Queen's University

MTHE 493: Thesis Report

Optimizing Insulin Treatment for Type I Diabetics: A Reinforcement Learning Approach

Supervised by Dr. Serdar Yüksel

Group S2

Benjamin Armstrong (20097006)
Daniel Martin (20182047)
Ryan Simpson (20165313)
Emma Stickley (20100190)

April 2023

Abstract

Type I diabetes is an autoimmune disease preventing the pancreas from generating the insulin required for the bodily control of glucose levels. This thesis approaches this issue by applying a reinforcement learning approach to control insulin dosing policies for patients, adapting to the unique characteristics of their lifestyles and physiology. Off-policy Q-learning is applied to a patient simulation, using an augmented version of the Bergman Minimal Model to mathematically describe blood glucose-insulin dynamics. A near-optimal Q-table is trained and used as a starting point for individual patient treatment. On-policy Q-learning is then applied to personalize the dosing policy to specific patients.

Simulated patient data is generated, accounting for variations in insulin resistance, time of meals, and nutritional composition of meals. Results show strong performance in maintaining BGL, avoiding hypo/hyperglycemic episodes, and precision in insulin delivery. The proposed method allows for patients to manually input meal signalling to the controller for improved dosing accuracy and BGL management.

Future work includes analyzing the impact of sleep, exercise and stress on BGL. Furthermore, extending this Q-learning approach to other medical conditions could yield promising results. This research contributes to the development of personalized insulin dosing software, aiming to improve patient quality of life whilst complying with healthcare ethics and engineering guidelines.

Acknowledgement

We would like to thank Dr. Serdar Yüksel for his supervision of this thesis. His knowledge and guidance were crucial to the success of the project, and his encouragement fostered our enthusiasm.

Table of Contents

List of Figures

List of Tables

1 Background

1.1 Introduction

Diabetes Mellitus is a chronic condition that inhibits the body from properly regulating blood-glucose levels without interference [1]. According to 2023 reports from the World Health Organization, approximately 422 million people around the world are living with diabetes and 1.5 million deaths are directly caused by diabetes each year [2]. As a result, access to affordable and precise treatment plans are necessary for the survival of diabetes patients [2].

Specifically, type I diabetes accounts for approximately 9 million cases of diabetes globally [2]. Type I diabetes is a chronic, autoimmune disease that results in the pancreas producing little to no insulin on its own [3]. This lack of insulin causes elevated blood-glucose levels, also known as hyperglycemia, and hence the administration of exogenous insulin is required to dissipate glucose from the bloodstream and help it enter the body's cells [3] [4]. Type I diabetes is thought to be caused by an auto-immune process that destroys insulin-producing beta cells in the pancreas [1]. The reason that this auto-immune process occurs remains unknown, and hence preventative measures for the development of type I diabetes do not exist [2].

Some patients with type I diabetes both lack insulin production and struggle with insulin resistance, which is a characterizing feature of type II diabetes [5] [6]. Insulin resistance occurs when tissues that are targeted by insulin have a reduced response to its presence; in other words, a particular patient is considered to have insulin resistance when their insulin-targeted tissues are not as sensitive to insulin as they normally should be [5]. The result is that, for that specific patient, a greater than normal amount of insulin is required to generate any desired physiological reaction [5]. Unlike type I diabetes, the development of insulin resistance can be attributed to a combination of genetic predisposition and lifestyle choices; examples include age, sex, ethnicity, hormone production, weight, diet, and exercise habits [5].

1.2 Current Insulin Delivery Methods

Currently, two main methods of insulin treatment are commercially available: multiple daily injections and continuous insulin infusion by a pump [7]. The goal of both treatment methods is to keep blood-glucose levels within a target range for as much time as possible and prevent periods of low blood-glucose levels and high blood-glucose levels (also known as hypoglycemia and hyperglycemia, respectively) [7].

The highest priority of current insulin therapy techniques is to prevent hypoglycemia, as it is often associated with severe consequences in a short period of time [8]. Hypoglycemia is categorized into three severities:

- **Mild** hypoglycemia is associated with blood glucose levels below 70 mg/dL, and can usually be recognized and addressed by the patient themself [8]. Corresponding symptoms may include sweating and palpitations [8].
- **Moderate** hypoglycemia is associated with blood glucose levels below 55 mg/dL, and might (but not always) be recognized and addressed by the patient themself [8]. Corresponding symptoms may include those affiliated with mild hypoglycemia, as well as cognitive dysfunction [8].
- **Severe** hypoglycemia is associated with blood glucose levels below 40 mg/dL, and requires medical attention [8]. Corresponding symptoms may include unconsciousness, seizures, brain damage, coma, and death [8].

Hyperglycemia is less likely to cause immediate consequences, however leaving high blood-glucose levels untreated for an extended period of time can result in long-term health implications [9]. A patient is typically considered to be hyperglycemic if:

- **While fasting**, their blood-glucose level is above 125 mg/dL [10].
- **After eating a meal**, their blood-glucose level spikes above 180 mg[10].

Untreated hyperglycemia can result in severe damage to blood vessels and consequently impair functioning of major organs such as the eyes, kidneys, nerves, and heart [9].

1.2.1 Insulin Injection Therapy

Daily insulin injections serve as one of the most common forms of insulin therapy. In this method, a syringe (single-use) or an insulin pen (can be single-use or multi-use) is used to administer insulin through needle [11]. Insulin is injected into the subcutaneous tissue, which is the layer of fat tissue under the skin and above muscle, with common injection sites being the abdomen, arm, thigh, and buttocks [12]. For each patient, a physician will assess their glucose levels, diet, and lifestyle to determine the type and dosage of insulin required as well as a necessary injection frequency [11].

Generally, a patient will inject bolus insulin (or fast-acting insulin) at or prior to mealtimes in order to combat spikes in blood glucose that occur after they eat a meal [11]. Then, the patient will often inject basal insulin (or long-acting insulin) one to two times per day in order to regulate blood-glucose levels between mealtimes and overnight [11]. Since the routine of insulin injections is completely patient-controlled (unless the patient is hospitalized), blood glucose levels must be closely monitored by the patient to ensure hypoglycemia is avoided. This can be done by:

1. Using a continuous glucose monitor (CGM). A CGM is a small, disposable sensor that is typically worn on the abdomen or the arm; it has a small needle that penetrates the skin and remains in the subcutaneous tissue in order to measure blood-glucose levels every five minutes [13].

2. A finger-prick blood test. A finger-prick blood test involves obtaining a small blood sample by a finger prick, and using an external meter to read blood glucose levels.

1.2.2 Treatment by Continuous Insulin Infusion

In North America, the most advanced form of insulin therapy is the use of closed-loop insulin delivery systems; these are often referred to as artificial pancreas systems [7]. The artificial pancreas stores insulin in a pump, which is a small device used to administer insulin through a catheter placed into the subcutaneous tissue of the abdomen or arm [7] [11]. In artificial pancreas systems, insulin pumps are used in tandem with CGM systems and separate electronic devices or smartphones are used to monitor and control both blood-glucose levels and insulin dosing [7] [11]. Figure 2 shows an insulin pump and CGM system being used alonside each other.

Figure 1: Continuous insulin administration system with an insulin pump and a CGM device [14].

The development of closed-loop insulin delivery systems has been in motion since the 1970s, and since then systems with six different levels of automation have been studied (Figure 2) [7]. Two main types of these artificial pancreas systems are commercially available today: low-glucose suspend (LGS) systems and hybrid closed-loop systems [7].

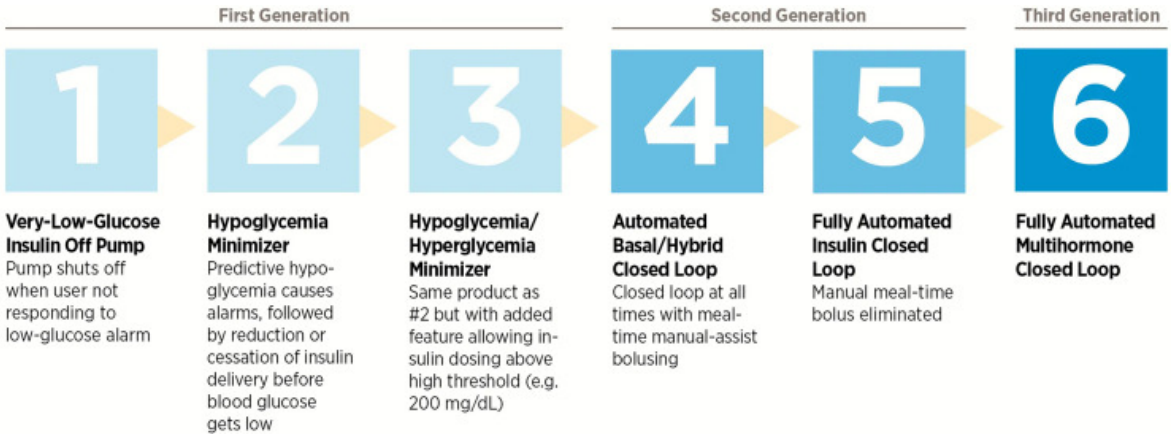

Figure 2: Categorization of closed-loop systems by six levels of automation [7].

Low-glucose suspend (LGS) systems have the ability to stop insulin administration when the patient's blood-glucose level falls below a pre-determined threshold, and can do so without user-input [7]. Further refinement of LGS systems resulted in the development of predictive low-glucose suspend (PLGS) systems, which use an algorithm to predict future hypoglycemia [7]. Hence, PLGS systems can suspend insulin administration prior to blood-glucose levels falling below the patient's pre-set threshold [7]. While LGS and PLGS systems are adept at handling hypoglycemia, they fail to address hyperglycemic spikes that occur after the patient eats a meal [7].

Hybrid closed-loop systems aim to minimize both hypoglycemia and hyperglycemia by continuously adjusting basal insulin and administering corrective bolus insulin based on the patient's current blood-glucose reading [7]. Even though these systems rely on algorithms to fine-tune basal and bolus insulin doses, they are not considered to be "fully" closed-loop since the patient is still required to manually input a bolus insulin dose after they eat a meal [7]. Bolus insulin doses must be calculated by the patient based on the carbohydrate content of the meal, as well as their current blood-glucose level, and entered into the electronic device or smartphone controlling the pump [7]. Numerous studies have shown that the use of hybrid closed-loop insulin delivery systems increases the time that a patient spends within their target blood-glucose range, however this is purely dependent upon their diligence and accuracy in inputting bolus doses [15] [16].

Both LGS and hybrid closed-loop systems rely on three different algorithms to control automated insulin administration [7].

- **Model predictive control (MPC)** algorithms use a model of the patient's glucose dynamics to predict blood-glucose fluctuations [7] [17]. The patient's insulin infusion rate is adjusted based on what dose will minimize the difference between the model-predicted blood-glucose level and the patient's blood-glucose target [17].

- **Proportional-integral-derivative (PID)** algorithms adjust insulin infusion rate by accounting for three different measurements: the difference between the current and target blood-glucose levels (proportional control), the area under the curve between the current and target blood-glucose levels (integral control), and the rate of change of observed blood-glucose levels (derivative control) [7] [17].
- **Fuzzy logic** algorithms control insulin infusion rate based on a set of parameters or rules that are meant to mimic a practitioner's diagnosis [7] [17]. These predetermined rules are based on practical knowledge and experience of diabetes practitioners [17].

Multiple studies have shown that artificial pancreas systems employing MPC and fuzzy logic are more successful in keeping the patient within their target range as compared to systems that use PID control [17] [18]. However, many patients that use artificial pancreas systems will frequently deactivate the algorithm since its lack of robustness creates fear that it will fail and cause hypoglycemia [7].

1.3 Development of Fully Closed-Loop Insulin Delivery

The development of a fully closed-loop artificial pancreas has become a prevalent topic in insulin therapy research. Fully closed-loop systems are meant to completely automate the insulin administration process by eliminating the need for patients to manually input mealtime boluses; the ultimate goal of these systems is to provide type I diabetes patients with treatment that is more convenient and less invasive as compared to traditional methods [7]. A 2017 meta-analysis on artificial pancreas systems reported that the majority of fully closed-loop systems being trialed rely on MPC algorithms to regulate insulin administration, however PID and fuzzy logic algorithms have been used in some cases [18].

The largest concern in fully closed-loop insulin delivery systems is hyperglycemic spikes following mealtimes, since the system no longer has information on the timing and carbohydrate content of the patient's meals [7]. This phenomenon may result in delayed administration of high insulin concentrations, hence leading to hypoglycemia [7]. Early trials have demonstrated that fully closed-loop systems are sufficient at minimizing hypoglycemia, however they lack robustness in responding to post-meal hyperglycemia; the trials showed that the performance of fully closed-loop systems significantly improves when a patient manually inputs meal boluses [7] [19]. Researchers have attempted to improve post-meal insulin responses in fully closed-loop systems by using algorithms to detect unannounced meals, however future studies are required to determine their feasibility [20] [21]. These algorithms are meant to predict when a patient is eating a meal, as well as the carbohydrate content of that meal [20] [21].

Currently, the only fully closed-loop insulin delivery system that is commercially available is restricted to use in Japan [7] [22]. This system is a bedside device that both

measures blood-glucose levels and administers insulin intravenously, and hence it has only been approved for supervised use in hospitals for a limited three-day period [22].

1.4 Using Reinforcement Learning Methods in Insulin Dosing

In recent years, the use of reinforcement learning in insulin treatment has become an emerging research topic. Previous works have attempted to employ both model-based and model-free reinforcement learning algorithms in developing optimal insulin care policies to be used in closed-loop insulin delivery systems. This research demonstrates that reinforcement learning algorithms may be a useful tool in recommending accurate and safe insulin doses for type I diabetes patients, however further testing and investigation of these methods is required before they can be implemented in-practice.

1.4.1 Modelling Blood-Glucose-Insulin Dynamics

The concept of using reinforcement learning algorithms in insulin treatment is still relatively new, and as a result it has not yet been deemed safe to test these algorithms on real patients. Instead, researchers rely on simulation environments in which they model "test patients" to trial their algorithms on. In simulating a patient, it is important to have an accurate model of their blood-glucose-insulin dynamics in order to understand how a specific insulin dose will impact their blood glucose level. The human body is a very complex system, and hence a unique patient's blood-glucose-insulin dynamics cannot be modelled exactly. However, researchers have employed various methods as an attempt to approximate what impact insulin will have on a specific patient's blood-glucose level.

Type I Diabetes Patient Datasets
Some studies choose to initialize test patients based on type I diabetes patient datasets. These datasets include information regarding various patient's blood-glucose levels and corresponding insulin dosages over time. They may also include additional information regarding the patient's meal habits, stress levels, and sleep habits [23]. When generating a blood-glucose-insulin model based on patient data, having a large dataset is more beneficial as it will provide more diversity in patient responses. Examples type I diabetes patient datasets include the OhioT1DM dataset from Ohio University and the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) dataset [23] [24].

Type I Diabetes Simulators
Many studies choose to use type I diabetes simulators in order to construct virtual patient environments. Two main simulators are used in diabetes research.

1. **The AIDA Simulator** is an interactive type I diabetes simulator that was created for educational purposes [25]. It can simulate one full day of blood-glucose

levels for a test patient, given the following information: weight, carbohydrate content of six meals, and timing and dosage of four insulin treatments [25].

2. **The UVA/PADOVA Simulator** is an FDA-approved type I diabetes simulator [26]. It is equipped with a database of pre-defined patients, and can simulate the blood-glucose-insulin dynamics for a given amount of time and for a given meal plan (timing and carbohydrate content of meals) [26].

Mathematically Modelling the Human Metabolic System
In some studies, researchers choose to personally define a unique set of test patients and employ a mathematical model to determine their blood-glucose-insulin dynamics. A well-known model in the type I diabetes research community is the Bergman Minimal Model (BMM) [27]. The BMM is a simplified model of three differential equations that is used to determine a patient's next blood-glucose level based on the following current measurements: blood-glucose level, gut carbohydrate content, and previous insulin dose administered [27]. The model is presented in Appendix A

1.4.2 Previous Work Done: Value-Iteration

Shifrin et al. used a value iteration algorithm (model-based) to generate an optimal insulin care policy to be used in hybrid closed-loop insulin delivery systems [9]. The NIDDK dataset was used to define a set of test patients, and the blood-glucose-insulin dynamics of each test patient was modelled using the AIDA simulator [9]. The following information was used to estimate a patient's future blood-glucose-level (every five minutes):

- The patient's current blood-glucose level [9].
- Whether or not the patient is eating a meal, and if so the estimated carbohydrate content of that meal (limited to six meals per day) [9].
- Whether or not insulin is being administered, and if so the given dose (limited to four treatment points per day) [9].

The algorithm used a reward function, that was individualized to each patient, based on the patient's current and target blood-glucose values [9]. The resulting optimal care policy allowed test patients to remain within their target blood-glucose ranges more than 80% of the time; this is greater than the amount of time the test patients remained within their target blood-glucose ranges when relying on conventional treatment methods [9].

1.4.3 Previous Work Done: Q-Learning

Javad et al. used a Q-learning algorithm (model-free) to generate an optimal insulin treatment policy for a set of test patients, and compared this policy to the actual

insulin dosage that was prescribed to each patient by a physician [28]. The algorithm was trained on patient data from the Massachusetts General Hospital, and used the following information to learn an optimal insulin care policy for each patient:

- The patient's blood-glucose levels [28].
- The patient's body mass index [28].
- The patient's average weekly engagement in physical activity [28].
- The patient's alcohol consumption per week [28].

The algorithm used a reward function that was based on the patient's current and target blood-glucose values alone [28]. The study found that the optimal insulin dosage given by Q-learning matched the physician's diagnosis in 88% of cases, however a limitation was that meal consumption was not considered which is known to have a significant impact on a patient's blood-glucose-insulin dynamics [28].

1.4.4 Previous Work Done: Q-Learning and Deep Learning

Shi et al., as well as Emerson et al., studied the combination of Q-learning and deep learning (model-free) in optimizing insulin therapy for a set of test patients [29] [30]. The algorithm used by Shi et al. was trained on a the OhioT1DM dataset [29]. In the experiment, each day was divided into discrete one-hour intervals; hence, every hour a decision was made as to what dose of insulin should be administered, or if no insulin should be administered [29]. The algorithm used the following information to learn an optimal insulin care policy for each patient:

- The patient's average blood glucose level per hour [29].
- An estimate of the carbohydrate content in any meals the patient eats per hour [29].
- The intensity of exercise the patient completes per hour [29].

The algorithm used a reward function based on the patient's projected blood-glucose level at a given time [29]. The performance of the resulting optimal care policy seemed to decrease over time, which was most likely a result of limitations in the patient dataset; the algorithm was only trained on data from six patients, which most likely resulted in some patient states not being explored during training [29].

Meanwhile, the algorithm employed by Emerson et al. was trained using nine virtual patients who's blood-glucose-insulin dynamics where modelled by the UVA/PADOVA simulator [30]. The goal of their experiment was to use Q-learning and deep learning to improve blood-glucose control within hybrid closed-loop insulin delivery systems [30]. The following information was used by the algorithm in order to determine how a patient's insulin infusion rate should be updated (every thirty minutes):

- A history of the patient's ten most recent glucose measurements (sampled at thirty-minute intervals) [30].
- An estimation of the carbohydrate content in any meals the patient has eaten within the last thirty minutes [30].

The reward function used by the algorithm was based on the patient's current and target blood-glucose levels [30]. The results of this study show that employing Q-learning and deep learning in hybrid closed-loop insulin delivery systems can increase the amount of time a patient spends in their target blood-glucose range as compared to using PID algorithms [30]. However, the success of this study is dependent on carbohydrate contents being reported accurately and at every meal; the performance of the optimal insulin care policy greatly decreased when meal uncertainties or inaccuracies were considered [30].

2 Design Process

2.1 Problem Statement

Patients suffering from type I diabetes require manual insulin administration to maintain their blood glucose levels within a safe range. Industry-standard insulin dosing policies predominately use a generalized approach for all patients. This decision disregards individual lifestyle factors such as exercise, appetite and physiological characteristics. The generalized approach provides patients with safe blood glucose levels but provides a sub-optimal policy which can increase the risk of long-term complications.

The purpose of this research is to employ RL techniques to provide patients with a personalized insulin dosing policy adapted to their lifestyles. This policy will iterate over time, reducing the risk of patients contracting hyper- and hypoglycemia. The proposed algorithm will adjust policy based on a continuous source of bio-signal data, including blood levels and rate of insulin administration. This model can be implemented into existing insulin delivery devices and CGMs. The limits of current insulin dosing practices will be addressed, ultimately leading to improved long-term health, reduced healthcare costs and improved quality of life for patients suffering from type I diabetes.

2.2 Design constraints

2.2.1 Safety and Regulatory Considerations

When using machine learning for policy optimization of insulin dosing in type I diabetes patients, health and safety considerations must be taken into account. The following concerns address the simulated patients used in this research, as well as further considerations for possible patient trial and testing.

1. **Patient safety:** The well-being of patients is of the highest priority. Maintaining glucose levels within safe ranges takes president over advancements in algorithmic development and testing.

2. **Hardware compatibility:** The dosing policy must comply with current standards of type I diabetes medical equipment. The current standard for CGMs distributed to patients is a data sampling rate of 5 minutes per reading [13], hence this frequency will be used for simulated BGL observations.

3. **Patient ease of use:** Improvements in policy optimization at the cost of patient accessibility may render solutions redundant. These methods could include relying upon other data acquisition hardware, patient monitoring of glucose intake, or any other attempts to gain additional information at the cost of patient convenience. The purpose of researching RL applications for insulin infusion is to increase ease of use while maintaining compliance with safe glucose levels.

4. **Algorithmic development and validation:** Rigorously testing and validating the Q-learning algorithm using simulated and retrospective data is necessary before considering applications to real-world environments. The algorithm must be robust to edge-cases, uncertainties and all possible variations in patient responses before any clinical trials can take place.

5. **Clinical trials and implementation:** Rigorous clinical trials are necessary to evaluate the safety and efficacy of the insulin dosing policy. Following federal medical regulations and ethical guidelines is necessary. In Canada, medical technology prototypes must adhere to the guiding principles of good machine-learning practice for medical equipment in Canada [31].

2.2.2 Triple Bottom Line Considerations

1. **Societal Factors**

Type I Diabetes is a leading cause of death in many countries. Among Canadians, type I diabetes results in 19.9 deaths per 100,000 from 2000-2020 [32] Providing patients with higher-quality care reduces mortality. Furthermore, poor control of glucose levels can lead to hyper- and hypoglycemia, which can induce conditions such as dizziness, heart disease, stroke, and many other life-threatening long-term complications [10] [33]. Furthermore, dealing with type I diabetes requires patients to make several lifestyle changes that impede their quality of life. Restrictions include constant monitoring of blood sugar levels, several dietary restrictions, mandated exercise to maintain healthy body weight, and mitigating sources of stress [34]. Accurate control of blood sugar levels grants access to freedoms many take for granted, providing patients with a source of empowerment in their life.

2. **Environmental Factors**
Optimizing the dosing policy allows for more efficient use of insulin, in turn, reducing environmental impact. Many issues are caused by the manufacturing of insulin and other pharmaceutical products such as water contamination, energy consumption, landfill waste products, and carbon emissions [35]. Ecological harm associated with the manufacturing and transport of insulin can be alleviated by minimizing the consumption waste of pharmaceutical products.

3. **Economic Factors**
The average Canadian suffering from Diabetes spends approximately $725 on insulin annually [36]. Effective use of insulin reduces costs for patients and decreases the rate of hospitalization. In turn, this improvement will alleviate the burden on the healthcare system, allowing resources to be allocated to other pressing issues.

2.3 Patient simulation

Implementation of reinforcement learning requires a large quantity of historical data. Due to privacy concerns, datasets tracking CGM readings and insulin infusion for specific patients require approval for release. Due to this lack of publicly available data tabulating the individual dosing history for type I diabetes Patients, a mathematical model is employed to simulate patients.

2.3.1 Expanded Bergman Minimal Model

The Expanded Bergman Minimal Model (EBMM) is a nonlinear dynamic system model (based on the Bergman Minimal Model shown in Appendix A) describing the relationships between meal intake, blood glucose, and insulin infusion. This model has been used widely in the field of diabetes research to study and predict the behaviour of insulin-glucose regulation [37].

This expanded BMM adds several subsystems to better describe the behaviour of gut glucose and two subcutaneous insulin compartments. The model is a nonlinear dynamic system of six differential equations with the following variables.

- $G(t)$: Blood glucose concentration (mg/dl)
- $I_R(t)$: Plasma insulin concentration (mU/L)
- $I(t)$: Remote insulin action (mU/L)
- $S_1(t)$: Subcutaneous insulin compartment 1 (mU)
- $S_2(t)$: Subcutaneous insulin compartment 2 (mU)
- $G_g(t)$: Gut glucose concentration (mg/dl)

- $D(t)$: Meal intake (g/min)
- $u(t)$: Insulin infusion (mU/min)

Variables are related by a set of nonlinear differential equations representing the change in model variables over time:

$$\begin{aligned}
\frac{dG}{dt} &= -p_1(G - G_B) + -s_i I_R G + \frac{f k_{abs} G_G}{v_g} + \frac{fD}{v_g} \\
\frac{dI_R}{dt} &= -p_2(I - X) \\
\frac{dI}{dt} &= -s_E I + u \\
\frac{dS_1}{dt} &= u(t) - kemp_{S_1} \\
\frac{dS_2}{dt} &= -k_{emp}(S_2 - S_1) \\
\frac{dG_g}{dt} &= k_{emp} S_2 - k_{abs} G_G
\end{aligned} \tag{1}$$

Model parameters are defined as follows:

- $g_b = 291$
- $p_1 = 0.0317$
- $p_2 = 0.0123$
- $s_i = 0.029$
- $k_e = 0.09$
- $k_{abs} = 0.012$
- $k_{emp} = 0.18$
- $f = 0.8$
- $v_i = 12$
- $v_g = 12$

2.3.2 Simulating Patient Individuality

The EBMM is designed to model the characteristics of an average patient. To simulate the characteristics and lifestyle of individual patients, we consider variations in the time of meals, size of meals, and insulin resistance. This approach provides the reinforcement learning algorithm with more unique environments in which it will interact and adapt to the characteristics of different patients, eventually developing different insulin dosing policies.

Insulin resistance: Insulin resistance is caused by the reduced ability of the patient's cells to respond to insulin, a very common condition varying in severity among many type I diabetes patients [38]. Insulin resistance was modelled by adjusting the value of p_2 which adjusts the rate of change of I_R, subsequently causing the blood glucose to have variable sensitivity to insulin infusion. By generating patients with different levels of insulin resistance, the RL algorithm learns policies for patients with varying insulin sensitivity.

Meal Times and Quantities: Variance in lifestyle means patients do not all follow similar meal habits. Specifically, the size and time of consumption varies on a patient-level basis. Randomized meal times and quantities are introduced into the patient simulation model for this discrepancy. Patients are assumed to have 3 large meals spread throughout the day, along with a random chance of having an intermittent snack. This model was determined to represent the standard habits of human behaviour, while providing a component of unpredictability.
Figure 3 shown below depicts 10 possible meal distributions over the course of a day.

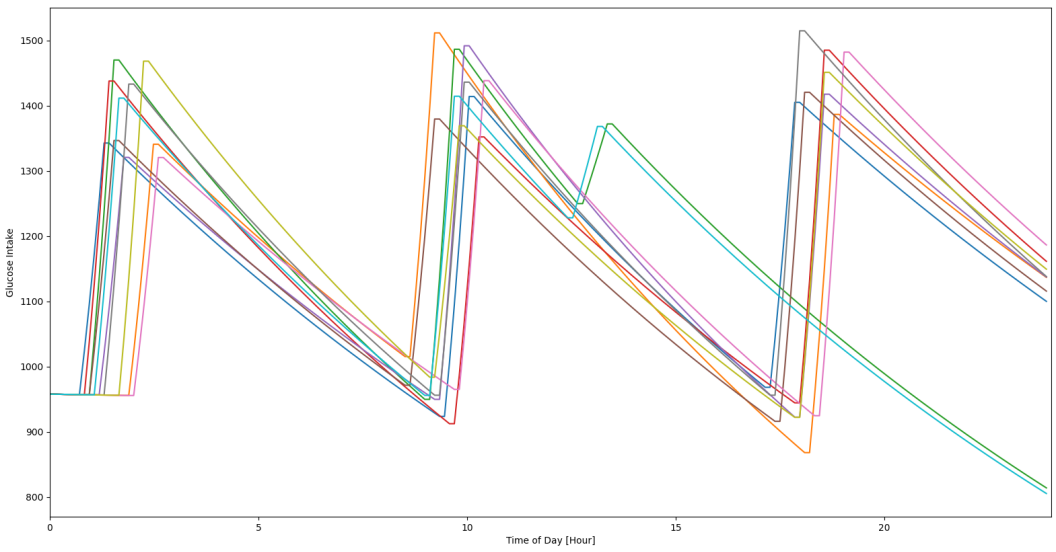

Figure 3: Sample meal distributions over the course of a day

2.4 Stochastic Control-Policy Optimization

2.4.1 Insulin Dosing as a Partially Observed Markov-Decision Process

A Partially Observed Markov Decision Process (POMDP) models situations in which outcomes are dependent on the actions of the agent. In the context of insulin dosing, the problem can be formed as an POMDP.

- **State** (X)**:** States in the POMDP represent all information in the patient that is relevant to the patient's insulin-glucose dynamics.
- **Measurement** (Y)**:** Measurements in the POMDP are represented by the observation of a patient's blood glucose level, as well as other relevant factors such as previous glucose readings, and previous insulin infusions. Measurements determine which variables are considered when choosing an action, hence is crucial to the performance of the dosing policy.
- **Action** (U)**:** Actions correspond to possible rates of insulin infusion administered to the patient.
- **Transition probabilities** ($P_u(X, X')$)**:** Transition probabilities describe the likelihood of transitioning from one state to another, as the outcome of an action. In the context of insulin dosing, transition probabilities are unknown, as they are influenced by several unknown factors such as the patient's metabolism, exercise habits, carbohydrate intake, and various patient-specific conditions which influence insulin-glucose dynamics within the body.
- **Reward** ($R(X)$)**:** The reward function describes the desirability of a given state. For insulin dosing, a reward function must encourage actions that maintain the patient's BGL close to the target, penalizing actions which result in BGL reaching levels at which hypo- and hyperglycemia occur.

These items Modeling the decision-making process of insulin dosing given a set of information in an POMDP allows for the application of reinforcement learning techniques. Please refer to the Appendix B for a more detailed overview of stochastic control and POMDPs, within the context of the insulin-glucose problem.

2.4.2 Overview of Policy Optimization Models for Insulin Dosing

Many different policy optimization techniques are suitable for MDPs. Different solutions have benefits and drawbacks depending on the nature of the problem.

Proportional-Integral-Derivative (PID) Controllers: PID Controllers are widely used in many feedback control mechanisms. Applied to the context of insulin dosing, PID controllers can maintain BGL within the target range by adjusting insulin infusion based on error from the desired glucose level. Unfortunately, the constraint of periodic

5-minute measurements limits the effectiveness of PID, as low temporal resolution data prevents Integral and Derivative terms from taking full effect, thus it cannot provide sufficient response to rapid changes in glucose levels.

Model-Predictive Control (MPC): MPC utilizes a system model to predict future states, then determines an optimal action accordingly. This control mechanism can achieve accurate results but requires a mathematical model for the system. While patients are being simulated using a mathematical model, the RL algorithm does not have access to specific mathematical models for each patient. Furthermore, MPC is suited for use on linear system models [38]. The EBMM utilizes non-linear system dynamics. Hence, this form of model-based learning is not suitable for the purposes of this project.

Value Iteration and Policy Iteration: Value iteration and policy iteration are dynamic programming methods used for solving MDPs. These optimization techniques require complete knowledge of the transition kernel. Accurately modelling the biological processes of patients is infeasible, hence a form of model-free learning is required.

Q-Learning: Q-Learning is a model-free RL algorithm that learns the quality of each action at a given state without access to transition probabilities of the MDP. Q-Learning updates the Q-function based on the observed rewards of state-action transitions. It does not require a mathematical model and is capable of constantly updating policy based only on real time-data. These features make Q-Learning a promising approach for offering personalized, adaptive, and effective dosing policies on a patient-specific basis.

2.4.3 Q-Learning

Q values for all state-action pairs (x_t, u_t) are stored in the Q-table. Values are recursively defined using the Bellman equation:

$$Q_t(x_t, u_t) = (1 - \alpha_t(x_t, u_t)Q_t(x_t, u_t) + \alpha_t(x_t, u_t)(r(x_t, u_t) + \gamma \max_a Q_t(x_{t+1}, a)) \quad (2)$$

Where

- R is the reward function. This determines the desirability of a given state-action pair.
- γ is the discount factor, used to quantify the importance of future results. For the purposes of our simulation, future results are valued highly, as safe glucose levels must be preserved at all times.
- $\alpha \in (0, 1]$ is the learning rate. The learning rate determines how new information affects the Q function. High learning rates allow the Q-function's policy to change

rapidly in response to recent results, while low learning rate makes the learning process more stable. Learning rate plays a crucial role in the development of optimal policy. Q values may converge, or continuously change depending on how the α is defined.

- λ is the exploration rate. The exploration rate determines an agent's tendency to try new actions, which may lead to discovering better policies. This off-policy decision-making ensures all possible actions are explored. At a given state x_t, the probability of choosing an action $u_t \in U$ is defined as:

$$P(a = u_t | x_t) = f(x_t, \lambda)$$

for some measurable f.

2.4.4 Design Criteria

To evaluate the efficacy of solutions, the following design criteria must be considered.

1. **Maintain BGL within target range:** The primary objective of the control system is to ensure the patient's blood glucose levels are maintained within a safe range, minimizing the risk of side effects and long-term complications.

2. **Avoid hypoglycemic episodes:** Hypoglycemia occurs when blood glucose levels fall below safe levels. This issue can cause dizziness, cognitive impairment, and fainting. Hence, the algorithm must avoid actions causing hypoglycemia.

3. **Avoid hyperglycemic episodes:** Hyperglycemia occurs when blood glucose levels rise above safe levels, causing kidney damage, nerve damage, and cardiovascular diseases. It is crucial for the algorithm to select actions that avoid hyperglycemia.

4. **Average optimal BGL:** The control system should maintain blood glucose levels around an optimal setpoint.

5. **Robust to random meal timing and amounts:** The control system must account for variance in the patient's lifestyle, eating meals at irregular times and varying nutritional value in order to effectively regulate glucose levels while allowing for freedom of lifestyle.

6. **Robust to random snacks:** Similar to meals, the control system must adapt to unpredictable snacking habits throughout the day, which affects glucose levels.

7. **Fully automated system:** The control system ideally should operate without any manual intervention or data input from the patient, to minimize the burden of living with diabetes.

8. **Adapting to individual patient dynamics:** Each person's physiological response to insulin and blood glucose is unique. The model must be adaptable to these changes and be able to provide the optimal dosage regardless of patient dynamics.

9. **Efficiently converges to optimal dosages:** The algorithm should adapt quickly based on patient-specific data, converging to optimal policy within a reasonable time frame. This is significant as a patient's dynamics can change over a short time frame and the model must adapt quickly to these changes.

By evaluating the development of algorithmic solutions against these design criteria, a solution can be developed which will minimize the risk of diabetes while maximizing the quality of life for patients.

3 Solution Design

With the chosen simulation model and policy optimization technique decided upon, a full integrated solution was designed with all technical and ethical constraints considered. The high-level outline of the final design is summarized in the following:

- Apply Q-learning to general patient simulated using (1)
- Save Q-table learned in the previous, this will be the near-optimal policy that subsequent patients will be initialized with
- Apply on-policy Q-learning to individual patients

3.1 Mathematical Design of Solution

When tuning the Q-learning algorithm to best meet the design requirements, it was important to first perform a rigorous analysis of the Q-learning algorithm when applied to partially observed models.

3.1.1 Off-Policy Q-learning for General Patient

When applying Q-learning to a general patient, it is important for the Q-table to converge to some Q*, in order to obtain a general optimal policy. In order to ensure convergence of the Q-learning algorithm implemented, the following assumptions were made regarding the problem setup and algorithm implemented:

Assumptions for Q-learning:

1. The one stage reward function R is bounded and continuous
2. The stochastic kernel $\tau(\cdot|x,u)$ is weakly continuous in $(x,u) \in \mathbb{X} \times \mathbb{U}$

3. The action space, $\mathbb{U}$ is compact

4. The state space, $\mathbb{X}$ is compact

5. $y = \rho(x)$, where ρ is the nearest-neighbor quantizer for the set $\mathbb{X}_{measured}$ (the state fed to the controller)

6. $\alpha_t(y,u) = \frac{1}{1+\sum_{k=0}^{t} 1_{\{Y_k=y, U_k=u\}}}$

7. Under the exploration policy, P, the distribution of state processes converges to the invariant measure π_P weakly

8. During the exploration phase, every observation-action pair (y,u) is visited infinitely often

Using the assumptions above as constraints, the problem was formulated as a POMDP, and the Q-learning algorithm was designed accordingly. The precise formulation of the problem as a POMDP is given in the following.

- The state space, $\mathbb{X}$ is defined as the set of values of the form $X = (G, I_R, I, G_g, S_1, S_2, t, t_{meals})$, where the first 6 terms are the state variables in the dynamics (1), t is the time of day, (in increments of 5 minutes), and $t_{meals} = 1_{\{eating\}}$ is an indicator variable that is equal to 1 when the patient is eating a meal. The time of day and time of meals is included in the state because, due to diet habits that govern when meals are typically eaten, the distribution of state processes are not time-invariant, and so by including these values in the state this can be remedied.

- The action space, $\mathbb{U}$, is the set of possible doses of insulin that can be administered. In addition to being compact, this set should have finite elements, so that every observation-action pair, (y,u), may be visited infinitely often. Thus, $\mathbb{U}$ is set to a finite number of values in the range of $\{[0,10] \cap \mathbb{R}\}$. Subject to these constraints, $\mathbb{U}$ is treated as a tun-able parameter.

- The space $\mathbb{X}_{measured}$, the space of state-values that are available to the controller is defined to be the set of information variables of the form $I = (G_{[t-N,t]}, U_{[t-N,t-1]}, \tau = h(t, t_{meals}))$. Here, $N \in \mathbb{N}$, and $\tau = h(t, t_{meals})$ is a variable which keeps track of the time which has elasped since the most recent meal was eaten. The value N, and the decision to include τ are treated as tun-able parameters.

- The observation space, $\mathbb{Y}$ is the set of observations that are used by the controller to make decisions (ie. the observations seen by the Q-table). The set $\mathbb{Y}$ contains all values of the form $Y = \rho(I) = (\rho(G_{[t-N,t]}), U_{[t-N,t-1]}, \tau)$. Here $\rho : \mathbb{R} \to \{[0,240] \cap \mathbb{R}\}$ is the nearest-neighbor quantizer. The number of bins in the quantizer is viewed as a tun-able parameter.

See Appendix B.2 for more details of how the POMDP is formulated. Now, the following are the parameters used for Q-learning:

- The reward function, R, is defined as:

$$R(x_t, u_t) = R(g_t) = \begin{cases} -1000, & (g_{hypo} - 45) \leq g_t < (g_{hypo} - 25) \\ \frac{-1000}{(-25)^2} \cdot (g_t - g_{hypo})^2, & (g_{hypo} - 25) \leq g_t < g_{hypo} \\ (g_t - g_{hypo})^3, & g_{hypo} \leq g_t < g_{target} \\ \frac{(500 - (g_{target} - g_{hypo})^3)}{(g_{hyper} - g_{target})^2} \cdot (g_t - g_{target})^2 + (g_{target} - g_{hypo})^3, & g_{target} \leq g_t < g_{hyper} \\ \frac{-1000}{75^3} \cdot (g_t - g_{hyper})^3 + 500, & g_{hyper} \leq g_t < (g_{hyper} + 75) \\ -500, & (g_{hyper} + 75) \leq g_t \leq (g_{hyper} + 95) \end{cases} \tag{3}$$

  where g_t is the BGL at time t.

- γ, the discount factor, was experimentally determined to be 0.9

- A decreasing learning rate α_t is implemented to ensure convergence of Q-values. α_t is defined as follows.

$$\alpha_t(y, u) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{Y_k = y, U_k = u\}}} \tag{4}$$

- The exploration policy, P, is defined as:

$$P(u_t = a | x_t) = \frac{e^{-\lambda Q[x_t, a]}}{\sum_{u \in U} e^{-\lambda Q[x_t, a]}}$$

  where $\lambda = 0.9$ is the exploration rate.

Theorem 3.1. *Assume that the transition kernel $\tau(\cdot|x, u)$ possesses the weak-feller property for $(x, u) \in \mathbb{X} \times \mathbb{U}$, and that the assumptions in 3.1.1 hold. Then, with the POMDP defined in 3.1.1, the Q-learning algorithm with the parameters as defined also in 3.1.1 will be near-optimal for the problem with I as its' state, as the quantization rate goes to ∞.*

Proof. With this POMDP, and under these assumptions, the result follows directly from [39].

□

3.1.2 On-Policy Q-learning for Individual Patient

For the case in which Q-learning is being applied to an individual patient that may have slightly different dynamics than the general patient, it is not necessary to have convergence. In fact, it is more desirable to not have convergence, since then the controller will remain robust to physiological changes over time. Additionally, it is not ethical to use a random exploration policy, since the patient should always be given the perceived optimal dose of insulin, even during training. The assumptions and constraints given in the pervious section, therefore, do not apply. The formulation of the POMDP for the case of an individual patient remains unchanged from the general one, although in this case the learning rate α_t, and the exploration policy P are changed. The following are the changes made:

- The learning rate, α_t, is given by $\alpha_t = \alpha = 0.1$
- The exploration policy, P, is given by:

$$P(u_t = a|x_t) = \begin{cases} 1 & a = \arg\max_{u \in \mathbb{U}}(Q[x_t, u]) \\ 0 & otherwise \end{cases}$$

3.2 Design Architecture

Prior to software implementation and simulation tuning, a system architecture was designed in order to capture all necessary capabilities and dependencies, and to visualize the flow of data. The system architecture was designed within the context of current technological capabilities, and so it is assumed that a Continuous Glucose Monitor (CGM) is used to collect and deliver BGL readings every 5 minutes from the patient to the controller. The system architecture is shown below:

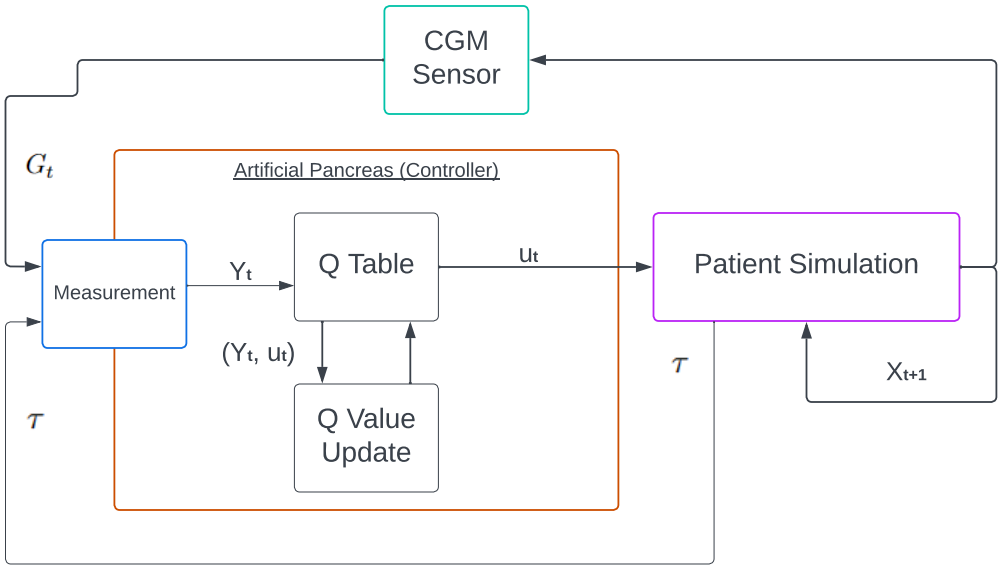

Figure 4: Q-table development framework.

During this simulation, parameters free of additional noise or insulin resistance were used. This method ensures the generalized Q-table provides an optimal dosing policy for the average patient, which can then be fine-tuned after sufficient data acquisition. When developing the generalized policy, the algorithm is not deployed in patients, thus no lives are at risk.

Now, a system architecture was also developed for the broader goal of initializing a controller with a general optimal Q-table for an individual patient and then allowing this controller to adjust its' policy to fit the needs of that patient using an 'on-policy' version of Q-learning. The system architecture for this process is shown below:

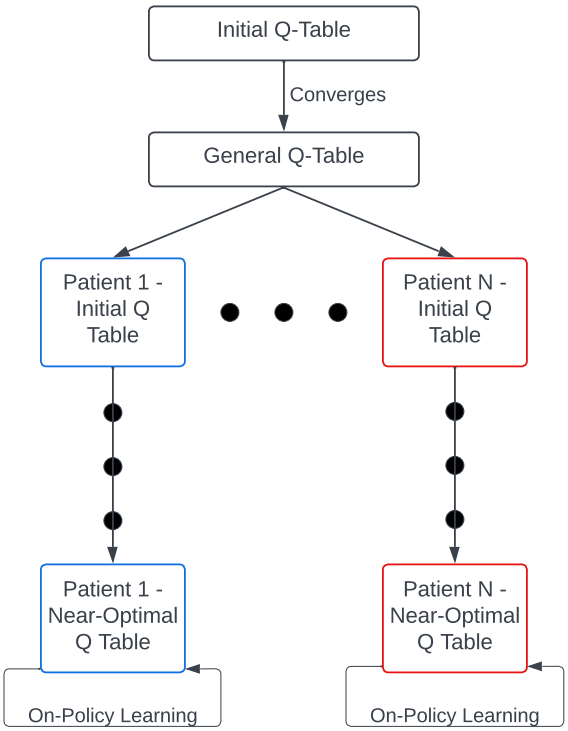

Figure 5: High-Level Architecture for Individualized On-Policy Learning Model

As can be seen in the architecture above, the Q-learning algorithm used to develop the general optimal Q-table must converge in order to obtain a static Q-table that can be saved and thus be available to load into the controller of individual patients when initialized. In contrast, the near-optimal Q-table that is learned for the individual patient should **not** converge. This is because the controller should remain flexible and robust over time, to reflect the fact that the patients' physiology may vary over time and the controller must adjust accordingly.

4 Implementation and Results

The final solution was implemented and tested in Python according to the system architecture outlined in the previous section. Python was used because its' object-oriented paradigm allowed for logical and intuitive development, and because the extensive packages and resources simplify computation and provide excellent means to visualize results. A patient class was created which contains all relevant information pertaining to the patients' glucose-insulin system, including the values and dynamics of all the

state variables in (1), the patients' reward function, and the meals that the patient will eat on a given day. Patients were then created and used as the framework in which to implement the architectures in Figure 4 and Figure 5.

4.1 Algorithm Pseudo-Code

The algorithmic implementation of the architectures in Figure 4 and Figure 5 will be presented separately, although it should be noted that Algorithm 2 depends on Algorithm 1. It should also be noted that not all code will be presented in this section, in order to maintain readability and to avoid redundancy. Simple functions and processes such as plotting, initializing, and certain computations are omitted.
Now, the pseudo-code for the software implementation of Figure 4 is shown below:

Algorithm 1 Q-learning Pseudo-Code (Q-Value Update)
Arguments: Q-table, Patient (object which includes X_t, τ, meals, time, Reward()), U_t, α, γ, λ

$\tilde{X}_t \leftarrow Quantize(X_t['BGL'])$
$Y_t \leftarrow (\tilde{X}_t, \tau)$
procedure SIMULATE ACTION(Arguments: Patient, U)
  if time % length(meals) = 0 **then**
    meals $\leftarrow$ Randomize meals (randomize full day of eating)
  end if
  procedure SOLVE IVP(Arguments: X_t, U_t, meal, time step)
    Solve Initial Value Problem for differential equations (1) with initial value of X_t, action = U_t, meal = meals[time], time step = 5 mins
    return Solution of IVP over time step (5 mins)
  end procedure
  $X_{t+1} \leftarrow$ Solution of IVP at end time (5 mins)
  if $time < t_{breakfast}$ **then**
    $\tau \leftarrow -1$
  else
    $\tau \leftarrow (time - last)$, where $last = min(t \in \{t_{breakfast}, t_{lunch}, t_{dinner}\} : t < time)$
  end if
end procedure
$maxQ \leftarrow max(Q[X_{t+1}, U] : U \in \mathbb{U})$
$Q_{t+1}[Y_t, U_t] \leftarrow (1 - \alpha)Q_t[Y_t, U_t] + \alpha(Reward(X_{t+1}) + \gamma(maxQ))$
for $U \in \mathbb{U}$ **do** $P(U_{t+1} = U) = \frac{e^{-\lambda * Q[X_{t+1}, U]}}{\sum_{U \in \mathbb{U}} e^{-\lambda * Q[X_{t+1}, U]}}$
end for
return $Q_{t+1}, X_{t+1}, U_{t+1}$

Now, the software implementation of Figure 5 is shown in the following. The first of the following two algorithms (2) is the pseudo-code implementation of the 'off-policy'

Q-learning algorithm used to obtain a general optimal Q-table for an average patient. The second of the two algorithms (3) is the pseudo-code implementation of the 'on-policy' Q-learning algorithm used to adjust the general optimal Q-table based on the needs of the specific patient it is being used for. The second algorithm (3) is also the one used for testing and plotting of results achieved using Algorithm 2.

Algorithm 2 Q-learning Pseudo-Code (Training)

```
Patient ← Initialize with target state values (X_0)
t ← 0
while t < MaxIterations do
    Q_{t+1}, X_{t+1}, U_{t+1} ← qValUpdate(Q_t, Patient, U_t, α, γ, λ)
    t ← t + 1
    if X_{t+1}['BGL'] > 240 then
        Patient ← Initialize with target state values
    end if
end while
Controller ← Optimal Policy (save Q-table)
```

Algorithm 3 Q-learning Pseudo-Code (Simulation)

```
Patient ← Initialize with target state values (X_0)
Q ← General Optimal Q-table
t ← 0
while t < EndSimulation do
    Q_{t+1}, X_{t+1}, U_{t+1} ← SimulationUpdate(Q_t, Patient, U_t, α, γ, λ)
    t ← t + 1
end while
Plot Results (BGL and Insulin vs Time)
```

In the above pseudo-code implementation (Algorithm 3), we do not call the function $qValUpdate$, as in Algorithm 2. Instead, we call a function named $SimulationUpdate$. This function is nearly the same as $qValUpdate$ (which is shown in Algorithm 1), except that instead of using a random exploration process to choose the next action, the controller will always select the action corresponding to the highest reward. Therefore, the output U_{t+1} from $SimulationUpdate$ will always be equal to $\max U = \arg\max(Q[X_{t+1}, U] : U \in \mathbb{U})$. This is what makes Algorithm 3 an 'on-policy' implementation of Q-learning. The controller must learn about the physiological differences in a specific patient while always choosing to take what it believes to be the optimal action. This a much more ethical and safe implementation of Q-learning for an actual patient, since it can be very dangerous to knowingly administer the wrong dose of insulin for the sake of fine-tuning the controller's policy. There may be some loss in optimality with this 'on-policy' approach, since the controller is not able to

fully explore the environment and so not every state-action pair is reached infinitely many times, which is an assumption in the proof of the convergence of Q-learning [40]. It's therefore possible that at some state, there exists a dose of insulin that is indeed optimal for this specific patient, but that the controller never learns of this optimal dose at this state because it is a poor action to take for the general patient. There is, therefore, a trade-off between long-term optimality and short term safety, and based on the design criteria it was decided that short-term safety should not be sacrificed. It is also possible that the optimal dose for the general patient is actually quite dangerous for some specific patient. It is assumed in the scope of this project, however, that the optimal dose for the general patient is the 'best guess' of the optimal dose for any specific patient without any prerequisite patient-specific training.

4.2 State space analysis

The design solution was implemented and tested, while varying the measurement space, in order to fine-tune the model. In this way, a comprehensive understanding of the effect of including certain information to the controller on the performance was acquired. Recall that the information available to the controller must be of the form $Y_t = (\rho(G_{[t-N,t]}), U_{[t-N,t-1]}, \tau)$, for some $N \in \mathbb{N}, N < t$. G_t, U_{t-1}, may be omitted as well.

4.2.1 Periodic Meals

Firstly, the model was implemented and tested on a general patient who sticks to an extremely strict diet - eating the same meals for breakfast, lunch, and dinner, and at the same exact times, every day. Although type I diabetics do typically maintain a strict diet, this idealization is certainly an extreme, and so the merit of the results in this section are mainly to provide easily interpret-able results that guide the solution when randomized meals are introduced. Below the results are shown for the implementation of the model when only giving the controller access to the patients' current (quantized) BGL. The figure shows the BGL and the insulin of the patient over 300 time steps (5 mins each), with the insulin controlled using the optimal policy after training.

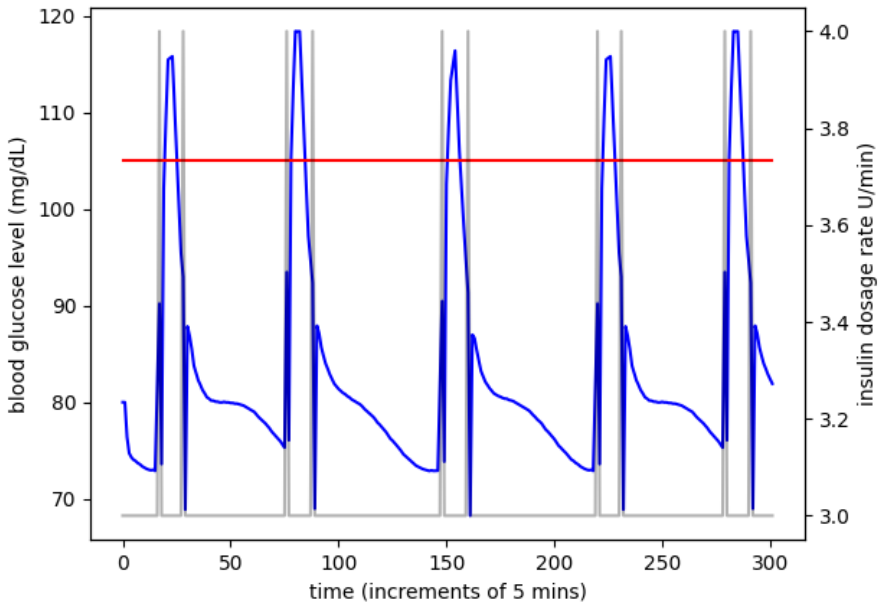

Figure 6: BGL (blue) and Insulin Administered (grey) vs time with $Y_t = \rho(G_t)$, Threshold BGL (red) overlayed.

There are 5 meals shown in this figure, and as can be seen clearly, the controller does not meet the requirements, as the BGL repeatedly spikes above the threshold level of 105 mg/dL each time a meal is eaten. This indicates that the current BGL is not enough information for the controller to accurately predict BGL spikes due to meals, and so it is not able to preemptively administer an increased dose around meal time. Q-learning is an effective algorithm to find an optimal **stationary** policy, which means that the policy does not change over time. In the case where the controller only has access to the BGL and no information about time or about meals, a stationary policy is unlikely to be effective because the controller must use the same policy whether it is meal time or not. It would be ideal for the controller to learn to recognize meals based purely on BGL, although it seems that it has failed to do so, as seen in Figure 6, and so it acts as if it is near the target value the majority of the time.

In order to improve upon this, it is crucial to adjust the state space so that the controller can capture information about time and meals so that the stationary policy achieved using Q-learning can mimic a non-stationary policy which changes over time. This can, equivalently, be thought of as adjusting the state-space so that the process has an invariant distribution. In order to maintain autonomy, it is ideal to design a model that can recognize meal times without user input, thus without including the time of meal into the measurement space. In order to achieve this, a finite window of history was included in the measurements available to the controller. The hope is that,

given information of previous BGL and previous insulin doses, the controller can learn to detect incoming meals based on the trend of recent data. This is also in accordance with theoretical research on Q-learning for POMDPs, as seen in [41], which proves near-optimal convergence of finite-history Q-learning for POMDPs, under certain filter stability conditions. It was then important to decide how much history to include, ie. what value of $N \in \mathbb{N}$, where $Y_t = (\rho(G_{[t-N,t]}), U_{[t-N,t-1]})$. In order to do this, the system was linearized around the target value of BGL, and the minimum number of time steps before a dose of insulin was seen in the BGL measurement was computed.

The blood-glucose-insulin model presented in 1 was linearized around the following steady state values:

- $G = 80$
- $I_R = 30$
- $I = 30$
- $S_1 = 17$
- $S_2 = 17$
- $G_g = 250$

and constructed as the generalized state-space representation shown in 5.

$$\begin{aligned} X_{t+1} &= AX_t + Bu_t \\ Y_t &= CX_t \end{aligned} \tag{5}$$

Parameter values for the average patient were used in the linearization. The matrix A is given by the Jacobian of X_{t+1}, evaluated at the steady-state values and the generalized patient parameters. The matrix C is a six-dimensional row vector with 1 as the first entry and 0 as the remaining entries, representative of the fact that state variable, G is the only observed element of the state. The resulting A and C matrices are given in Appendix C.

Then, expression 6 was evaluated for $k \in (0, 1, 2, 3, 4, 5, 6)$.

$$CA^k \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{6}$$

The six-dimensional column vector shown in expression 6 represents the fact that state variables I and S_1 are directly impacted by the control input. By the concept of observability, the first value of k resulting in a non-zero solution indicates how many time steps it takes for the impacts of the control input to appear in the measurement. It was determined that the first non-zero solution to expression 6 appears when $k = 2$. This result indicates that the controller must have access to at least two previous measurements and actions before any significant improvement can be seen in the controller's ability to predict meal disturbances.

Now, given the discussion above, the measurement space was altered to include the insulin administered in the previous two time steps. The measurement was then given by $Y_t = (\rho(G_t), U_{t-1}, U_{t-2})$. The results are shown in the following Figure:

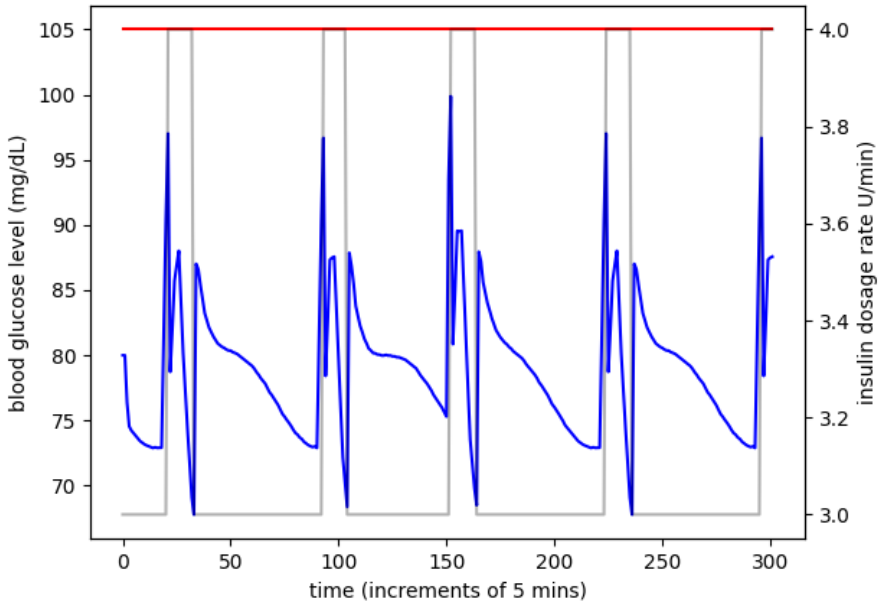

Figure 7: BGL (blue) and Insulin Administered (grey) vs time with $Y_t = (\rho(G_t), U_{t-1}, U_{t-2})$, Threshold BGL (red) overlayed

As can be seen in Figure 7, the performance of the controller does indeed meet specifications after expanding the measurement space. The BGL clearly stays below threshold the entire time, and it averages around the target value of 80. It should be noted, however, that this expansion of the measurement space does increase the complexity of the algorithm, as the memory requirement of the Q-table is n^2 times larger than before, where n is the number of elements in the action space. In this case, there are 10 elements in the action space, and so the Q-table is 100 times larger. This also means that the algorithm takes much longer to converge, since Q-learning requires each

state-action pair to be visited infinitely often, and expanding the measurement space means there are more state-action pairs to visit.
Also, it should be noted that, while this controller does meet specifications when the patient eats the same meals at the same time everyday, this is not sufficient to meet the design criteria of the project, as the controller must be robust to randomized meals, since it is inconvenient and unrealistic for a patient to stick to a diet this strict.

Now, given that the performance of the controller improved when history of actions was introduced, it was important to test whether even more information could improve the performance further. To this end, the BGL from the previous time step was also included into the measurement space. Therefore, the new measurements were given by $Y_t = (\rho(G_t), \rho(G_{t-1}), U_{t-1}, U_{t-2})$. The results can be seen in the following Figure:

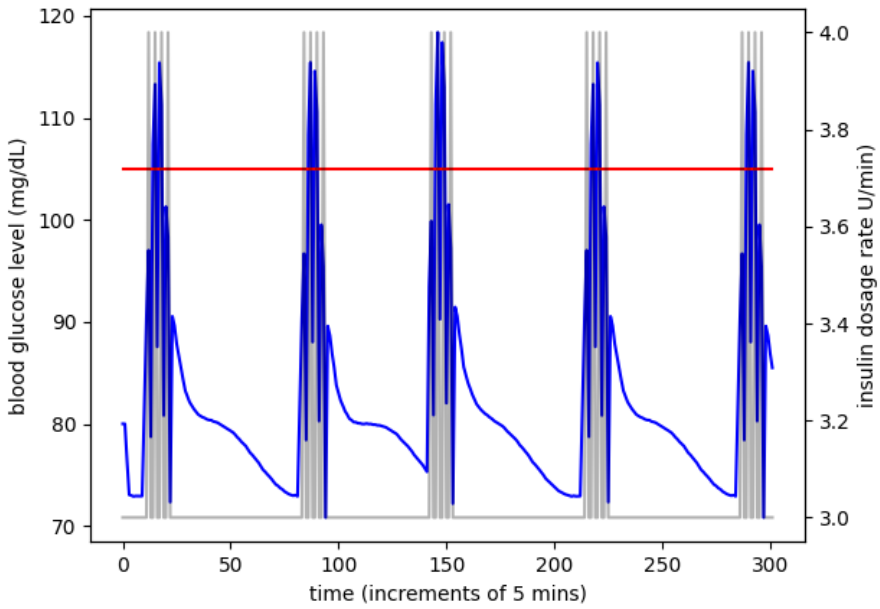

Figure 8: BGL (blue) and Insulin Administered (grey) vs time with $Y_t = (\rho(G_t), \rho(G_{t-1}), U_{t-1}, U_{t-2})$, Threshold BGL (red) overlayed

As can be seen in Figure 8, the performance of the controller actually worsened. This is because the added memory complexity of the Q-table became too much for the algorithm to hit each state-action pair enough to converge in a reasonable amount of time. Here, the state space for the BGL contains 50 quantization bins in the range $[0, 240]$, and so the memory requirement for the Q-table is $5,000$ times larger than the requirement when only the current BGL is used. The quantization was decreased as low as 20, without much improvement. Therefore, since not every state-action pair is being visited infinitely often, there are some state-action pairs which do not have

accurate Q-values, and so in this case the controller administers the default dose of 3 U/min (this default was decided upon based on learned results, see Figure 6). This behaviour is seen during meal spikes in Figure 8, in which the insulin repeatedly drops down to the default throughout the duration of the spike. Upon further examination, it seems that the *intended* policy of the controller in Figure 8 is actually the same as that seen in Figure 7, in which only the two past actions were included in the measurement space. This implies that there was no obvious performance gain to expanding the measurement space further, and so due to the increased memory complexity, it was decided that there was no benefit to expanding the measurement space further.

4.2.2 Randomized Meals

Now that a controller was designed that performed well when implemented in a general patient which ate the same meals at the same time everyday, it was necessary to implement the controller with randomized meals, and to improve it where needed. Recall Section 2.3.2 where the randomization of meals is described, and Algorithm 1 for the implementation of randomized meals within the software environment.

The performance of the controller which has access to the measurement $Y_t = (\rho(G_t), U_{t-1}, U_{t-2})$, when implemented with randomized meals can be seen in the following Figure:

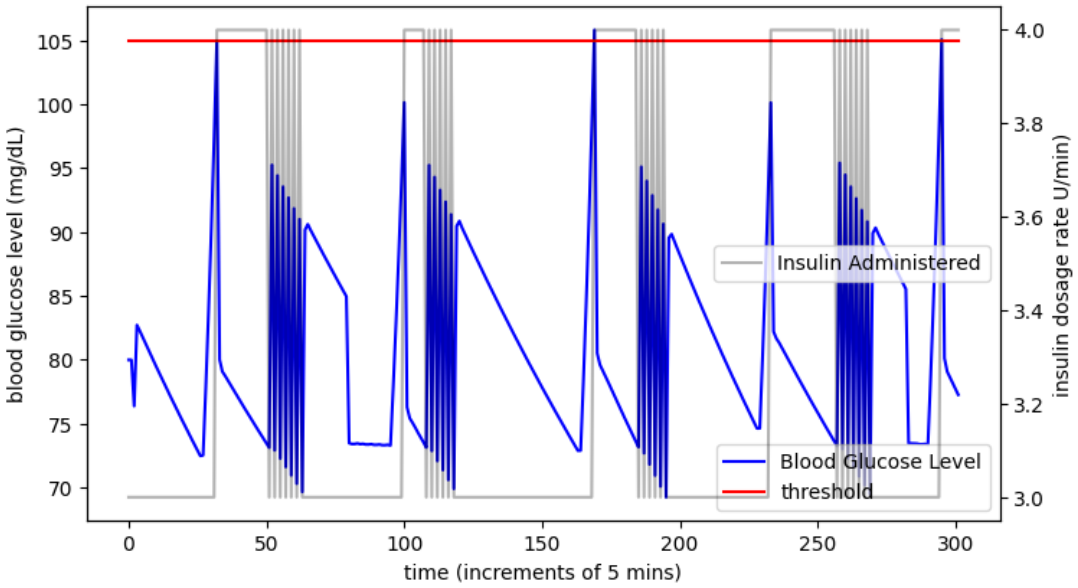

Figure 9: BGL (blue) and Insulin Administered (grey) vs time with random meals and with $Y_t = (\rho(G_t), U_{t-1}, U_{t-2})$, Threshold BGL (red) overlayed

As can be seen in Figure 9, the controller no longer meets specifications, as the BGL spikes above the threshold value during some meals. This indicates that the desgined controller is not robust enough to handle differences in meal times and carbohydrate amounts.

In order to improve the performance of the controller when implemented with randomized meals, the measurement space was again altered. In this final iteration, the measurement space was changed to only include the time since the last meal was consumed. Therefore, the measurement is given by $Y_t = \tau$. The idea is that all of the relevant information that the controller gains by observing the BGL and previous actions is to give it a better estimate of when a meal is being consumed. By simply including the time since last meal in the measurement, the controller would have a more accurate estimate of when meals would / have been consumed, and it greatly decreases the memory requirement of the Q-table. Due to this decrease in the size of the Q-table from the condensed measurement space, it was then feasible to increase the granulation of the action space within the range [0,10]. Therefore, the action space was granulated to contain 50 possible doses in the range [0,10], where only 10 possible doses were included previously. The results of this simulation can be seen in the following Figure:

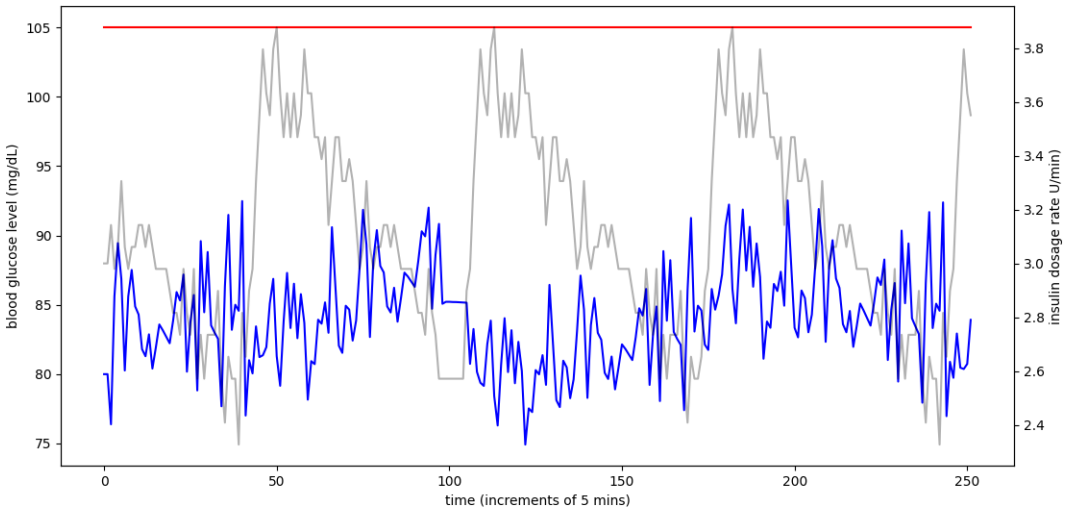

Figure 10: BGL (blue) and Insulin Administered (grey) vs time with random meals and with $Y_t = \tau$, Threshold BGL (red) overlayed

As can be seen in Figure 10, the performance of this controller is very good. The BGL never exceeds the threshold, and it stays very close to the target value at all times. There is, however, a trade-off in autonomy required to meet these performance

specifications, since in this case the patient would be required to manually signal to the controller that a meal is being consumed. This introduces risk since the controller is susceptible to human error due to incorrect meal signalling, and it may be inconvenient to some patients.

4.2.3 On-Policy Learning for Individual Patients

Now, recall the broader goal of the project, to design a controller that can be initialized in an individual patient with a general optimal Q-table, and then learn to adjust its' policy to achieve near-optimal performance by using an on-policy version of Q-learning. See the system architecture 5, and the Algorithm 3 for the design and implementation of this feature.
In order to test this feature, a patient was created that is physiologically different than the patient that the general Q-table was trained on. This physiological difference is reflected in the insulin resistance of the patient. The Bergman parameter p_2 (see (1)) was decreased in order to model a patient that is less sensitive to insulin than the typical patient. This patient was then initialized with a controller that utilized the general optimal Q-table as obtained from training with the off-policy Q-learning algorithm. Using an on-policy version of Q-learning, the controller was able to learn of this difference in insulin resistance, and subsequently stabilize the patient back around the target BGL. The results can be seen in the following Figure:

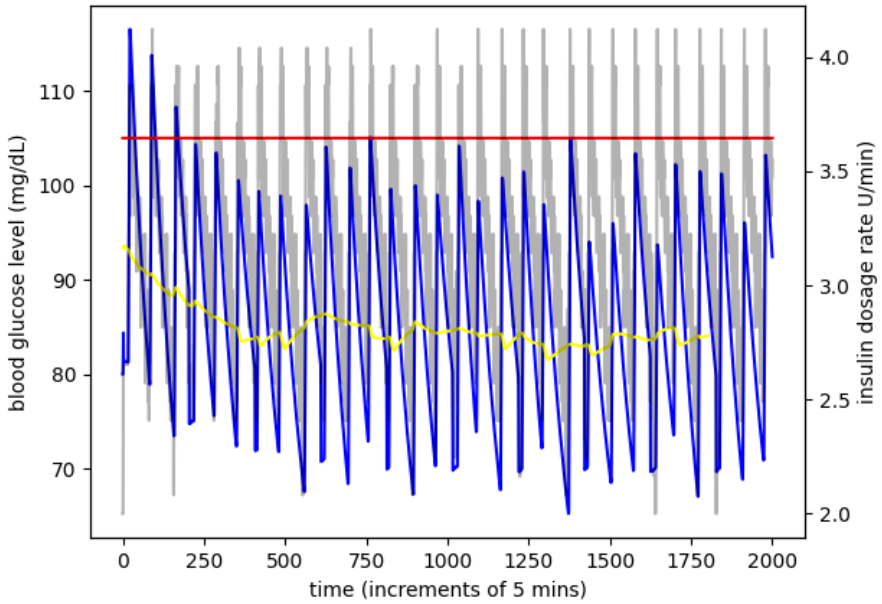

Figure 11: On-Policy Learning for Patient with Increased Insulin Resistance, Initialized with General Optimal Q-table. BGL (blue) and Insulin Administered (grey) vs time with random meals and with $Y_t = \tau$, Threshold BGL (red), moving avg. of BGL (yellow) overlayed

As can be seen in Figure 11, the patient's BGL is initially too high, since the administered insulin is not as effective in lowering the patient's BGL. The controller is able to effectively stabilize the patient around the target value, as can be seen by the moving average (in yellow) approaching 80 mg/dL. Note that the performance of the controller in Figure 11 is not as good as in Figure 10. This is because this version of the Q-learning algorithm does not use a random exploration policy, and so the Markov Chain induced by the controller's policy is not irreducible and recurrent, meaning not every state-action pair is visited infinitely often.

5 Design Evaluation

5.1 Performance Against Design Criteria

The final solution (seen in Figure 10) was evaluated against the design criteria from Section 2.4.4 and given a final evaluation score of 10/11, as seen in the table below:

Criteria	**Metric Met (Y/N)**
Maintained BGL Within Target Range	Y
Avoids Hypoglycemic Episodes	Y
Avoids Hyperglycemic Episodes	Y
Averages Around Optimal BGL	Y
Robust to Random Meal Times	Y
Robust to Random Meal Amounts	Y
Robust to Random Snack Times	Y
Robust to Random Snack Amounts	Y
Fully Automated System	N
Adaptability to Individual Patient Dynamics	Y
Efficiently Converges to Optimal Dosages	Y
Overall Score	10/11
Average Reward	20.03695

Table 1: Evaluation of the Final Solution Against the Design Criteria.

An individuals health was the primary consideration when evaluating potential solutions and determining the final solution. Consistently, the final solution stayed well within the safe ranges for blood glucose levels achieving a minimum of 75mg/dL and a maximum of 95mg/dL. These metrics were achieved while simulating randomized meals, meal times, snacks, and snack times. An individual is able to go throughout their day with the freedom to eat at different times and different amounts without maintaining a specific regime that is required in most modern methods. However, one of the goals was to ensure that the system could perform insulin administration automatically without requiring any user interference. When times of the meals were not included in the state space, the model performed effectively but was still prone to reaching BGL's outside the target range. Though the levels reached are still generally safe for the patient, continued occurrences of these can lead to long term health complications. It was determined that the design requirement for a fully automated system could be forfeited to ensure patients safety throughout the day. The design criteria evaluation for the fully automated system can be seen in Table 2, which also received a 10/11 based on the design metrics (see below).

Criteria	**Metric Met (Y/N)**
Maintained BGL Within Target Range	N
Avoids Hypoglycemic Episodes	Y
Avoids Hyperglycemic Episodes	Y
Averages Around Optimal BGL	Y
Robust to Random Meal Times	Y
Robust to Random Meal Amounts	Y
Robust to Random Snack Times	Y
Robust to Random Snack Amounts	Y
Fully Automated System	Y
Adaptability to Individual Patient Dynamics	Y
Efficiently Converges to Optimal Dosages	Y
Overall Score	10/11
Average Reward	-243.1582

Table 2: Evaluation of the Fully Automated Solution Against the Design Criteria.

Using the reward function (Equation 3), it was found that the average reward was 20.03695 for the final solution. This is a substantial improvement over the fully automated system which had an average reward of -243.1582. A fully automated system also required a larger Q-table, this meant a longer period of time would be required for the table to adapt to an individual's parameters. During the adaptation phase, a patient is more prone to reaching BGLs outside the target range. For the final solution, the personalized table that was generated from the generalized Q-table exhibited rapid convergence based on the new physiological parameters. Within 31.25 hours, the table established a near-optimal dosage regime for the patient. The efficient convergence of the solution ensures that the On-Policy Learning will adapt as the individual's parameters shift each day. During the convergence, it should be noted that the first 3 meals resulted in BGLs above 105. These levels are safe for the patient and only become dangerous over extended periods of time.

5.2 Limitations

While the design satisfied the majority of the design criteria, there are a number of limitations of the current model. The main limitation is the requirement to signal the system when a meal is being consumed. This prevents the system from being fully automated. The simulated blood glucose values used in this report are from a simplified model that does not take into consideration many of the factors that can affect an individual's BGLs and their physiological parameters. Factors that should be included in future tests to ensure the system is robust to BGL changes are exercise, sleep, stress, medications, alcohol consumption, fluid intake, and illnesses. These can

result in the body's physiological response to insulin and blood glucose changing from their normal parameters. Prior to implementation, these factors should be included in the simulations to ensure the safety of the individual. These limitations are further explored in Section 7.

6 Implication of Results

6.1 Patient Implications

Concluding that Q-Learning can be used as an effective tool for determining the optimal insulin dosages for an individual has wide-ranging implications. Results suggest an improved method of delivering continuous insulin infusion without the occurrences of hyperglycemia or hypoglycemia that many modern methods of insulin delivery currently struggle to avoid. Increased precision in insulin dosages results in a lower variance of blood glucose levels preventing glycemic variability from disrupting the normal physiological processes of the human body which can lead to many long-term complications.

The proposed model suggests enhanced patient convenience over current methods that often require manual calculations and inputs. These inconveniences often disrupt an individual's ability to enjoy a normal lifestyle without the constant threat of diabetes weighing on their minds. The proposed model requires only that the user input the time of meals. This could be achieved by using a button on the device to notify the system of a meal. As seen in the results, no action is required to notify the system of smaller snacks. Similarly, the individual is able to eat a wider range of meals and is no longer restricted to a prescribed routine.

Providing personalized treatment to individuals is important given the complex dynamics of the human body. The model has been shown to adapt efficiently from the general Q-Table to the personalized table. This signifies an important element in designing a system that can be universal to a wide range of patients without requiring significant manual modifications. Patient dynamics can change day by day as previously discussed. Providing a model that can adapt to these changes ensures that individuals need not be concerned by changes in their lifestyle.

6.2 Implications on Health Care Ethics

Implementing an automatic insulin infusion pump requires an understanding of the key regulations and guidelines in Canada. The system is classified as a Class III medical given the severe complications that can occur if incorrect dosages are administered [42]. "Good Manufacturing practices guide for drug products" must be followed to ensure

the quality and manufacturing are completed in accordance with established guidelines within Canada [43]. This ensures that the chance of defects and complications is minimized, prioritizing the individual's safety.

6.3 Implications on Engineering Ethics

As stated in the Professional Engineers Act, R.S.O. 1990, c. P.28 Section 77-2-i, "A practitioner shall, regard the practitioner's duty to public welfare as paramount" [44]. All decisions should be made with patient safety, regulatory compliance, and patient privacy as paramount considerations.

7 Future Work

7.1 Impact of Sleep on Blood Glucose

Sleep is a critical element in maintaining a healthy body. However, many challenges arise as a result of sleep for people with diabetes. The first concern that should be further analyzed is the impact of the dawn phenomenon. This refers to a rise in blood glucose levels that occurs around 3:00 am to 8:00 am which can result in hyperglycemia. The phenomenon occurs in approximately 50% of people with diabetes [45]. During this period, average blood glucose levels can rise by approximately 10-20mg/dL [46]. It has been shown in previous studies that continuous insulin infusion is more effective at controlling glucose levels during the dawn phenomenon [46]. Incorporating the phenomenon in future models will be important in ensuring the safety of patients by preventing hypoglycemic events during sleep.

Second, it has been found that people with type I diabetes struggle to maintain a regular and sufficient amount of sleep [47]. This is a result of an increase in occurrences of sleep apnea and shorter sleep duration caused by irregularities in blood pressure. Getting less than 7 hours of sleep per night has been shown to increase an individual's insulin resistance [48]. The effectiveness of the generated Q-table for any individual is then impacted as it will no longer provide the optimal dose as the patient's physiological dynamics have changed. Implementation of a method to predict the change in insulin resistance as a result of sleep deprivation would improve the robustness of any implemented system. However, an effective method for measuring the change in insulin resistance relative to the amount of sleep an individual gets has not yet been found.

Finally, type I diabetes has been shown to result in a decreased awakening response to hypoglycemia [47]. This poses a significant danger to individuals as it prevents

them from managing their blood glucose levels prior to reaching dangerous levels. As a result, further testing should be done to ensure that the system does not administer too much insulin during sleep. Another possibility would be to use a duel-hormone system instead of an insulin infusion system. This will be discussed further in Section 7.4.

7.2 Impact of Exercise on Blood Glucose

Understanding the changes in blood glucose levels during and after exercises is important in ensuring the model can appropriately manage an individual's levels. During exercise, muscles in the individual's body utilize a greater-than-average amount of glucose in order to contract. It has been found that consistently working out can result in a decrease in average blood glucose levels by approximately 22mg/dL [49]. Similarly, while sleeping after a workout, the body experiences muscle hypertrophy which consumes a larger supply of glucose than usual. This can potentially lead to the individual experiencing hypoglycemia. However, some individuals bodies react differently and experience the opposite effect where there blood glucose levels increase as a result of stress hormones being released [50]. Given the variety of effects individuals might experience as a result of exercise, further research should be done to ensure the system is robust to the effects of exercising. The measurements of the system should be altered to reflect this, for example by including a heart-rate sensor into the system. Implementation of a duel-hormone system should also be considered as a potential solution to this concern. This is discussed in further detail in Section 7.4.

7.3 Impact of Stress on Blood Glucose Levels

Further research should be done to ensure the system is robust to changing levels of stress within the individual. During times of stress the body produces more cortisol, glucagon, epinephrine, and nor-epinephrine which result in increasing blood glucose levels for the patient [51]. A simplistic mathematical model on the impact of stress on blood glucose levels that can be used to test the system can be found in "Rudimentary Model of Glucose Response to Stress" by Nasha Rios-Guzman [52]. This model should only be used as a baseline for further research as it does not encompass the full scope of the impacts stress has on the human body.

7.4 Implementation of a Duel-Hormone System

One of the biggest limitations of the proposed system is that it only allows for insulin infusion. As a result, the system struggles to deal with hypoglycemic episodes without the individuals interference. Further research should be done to test the effectiveness of Q-Learning with a duel-hormone system (DHS) instead of a simple single-hormone

system. A DHS is an artificial pancreas that has been researched for type I diabetes and has shown promising results for helping prevent hypoglycemia [53]. It is able to infuse both insulin and glucagon. Where insulin is administered to help lower blood glucose levels, glucagon is instead used to raise blood glucose levels. This would be especially effective in dealing with hypoglycemia while sleeping since the individual is unable to manually raise there blood glucose levels with food without waking up.

7.5 Other Medical Implementations

This paper has shown promising results in the use of Q-Learning for finding optimal insulin dosages for diabetic patients. Further research should be done into other medical conditions that would benefit from using Q-Learning to find the optimal dosages. The following conditions use continuous infusion pumps that could benefit from using Q-Learning to find optimal dosages: Vasopressors (blood pressure control) [54], Sedatives (central nervous system depressants), and Antibiotics (treat bacteria infections). Research has been done to test the viability of reinforcement learning for finding optimal dosages for these conditions but insufficient research has been done using a Q-Learning approach [54].

References

[1] A. Sapra and P. Bhandari, "Diabetes mellitus," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK551501/.

[2] WHO. "Diabetes." (2023), [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[3] L. A. DiMeglio, C. Evans-Molina, and R. A. Oram, "Type 1 diabetes," *Lancet (London, England)*, vol. 391, no. 10138, pp. 2449–2462, Jun. 16, 2018, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(18)31320-5. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6661119/.

[4] CDC. "What is type 1 diabetes?" Centers for Disease Control and Prevention. (Mar. 11, 2022), [Online]. Available: https://www.cdc.gov/diabetes/basics/what-is-type-1-diabetes.html.

[5] M. Wolosowicz, B. Lukaszuk, and A. Chabowski, "The causes of insulin resistance in type 1 diabetes mellitus: Is there a place for quaternary prevention?" *International Journal of Environmental Research and Public Health*, vol. 17, no. 22, p. 8651, Nov. 2020, ISSN: 1661-7827. DOI: 10.3390/ijerph17228651. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7700208/.

[6] ADA. "Insulin resistance and type 1 diabetes — ADA." (2023), [Online]. Available: https://diabetes.org/diabetes/type-1-diabetes/can-you-have-insulin-resistance-and-type-1-diabetes.

[7] S. Templer, "Closed-loop insulin delivery systems: Past, present, and future directions," *Frontiers in Endocrinology*, vol. 13, p. 919 942, Jun. 6, 2022, ISSN: 1664-2392. DOI: 10.3389/fendo.2022.919942. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9207329/.

[8] S. Kalra, J. J. Mukherjee, S. Venkataraman, *et al.*, "Hypoglycemia: The neglected complication," *Indian Journal of Endocrinology and Metabolism*, vol. 17, no. 5, pp. 819–834, 2013, ISSN: 2230-8210. DOI: 10.4103/2230-8210.117219. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3784865/.

[9] M. Shifrin and H. Siegelmann, "Near-optimal insulin treatment for diabetes patients: A machine learning approach," *Artificial Intelligence in Medicine*, vol. 107, p. 101 917, Jul. 1, 2020, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2020.101917. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365719305640.

[10] M. Mouri and M. Badireddy, "Hyperglycemia," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK430900/.

[11] CDC. "4 ways to take insulin," Centers for Disease Control and Prevention. (Mar. 25, 2021), [Online]. Available: https://www.cdc.gov/diabetes/basics/type-1-4-ways-to-take-insulin.html.

[12] A. G. Tsadik, M. T. Gidey, B. T. Assefa, *et al.*, "Insulin injection practices among youngsters with diabetes in tikur anbesa specialized hospital, ethiopia," *Journal of Diabetes and Metabolic Disorders*, vol. 19, no. 2, pp. 805–812, Jun. 16, 2020, ISSN: 2251-6581. DOI: 10.1007/s40200-020-00565-6. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7843670/.

[13] NIDDK. "Continuous glucose monitoring - NIDDK," National Institute of Diabetes and Digestive and Kidney Diseases. (Jun. 2017), [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/managing-diabetes/continuous-glucose-monitoring.

[14] "Blood sugar testing: Why, when and how," Mayo Clinic. (2022), [Online]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/in-depth/blood-sugar/art-20046628.

[15] S. K. Garg, S. A. Weinzimer, W. V. Tamborlane, *et al.*, "Glucose outcomes with the in-home use of a hybrid closed-loop insulin delivery system in adolescents and adults with type 1 diabetes," *Diabetes Technology & Therapeutics*, vol. 19, no. 3, pp. 155–163, Mar. 1, 2017, ISSN: 1520-9156. DOI: 10.1089/dia.2016.0421. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5359676/.

[16] G. P. Forlenza, O. Pinhas-Hamiel, D. R. Liljenquist, *et al.*, "Safety evaluation of the MiniMed 670g system in children 7–13 years of age with type 1 diabetes," *Diabetes Technology & Therapeutics*, vol. 21, no. 1, pp. 11–19, Jan. 1, 2019, ISSN: 1520-9156. DOI: 10.1089/dia.2018.0264. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350071/.

[17] C. K. Boughton and R. Hovorka, "Automated insulin delivery in adults," *Endocrinology and metabolism clinics of North America*, vol. 49, no. 1, pp. 167–178, Mar. 2020, ISSN: 0889-8529. DOI: 10.1016/j.ecl.2019.10.007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6986367/.

[18] A. Weisman, J.-W. Bai, M. Cardinez, C. K. Kramer, and B. A. Perkins, "Effect of artificial pancreas systems on glycaemic control in patients with type 1 diabetes: A systematic review and meta-analysis of outpatient randomised controlled trials," *The Lancet Diabetes & Endocrinology*, vol. 5, no. 7, pp. 501–512, Jul. 1, 2017, Publisher: Elsevier, ISSN: 2213-8587, 2213-8595. DOI: 10.1016/S2213-8587(17)30167-5. [Online]. Available: https://www.thelancet.com/journals/landia/article/PIIS2213-8587(17)30167-5/fulltext.

[19] G. P. Forlenza, F. M. Cameron, T. T. Ly, *et al.*, "Fully closed-loop multiple model probabilistic predictive controller artificial pancreas performance in adolescents and adults in a supervised hotel setting," *Diabetes Technology & Therapeutics*, vol. 20, no. 5, pp. 335–343, May 1, 2018, ISSN: 1520-9156. DOI: 10.1089/dia.2017.0424. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5963546/.

[20] S. Samadi, M. Rashid, K. Turksoy, *et al.*, "Automatic detection and estimation of unannounced meals for multivariable artificial pancreas system," *Diabetes Technology & Therapeutics*, vol. 20, no. 3, pp. 235–246, Mar. 1, 2018, ISSN: 1520-9156. DOI: 10.1089/dia.2017.0364. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5867514/.

[21] C. Mosquera-Lopez, L. M. Wilson, J. El Youssef, *et al.*, "Enabling fully automated insulin delivery through meal detection and size estimation using artificial intelligence," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–7, Mar. 13, 2023, Number: 1 Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-023-00783-1. [Online]. Available: https://www.nature.com/articles/s41746-023-00783-1.

[22] T. Namikawa, M. Munekage, T. Yatabe, H. Kitagawa, and K. Hanazaki, "Current status and issues of the artificial pancreas: Abridged english translation of a special issue in japanese," *Journal of Artificial Organs: The Official Journal of the Japanese Society for Artificial Organs*, vol. 21, no. 2, pp. 132–137, Jun. 2018, ISSN: 1619-0904. DOI: 10.1007/s10047-018-1019-4.

[23] C. Marling and R. Bunescu, "The OhioT1dm dataset for blood glucose level prediction: Update 2020," *CEUR workshop proceedings*, vol. 2675, pp. 71–74, Sep. 2020, ISSN: 1613-0073. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7881904/.

[24] "NIDDK central repository homepage." (2023), [Online]. Available: https://repository.niddk.nih.gov/home/.

[25] E. D. Lehmann, "AIDA–a computer-based interactive educational diabetes simulator," *The Diabetes Educator*, vol. 24, no. 3, pp. 341–346, 348, 1998, ISSN: 0145-7217. DOI: 10.1177/014572179802400309.

[26] B. P. Kovatchev, M. Breton, C. D. 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 (Online)*, vol. 3, no. 1, pp. 44–55, Jan. 2009, ISSN: 1932-2968. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2681269/.

[27] S. Lynch and B. Bequette, "Estimation-based model predictive control of blood glucose in type i diabetics: A simulation study," in *Proceedings of the IEEE 27th Annual Northeast Bioengineering Conference (Cat. No.01CH37201)*, Apr. 2001, pp. 79–80. DOI: 10.1109/NEBC.2001.924729.

[28] M. Oroojeni Mohammad Javad, S. O. Agboola, K. Jethwani, A. Zeid, and S. Kamarthi, "A reinforcement learning–based method for management of type 1 diabetes: Exploratory study," *JMIR Diabetes*, vol. 4, no. 3, e12905, Aug. 28, 2019, ISSN: 2371-4379. DOI: 10.2196/12905. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6737889/.

[29] C. Shi, S. Luo, Y. Le, H. Zhu, and R. Song, *Statistically efficient advantage learning for offline reinforcement learning in infinite horizons*, Jul. 26, 2022. DOI: 10.48550/arXiv.2202.13163. arXiv: 2202.13163[cs,stat]. [Online]. Available: http://arxiv.org/abs/2202.13163.

[30] H. Emerson, M. Guy, and R. McConville, *Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes*, version: 1, Apr. 7, 2022. DOI: 10.48550/arXiv.2204.03376. arXiv: 2204.03376[cs]. [Online]. Available: http://arxiv.org/abs/2204.03376.

[31] H. Canada. "Good machine learning practice for medical device development: Guiding principles." Last Modified: 2021-10-26. (Oct. 27, 2021), [Online]. Available: https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/good-machine-learning-practice-medical-device-development.html.

[32] "Deaths and age-specific mortality rates, by selected grouped causes." (2022), [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039201.

[33] CDC. "Low blood sugar (hypoglycemia)," Centers for Disease Control and Prevention. (Mar. 25, 2021), [Online]. Available: https://www.cdc.gov/diabetes/basics/low-blood-sugar.html.

[34] "Type 1 treatment," DiabetesCanadaWebsite. (2023), [Online]. Available: https://www.diabetes.ca/about-diabetes/type-1/treatment.

[35] A. Pfützner, P. B. Musholt, B. Malmgren-Hansen, N. H. Nilsson, and T. Forst, "Analysis of the environmental impact of insulin infusion sets based on loss of resources with waste," *Journal of Diabetes Science and Technology*, vol. 5, no. 4, pp. 843–847, Jul. 1, 2011, ISSN: 1932-2968. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3192587/.

[36] T. Schneider, T. Gomes, K. N. Hayes, K. J. Suda, and M. Tadrous, "Comparisons of insulin spending and price between canada and the united states," *Mayo Clinic Proceedings*, vol. 97, no. 3, pp. 573–578, Mar. 2022, ISSN: 1942-5546. DOI: 10.1016/j.mayocp.2021.11.028.

[37] R. Gillis, C. C. Palerm, H. Zisser, L. Jovanovic, D. E. Seborg, and F. J. Doyle, "Glucose estimation and prediction through meal responses using ambulatory subject data for advisory mode model predictive control," *Journal of Diabetes Science and Technology*, vol. 1, no. 6, pp. 825–833, Nov. 1, 2007, Publisher: SAGE

Publications Inc, ISSN: 1932-2968. DOI: 10.1177/193229680700100605. [Online]. Available: https://doi.org/10.1177/193229680700100605.

[38] CDC. "The insulin resistance–diabetes connection," Centers for Disease Control and Prevention. (Jun. 20, 2022), [Online]. Available: https://www.cdc.gov/diabetes/basics/insulin-resistance.html.

[39] A. D. Kara, N. Saldi, and S. Yüksel, *Q-learning for mdps with general spaces: Convergence and near optimality via quantization under weak continuity*, 2021. arXiv: 2111.06781 [cs.LG].

[40] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[41] A. D. Kara and S. Yüksel, "Convergence of finite memory q-learning for pomdps and near optimality of learned policies under filter stability," *CoRR*, vol. abs/2103.12158, 2021. arXiv: 2103.12158. [Online]. Available: https://arxiv.org/abs/2103.12158.

[42] H. Canada. "Notice: Changes to evidence requirements for class II and class III infusion pumps." Last Modified: 2021-03-18. (Aug. 13, 2018), [Online]. Available: https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/activities/announcements/notice-evidence-infusion-pumps.html.

[43] H. Canada. "Good manufacturing practices guide for drug products (GUI-0001)." Last Modified: 2022-11-07. (Feb. 28, 2018), [Online]. Available: https://www.canada.ca/en/health-canada/services/drugs-health-products/compliance-enforcement/good-manufacturing-practices/guidance-documents/gmp-guidelines-0001/document.html.

[44] "Law document english view," Ontario.ca. (Jul. 24, 2014), [Online]. Available: https://www.ontario.ca/laws/view.

[45] "Dawn phenomenon," Cleveland Clinic. (2023), [Online]. Available: https://my.clevelandclinic.org/health/diseases/24553-dawn-phenomenon.

[46] T. B. O'Neal and E. E. Luther, "Dawn phenomenon," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK430893/.

[47] S. S. Farabi, "Type 1 diabetes and sleep," *Diabetes Spectrum : A Publication of the American Diabetes Association*, vol. 29, no. 1, pp. 10–13, Feb. 2016, ISSN: 1040-9165. DOI: 10.2337/diaspect.29.1.10. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4755454/.

[48] CDC. "Sleep for a good cause," Centers for Disease Control and Prevention. (Jul. 28, 2022), [Online]. Available: https://www.cdc.gov/diabetes/library/features/diabetes-sleep.html.

[49] O. P. Adams, "The impact of brief high-intensity exercise on blood glucose levels," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. 6, pp. 113–122, Feb. 27, 2013, ISSN: 1178-7007. DOI: 10.2147/DMSO.S29222. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3587394/.

[50] N. S. Tyler, C. Mosquera-Lopez, G. M. Young, J. El Youssef, J. R. Castle, and P. G. Jacobs, "Quantifying the impact of physical activity on future glucose trends using machine learning," *iScience*, vol. 25, no. 3, p. 103 888, Feb. 8, 2022, ISSN: 2589-0042. DOI: 10.1016/j.isci.2022.103888. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8889374/.

[51] R. S. Surwit, M. S. Schneider, and M. N. Feinglos, "Stress and diabetes mellitus," *Diabetes Care*, vol. 15, no. 10, pp. 1413–1422, Oct. 1, 1992, ISSN: 0149-5992. DOI: 10.2337/diacare.15.10.1413. [Online]. Available: https://doi.org/10.2337/diacare.15.10.1413.

[52] University of South Florida and N. Rios-Guzman, "Rudimentary model of glucose response to stress," *Undergraduate Journal of Mathematical Modeling: One + Two*, vol. 8, no. 1, Sep. 2017, ISSN: 2326-3652. DOI: 10.5038/2326-3652.8.1.4887. [Online]. Available: http://scholarcommons.usf.edu/ujmm/vol8/iss1/6/.

[53] M. Infante, D. A. Baidal, M. R. Rickels, *et al.*, "Dual-hormone artificial pancreas for management of type 1 diabetes: Recent progress and future directions," *Artificial organs*, vol. 45, no. 9, pp. 968–986, Sep. 2021, ISSN: 0160-564X. DOI: 10.1111/aor.14023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9059950/.

[54] M. Mollura, C. Drudi, L.-W. Lehman, and R. Barbieri, "A reinforcement learning application for optimal fluid and vasopressor interventions in septic ICU patients," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2022, pp. 321–324, Jul. 2022, ISSN: 2694-0604. DOI: 10.1109/EMBC48229.2022.9871055.

[55] S. Yüksel, "Optimization and control of stochastic systems," *Queen's University Department of Mathematics and Statistics*, Jan. 2023.

Appendix

A The Bergman Minimal Model (BMM)

The Bergman Minimal Model (BMM) discussed in Section 1.4.1 is given below:

$$\begin{aligned} \frac{dG}{dt} &= -P_1 G - X(G + G_b) + D(t) \\ \frac{dX}{dt} &= -P_2 X + P_3 I \\ \frac{dI}{dt} &= -n(I + I_b) + \frac{U(t)}{V_1} \end{aligned} \tag{7}$$

where:

- G = blood-glucose concentration
- X = insulin in the remote compartment
- I = blood-insulin concentration
- $D(t)$ = meal glucose disturbance
- $U(t)$ = exogenous insulin infusion rate
- G_b = basal blood-glucose concentration
- I_b = basal blood-insulin concentration

and (P_1, P_2, P_3, V_1, n) are patient-specific parameters for a type I diabetic [27].

B Stochastic Control and Formulation of POMDP

B.1 Definitions

[55] Definition: Partially Observed Markov Decision Process (POMDP) A partially Observed Markov Decision Process is a stochastic control problem of the form:

$$x_{t+1} = f(x_t, u_t, \omega_t)$$

$$y_t = g(x_t, \nu_t)$$

Where x_t is the $\mathbb{X}$-valued state, u_t is the $\mathbb{U}$-valued control action, and y_t is the $\mathbb{Y}$-valued measurement process. The terms (ω_t, ν_t) are i.i.d noise processes and they are independent of each other. The transition kernel τ, where:

$$\tau(A|x, u) = P(x_{t+1} \in A | x_t = x, u_t = u)$$

is the conditional probability measure induced by f.

[55] Definition: Deterministic Admissible Control Policy A Deterministic Admissible Control Policy, γ, is a sequence of functions $\gamma = (\gamma_0, \gamma_1, \cdots, \gamma_t, \cdots)$, such that $u_t = \gamma_t(y_{[0,t]}, u_{[0,t-1]}) = \gamma_t(I_t)$.

[55] Definition: Deterministic Stationary Policy A Deterministic Stationary Policy is a policy that does not change with time, ie. a policy of the form $\gamma = (\gamma, \gamma, \cdots)$ $= (f, f, \cdots)$ for some measurable f. Here $u_t = f(I_t), \forall t \in \mathbb{N}$.

B.2 Formulation of POMDP

Recall the description of the various spaces and variables of the POMDP of the problem in section 3.1. In light of the definition above, the following is the representation of the glucose-insulin problem:

$$X_{t+1} = f(X_t, U_t, M_t)$$

$$Y_t = g(X_t)$$

Where M_t are the meals eaten by the patient, which is a random process. f is the dynamics of the system, given by 1, and $g(X_t) = \rho(I_t)$, where I_t is a function of X_t. The solution to this problem is to find an optimal policy γ which satisfies:

$$J_\beta(\gamma, X_0) = max_{\gamma \in \Gamma_A} E[\sum_{t=0}^{\infty} \beta^t R(X_t)]$$

where R is the reward function defined in section 3.1, and where X_t is affected by the policy through f, where $U_t = \gamma(Y_t)$. This is called a discounted optimal control problem.

C Linearization about Steady State

The A and C matrices obtained in the linearization discussed in Section 4.2.1 are given below:

$$\mathrm{A} = \begin{bmatrix} -0.9017 & -2.3200 & 0 & 0 & 0 & 0.0008 \\ 0 & -0.0123 & 0.0123 & 0 & 0 & 0 \\ 0 & 0 & -0.0900 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.0180 & 0 & 0 \\ 0 & 0 & 0 & 0.0180 & -0.0180 & 0 \\ 0 & 0 & 0 & 0 & 0.0180 & -0.0120 \end{bmatrix} \quad \mathrm{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$