00:00:00" value="1.2"/>
        ...
        <event ts="2021-11-01 18:30:00" value="0.9"/>
    </basal>
    <temp_basal>
        <temp_basal>
            <event ts="begin=2021-10-29 11:36:37" ts_end="2021-10-29 11:36:37" type="square dual" dose="3.6" bwz_carb_input="16"/>
            ...
            <event ts="begin=2021-11-07 19:43:37" ts_end="2021-11-07 11:36:37" type="normal" dose="3.2" bwz_carb_input="12"/>
        </temp_basal>
    </temp_basal>
    <bolus>
        <event ts="begin=2021-10-29 11:36:37" ts_end="2021-10-29 11:36:37" type="square dual" dose="3.6" bwz_carb_input="16"/>
        ...
        <event ts="begin=2021-11-07 19:43:37" ts_end="2021-11-07 11:36:37" type="normal" dose="3.2" bwz_carb_input="12"/>
    </bolus>
    <heart_rate>
        <basis_heart_rate>
            <event ts="2021-10-28 00:01:00" value="100"/>
            ...
            <event ts="2021-11-08 00:21:00" value="79"/>
        </basis_heart_rate>
    </heart_rate>
    ...
    ...
</patient>
```

Figure S1: A pseudo example of patient’s record in the OhioT1DM dataset, showing the structure of the tabular data.
Figure S2: **External inputs extracted from patients’ record in the OhioT1DM dataset.** \(u_1\) denotes insulin infusion; \(u_2\) denotes carbohydrate intakes; \(u_3\) denotes exercise intensity.

---

**Algorithm 1 Batch constrained Q-learning algorithm.**

**Require:** Batch \(B\), horizon \(T\), target network update rate \(\tau\), mini-batch size \(N\), max perturbation \(\Phi\), number of sampled actions \(n\), minimum weighting \(\lambda\). Initialize Q-networks \(Q_{\theta_1}, Q_{\theta_2}\), perturbation network \(Q_{\epsilon}\), and VAE \(G_{\omega} = \{E_{\omega_1}, D_{\omega_2}\}\), with random parameters \(\theta_1, \theta_2, \phi, \omega\), and target networks \(Q_{\theta_1}', Q_{\theta_2}', Q_{\epsilon}',\) with \(\theta_1' \leftarrow \theta_1, \theta_2' \leftarrow \theta_2, \phi' \leftarrow \phi, t \leftarrow 0\).

**while** \(t < T\) **do**

Sample mini-batch of \(N\) transitions \((s, a, r, s')\) from \(B\)

\[\mu, \sigma = E_{\omega_1}(s, a), \tilde{a} = D_{\omega_2}(s, z), z \sim N(\mu, \sigma)\]

\[\omega \leftarrow \text{argmin}_\omega \sum (a - \tilde{a})^2 + D_{KL}(N(\mu, \sigma)||N(0, 1))\]

Sample \(n\) actions: \(\{a_i \sim G_{\omega}(s')\}_{i=1}^n\)

Perturb each action: \(\{a_i = a_i + \epsilon_{\phi}(s', a, \Phi)\}_{i=1}^n\)

Sate value target \(y\)

\[\theta \leftarrow \text{argmin}_\theta \sum (y - Q_{\theta}(s, a))^2\]

\[\phi \leftarrow \text{argmax}_\phi \sum Q_{\theta_1}(s), a + \epsilon_{\phi}(s, a, \Phi), a \sim G_{\omega}(s)\]

Update target networks: \(\theta_1' \leftarrow \tau \theta + (1 - \tau)\theta_1'\)

\[\phi' \leftarrow \tau \phi + (1 - \tau)\phi'\]

\(t \leftarrow t + 1\).

**end while**
Hyper-parameters used in the deep offline reinforcement learning.

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizer</td>
<td>Adam (59)</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>Batch Size</td>
<td>512</td>
</tr>
<tr>
<td>Normalized Observations</td>
<td>Yes</td>
</tr>
<tr>
<td>Gradient Clipping</td>
<td>No</td>
</tr>
<tr>
<td>Discount Factor</td>
<td>0.99</td>
</tr>
<tr>
<td>Target Update Rate ($\tau$)</td>
<td>0.005</td>
</tr>
<tr>
<td>Exploration Policy</td>
<td>$N(0,0.1)$</td>
</tr>
</tbody>
</table>

Table S2: Hyper-parameters used in the deep offline reinforcement learning.
Figure S3: Prediction of 6 state variables in the Roy-Parker model using SBINNs patient with ID 559.
Figure S4: Prediction of 6 state variables in the Roy-Parker model using SBINNs patient with ID 563.
Figure S5: Prediction of 6 state variables in the Roy-Parker model using SBINNs patient with ID 570.
Figure S6: Prediction of 6 state variables in the Roy-Parker model using SBINNs patient with ID 575.