



UiT The Arctic University of Norway

Faculty of Science and Technology

Department of Computer Science

Glucose Regulation for In-Silico Type 1 Diabetes Patients Using Reinforcement Learning

Miguel Ángel Tejedor Hernández

A dissertation for the degree of Philosophiae Doctor – February 2021



Abstract

Type 1 diabetes is a metabolic disorder characterized by high blood glucose levels as a consequence of deficiency of the hormone insulin. This condition leads to acute complications, damaging several organs and tissues throughout the patient's body. Despite years of research and clinical trials, no cure for type 1 diabetes exists yet, requiring lifelong treatment by external insulin administration.

However, new technologies have impacted current research for type 1 diabetes, changing how the disease is treated and leading to vast improvements in patient's quality of life. Among others, the artificial pancreas for automatically regulating blood glucose levels has gained importance in recent years, becoming the holy grail of the diabetes research.

Furthermore, the artificial pancreas has opened doors for new research fields, and recent advances are focused on automated insulin delivery systems for blood glucose control. This has resulted in the application of machine learning techniques competing with traditional control approaches. Concretely, reinforcement learning methods have emerged as a promising and personalized solution for the blood glucose regulation problem in type 1 diabetes. This thesis explores the use of reinforcement learning methods as a control algorithms in the artificial pancreas system.

The first work of this thesis presents a systematic review of reinforcement learning methods for diabetes blood glucose control. Specifically, the effort is dedicated to recognize the challenges and the opportunities for reinforcement learning as the control algorithm in the artificial pancreas system. An exhaustive literature search is performed to analyze the state-of-the-art in the application of reinforcement learning approaches in diabetes blood glucose regulation, while identifying the existing problems in the research field.

A main motivation for the second work is to take advantage of the external knowledge from the diabetes disease and include this relevant information in the reinforcement learning framework. Concretely, diabetes domain knowledge about the well known hazardousness of the low blood glucose levels is taken into account when designing the reward function for the reinforcement learning algorithm.

Next, the use of reinforcement learning algorithms as an alternative approach

to the traditional control methods used in the artificial pancreas system is explored in the third work of this thesis. Concretely, a policy gradient approach called trust region policy optimization is suggested as an alternative to traditional model predictive control methods widely used for the blood glucose control task.

The last work of this thesis introduces a food recommendation system to prevent hazardous low blood glucose levels during physical activities in patients with type 1 diabetes. This system lays the basis for the inclusion of a reinforcement learning agent to automatically calculate the optimal amount of food required to safely exercise.

Acknowledgments

I really think that after *Luigi's* work the acknowledgments section should be removed from every thesis, because he did such a great job that makes the rest of us look bad! Anyway, I will try to do my best.

I would like to start by thanking *Fred*. You trusted me giving me the opportunity to come to Norway, do my PhD, and be part of the UiT Machine Learning Group, changing my life for the better. Thank you for all your support, assistance, and help from the very first day to the very last one.

If I managed to survive and finish my PhD it was thanks to you, *Jonas*, who helped me keeping always *positive mental attitude*. You have been accompanying me on my PhD journey from the beginning, and I know I can count on you in the future. During the last four years, you have not been just my partner at work, but you have also become my *friend*. I am pretty sure that people who know you will agree with me that you are that kind unforgettable person that everyone wants to keep in his/her life. I will always remember your jokes and your ability to be kidding most of the time, and of course I will miss your drawing skills. I am very happy and proud to be your first PhD student, so thanks you a lot, because without you, I would not be writing these lines today.

I would also like to express my gratitude to *Gunnar* and my other cosupervisors for all your support and good discussions. You have always assisted me every time I needed any help.

Ilkka and *Phuong*, both of you complete our *diabetes team*. Thank you very much for helping me when I needed it. All your invaluable contributions have guided me to finish my PhD. In addition, I would like to thank *Ashenafi* for the very nice discussions and for giving me some good inputs when I more needed them.

I think *Thomas* also deserve a special mention here. Thank you very much for helping and assisting me with absolutely every question I had for you. You always amaze me with your endless knowledge and your willingness to help others. Without your invaluable guidance, our group would be lost in technical difficulties. Thank you also so much for our informal talks about our common passions, I really enjoy them!

Thank to my other coauthors, especially *Anas El Fathi*, for your fruitful discussions and for sharing all your incalculable diabetes knowledge. I would also like to thank *Ahmad* and his McGill Artificial Pancreas Lab for welcoming me into their research group and exposing me to the clinical side of the diabetes research.

I would like to thank everyone at the UiT Machine Learning Group. I have enjoyed the time I spent with each of you. I really appreciate how we support and help each other. I really look forward to having more sauna and bingo sessions in the future, while presenting *the holy grail* of the diabetes research. UiT Machine Learning Group, I do not know if we are the best in the machine learning part (probably yes!), but I am sure we are the best as a group!

I am grateful to my committee members, *Kezhi*, *Kjersti*, and *Chiara*, who have spent their time and efforts to read this thesis.

I would like to especially thank my parents, *Miguel* and *Sandra*, for their constant support. I really appreciate all your effort and love in taking care of me. I love you both and I feel lucky to have you both as my parents. Thank you for everything. I would also like to thank the rest of my family and friends for their support.

When I arrived to Tromsø some years ago to do my PhD I could not imagine how my entire life was about to change, and now I could not imagine my life without you, *Yanina* and *Martina*. Thanks to you, my decision of coming here has become the best decision of my life. Thanks for standing by me during those years, especially during these stressful months, taking care of me day by day. I really appreciate all your support and patient, encouraging me to keep going and do my best, while at the same time helping me to disconnect from work. *Juntos somos fuertes y formamos un gran equipo, TATO.*

Takk takk!

Miguel Ángel Tejedor Hernández,
Tromsø, February 2021

Contents

Abstract	i
Acknowledgments	iii
List of Figures	ix
List of Abbreviations	xi
1 Introduction	1
1.1 Challenges and opportunities	1
1.2 Objectives	3
1.3 Brief summary of papers	3
1.4 Thesis organization	4
I Background theory and methodology	7
2 Diabetes Mellitus	9
2.1 Glucose-insulin dynamics	11
2.2 Current state-of-the-art in diabetes treatments	14
2.2.1 Basal-bolus insulin regimen	14
2.2.2 Insulin pump	16
2.2.3 Continuous glucose monitor	17
2.2.4 Artificial pancreas	18
3 In-silico diabetic patients simulation	25
3.1 Bergman’s minimal model	26
3.2 Hovorka’s model	27

3.3	UVA/Padova model	27
3.3.1	Breton’s physical activity model	28
4	Reinforcement learning	31
4.1	Markov decision processes	33
4.2	The Agent-environment interface	33
4.2.1	Markov decision property	34
4.3	Goals and rewards	35
4.4	Returns and episodes	35
4.5	Policies and value functions	36
4.5.1	Optimal policies and value functions	38
4.6	Exploration-exploitation dilemma	38
4.7	Value-based methods	39
4.8	Policy gradient methods	40
4.8.1	REINFORCE	42
4.8.2	Trust region policy optimization	42
4.9	Deep reinforcement learning	43
4.9.1	Neural networks	44
II	Summary of research	49
5	Research publications	51
5.1	Paper summaries	51
5.2	Other publications	55
6	Concluding remarks	57
6.1	Limitations	58
6.2	Future work	59
III	Included papers	63
7	Paper I	65
8	Paper II	81
9	Paper III	89

10 Paper IV	113
Bibliography	129

List of Figures

1.1	The papers included in Chapters 7 to 10 are accordingly placed in the proposed taxonomy of this thesis.	5
2.1	Complications related to uncontrolled diabetes [1].	10
2.2	Insulin and glucagon hormones are secreted by the pancreas in response to blood glucose levels, but in opposite fashion [2–4].	12
2.3	Glucose tolerance test: healthy and diabetic subjects [5].	13
2.4	Self-managed blood glucose control. Blood glucose concentrations are measured by the patient using manual finger-prick or a CGM device. The patient decides the amount of insulin required for blood glucose regulation based on the measured glucose values.	16
2.5	Blood glucose management based on the artificial pancreas.	20
2.6	Conceptual overview of the PID controller used in T1D treatment and control [6].	21
2.7	Conceptual overview of the MPC strategy used in T1D treatment and control [7].	22
3.1	Schematic representation of the Bergman minimal model [8].	26
3.2	Overview of the Hovorka model model [9].	27
3.3	Description of the UVA/Padova model [10].	29
4.1	Reinforcement learning framework.	32
4.2	MDP of the interaction between the agent and its environment [11].	34

4.3	Neural network policy parameterization. The neural network maps the state to the policy parameters, where θ are the weights of the neural network. The output is an action sampled from the parameterized policy [12].	44
4.4	Basic neural network architecture [13].	45
4.5	Neuron from neural network. Figure adapted from [14].	46
4.6	A Bayesian neural network with random weights instead of fixed. Figure adapted from [15].	47
5.1	Number of publications found in the literature review from 2009 to July 2019 related to RL application in blood glucose regulation for diabetic patients.	52

List of Abbreviations

CGM Continuous Glucose Monitor.

MDP Markov Decision Process.

MPC Model Predictive Control.

PID Proportional-Integral-Derivative.

RL Reinforcement Learning.

T1D Type 1 Diabetes Mellitus.

TRPO Trust region policy optimization.

Chapter 1

Introduction

1.1 Challenges and opportunities

Diabetes mellitus impairs the body's ability to produce and use insulin, resulting in life-threatening complications as a consequence of chronic high blood glucose levels. Diabetes is one of the leading causes of death around the world [16]. This condition produces the second biggest negative total effect on reducing life expectancy worldwide [17], with people suffering from diabetes having a 2–3 folds risk of mortality [18]. Global incidence, prevalence, and death associated with diabetes were 22.9 million, 476.0 million, and 1.37 million in 2017, with a projection to 26.6 million, 570.9 million, and 1.59 million in 2025, respectively [19]. Therefore, diabetes imposes a heavy global burden on public health and socioeconomic development, and it is considered one of the largest global public health concerns [20, 21].

In the case of type 1 diabetes (T1D), the body loses its insulin production capabilities, requiring the patient to follow a strict personalized protocol of food intake, subcutaneous insulin administration as a treatment for the high blood glucose levels, and exercise. Diabetes research activities have experienced an extensive acceleration as a consequence of recent technological advances in sensor technologies and wearable devices, and the increased processing power in mobile phones [22, 23]. These new technologies have boosted the development of an artificial pancreas for automated insulin treatment, improving blood glucose regulation while favoring patients' quality of life

and independence [24, 25]. The artificial pancreas configuration consists of a continuous glucose monitor (CGM) tracking blood glucose levels, an insulin pump, and a control algorithm to translate changes in blood glucose concentrations into optimal insulin doses.

The control algorithm represents the key component of the artificial pancreas system, since maintaining normoglycemia is a challenging task in the treatment of diabetes. Traditional controllers such as model predictive control (MPC) methods used in blood glucose regulation assume a perfect model of the complex glucose-insulin regulatory system, so the patients get exposed to harmful situations when facing external events not captured by the model. Another current approach is purely reactive method, such as proportional-integral-derivative (PID) controllers. These algorithms react only to current glucose values and they are unable to respond fast enough, especially during meals. Therefore, adaptive, flexible, and automated insulin delivery algorithms able to deal with unpredictable events while providing personalized control for the patients are beyond the state-of-the-art in the blood glucose regulation problem [26, 27].

Among diabetes research fields, the inclusion of artificial intelligence solutions has allowed the application of machine learning and data mining techniques in T1D [28, 29], of which blood glucose prediction appears as the most popular focus [30]. This new scenario has led to the development of blood glucose control strategies as one of the most important issues during the last years [31], becoming an active research area approached from many different angles by a large number of scientists in different fields. At this point, reinforcement learning (RL) algorithms emerge as a highly promising approach to handle the disadvantages associated with traditional blood glucose control strategies [26], gaining increased attention in the artificial pancreas research [27, 32–40].

RL methods provide an adaptive and personalized solution to calculate optimal insulin doses in the artificial pancreas system, reacting to the immediate needs of the patients while at the same time adapting to underlying behavioral patterns. In comparison with other traditional methods, model-free RL approaches do not require a detailed description of the glucose-insulin regulatory system. However, there are challenges related to the RL application. These methods are not very efficient in terms of data, usually requiring a large amount of data during training. Finally, RL algorithms are not well

suitable to problems with inherent delayed actions, which might be a problem in the blood glucose control task because of the delayed action's effect caused by the use of subcutaneous insulin infusion [26, 27].

1.2 Objectives

The objective of this thesis is to develop control algorithms to automatically adjust insulin delivery based on data from both, the CGM and the insulin pump, to improve diabetes management in hybrid closed-loop artificial pancreas systems for T1D patients. Concretely, this work explores the use of RL algorithms as an alternative approach to the traditional control methods used in the artificial pancreas system for the blood glucose control task. Specifically, the effort is dedicated to recognize the challenges and the opportunities in the artificial pancreas system, analyze the state-of-the-art in diabetes blood glucose control using RL approaches, identify the existing problems, and provide solutions based on RL.

1.3 Brief summary of papers

The following papers are included in this thesis:

- (I) Miguel Tejedor, Ashenafi Zebene Woldaregay and Fred Godtlielsen, "**Reinforcement learning application in diabetes blood glucose control: A systematic review**," *Artificial Intelligence in Medicine*, vol. 104, 2020.
- (II) Miguel Tejedor and Jonas Nordhaug Myhre, "**Controlling Blood Glucose For Patients With Type 1 Diabetes Using Deep Reinforcement Learning – The Influence Of Changing The Reward Function**," *Proceedings of the Northern Lights Deep Learning Workshop*, vol. 1, pp. 1-6, 2020.
- (III) Jonas Nordhaug Myhre, Miguel Tejedor, Ilkka Kalervo Launonen, Anas El Fathi and Fred Godtlielsen, "**In-Silico Evaluation of Glucose Regulation Using Policy Gradient Reinforcement Learning for Patients with Type 1 Diabetes Mellitus**," *Applied Sciences*, vol. 10, no. 18, 2020.

- (IV) Phuong Ngo, Miguel Tejedor, Maryam Tayefi, Taridzo Chomutare and Fred Godtlielsen, ”**Risk-Averse Food Recommendation Using Bayesian Feedforward Neural Networks for Patients with Type 1 Diabetes Doing Physical Activities**,” *Applied Sciences*, vol. 10, no. 22, 2020.

Paper I. In this paper we perform an exhaustive literature review to evaluate the state-of-the-art of RL approaches to design blood glucose control algorithms for diabetic patients, critically analyzing relevant articles in the research field. Therefore, this paper lays the basis for future research work, supporting the rest of the papers included in this thesis.

Paper II. In this paper, a hand-designed reward function including external knowledge from the diabetes disease is designed, evaluating the influence of changing the reward function in the blood glucose control task for T1D patients.

Paper III. This paper tests and evaluates a RL approach based on deep reinforcement learning, in which deep learning is used for learning feature representations that in the traditional framework are usually hand-engineered. In addition, the deep RL algorithm is compared with the state-of-the-art in blood glucose control algorithm for T1D patients.

Paper IV. This paper presents a food recommendation system based on Bayesian neural networks for diabetic patients doing physical activities, reducing the risk of hypoglycemia during exercise. This system is conceived to serve as a preliminary stage for a RL agent optimizing the recommended food sizes.

Figure 1.1 shows where the presented papers fit in the overviewing picture of this thesis.

1.4 Thesis organization

A summary of the content of this thesis is provided below, including background theory, simulation tools, proposed methodologies, resulting publications, and final remarks.⁷

Chapter 2 introduces the blood glucose control problem and presents an overview of the current solutions to glucose regulation in T1D.

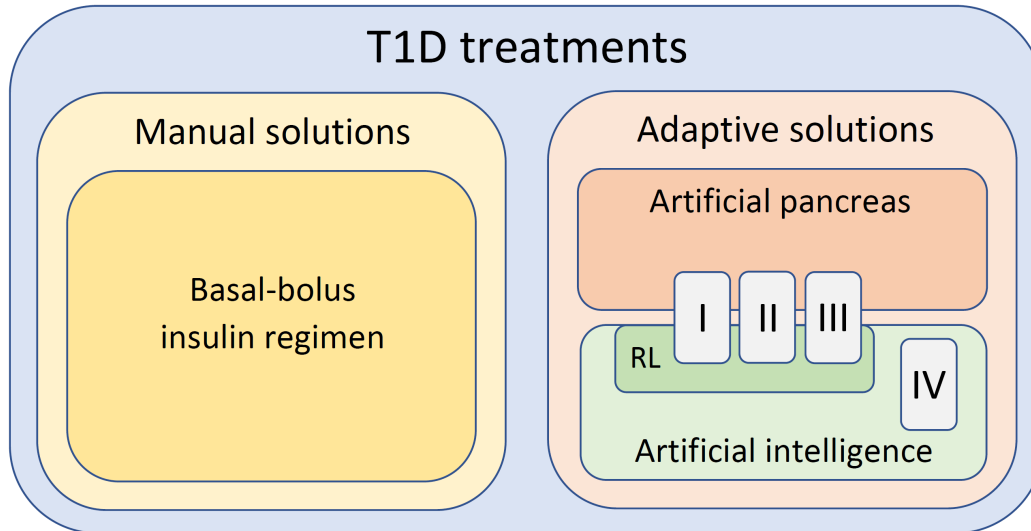


Figure 1.1: The papers included in Chapters 7 to 10 are accordingly placed in the proposed taxonomy of this thesis.

Chapter 3 presents the main physiological models used in the T1D research field to generate simulated data.

Chapter 4 presents the basics, weaknesses, and strengths of RL and its application in diabetic blood glucose control, putting particular stress on policy gradient methods in a deep RL approach.

Chapter 5 summarises the scientific contributions accomplished during this research work.

Chapter 6 provides some concluding remarks and a discussion on future research directions.

Chapters 7 to 10 report the publications included in this thesis.

Part I

Background theory and methodology

Chapter 2

Diabetes Mellitus

Diabetes Mellitus is characterized by a metabolic disorder that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. This results in chronic high blood glucose levels, leading to long-term damage, dysfunction and failure of various organs such as those summarized in figure 2.1 [41, 42]. According to the International Diabetes Federation approximately 1 in 11 adults has diabetes, which means 463 million people worldwide suffered from these conditions in 2019 [43]. This represents 9.1 % of the adult population, while trends suggest the rate would continue to rise [19]. Furthermore, diabetes at least doubles a person's risk of early death, resulting in approximately 1.7 million deaths directly attributed to diabetes each year, while 10 % of global health expenditure is spent on diabetes (USD760 billion) [44]. Because of the high incidence and prevalence of diabetes, the share of research devoted to the disease is continuously increasing [45].

There exist three main types of diabetes: T1D, a chronic condition in which the pancreas produces little or no insulin by itself and the patient requires daily insulin administration, Type 2 Diabetes Mellitus, which occurs when the body becomes resistant to insulin or does not produce enough insulin, and gestational diabetes, produced by high blood glucose levels during pregnancy. All of them require continuous management from patients and physicians in order to avoid complications, which eventually may be disabling or even life-threatening [41].

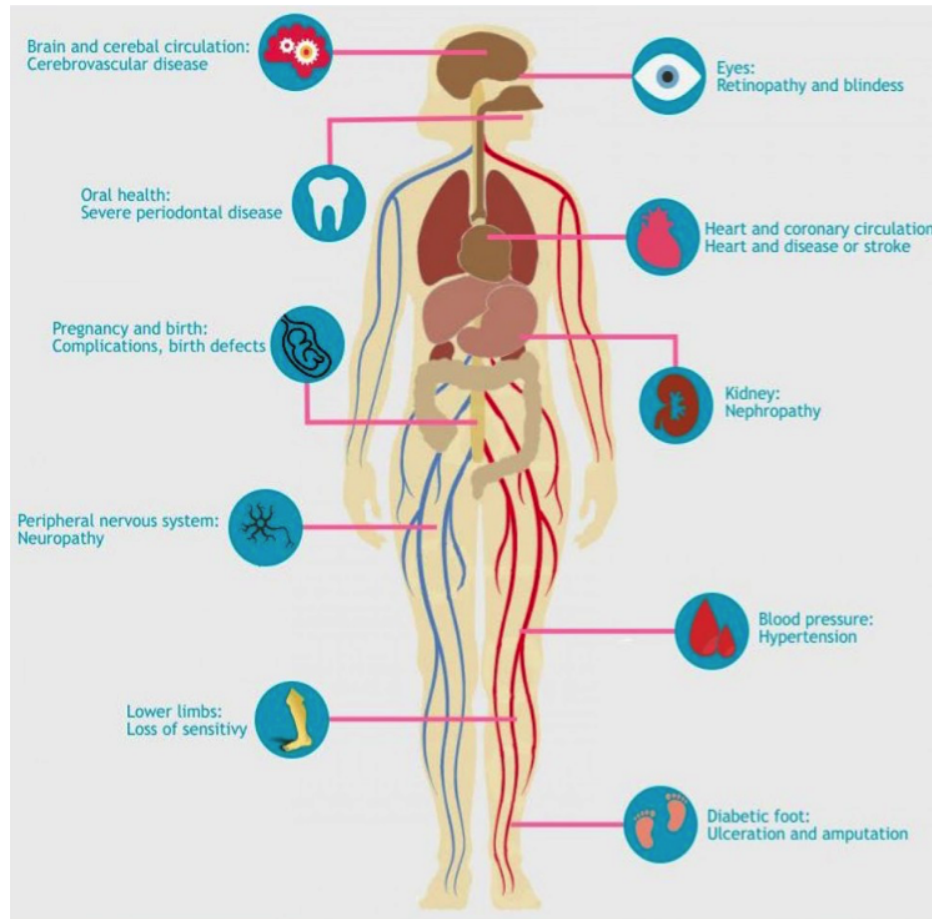


Figure 2.1: Complications related to uncontrolled diabetes [1].

2.1 Glucose-insulin dynamics

The human body is dependent on keeping of blood glucose levels in a very narrow normoglycemic range in order to ensure normal body function. Insulin and glucagon are the hormones produced by the pancreas to regulate blood glucose levels. Disturbances in the interplay of the hormones involved may lead to metabolic disorders such as diabetes, whose medical costs, prevalence and comorbidities take on a dramatic scale [46].

Figure 2.2 shows the relationship between insulin and glucagon, with the pancreas serving as the central player in the tight control task [4]. Blood glucose levels are regulated by the pancreas secreting the blood sugar-lowering hormone insulin and its opposite glucagon [3]. High blood glucose concentration stimulates the *insulin* secretion by the beta cells of the pancreas while inhibiting glucagon secretion. Conversely, low blood glucose concentration stimulates the *glucagon* secretion by the alpha cells of the pancreas while inhibiting insulin secretion, although there is always a low level of insulin secreted by the pancreas [47]. In response to insulin, the cells absorb glucose from the bloodstream, lowering the high blood glucose levels into the normal range. Similar to insulin, the glucagon counterpart works in the opposite way, mainly influencing the liver cells to release the stored glucose into the bloodstream, increasing the low blood glucose levels into the normal range [2].

Glucose homeostasis is the balanced and opposing actions of insulin and glucagon by the pancreas, accomplishing the preservation of blood glucose levels within a range of 4-6 mmol/L (70-110 mg/dL) [46]. Low blood glucose (hypoglycemia) is when the blood sugar concentration is below 4 mmol/L (70 mg/dL), while high blood glucose (hyperglycemia) is defined as values above 10 mmol/L (180 mg/dL). In the case of T1D, desirable blood glucose levels before meals are defined to be between 4 and 7 mmol/L (70-126 mg/dL), with values under 9 mmol/L (162 mg/dL) as the target after meals [48]. Figure 2.3 shows the results from a glucose tolerance test where the blood glucose values from a healthy subject and a diabetic subject are compared. In this test, oral glucose is given to the subjects and blood samples are taken afterward to determine blood glucose clearance. This test is usually used in diabetes diagnosis, since diabetic blood glucose rises to hyperglycemic values due to the lack of insulin.

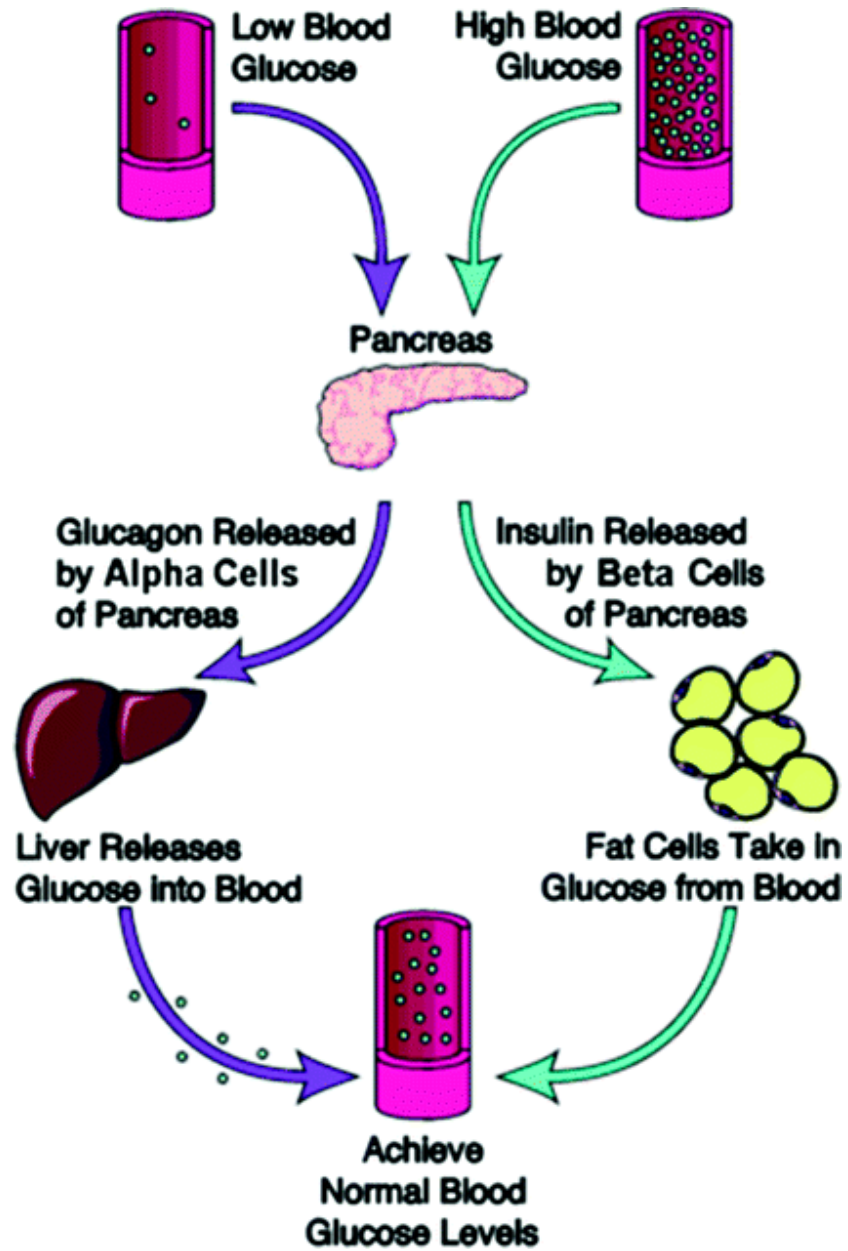


Figure 2.2: Insulin and glucagon hormones are secreted by the pancreas in response to blood glucose levels, but in opposite fashion [2–4].

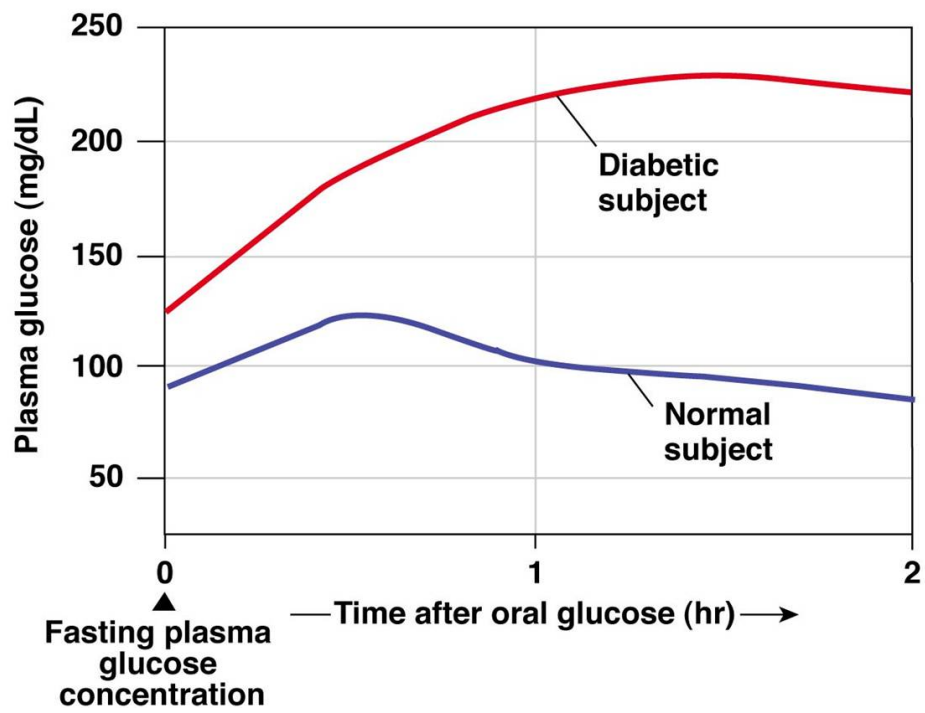


Figure 2.3: Glucose tolerance test: healthy and diabetic subjects [5].

Diagnosis of blood sugar conditions are determined by the insulin and glucagon secretion from the pancreas [2]. In this regard, *T1D* is the autoimmune destruction of insulin-producing beta cells in the pancreas, resulting in an increase in blood glucose that over time leads to the damage of various organ systems.

2.2 Current state-of-the-art in diabetes treatments

Self-treatment of T1D mainly involves multiple glucose level measurements throughout the day using manual finger-prick or a CGM with a glucose sensor embedded in the subcutaneous tissue (described below in section 2.2.3) [49]. In addition, administration of insulin via multiple daily insulin injections or through a pump providing a continuous subcutaneous insulin infusion is required (described below in section 2.2.2) [50]. In combination with this, a physician will design a treatment plan in collaboration with the diabetic patient, self-administering insulin according to the monitored blood glucose concentrations [51].

Due to the demands of everyday life and the fact that patients to a large degree are responsible for treating themselves, the decisions related to the insulin treatment are thus based partly on hard calculations, personal and medical experience, rules of thumb, and, in some cases, just pure guesswork. Although this results in effective treatment when done correctly, it is extremely time-consuming and a constant burden for the patients [52]. Therefore, during their daily life the patients have to deal with many difficulties, while T1D management becomes a really challenging task for them [53]. Even with a due amount of vigilance, many patients may still suffer significant diabetes-associated complications [54].

Current approaches in diabetes treatment are discussed in the following sections.

2.2.1 Basal-bolus insulin regimen

Basal-bolus insulin therapy is an insulin treatment in which patients separately inject a combination of different insulins (basal and bolus) to regulate

their blood glucose concentrations. The insulins are administered via subcutaneous injections in the fatty tissue just below the skin. This implies a delay in the insulin's action compared to the natural insulin secretion from the pancreas. In addition, diabetic patients on a basal-bolus regimen need to monitor whether the correct insulin doses are being administered by regularly measuring their blood glucose levels throughout the day [55].

Basal insulin is a long-acting insulin to moderate blood glucose when not eating, keeping glucose levels stable through periods of fasting, while allowing the cells to convert sugar into energy more efficiently. Patients usually inject basal insulin once or twice a day to keep fasting blood glucose levels consistent, since it reaches the bloodstream several hours after injection and remains effective for up to 24 hours.

Bolus insulin is a short-acting insulin most often given in higher doses, with faster action, but shorter-lived effect on blood glucose levels than basal insulin. It begins working in about 15 minutes or less, peaks in about 1 hour, and remains in the bloodstream for up to 2 to 4 hours. Typically, diabetic patients inject bolus insulin around mealtimes to quickly reduce the impact of high blood glucose concentrations resulting from dietary glucose. Therefore, carbohydrate counting is one of the diabetic patient responsibilities, adjusting the amount of insulin they need to cover the carbohydrate content of their meals [56]. Furthermore, the short-acting insulin is also used to administer correction bolus when blood glucose is high. For example, if the patient made a carbohydrate counting error during the last meal, underestimating the amount of carbohydrates intake, and so administering a not big enough meal bolus; this patient will need to take a correction bolus to reduce the high blood glucose and mitigate the hyperglycemia.

While a basal-bolus regimen allows for a flexible lifestyle regarding the amount of food eaten and timing of meals [57], this approach involves more work on the patient part. Moreover, unless patients are insulin pump users, basal-bolus treatment involves taking multiple injections every day, which might be problematic for some people since adapting to this routine might provide emotional and social challenges. For example, children at school following basal-bolus regimen need to feel comfortable with injecting insulin at meal times. An insulin pump is a device to deliver insulin either automatically or in response to instructions given by the patient. While diabetic patients usually take basal and bolus insulins via injections, insulin pumps work sim-

ilarly, and many patients prefer to use pumps instead of manual injections. Indeed, doctors now tend to recommend devices that provide better life quality instead of basal-bolus injections [58]. Therefore, the basal-bolus regimen is becoming less and less frequent among diabetic patients. Approximately 30-40 percent of T1D patients are using insulin pump and glucose sensor technologies, avoiding the need for daily injections [59]. However, switching between basal and bolus insulin doses at strategic times is the foundation for newer and automated diabetes care technologies [60]. This traditional and manual basal-bolus blood glucose control strategy is shown in figure 2.4.

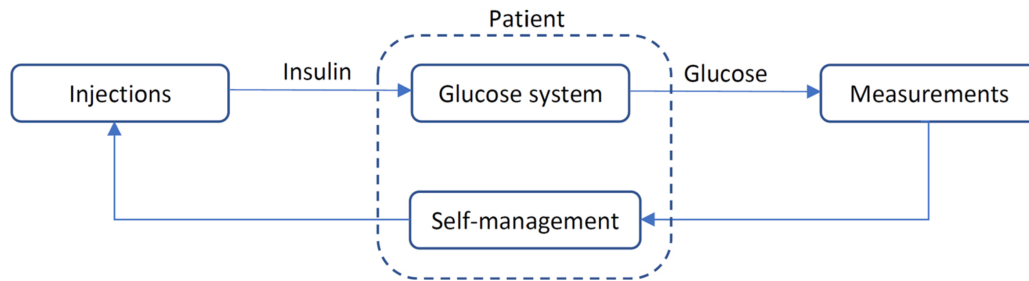


Figure 2.4: Self-managed blood glucose control. Blood glucose concentrations are measured by the patient using manual finger-prick or a CGM device. The patient decides the amount of insulin required for blood glucose regulation based on the measured glucose values.

2.2.2 Insulin pump

Advances in healthcare technologies have allowed diabetic patients to use automatic insulin pumps and CGMs, reducing the number of basal-bolus regimen users while avoiding the need for multiple daily injections throughout the day [61]. Therefore, pumps have rapidly become the mainstream alternative to insulin injections, since diabetic patients have more access to insulin pumps in recent years [62].

Insulin pumps are continuous subcutaneous insulin infusion systems administering a steady and measured insulin dose performing as basal insulin, while increasing the insulin dose to work as a meal bolus when needed [63]. Insulin therapy may become less disruptive and timing-dependent when using pumps, but the patient still have to perform carbohydrate counting and requests the pump to increase insulin dose at mealtime [64].

Typically, the patient is wearing the pump at all times, delivering insulin through a plastic tube with a cannula inserted under the patient's skin at the end of the infusion set. However, patients tend to take off the pump during reduced or removed clothing activities, such as swimming, washing, and sexual intercourse, since pumps might be cumbersome [65]. Nonetheless, patients and relatives generally report high levels of satisfaction and no social difficulties associated with the use of insulin pumps [66].

Insulin pumps and CGM devices have shown good performance reducing hypoglycemia risk while improving glycemic control, demonstrating to be clinically valuable [59]. In addition, most patients report better ability to participate in social activities while improving overall lifestyle flexibility [67]. Therefore, modern insulin pumps provide clinically meaningful benefits improving life quality of diabetic patients, with smart pumps recording blood glucose data and reporting directly to the doctors, making administration easier thanks to connections with phone apps [68].

Despite the improved life quality, the pumps are not without problems, and some patients report having experienced downsides to using the insulin pump [69]. Issues such as insulin infusion errors because of insulin infusion set blockage, insulin stability, infusion site problems, user error, pump failure, or a combination of these, might occur even when using state-of-the-art insulin pumps, exposing users to significant hazards [64].

2.2.3 Continuous glucose monitor

Another device that has changed diabetes management along with the insulin pump is the CGM [70]. This compact medical system continuously monitors patient's subcutaneous blood glucose levels (usually every 5 minutes) using a sensor with a cannula penetrating in the adipose tissue [71]. This causes a delay associated to the blood glucose measurements, since CGM systems measure glucose in interstitial fluid but not in blood [72].

Patients using CGM devices report improved life quality [73], reducing risks of hypoglycemia and hyperglycemia, as well as glycemic variability [74]. Despite improved glycemic profiles, CGM users report burdens such as the cost, untrusted readings, pain, time consumed, and, to a lesser extent, cutaneous complications [75, 76]. In total, the patients using CGM describe more benefits and less burdens when comparing with those who are not using a

CGM [77].

2.2.4 Artificial pancreas

Recent technological advances and improvements in diabetes treatment equipment have resulted in the development of the *artificial pancreas*, emerging as a new approach for treating diabetes [78–80]. The successful development of an artificial pancreas combines three main elements: a CGM continuously monitoring blood glucose levels, an insulin pump delivering insulin doses, and a control algorithm calculating insulin doses administered by the pump in response to the blood glucose concentrations measured by the CGM [81]. This framework shown in figure 2.5 can be further extended to a broader scope resulting in a complete mHealth system, using wearables devices for health services and data collection [82]. The system would supervise the healthcare plan while monitoring the patient physiological status, thereby including additional relevant information for diabetes care, such as food intake, physical activity, stress level and infections [83].

There are three main classes of insulin delivery systems: open-loop, closed-loop and hybrid closed-loop. In the open-loop method, the patients manually adjust and administer insulin doses throughout the day [84], which corresponds to the basal-bolus insulin regimen previously describes in section 2.2.1. Conversely, the closed-loop delivery systems keep the user involvement in blood glucose control to a minimum, corresponding to the artificial pancreas idea. Ideally, a closed-loop blood glucose controller would be able to automatically calculate and deliver proper insulin doses in real time based exclusively on information from patient’s measurements, regardless of the situation, and adapting to the user’s lifestyle [85]. Finally, in the hybrid closed-loop setup the control algorithm is able to automatically increase and decrease pump’s basal insulin delivery attempting to keep glucose concentrations within a desirable range, while meal insulin boluses are still the patient’s responsibility and carbohydrate intake information has to be provided to the system [86].

The hybrid closed-loop setup requires the patients to estimate the ingested amount of carbohydrates during meals, which is a daily challenging task and prone to human errors [87]. The scientific community is well aware of the carbohydrate counting adversities, and the true effect of these errors is still a topic of debate. Kawamura et al. [88] found that meals with small amounts

of carbohydrate tended to be overestimated, while Vasiloglou et al. [89] found that larger meals led to larger estimation errors. Moreover, Deeb et al. [90] report that carbohydrate-counting errors are not correlated with meal size, while Reiterer et al. [87] note that glycemic control is more negatively affected by random carbohydrate counting errors than systematic bias errors. Therefore, these under- and over-estimated amounts of carbohydrates lead to undesirable postprandial hyperglycemia and hypoglycemia, respectively, as a consequence of inaccurate bolus insulin doses. In an attempt to mitigate this problem, the hybrid closed-loop systems temporarily change the basal insulin rate with the purpose of compensating carbohydrate counting errors.

Nonetheless, the artificial pancreas is the most promising solution for T1D patients, with multiple studies reporting safety and effectiveness in improving glycemic control and proportion of time spent in the target glucose concentration range when using artificial pancreas systems [59, 91–98]. Currently, the three commercial available artificial pancreas systems, the Tandem t:slim X2 [99], the CamAPS FX DanaRS [100], and the Medtronic 670G [101] (the next generation Medtronic 780G is expecting to commercial launch within this year 2021) [102], as well as several do-it-yourself systems, see e.g., [95], and academic systems, e.g., [103], are all hybrid closed-loop systems.

The artificial pancreas blood glucose control framework is shown in a flowchart in figure 2.5. This is a closed-loop system in which the control algorithm calculates the proper insulin dosage based on glucose concentrations measured by the CGM [104]. The insulin pump delivers the needed amount of insulin determined by the controller, affecting glucose system and changing blood glucose level. A new insulin dosage is calculated and applied based on the previous changes produced in the blood glucose concentration. This process implies that only information measured from the patient is used to make decisions by the controller, without knowledge of external data [85].

Roadblocks in the artificial pancreas

In recent years CGMs and insulin pumps have experienced rapid technological developments, while state-of-the-art dosage algorithms still requires regular intervention by the patient and/or caregiver. There exists a delay in the insulin's action as a consequence of the subcutaneous insulin administration in comparison with the normal insulin secretion from the pancreas. In addition, blood glucose values from CGM are also delayed on time. Moreover,

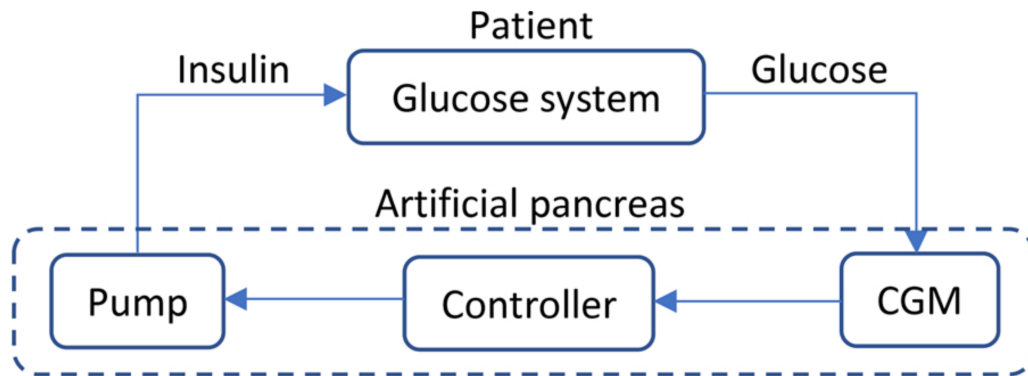


Figure 2.5: Blood glucose management based on the artificial pancreas.

patient-specific parameters variation is caused by dynamic factors complicating the control process. Particularly, the effect of physical activity on insulin and blood glucose dynamics is especially difficult to model and it is a major source of hypoglycemia [105]. A simple reactive controller translating momentary data streaming from the CGM into instructions for the insulin pump is not able to keep blood glucose levels in range after meals. Therefore, it becomes impossible to fully mimic the dynamic and person dependent control of blood glucose levels performed by beta cells in the pancreas.

In the blood glucose control research field, there have been investigations into fuzzy logic [106], and more recently techniques from machine learning and statistics [107, 108]. Fuzzy logic are reactive systems of if-else statements to determine the timing and dosage of insulin, often developed in collaboration with caregivers [109]. However, there are currently two dominant artificial pancreas controller algorithm paradigms, namely PID control [101, 110], and MPC [111, 112]. A meta-analysis of the clinical data obtained in studies performed using these approaches is conducted in [113]. In what follows, we discuss state-of-the-art closed-loop controller algorithms, mostly hybrid systems, for T1D.

Proportional-integral-derivative

A PID controller is a reactive control loop system employing feedback by measuring the output variable and adjusting the input according to the error value, which is estimated as the difference between the desired set point and the measured output variable. Then the controller applies a correction

based on proportional, integral, and derivative responses and sums those three components to compute the output [114]. Therefore, a PID controller estimates the amount of insulin required to minimize a weighted sum of these three terms, which the proportional term referring to the difference between actual and desired blood glucose concentration, the integral term referring to the accumulation of this difference over time, and the derivative term referring to the proportional change rate [115]. This kind of control algorithms are considered one of the most used techniques in the artificial pancreas framework [113]. Figure 2.6 shows the working flowchart of the PID controller.

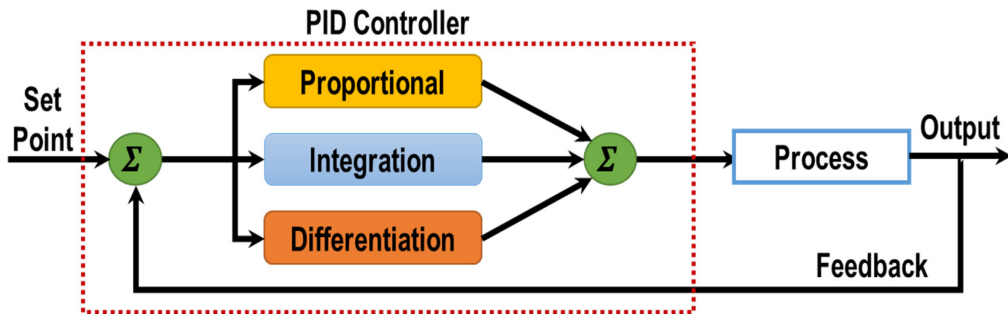


Figure 2.6: Conceptual overview of the PID controller used in T1D treatment and control [6].

The Medtronic 670G system uses a PID controller with insulin feedback to continuously calculate insulin doses based on CGM levels [101], while other studies have been performed to show the feasibility of this approach [110, 116]. However, insulin delivery systems utilizing PID controllers have demonstrated susceptibility to late postprandial hypoglycemia. This is because of the delays in insulin absorption associated with the subcutaneous route of delivery, which inevitably lead to large postprandial glucose excursions [117, 118].

A comparison between self-managed control by the patient, a PID controller, and RL methods is conducted in [27]. From this study, RL algorithms were able to outperform traditional approaches under certain circumstances, although they do not outperform the PID controller across all settings.

Model predictive control

MPC is a proactive method to control a process while satisfying a set of constraints. This approach relies on dynamic mathematical models of the process to predict future behaviour. A mathematical optimization algorithm calculates the optimal process inputs using the predictions from the model in order to optimize future behaviour of selected variables in the process. Once the current prediction horizon is optimized, the controller implements only the first step of the control strategy, and the optimization process is repeated starting from the new current state. This capability to anticipate future events is the main advantage of MPC controllers, since PID methods do not have the ability to predict [119]. Figure 2.7 shows the working flowchart of the MPC controller.

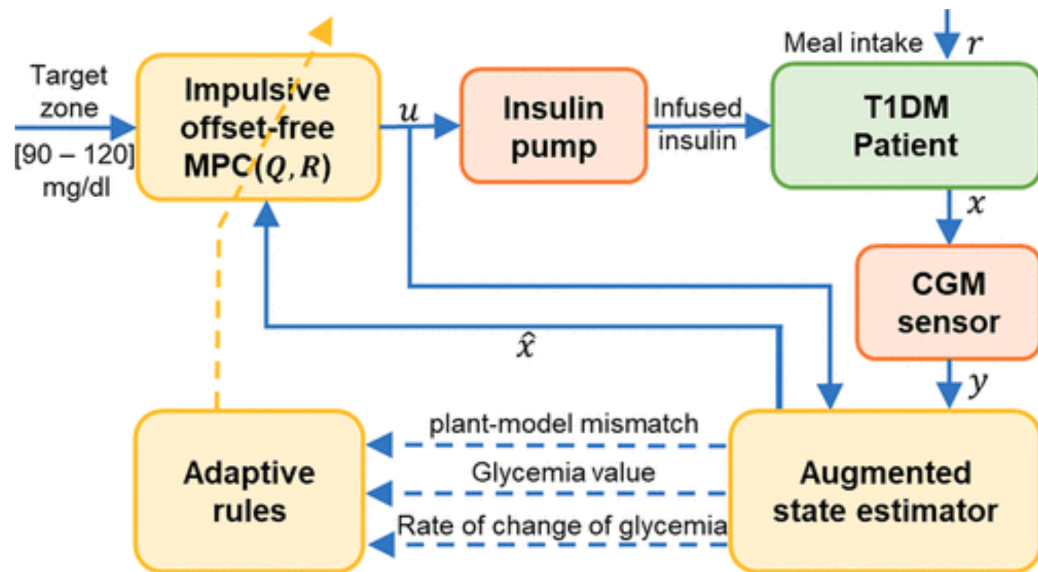


Figure 2.7: Conceptual overview of the MPC strategy used in T1D treatment and control [7].

MPC is one of the major options for blood glucose control in T1D, where glucose predictions are based on factors such as food intake, insulin delivery and previous blood glucose values [120]. In this scenario, the controller recommends an optimized sequence of changes in the basal insulin rate to minimize the difference between the predicted glucose curve from the model and the target glucose level [121]. Afterwards, the basal insulin rate is updated fol-

lowing the first of the suggested sequence of actions and the whole process is repeated. The goodness of the patient-specific parameters model is crucial to the algorithm's performance, because MPC approaches assume perfect knowledge of the true underlying model. This is one of the disadvantages related to conventional MPC controllers, since models checking can be difficult in reality. In addition, real-time algorithm update might be time-costly and so impractical, considering that accurate model parameter estimation may require large sample sizes. Furthermore, these control methods suffer from lack of flexibility to external perturbations not captured by the models, such as abnormal food intake or physical activity, because MPC strategies are model-driven rather than data-driven techniques. Therefore, these algorithms are somewhat limited to compensate for the incomplete glucose-insulin regulatory models used in the artificial pancreas application [26].

Chapter 3

In-silico diabetic patients simulation

Clinical trials are necessary for final validation of the artificial pancreas systems. However, *in-silico* evaluation through computer simulation is essential as a preliminary stage to establish robustness and limitations of insulin infusion algorithms. Simulated data accelerate the development of blood glucose controllers, alleviate the need for human or animal testing, and reduce both cost and ethical questions related to clinical trials. Actually, several in-silico evaluations should be performed to design, evaluate and verify the effectiveness of the controller before the actual clinical study [122]. Therefore, a model of underlying dynamics is necessary in order to develop control algorithms able to successfully connect a CGM and an insulin pump [122]. Furthermore, to perform evaluation experiments on diabetic patients may be neither possible, appropriate, convenient nor desirable, since some of these experiments cannot be done at all or are too difficult, dangerous and not ethical [123]. In addition, different countries have different procedures and regulatory conditions, which complicates the situation further.

There exist mainly three physiological models in the T1D research field, namely the Bergman minimal model [124], the Hovorka model [111], and the UVA/Padova model [10, 123]. This chapter introduces these main diabetes models from the literature.

3.1 Bergman's minimal model

The minimal model is the simplest model of the glucose–insulin homeostasis, which was proposed by Bergman and collaborators in the late seventies [125]. This is a simplified two-compartment linear model consisting of two differential equations, describing the dynamics of the plasma glucose uptake in response to the insulin concentration, and the pancreatic insulin release in response to the glucose stimulus [126]. Despite its simplicity, the minimal model glucose kinetics is still widely used in diagnosis as a clinical tool to calculate insulin sensitivity index [127]. However, this model does not consider the significant delays associated neither with the subcutaneous insulin infusion, nor the subcutaneous blood glucose measurements. The original minimal model includes one virtual patient [125]. A schematic representation of the Bergman minimal model is shown in figure 3.1.

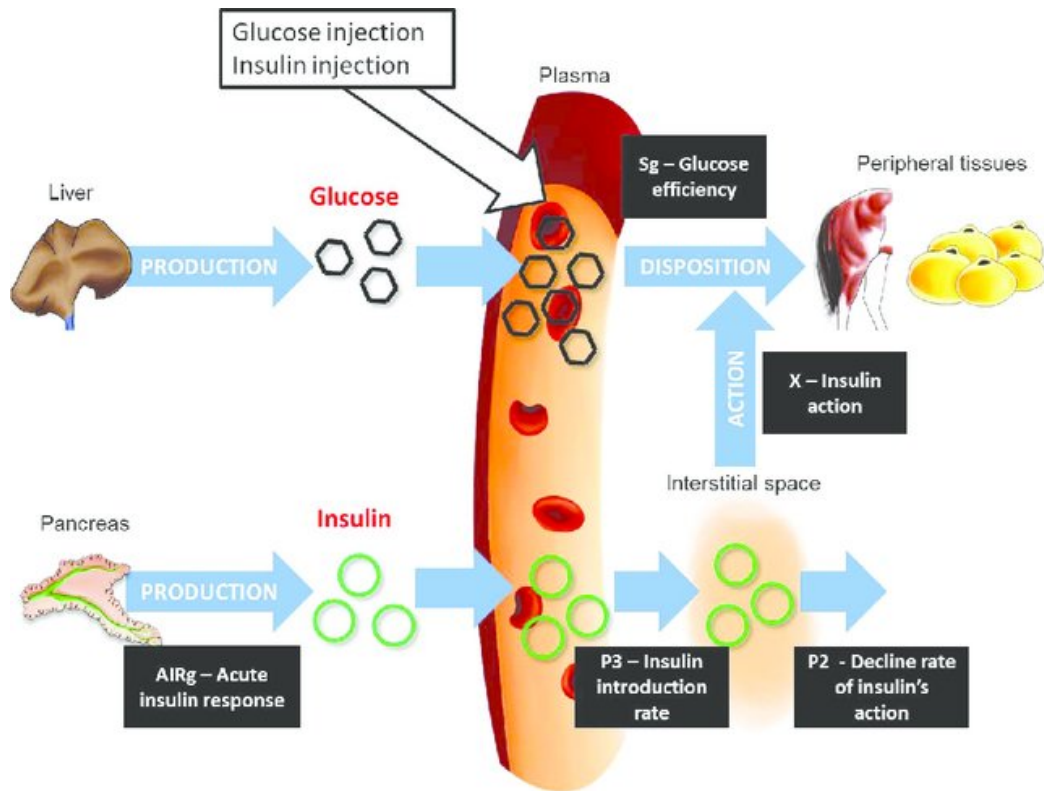


Figure 3.1: Schematic representation of the Bergman minimal model [8].

3.2 Hovorka's model

This model was developed by the Hovorka research group at Cambridge [128]. In this model, the glucose-insulin regulatory system is described by five submodels: two external compartments describing subcutaneous insulin absorption and interstitial glucose kinetics, and three internal compartments describing insulin action, glucose kinetics and glucose absorption from the gastrointestinal tract [129]. Unlike the minimal model, the Hovorka model includes delays related to subcutaneous insulin pump delivery and subcutaneous glucose measurements. Although the original Hovorka model includes one virtual patient, it is possible to simulate a virtual population by sampling model parameters from informed probability distributions, assigning a unique set of parameters to each individual [9, 121]. In addition to the inter-individual variability represented by the virtual population, intra-individual variability of the glucoregulatory response is represented by time-varying selected model parameters, which is an important advantage of this model [122]. An overview of the Hovorka model is shown in figure 3.2.

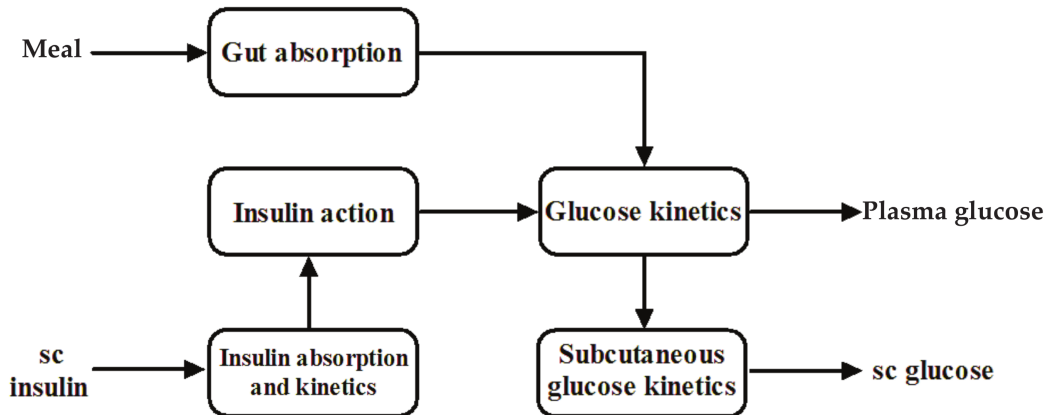


Figure 3.2: Overview of the Hovorka model model [9].

3.3 UVA/Padova model

The UVA/Padova model was developed through research efforts at the Universities of Padova and Virginia [123]. This model divides the glucose-insulin regulatory system into three external compartments describing subcutaneous

glucose, insulin and glucagon kinetics, and seven internal compartments describing the dynamics of glucose kinetics, insulin kinetics, glucagon kinetics and secretion, glucose rate of appearance, endogenous glucose production, glucose utilization, and renal excretion [10]. Similar to the Hovorka model, the UVA/Padova model also incorporates delays due to the subcutaneous glucose measurements and insulin administration, allowing more realistic simulations by adding models of CGMs and insulin pumps. The distributed version of the model has been validated by ten children patients, ten adolescents patients, and ten adults patients, while a more elaborated version of the model provides a large cohort of 300 virtual patients: 100 children, 100 adolescents, and 100 adults [130]. This is the only model of the dynamics of the human metabolic glucose-insulin system approved by The United States Food and Drug Administration as a substitute for animal trials in the pre-clinical testing of certain control strategies in T1D [131], which is probably the main reason why this model is widely used in the diabetes research. A description of the UVA/Padova model is shown in figure 3.3.

3.3.1 Breton's physical activity model

An extension of the UVA/Padova model has been developed to include the effect of physical activity in the model [132,133]. This physical activity model changes the glucose-insulin dynamics to simulate exercise sessions by modifying the insulin-dependent utilization component in the glucose-utilization subsystem. Physical activity simulation is of utmost importance since exercise is a major source of hypoglycemia in diabetic patients and risk of hypoglycemia is a significant limiting factor of their blood glucose regulation [105,134]. However, this model is not validated against data and further studies are needed for validation.

MEAL MODEL

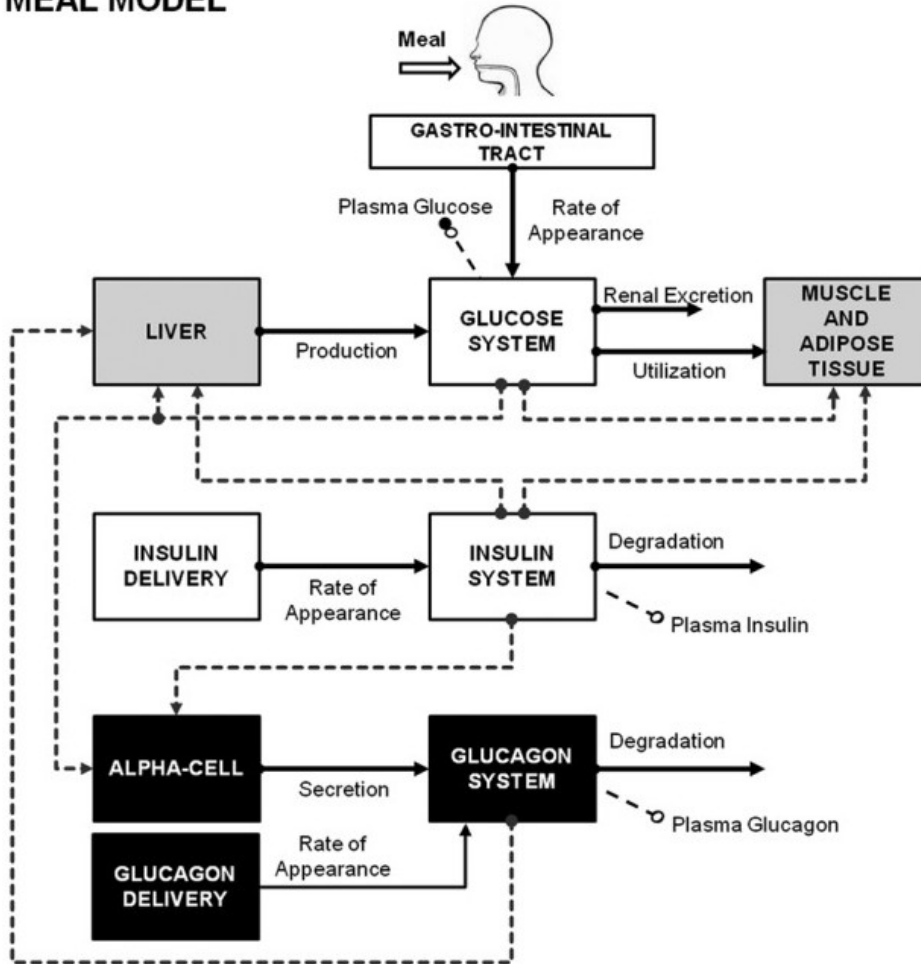


Figure 3.3: Description of the UVA/Padova model [10].

Chapter 4

Reinforcement learning

RL is a branch of machine learning based on the interaction between a decision making agent and an unknown environment, with the goal of training the agent to take actions that maximize its long term benefit [11]. At each decision time step, the agent takes an *action* for some given current *state* of the environment. As a consequence of this action, the *environment* reacts and transitions to a new state. The agent now receives a positive or negative reinforcement, a *reward* from the environment for the previously taken action. The RL framework is shown in figure 4.1, where the learner and decision maker is represented by the agent while the environment is what the agent interacts with, encompassing everything outside the agent [11]. The mapping of state to action is called the *policy*, which defines the behavior of the agent. The goal of RL is to learn an optimal policy that maximizes the amount of reward received over time, with the reward function defining the goal of the agent. In addition to the aforementioned RL elements, the *value function* indicates the total amount of reward expected by an agent when it starts from a given state and follows a given policy thereafter, specifying the long-term desirability of states. Similarly, the *action-value* function indicates the total amount of reward expected by an agent when it starts from a given state, takes a given action, and follows a given policy thereafter. Finally, some approaches use a model of the environment to predict future states and rewards, and are so called model-based methods [11].

In the RL blood glucose control task, the state space is a function of the interstitial glucose curve measured by the CGM. The agent is the controller (the

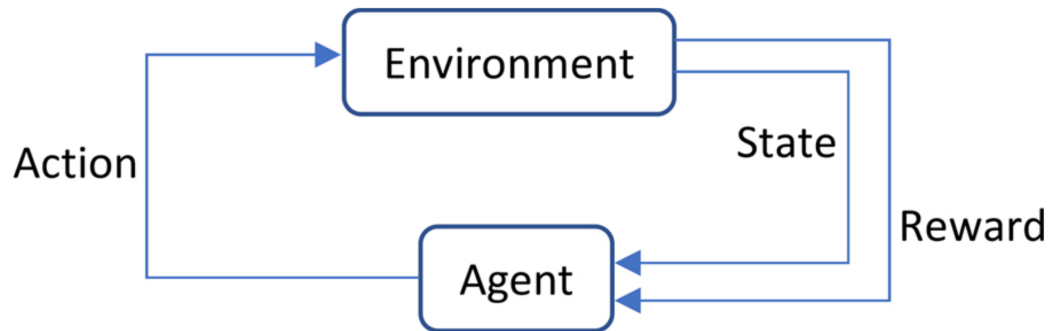


Figure 4.1: Reinforcement learning framework.

artificial pancreas), and its action space consists of insulin dosage amounts. Finally, the patient represents the environment, with the reward function measuring the discrepancy between ideal and actual glucose levels. In this research work only model-free RL approaches are considered, since it is not possible to know the true underlying model of the patient.

Several features of RL suggest high potential for the T1D control and management. First, RL is an appropriate solution for decision making processes with actions sequentially taken along a timeline, with those actions depending on the observed state, and with some notion of preferred states. These features are certainly present in the blood glucose control challenge.

In addition, RL algorithms do not require a detailed description of the environment unlike traditional control strategies. This is a very important factor in the diabetes application, since existing glucose-insulin models are inaccurate and do not catch the entire dynamics behind the glucose regulatory system.

Another advantage is that only data from the patient is used in the decision making process, leading to truly personalized recommendations since the controller continuously adapts and evolves with the user. This allows to introduce model-free and data driven algorithms that can enable another level of patient individualization, in contrast to many traditional control strategies where individual patient recommendations are based on an overall model fitted using a large dataset [26].

Finally, RL algorithms can control systems with delayed reward, which is one of the fundamental properties of these methods [135]. This implies that

an action in a state can still be considered to be good even if the immediate reward from taking that action is not considered good, since what matters for good behavior is to maximize the total reward in the long run. However, RL is not well suited to problems with delayed actions, since the agent expects that the state of the environment changes after an action is taken. This might be a problem in the blood glucose control task because of the delayed action's effect caused by the use of subcutaneous insulin infusion, since actions' effects manifest at later points in time than the actions inducing them. It is necessary to take this action delay into account during the design of the control process, even though this issue can be mitigated through the use of faster acting insulins [136]. This is because a RL application assumes an underlying Markov decision process (MDP), which is explained in further detail in the following sections. In addition, the existing delay in the blood glucose values introduced by the subcutaneous CGM measurements needs to be considered when facing the blood glucose control problem [72]. Additional convincing arguments for the use of RL in the T1D scenario are given in [26].

4.1 Markov decision processes

A RL problem can be formulated as a MDP, which is a formalization of sequential decision making [137]. The MDP framework is an abstraction of the goal-directed learning from an interaction problem. The MDP provides the mathematical framework for modeling the RL problem and make precise theoretical statements. This framework includes delayed reward, since actions influence future states and rewards instead of just immediate reward, creating the need to trade off immediate and delayed rewards [11]. A MDP is a stochastic process that satisfies the Markov property described below in section 4.2.1.

4.2 The Agent-environment interface

In a RL problem the agent is both the learner and decision maker continually interacting with its environment, which comprises everything outside the agent. This interaction is typically stated in the form of a MDP. Concretely, the interaction between the agent and its environment occurs at each time step, $t = 0, 1, 2, 3, \dots$, in which the agent perceives the state of the environ-

ment, $S_t \in \mathcal{S}$, and based on that representation selects an action, $A_t \in \mathcal{A}(s)$, where \mathcal{S} is the set of all states and $\mathcal{A}(s)$ is the set of all possible actions available in state s . As a consequence of its previously taken action, at the next time step the agent receives a numerical reward, $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$, and the environment moves to a new state, S_{t+1} . This process is represented in figure 4.2.

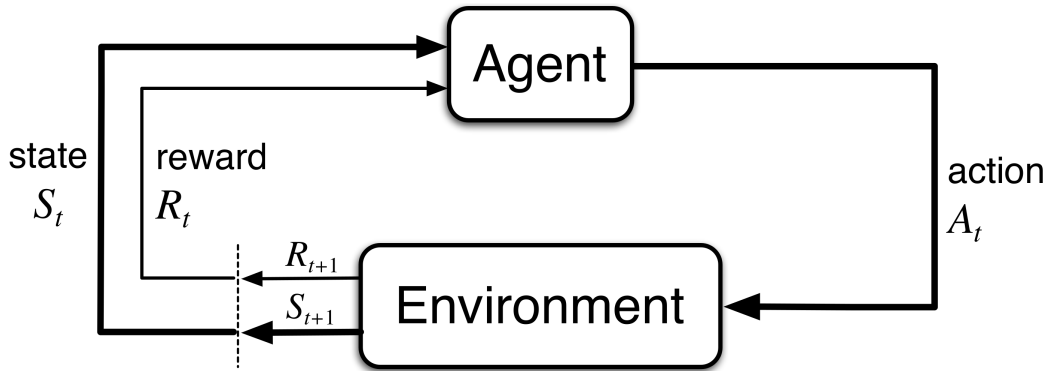


Figure 4.2: MDP of the interaction between the agent and its environment [11].

4.2.1 Markov decision property

In a MDP with finite number of states, \mathcal{S} , actions, \mathcal{A} , and rewards, \mathcal{R} , the discrete probability distributions of the random variables S_t and R_t depend only on the preceding state and action [138]. Therefore, for a particular state, $s' \in \mathcal{S}$, and a particular reward, $r \in \mathcal{R}$, the probability of those state and reward occurring at time t , given particular values of the preceding state and action is:

$$p(s', r | s, a) = \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}, \quad (4.1)$$

for all $s', s \in \mathcal{S}$, $r \in \mathcal{R}$, and $a \in \mathcal{A}$. The MDP dynamics is defined by the deterministic transition probability function p , specifying a probability distribution for each choice of s and a :

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r | s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s). \quad (4.2)$$

In a MDP, the environment’s dynamics are completely characterized by the probabilities given by p , with the probability of each possible state, S_t , and reward, R_t , depending only on the immediately preceding state, S_{t-1} , and action, A_{t-1} , and not on earlier states and actions. Accordingly, the state is said to have the Markov property, which refers to the memoryless property of a stochastic process [139].

4.3 Goals and rewards

At each time step, the agent receives a reward, $R_t \in \mathbb{R}$, from the environment. The goal of the agent is to maximize the cumulative long term reward it receives over time, i.e., the expected value of the cumulative sum of the received rewards. From the design point of view, the reward function is used to communicate to the agent what we want to achieve, but not how to achieve it, while the agent accomplishing our purpose through maximizing the provided rewards. Therefore, the reward function formalizes the goal of the agent, which is one of the most distinctive characteristics of RL [11].

The design of the reward function is one of the most critical part of any RL problem, since the success of the application is defined by how well the reward function formulates the goal of the problem [140].

4.4 Returns and episodes

The goal of the agent is to maximize the function of the sequence of discounted rewards defined as the *discounted return*:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad (4.3)$$

where $0 \leq \gamma \leq 1$ is the discount factor parameter. For γ values close to 0, the agent is focused on maximizing immediate rewards and discards the long-term return, while for γ values close to 1 the agent takes future rewards more into account. Therefore, the discount factor determines the present value of future rewards [11].

Note that equation (4.3) is defined for continuing tasks with final time step $T = \infty$, in which the agent continually interacts with its environment without time limit. However, this notation also works on *episodic* tasks where the interaction between the agent and its environment breaks into episodes with natural notion of a final time step T [11].

Equation (4.3) can be rewritten with a recursive relationship, since successive returns are related to each other:

$$\begin{aligned}
 G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots \\
 &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots) \\
 &= R_{t+1} + \gamma G_{t+1}
 \end{aligned} \tag{4.4}$$

Equation (4.4) is important for the theory and algorithms of RL, since this equation is used in the definition of the Bellman equations as described below in section 4.5.

4.5 Policies and value functions

The expected return is used to define the goodness of states and state-action pairs through the value functions, which in turn estimate the desirability of states, or actions given a state, for the agent [141]. Since the expected return depends on the actions taken by the agent, the value functions are defined with respect to the policy, which is a mapping from states to probabilities of selecting each possible action. The *policy*, $\pi(a|s)$, is the probability that $A_t = a$ if $S_t = s$ when at time t the agent is following the policy π . This defines a probability distribution over $a \in \mathcal{A}(s)$ for each $s \in \mathcal{S}$. RL methods specify how the policy of the agent changes as a result of its experience, while the value functions are essential to accurately assigning credit for long-term consequences to individual action selections [11].

The state-value function for policy π is the value of a state s under a policy π , i.e., the expected return when starting in s and following policy π thereafter,

and this value is defined by:

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right], \text{ for all } s \in \mathcal{S}, \quad (4.5)$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expected value of a random variable given that the agent follows policy π and t is any time step [11].

Similarly, the action-value function for policy π is the value of taking action a in state s under a policy π , i.e., the expected return of taking the action a , starting from state s , and following policy π thereafter, and this value is defined by:

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right]. \quad (4.6)$$

The value functions can be rewritten into Bellman equations following equation (4.4) and using recursive relationships to decompose these functions into two parts: the immediate reward plus the discounted future values [142]:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) [r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s']] \quad (4.7) \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a) [r + \gamma v_\pi(s')], \text{ for all } s \in \mathcal{S}, \end{aligned}$$

in which the unique solution to the equation is its value function $v_\pi(s)$. This Bellman equation (4.7) averages over all the possible states weighting each by its probability of occurring, stating that the value of the start state must equal the discounted value of the expected next state plus the reward expected along the way [11].

The Bellman equations are one of the central elements of many RL algorithms, since these equations form the basis to compute, approximate, and learn $v_\pi(s)$ [11].

4.5.1 Optimal policies and value functions

Bellman equations are used to find the state-value function and the action-value function of a given MDP. A RL problem is solved when the best way to behave in a MDP is learned, i.e., a policy that obtains the maximum possible long-term reward is found [11]. Consequently, a policy π is better than or equal to a policy π' if the expected return of policy π is greater than or equal to the expected return of policy π' for all states:

$$\pi \geq \pi' \iff v_\pi(s) \geq v_{\pi'}(s), \text{ for all } s \in \mathcal{S}. \quad (4.8)$$

An optimal policy is a policy that is better than or equal to all other policies. There may be more than one optimal policy in a MDP, but there exist always at least one optimal policy and all the optimal policies are denoted by π_* .

According to the different policies, there are many different value functions for a given MDP environment. However, all the optimal policies results in the same optimal value functions, which yield maximum value compared to all other value function from other policies. In this regard, the optimal state-value function,

$$v_*(s) = \max_{\pi} v_\pi(s), \text{ for all } s \in \mathcal{S}, \quad (4.9)$$

is defined as the expected return when starting in the state s and following the optimal policy π_* thereafter, maximizing the state-value function over all policies. Similarly, the optimal action-value function,

$$q_*(s, a) = \max_{\pi} q_\pi(s, a), \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s). \quad (4.10)$$

is defined as the expected return of taking the action a , starting from the state s , and following the optimal policy π_* thereafter, maximizing the action-value function over all policies.

4.6 Exploration-exploitation dilemma

Another distinctive characteristic of RL is the trade-off between exploration and exploitation, which is one of the challenges that arise during the algo-

rithm design [143]. The RL agent has to exploit its knowledge about previously taken actions to select actions that maximize the expected reward obtained on the one step. Moreover, the agent has to explore new actions in order to discover which actions yield the highest possible reward in the long run, gathering more data about the environment while learning a better policy [144]. Thereby arises the exploration-exploitation dilemma, since the agent has to exploit preceding experience while exploring to make better action selections in the future. However, the agent cannot explore and exploit at every action selection step, so first the agent must explore and try different actions to progressively exploit and favor actions that appear to be the best [145]. Consequently, the obtained reward is lower in the short run when exploring, but higher in the long run when exploiting the discovered best actions. In addition, each action must be tried many times in order to obtain a reliably estimate of its expected reward when the agent is facing a stochastic problem [11].

Since a balance between exploration and exploitation is required to solve a RL task, a simple approach is to follow a greedy policy taking the best action most of the time with a small probability ϵ of taking a random action. This is called an ϵ -greedy policy. The optimal solution to the exploration-exploitation dilemma has been intensively studied by the research community for many decades and remains unresolved [11].

4.7 Value-based methods

Several approaches have been proposed to reach the RL goal: learn the optimal policy. This have originated many RL methods with two main different variants: value-based methods and policy-based methods. Value-based methods are based upon temporal-difference learning to estimate the value function, v_π , and find an optimal policy [146]. Temporal-difference methods learn directly from raw experience without a model of the environment's dynamics, using bootstrapping to perform updates from the current estimate of the value function. Given some experience following a policy π , temporal-difference methods update the estimation $V(S_t)$ of the value function v_π for the state S_t occurring in that experience. At the next time step $t + 1$, the estimation $V(S_t)$ is updated using the observed reward R_{t+1} and the estimate

$V(S_{t+1})$:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)], \quad (4.11)$$

where $\alpha \in (0, 1]$ is a constant step-size parameter [11].

Two of the most popular algorithms based on temporal-difference learning are Sarsa and Q-learning, both based on equation 4.11 but learning an action-value function rather than a state-value function [147, 148].

4.8 Policy gradient methods

Unlike value-based methods, policy-based methods directly learn a *parameterized policy* that can select actions without consulting a value function to approximate an optimal policy π_* . The parameterized policy is defined as the probability of taking action a at time step t given that the environment is in state s at time step t with parameter $\boldsymbol{\theta}$:

$$\pi(a|s, \boldsymbol{\theta}) = \Pr\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\}, \quad (4.12)$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the policy's parameter vector. The policy can be parameterized in any way, provided $\pi(a|s, \boldsymbol{\theta})$ is differentiable with respect to its parameters, i.e., $\nabla \pi(a|s, \boldsymbol{\theta})$ exists and is finite for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$, and $\boldsymbol{\theta} \in \mathbb{R}^d$. In practice, a stochastic policy, $\pi(a|s, \boldsymbol{\theta}) \in (0, 1)$ for all s , a , and $\boldsymbol{\theta}$, is required to ensure exploration [11].

Policy gradient methods learn the policy parameter based on the gradient of some scalar performance measure, $J(\boldsymbol{\theta})$, with respect to the policy parameter. These methods maximize performance approximating gradient ascent in J :

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \widehat{\nabla J(\boldsymbol{\theta}_t)}, \quad (4.13)$$

where $\widehat{\nabla J(\boldsymbol{\theta}_t)} \in \mathbb{R}^d$ is a stochastic estimate whose expectation approximates the gradient of the performance measure with respect to its argument $\boldsymbol{\theta}_t$ [149].

The performance measure is usually defined as the value of the initial state s_0 :

$$J(\boldsymbol{\theta}) = v_{\pi_{\boldsymbol{\theta}}}(s_0). \quad (4.14)$$

This is equivalent to optimize the value of the initial state, where $v_{\pi_{\boldsymbol{\theta}}}$ is the true value function for $\pi_{\boldsymbol{\theta}}$, the policy determined by $\boldsymbol{\theta}$.

The gradient of the performance measure, $\nabla J(\boldsymbol{\theta})$, depends on the action selection and the distribution of the states, with both of them affected by the policy parameter $\boldsymbol{\theta}$. At the same time, the effect of the policy on the state distribution is a function of the environment, which is generally unknown [11]. The policy gradient theorem described later in this section solves this problem simplifying the gradient computation of the performance measure.

Policy-based methods offer some advantages over value-based methods. Policy-based methods are more effective than value-based methods in high dimensional action spaces. This is because value-based methods need to estimate the action-value function of each possible action, while policy-based methods directly adjust the policy parameters [150].

In problems with significant function approximation, the best approximate policy may be stochastic. Policy approximating methods can find stochastic optimal policies, while value-based methods have no natural way of finding them. This is because parameterized policies enable the selection of actions with arbitrary probabilities [11]. Therefore, policy-based approaches do not require any implementation of the trade-off between exploration and exploitation, since a stochastic policy ensure agent exploration.

Furthermore, policy gradient methods present stronger convergence guarantees than value-based methods. This is because in policy-based methods the action probabilities are function of the learned parameters of the policy and therefore change smoother than in value-based methods, where the action probabilities may change dramatically if an arbitrarily small change in the estimated action values results in a different action having maximal value [11]. Finally, policy parameterization may be a good way to include prior knowledge about the desired policy in the RL problem [11, 150].

The policy gradient theorem

The policy gradient theorem establishes that the gradient of the performance measure, $J(\boldsymbol{\theta})$, is proportional to the gradient of the policy itself:

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla \pi(a|s, \boldsymbol{\theta}), \quad (4.15)$$

where the distribution μ is the stationary distribution of the succeeding states of s_0 when following π . This theorem is one of the key points of policy gradient algorithms, providing an expression for the gradient of the performance measure, $\nabla J(\boldsymbol{\theta})$, with respect to the policy parameter, $\boldsymbol{\theta}$, without involving the derivative of the state distribution. This theorem is of great benefit, as it allows the use of any differentiable policy parameterization [11, 149].

4.8.1 REINFORCE

The policy gradient theorem allows the formulation of REINFORCE, a simple sample-based algorithm which lays the basis for most of the advanced policy gradient algorithms [151]. In REINFORCE, the agent uses samples S_t and A_t to update the policy parameter $\boldsymbol{\theta}$. The expectation of the sample gradient is equal to the actual gradient of the performance measure:

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right], \quad (4.16)$$

where G_t is the return. The policy parameter is updated following:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)} \quad (4.17)$$

where α is the step-size parameter. The complete derivation of the REINFORCE algorithm can be found in Sutton and Barto [11].

4.8.2 Trust region policy optimization

The current state-of-the-art in model free policy gradient algorithms is trust region policy optimization (TRPO) [152], and a simplified version of the same

algorithm called proximal policy optimization [153]. The idea behind TRPO is to improve training stability by limiting parameter updates so that the policy does not change too much at one step. The size of the policy update at each iteration is constrained using the Kullback-Leibler divergence [154], enforcing the distance between the old policy, $\pi_{\theta_{old}}$, and the new policy, π_{θ} , to be small enough:

$$\mathbb{E}_{s,a \sim \pi_{\theta_{old}}} [D_{KL}(\pi_{\theta_{old}}, \pi_{\theta})] \leq \delta, \quad (4.18)$$

where D_{KL} is the Kullback–Leibler divergence, θ is the policy parameters before the update, and δ is the bound on Kullback–Leibler divergence. An importance sampling estimator is introduced to compensate the mismatch between the old policy and the new policy. Concretely, importance sampling is used to estimate the values functions for the new policy π_{θ} , with samples previously collected from the old policy $\pi_{\theta_{old}}$ [155]. The performance measure is defined as:

$$J(\theta) = \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} q_{\theta_{old}}(s, a) \right], \quad (4.19)$$

where $q_{\theta_{old}}(s, a)$ is the action-value function.

TRPO aims to maximize the performance measure $J(\theta)$ subject to the *trust region constraint* defined in equation 4.18. TRPO guarantees a monotonic improvement over policy iteration. See Schulman et al. [152] for a complete description of the algorithm.

4.9 Deep reinforcement learning

Traditional RL algorithms cannot solve decision making problems with high dimensional state space, since carefully chosen hand-engineered feature representations are required [27, 156]. Deep RL combines RL with artificial neural networks to solve complex decision making tasks. The inclusion of neural networks in the RL framework allows the agent to take in very large inputs and make decisions from unstructured input data without manual engineering of the state space [157]. For value-based methods, the neural network is used to estimate the value function, while in policy-based methods the neural

network directly approximates the policy itself. An illustration of a policy-based approach is shown in figure 4.3, where the policy is represented as a neural network. In this case, the network takes the state as input and generates a probability distribution across the action space as output, mapping the state to the parameters of the policy.

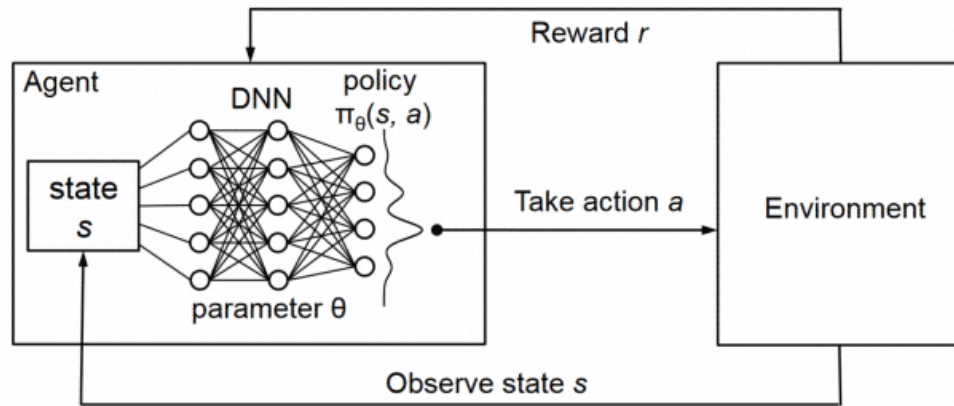


Figure 4.3: Neural network policy parameterization. The neural network maps the state to the policy parameters, where θ are the weights of the neural network. The output is an action sampled from the parameterized policy [12].

4.9.1 Neural networks

A neural network is a network of interconnected functions used to translate a data input into a desired output, learning to perform tasks by considering examples. Neural networks are composed of neurons organized into layers as shown in figure 4.4. There exist three main types of layers: input layer taking the initial data, hidden layers between input and output layers where the computation is done, and output layer producing the result for given inputs. Each layer takes and processes an input to produce an output which serves as the input for the next layer [158].

Each neuron in the input layer takes a single input feature from the data, for example one of the features in the state space. Then, each neuron is connected with each neuron from the next layer through synapses with particular weights, which are adjusted using training data. The weights represent the impact that each neuron has on the neuron from the next layer, amplifying

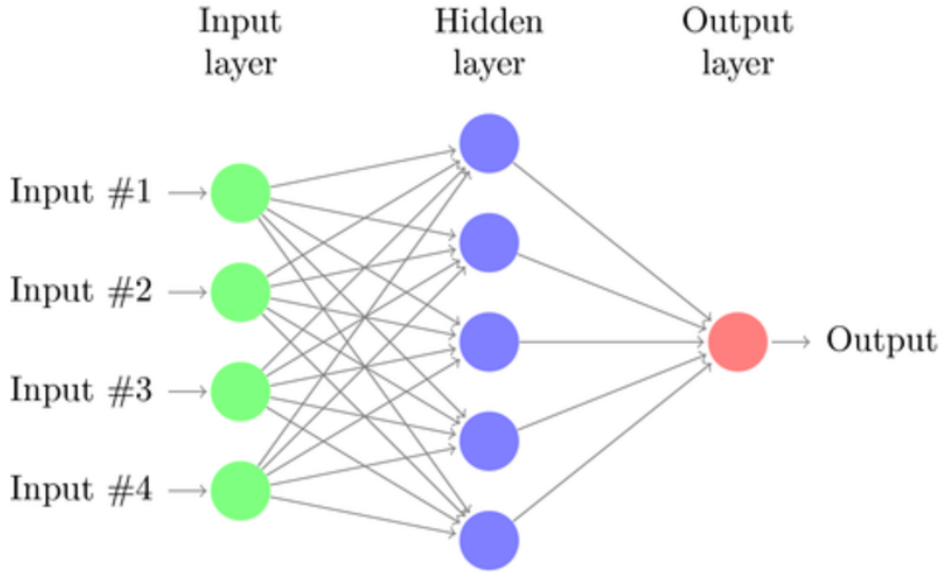


Figure 4.4: Basic neural network architecture [13].

or decreasing the output of the neurons after the learning process. An illustration of a neuron is shown in figure 4.5, where all the neurons from the previous layer are connected with it. The output values from the previous layer are multiplied by the weights assigned to the connections and summarized after that, so each neuron in the hidden and output layer consists of a weighted sum of the neuron's input values. The activation function defines how active this neuron will be based on the summarized value, with rectified linear unit (ReLU) function, $\phi(x) = \max(0, x)$, as one of the most commonly used activation function [159]. The additional node b is the bias, which is a constant that allows to shift the activation function. Bias adjusts the output along with the weighted sum of the inputs to the neuron, helping the model to get better fit for the given data [160]. The weights and biases are the learnable parameters of the neural network model. The output of the neuron k in the layer l with n inputs is defined by:

$$y_k^l = \phi \left(b_k^l + \sum_{i=0}^n w_{ki}^l y_i^{l-1} \right), \quad (4.20)$$

where ϕ is the activation function, b_k^l is the bias, and w_{ki}^l is the weight between

layer input y_i^{l-1} and layer output y_k^l [161].

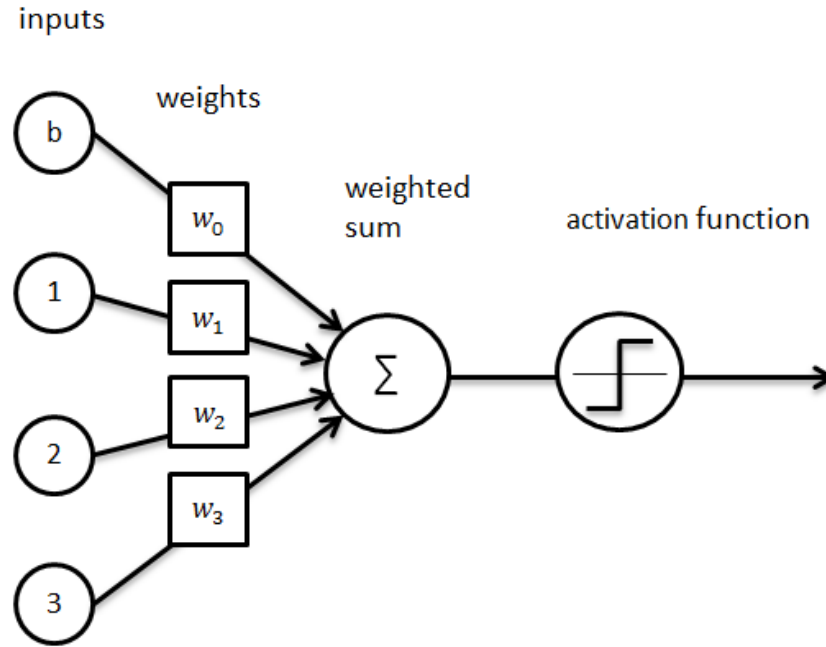


Figure 4.5: Neuron from neural network. Figure adapted from [14].

Bayesian neural networks

A particular type of neural networks are Bayesian neural networks, which are used in paper IV of this thesis to capture the uncertainties in the predicted risks of hypoglycemia and hyperglycemia. Traditional neural networks are trained ignoring any potential uncertainty in the proper weight values. Bayesian neural networks extend the standard neural networks with posterior inference, adding knowledge of confidence and certainty to the results [162]. In this Bayesian framework, the weight and bias parameters of the neural network are represented by probability distributions instead of discrete numbers [163]. Values for weights and biases are sampled from their probability distributions for each propagation through the network, and thus generating different output values [164]. The output values are then represented by a probability distribution on output values, with confidence and uncertainty information for each of the outputs. The Bayesian neural network concept is illustrated in figure 4.6.

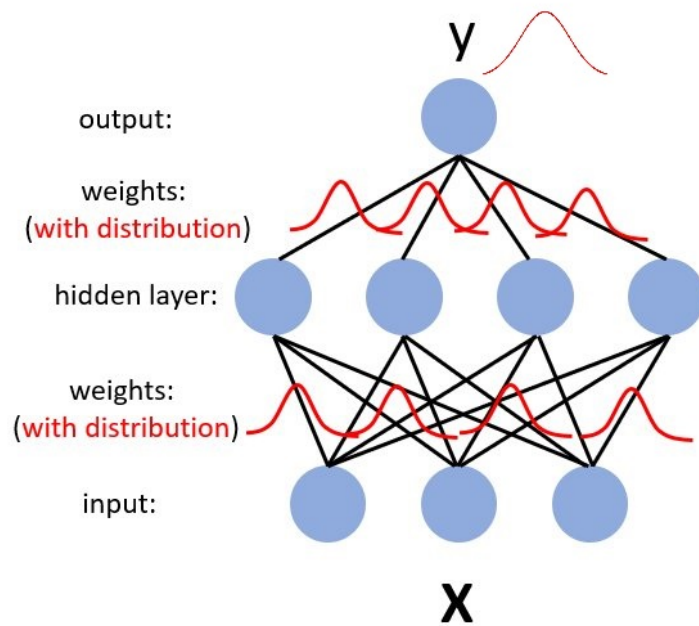


Figure 4.6: A Bayesian neural network with random weights instead of fixed. Figure adapted from [15].

Part II

Summary of research

Chapter 5

Research publications

This chapter offers a summary of the publications enclosed in this thesis as well as a list of the works that were not included.

5.1 Paper summaries

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

This paper performs an exhaustive literature review to evaluate the state-of-the-art of RL approaches to design blood glucose control algorithms for diabetic patients, critically analyzing relevant articles in the research field. Therefore, this paper lays the basis for future research work, supporting the rest of the papers included in this thesis.

The results suggest that the application of RL as a blood glucose controller in the artificial pancreas is still an emerging research field, since there exist few articles in the literature focused on glycemic regulation in diabetes using RL methods. However, the trend suggests that RL algorithms for blood glucose control tasks will be used more frequently in the coming years, since the use of these algorithms have recently increased in the diabetes research area as is shown in 5.1.

Furthermore, the reviewed literature stresses the importance of choosing a good reward function, which is crucial for the correct performance of the

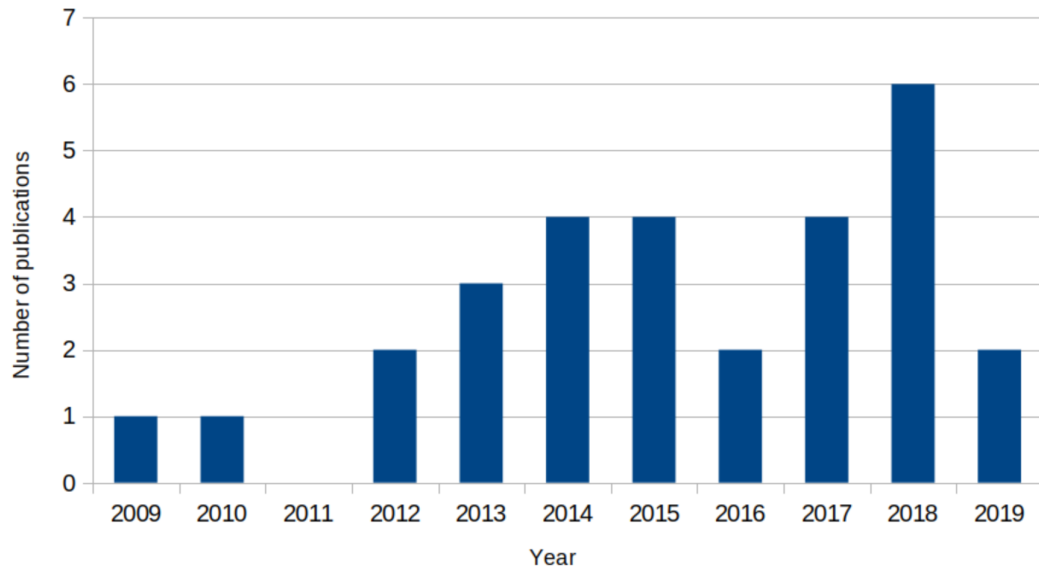


Figure 5.1: Number of publications found in the literature review from 2009 to July 2019 related to RL application in blood glucose regulation for diabetic patients.

algorithm. Therefore, a hand-designed reward function including external knowledge from the diabetes disease is designed in the paper II of this thesis, evaluating the influence of changing the reward function in the blood glucose control task for T1D patients.

Moreover, most of the traditional RL algorithms found in the literature review require carefully chosen feature representations. Therefore, in the paper III of this thesis a RL approach based on deep RL is tested, in which deep learning is used for learning feature representations that in the traditional framework are usually hand-engineered [156].

Finally, there are few papers in the reviewed literature which consider food intakes and physical activity as part of the control problem, despite clear influence of these factors in blood glucose levels. Therefore, a food recommendation system for diabetic patients doing physical activities is implemented in paper IV of this thesis. In future research work, this system will serve as a preliminary stage of a RL algorithm used to optimize the amount of food required to avoid hypoglycemic events during physical activities for patients with T1D.

Contributions by the author

- The initial idea was conceived by all the authors.
- The selection criteria and research questions were established by the second author and me.
- I performed the literature search.
- I analyzed the relevant papers found in the literature.
- The manuscript draft was written by me and edited in collaboration with the coauthors.

Paper II - Controlling Blood Glucose For Patients With Type 1 Diabetes Using Deep Reinforcement Learning – The Influence Of Changing The Reward Function

This paper evaluates the influence of changing the reward function when controlling blood glucose for T1D patients using deep RL. Concretely, the state-of-the-art TRPO algorithm described in section 4.8.2 is used for blood glucose control in this work, while in-silico patients are simulated using the Hovorka model described in section 3.2. In addition, in this paper two hand-designed asymmetric reward function are introduced by the authors, including external knowledge from the diabetes disease to give more penalty to hypoglycemic events. This design decision is a consequence of the importance of avoiding hazardous blood glucose levels reached during hypoglycemia.

The results show the impact on the overall performance of the RL algorithm when changing the reward function. Furthermore, this work shows that the inclusion of diabetes domain knowledge in the reward function reduces both hypoglycemic events and risk indices in general, improving the safety of the in-silico T1D patients.

Contributions by the author

- The initial idea was conceived by me, and further developed with input from the coauthor.
- I designed the proposed reward functions with input from the coauthor.

- I implemented the reward functions and the code used for running the experiments.
- I conducted all the experiments.
- I wrote the first draft of the manuscript and managed the subsequent editing process.

Paper III - In-Silico Evaluation of Glucose Regulation Using Policy Gradient Reinforcement Learning for Patients with Type 1 Diabetes Mellitus

This paper tests and evaluates a deep policy gradient RL algorithm for blood glucose regulation in in-silico T1D patients. Concretely, the state-of-the-art TRPO algorithm described in section 4.8.2 with the reward function proposed in paper II is used for blood glucose control in this work. The Hovorka model described in section 3.2 is used to simulate the in-silico patients. A comparison between TRPO and self-managed control by the patient is conducted in order to evaluate the RL agent performance. In addition, the TRPO algorithm is compared with the MPC approach described in section 2.2.4, which is considered the state-of-the-art in blood glucose control for T1D patients.

The experiments show that TRPO performs better than traditional approaches, while is able to compete with and sometimes outperform MPC in the blood glucose regulation task.

Contributions by the author

- The idea was conceptualized in joint collaboration between Jonas Nordhaug Myhre, Ilkka Kalervo Launonen and me, where some of the main contributions were my ideas.
- Code implementation was carried out by Jonas Nordhaug Myhre and me.
- I ran most of the experiments.
- I wrote most of the first draft of the manuscript in collaboration with Jonas Nordhaug Myhre, and edited in collaboration with the coauthors.

Paper IV - Risk-Averse Food Recommendation Using Bayesian Feedforward Neural Networks for Patients with Type 1 Diabetes Doing Physical Activities

In this paper, a method to select safe and optimal food amounts using Bayesian neural networks for T1D patients doing physical activities is implemented. This recommendation system is conceived to reduce the risk of hypoglycemia during physical activity in patients with T1D, since hypoglycemia is a significant limiting factor of blood glucose regulation in T1D patients and exercise is a major source of hypoglycemia. Bayesian neural networks are used to capture the uncertainties in the predicted risks of hypoglycemia and hyperglycemia.

The results show that the system is able to accurately predict blood glucose levels and therefore recommend food intakes to minimize the risk of hypoglycemia, presenting a potential direction for the future development of safe artificial intelligence methods.

Contributions by the author

- The idea was initially conceived by Phuong Ngo, and further developed with input from me.
- I implemented the UVA/Padova model described in section 3.3, the physical activity model introduced in section 3.3.1, and the CGM model.
- I implemented the risk and reward functions.
- I performed numerical simulations and generated all the data used to train the algorithms.
- Phuong Ngo and I wrote the first draft of the manuscript.

5.2 Other publications

The following papers and works were not included in this thesis:

5. Miguel Tejedor and Jonas Nordhaug Myhre, “**Including T1D knowledge in deep reinforcement learning reduces hypoglycemia**”,

- poster presented at: *International Conference on Advanced Technologies & Treatments for Diabetes*, Madrid, 2020.
6. Miguel Tejedor and Jonas Nordhaug Myhre, “**Controlling Blood Glucose For Patients With Type 1 Diabetes Using Deep Reinforcement Learning - The Influence Of Changing The Reward Function**”, poster presented at: *Northern Lights Deep Learning Conference*, Tromsø, 2020.
 7. Jonas Nordhaug Myhre, Miguel Tejedor, Ilkka Kalervo Launonen and Fred Godtlielsen, “**In-silico Evaluation of Trust Region Policy Optimization Reinforcement Learning for T1DM Closed-Loop Control**”, poster presented at: *International Conference on Advanced Technologies & Treatments for Diabetes*, Berlin, 2019.
 8. Jonas Nordhaug Myhre, Miguel Tejedor, Ilkka Kalervo Launonen and Fred Godtlielsen, “**In-silico Evaluation of Type-1 Diabetes Closed-Loop Control using Deep Reinforcement Learning**”, poster presented at: *Northern Lights Deep Learning Conference*, Tromsø, 2019.
 9. Phuong Dinh Ngo, Miguel Tejedor and Fred Godtlielsen, “**A Decision Support Tool for Optimal Control of Planet Temperature Using Reinforcement Learning**”, published in *17th Conference on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, Las Vegas, 2018.

Chapter 6

Concluding remarks

In this thesis, the blood glucose control problem for T1D patients is addressed using RL methods. To that end, this thesis presents a review and analysis of the state-of-the-art research on RL application in diabetes blood glucose control, contributing to disclose the gaps in this research field while establishing the guidelines for future research directions. However, at the beginning of this research project the application of RL techniques to blood glucose control in diabetes was still not explored in depth by the diabetes research community, while other approaches such as PID and MPC were widely used for this control task. For that reason, not so many relevant publications were found on the literature, although this trend slightly increased in recent years.

Despite only in-silico patients were used in this work, this project is conceived to face a real problem: blood glucose regulation in T1D patients. Consequently, ensure the safety of the patients takes priority over any other matter. In this regard, the thesis proposes a reward function which is designed including external knowledge from the diabetes disease, reducing hypoglycemic events, and improving the safety of the patients. Therefore, this thesis is meeting one of the most crucial and challenging tasks of applying RL methods, namely, the design of the reward function.

Part of the focus of this thesis was dedicated to the application of policy gradient RL algorithms in blood glucose regulation. Concretely, TRPO has been tested and evaluated for glycemic control in T1D patients. The results show promising results even when comparing to MPC controller, which is

considered the state-of-the-art of glucose regulation algorithms.

Despite the evidenced general and diabetes-specific health benefits of physical activities, many diabetic patients are still physically inactive. This is due to the fact that exercise is a major source of hypoglycemia in T1D patients, with the risk of hypoglycemia as a significant limiting factor of blood glucose regulation in diabetic patients. This dissertation also covers the physical activity problem in diabetes with the development of a food recommendation system designed to avoid hypoglycemic events during exercise, improving safety of patients when doing physical activities.

6.1 Limitations

RL algorithms are not well suited to problems with inherent delayed actions, which might be a problem in the blood glucose control task because of the delayed action's effect caused by the use of subcutaneous insulin infusion. RL approaches assume the world is Markovian, i.e., the environment describes a sequence of possible states in which the probability of moving to the next state depends only on the current state and the action taken. Therefore, given the current state and an action, the next state is conditionally independent of all previous states and actions; the state transitions satisfy the Markov property. This is not the case in blood glucose regulation for T1D patients, and therefore the environment is not a MDP. In the artificial pancreas framework, the actions are insulin infusions administered subcutaneously, and there exists a delay in the insulin action. Namely, given an action conformed by an insulin dose, the agent will expect a reaction from the environment, that is a reward and a transition to a new state as a consequence of the previous insulin administered. However, that will not happen since it will take some time until the insulin is absorbed from the subcutaneous tissue, and thus the effects of this action will not affect only to the next state, but to the consecutive future states. Therefore, the next state will not depend only on the current state and action taken, but also depends on previously taken actions, hence the Markov property is violated.

This limitation might be mitigated designing the state and action spaces in a smart way, since the delay in the insulin action depends on the type of insulin administered. For example, if used a short-acting insulin which starts working about 15 minutes after infusion, it would be enough to take

actions every 30 minutes to see part of the effect of each action to some extent, and thus work closer to a MDP. In addition, it is helpful to include information about previous insulin doses in the state space, so even when the blood glucose level does not change too much because of the insulin action delay, we will move to a new state because of the change in the amount of insulin included in the state space.

Another limitation of RL is the amount of data required by the algorithm, especially when using deep RL approaches. These methods are not very efficient in terms of data, since they usually require a large amount of experience during training to converge to a meaningful solution. That is the reason why RL methods work really well when solving video games and other simulated environments where it is feasible to get as much data as needed by just running more simulations. However, RL applications are more limited in the real world, where the access to the data is restricted and it is not easy or even possible to accumulate more experience. For example, the training process would be very limited in the blood glucose control application with real T1D patients, where action exploration might be dangerous for the patient. This limitation might be reduced by training the RL algorithm off-line from past historical data, accelerating the convergence process and thereby facilitating the clinical trials.

6.2 Future work

In this thesis only continuous state and action spaces on a model-free approach were tested. This decision was made based on the nature of the problem, in which a continuous data flow of CGM measurements and insulin infusions is expected, while at the end stage we are not able to know the model of the patient. A natural alternative to this work is to test discrete state and action spaces, simplifying the complexity of the problem. Moreover, it would be worth testing model-based approaches, using the model of the environment to predict future states and rewards. In this regard, it would be possible to learn the model of the environment using machine learning techniques, and thereby predicting future blood glucose values. These predicted blood glucose curves would be part of the state space and would be updated with every new action, alleviating the violation of the Markov property and turning the environment into a MDP. Furthermore, additional relevant in-

formation such as physical activity data might be included in the model for more accurate predictions.

State-of-the-art RL contains several research directions that can be explored. For example, the use of safe RL approaches would be a valuable contribution to the research field [165, 166]. These methods are used to explore only the safe areas of the action space, since performing action exploration on a real patient is very dangerous. Another interesting research direction would be the inverse RL, in which the reward function of an agent is inferred given its policy or observed behaviour [167, 168]. Moreover, in this thesis a general policy is used for the entire control task, so this policy learns how to behave in any possible situation. Hierarchical RL could be used to split the problem into smaller tasks, using different subpolicies for different situations such as food intakes or physical activity [169, 170].

This work only considers a single-hormone artificial pancreas system using insulin to regulate blood glucose levels. It would be interesting to evaluate how RL methods perform in a dual-hormone artificial pancreas system, where the insulin is commonly used together with glucagon in the blood glucose control task [171]. An alternative novel dual-hormone artificial pancreas system combines insulin and pramlintide, an injectable amylin hormone analogue drug for diabetes [172]. Pramlintide slows gastric emptying and suppresses glucagon secretion, alleviating carbohydrate counting and improving glucose control by reducing postprandial hyperglycemia [172]. This system can be further extended to a triple-hormone artificial pancreas where insulin, glucagon, and pramlintide are used in the blood glucose regulation process.

From the diabetes point of view, several factors such as physical activity, stress level, or infections, clearly influence the blood glucose level [173]. However, there exist few examples in the literature which include these factors as a part of the problem. Therefore, the inclusion of some of these factors in the blood glucose control task is a very important future research direction.

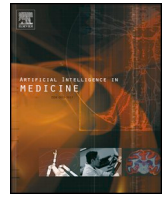
Regarding the food recommendation system, it would be interesting to include a RL agent to automatically optimize the recommended amount of food. In this updated version of the system, the action space would be comprised of the different amounts of food, while the state space would comprise blood glucose levels and physical activity information. Therefore, this agent would be able to automatically recommend the best amount of food to minimize the risk of hypoglycemia during exercise.

Finally, to perform evaluation experiments on diabetic patients may be dangerous and different countries have different execution procedures and regulatory conditions. This is particularly difficult in the case of RL, where a continuous interaction with the patient is needed in order to learn the correct amount of insulin for each situation. Therefore, there exist few clinical trials using RL approaches and it is necessary to perform more clinical trials in order to clinically validate the algorithms. This safety issue might be addressed by a nurse assisting the RL training process. In addition, it would be useful to explore the use of *off-policy* RL methods, in which the algorithms evaluate and improve a policy that is different from the policy that is actually used for action selection. This will allow to use a safer policy when taking actions, while exploring using a different policy.

Part III

Included papers

Paper I



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

Miguel Tejedor^{a,*}, Ashenafi Zebene Woldaregay^a, Fred Godtlielsen^b

^a Department of Computer Science, University of Tromsø-The Arctic University of Norway, Norway

^b Department of Mathematics and Statistics, University of Tromsø-The Arctic University of Norway, Norway

ARTICLE INFO

Keywords:

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

ABSTRACT

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

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

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

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

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

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

* Corresponding author.

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

<https://doi.org/10.1016/j.artmed.2020.101836>

Received 30 July 2018; Received in revised form 3 August 2019; Accepted 19 February 2020

0933-3657/ © 2020 Elsevier B.V. All rights reserved.

1. Introduction

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

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

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

2. Diabetes and blood glucose control using reinforcement learning

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

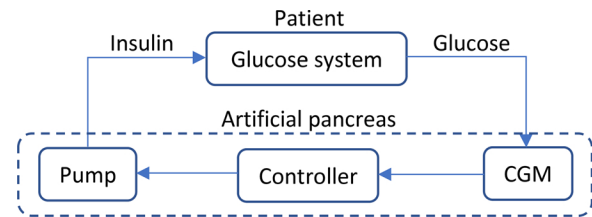


Fig. 2. Blood glucose management based on artificial pancreas.

manual BG control framework is shown in Fig. 1.

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

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

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

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

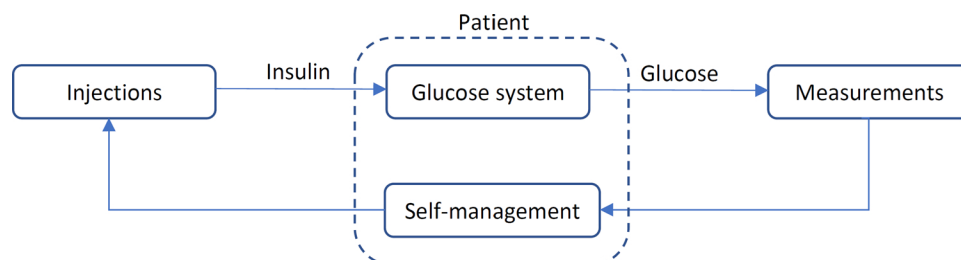


Fig. 1. Self-managed blood glucose control.

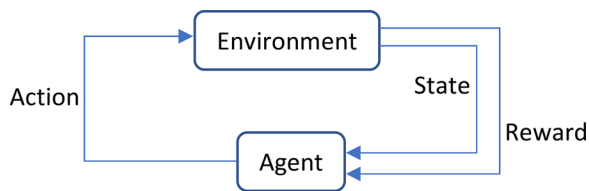


Fig. 3. Reinforcement learning framework.

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

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

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

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

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

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

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

3. Methods

The purpose of the review is to identify, assess and analyze the state-of-the-art RL algorithms and strategies focusing on its applications towards BG control in people with diabetes. As a result, a comprehensive literature search was conducted from 5th June 2019 to 3rd August 2019. The search was performed using different online databases such as ACM digital library, DBLP Computer Science Bibliography, Google Scholar, IEEE Xplore, Journal of American Medical Informatics Association (JAMIA), PubMed and ScienceDirect. Relevant papers were further

extracted from the reference lists of the selected articles. The search process covers a specified timeframe from 1990 to 2018 and considered peer reviewed journal articles and conference proceedings. The search was conducted using different combination of strings along with "reinforcement learning" including "artificial pancreas", "blood glucose control", "closed-loop in diabetes", "decision making in diabetes", "decision support in diabetes", "insulin infusion", "insulin pump" and "personalized feedback in diabetes". For the purpose of effective searching strategy, the search strings were combined using Boolean function such as "And" and "Or". During the search, relevant articles were identified by reviewing the title, keywords, and abstracts for a preliminary filter based on the inclusion and exclusion criteria. A full-text assessment was done on only articles that seemed relevant according to our inclusion and exclusion criteria. Information extraction were also done based on some structured predefined categories that is in line with our inclusion and exclusion criteria, which were defined based on discussions and brainstorming among the authors.

3.1. Inclusion and exclusion criteria

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

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

3.2. Data categorization and data collection

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

3.2.1. Subjects

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

3.2.2. Data sources

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

3.2.3. Preprocessing

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

3.2.4. RL approach

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

3.2.5. Class of RL

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

3.2.6. Exploitation versus exploration

This category encompasses the exploitation-exploration dilemma in

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

3.2.7. State space

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

3.2.8. Action space

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

3.2.9. Planning

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

3.2.10. Generalization approaches

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

3.2.11. Performance metrics or evaluation criteria

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

3.2.12. Model of optimal behavior

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

3.2.13. Reward function

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

3.3. Literature evaluation

Papers were evaluated based on the above predefined categories to evaluate the state-of-the-art approaches and strategies used in RL algorithms for BG control in people with diabetes. The first evaluation and analysis was done based on data characteristics including data sources, subjects and preprocessing approach. The second evaluation and analysis were conducted based on RL strategies including class of RL algorithms and its approaches. The third analysis was carried out based on exploitation versus exploration, to reveal the state-of-the-art approaches in solving the dilemma involved. The fourth evaluation and analysis was conducted based on state and action space including their respective nature and defining parameters. The fifth evaluation and analysis was carried out based on planning approaches employed during development. The sixth evaluation and analysis was conducted based on reward function used to learn the agent. Note that the number

of features extracted might exceed the number of reviewed articles since many features are reported in the literature. Therefore, the number of findings in each category might vary from the number of total studies included in the review, since more than one approach can be considered in the same article.

4. Results

4.1. Relevant literatures

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

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

4.2. Evaluation of literature

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

4.3. Data characteristics

4.3.1. Subjects

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

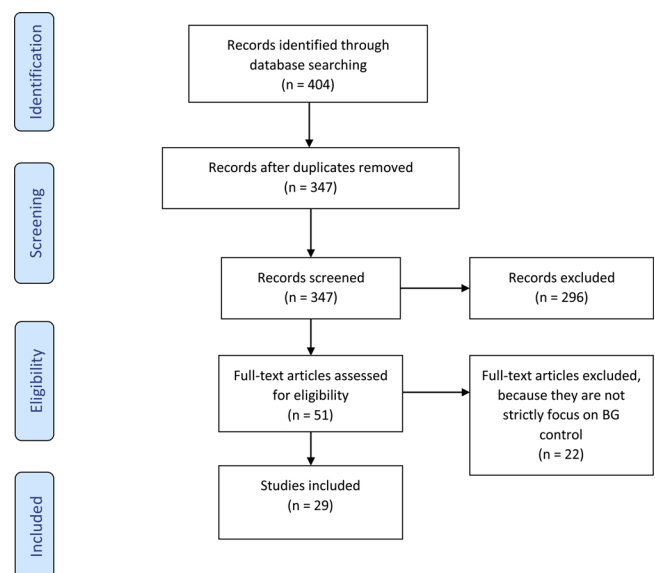


Fig. 4. Flow diagram of the process.

Table 1
Features extracted from the papers.

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

(continued on next page)

Table 1 (continued)

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

Table 2
Data sources used by the studies.

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

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

4.3.2. Data sources

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

4.3.3. Preprocessing

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

4.4. Reinforcement learning strategies

4.4.1. Class of reinforcement learning algorithms

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

Table 3
Preprocessing techniques used in the reviewed literature.

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

Table 4
Class of reinforcement learning algorithms.

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

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

4.4.2. Reinforcement learning approaches

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

4.5. Exploitation-exploration dilemma

In RL algorithm applications, exploitation-exploration dilemma is one of the most important constituents of the design choices. In this regard, Gaussian distribution function (24.25 %, 8/33) is the most popular choice, as shown in Table 6 below. BAL (12.12 %, 4/33) and ϵ -greedy policy (12.12 %, 4/33) are the second most important choices followed by greedy policy (9.09 %, 3/33) and exploration noise (9.09 %, 3/33). Least squares algorithm (6.06 %, 2/33) and uniform distribution (UN) (6.06 %, 2/33), are the fourth most popular choices followed by truncated gaussian (TG) (3.03 %, 1/33) and randomized

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

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

Table 6
Various design choices towards exploitation-exploration dilemma.

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

Table 7
Nature of the state space.

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

decision rule (3.03 %, 1/33). However, surprisingly (15.15 %, 5/33) of the studies either did not report their choices or did not consider it at all.

4.6. State and action spaces

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

4.6.1. State space

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

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

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

4.6.2. Action space

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

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

Table 8
State space defining parameters.

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

Table 9
Nature of the action spaces.

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

Table 10
Action space defining parameters.

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

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

4.7. Planning

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

4.8. Generalization approaches

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

4.9. Performance metrics or evaluation criteria

Various kinds of evaluation criteria have been used to measure the performance of the algorithm towards the specified goal as shown in [Table 13](#) below. In this regard, the most used approach is predefined

Table 11
Planning approaches.

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

Table 12
Generalization Approaches.

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

Table 13
Performance metrics or evaluation approaches.

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

Table 14
Model of optimal behavior.

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

target ranges (14/36, 38.90 %) followed by comparison with reference value (5/36, 13.89 %) and CVGA (4/36, 11.11 %). Low blood glucose index (LBGI) (3/36, 8.33 %) is the fourth most used approaches followed by on-line behavior monitoring (2/36, 5.55 %), high blood glucose index (HBGI) (2/36, 5.55 %), and total daily insulin (TDI) (2/36, 5.55 %). The sixth most used performance metrics are risk function (1/36, 2.78 %), mean amplitude of glucose excursion (MAGE) (1/36, 2.78 %), optimal insulin treatment policy (1/36, 2.78 %) and ability to reject the effect of meal disturbance and to overcome the variability in the glucose-insulin dynamics from patient to patient (1/36, 2.78 %).

4.10. Model of optimal behavior

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

4.11. Reward function

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

5. Discussion

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

Table 15
Reward functions.

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

(continued on next page)

Table 15 (continued)

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

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

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

use the value function to select actions [14]. Another much used generalization technique is GP regression, which is an interpolation method with the interpolated values modeled by a GP governed by prior covariances. Further information about GP in ML can be found in [52].

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

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

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

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

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

Another of the most important choices we must take during RL algorithm implementation is the definition of the state and action spaces. First of all, we found that most of these spaces are defined as continuous. As we mentioned above, this is because of the nature of the problem, in which we expected to have continuous BG measurements and continuous insulin infusion rate. Accordingly, in the BG control problem we will always have at least two information sources: BG level and insulin doses. Therefore, it is natural in the RL framework to relate that information with the states and actions respectively. There are various definitions of the state space in the reviewed literatures, all of them somehow related to the BG level. Concretely, most of the authors define the state space based only on the BG level, followed by these studies in which the states take into account not only the BG level, but also the insulin doses. Regarding the action space, there is only one study in which the actions are not based on the insulin doses [41]. In this paper, the authors take the actions choosing the best glycemic target under different circumstances, leaving the choice of agents and doses to achieve that target to the clinicians. It is worth to mention two articles in which not only the quantity of insulin is used as an action, but also the kind of insulin used [22], such as short-acting, intermediate-acting or long-acting, and even a combination of those

different insulins [35]. However, several additional factors affect the BG level such as CHO intake, PA, stress level, infections, etc [56]. This means that the use of this information is useful in order to face the BG control problem, so we expected to find this data as part of the state and action spaces. However, there are few papers in which for example CHO intakes and PA are included in the state space, although this information is really relevant for the algorithm and facilitates its operation. Furthermore, there is a lack of automatic CHO recording since in those cases this task relies on manual recording. In order to reduce the burden on the patient, as well as increase the objectivity during the control task, the combination of RL algorithms with meal detection algorithms such as [57,58] could be part of future perspectives in order to work in a fully closed-loop system. Concerning the action space, we found that despite the importance of the PA and CHO intakes, there is only one paper in the literature in which this value information is indeed taking into account as part of the actions [37]. This action space is formed by a hypothetical mHealth intervention where insulin injections are administered using an insulin pump while suggestions for food intake and PA are administered using a mobile app, considering all possible combinations of insulin injection, food intake, and PA.

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

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

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

6. Conclusion

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

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

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

Finally, traditional RL algorithms requires carefully chosen feature representations. Therefore, it would be interesting to test other RL approaches such as deep reinforcement learning [45], in which deep learning is used for learning feature representations, that in the

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

Declaration of Competing Interest

None.

Acknowledgements

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

Appendix A. Supplementary data

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

References

- [1] Diabetes. WHO; 2017 [cited 2018 25 June 2018]; Available from: <http://www.webcitation.org/719KGYXpa>.
- [2] International Diabetes Federation. IDF diabetes atlas. 8th edn Brussels, Belgium: International Diabetes Federation; 2017.
- [3] ADA. American diabetes association research programs 2018-01-17 [cited 2018 24 July 2018]; Available from: <http://www.webcitation.org/719lz6gfm>.
- [4] Kavakiotis I, et al. Machine learning and data mining methods in diabetes research. *Comput Struct Biotechnol J* 2017;15:104–16.
- [5] Oviedo S, et al. A review of personalized blood glucose prediction strategies for T1DM patients. *Int J Numer Method Biomed Eng* 2017;33(6).
- [6] Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res* 2018;20(5):e10775.
- [7] Lunze K, et al. Blood glucose control algorithms for type 1 diabetic patients: a methodological review. *Biomed Signal Process Control* 2013;8(2):107–19.
- [8] Bothe MK, et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev Med Devices* 2013;10(5):661–73.
- [9] Bekiari E, et al. Artificial pancreas treatment for outpatients with type 1 diabetes: systematic review and meta-analysis. *BMJ* 2018;361:k1310.
- [10] Hovorka R. Closed-loop insulin delivery: from bench to clinical practice. *Nat Rev Endocrinol* 2011;7(7):385–95.
- [11] Kumareshwaran K, Evans ML, Hovorka R. Closed-loop insulin delivery: towards improved diabetes care. *Discov Med* 2012;13(69):159–70.
- [12] Farmer Jr TG, Edgar TF, Peppas NA. The future of open- and closed-loop insulin delivery systems. *J Pharm Pharmacol* 2008;60(1):1–13.
- [13] Adibi SE. Mobile health: a technology Road map. Springer series in bio-/Neuroinformatics. 1 ed. Springer International Publishing; 2015.
- [14] Sutton RS, Barto AG. Reinforcement learning: an introduction. 1998.
- [15] McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;276–82.
- [16] De Paula M, Acosta GG, Martínez EC. On-line policy learning and adaptation for real-time personalization of an artificial pancreas. *Expert Syst Appl* 2015;42(4):2234–55.
- [17] Lehmann ED, Deutsch T. A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. *J Biomed Eng* 1992;14(3):235–42.
- [18] Akbari Torkestani J, Ghanaat Pisheh E. A learning automata-based blood glucose regulation mechanism in type 2 diabetes. *Control Eng Pract* 2014;26:151–9.
- [19] Daskalaki E, Diem P, Mougiakakou SG. An Actor-Critic based controller for glucose regulation in type 1 diabetes. *Comput Methods Programs Biomed* 2013;109(2):116–25.
- [20] Man CD, et al. The UVA/PADOVA type 1 diabetes simulator: new features. *J Diabetes Sci Technol* 2014;8(1):26–34.

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

Paper II

Controlling Blood Glucose For Patients With Type 1 Diabetes Using Deep Reinforcement Learning – The Influence Of Changing The Reward Function

Miguel Tejedor - miguel.tejedor@uit.no, Jonas Nordhaug Myhre - jonas.n.myhre@uit.no

Abstract

Reinforcement learning (RL) is a promising direction in adaptive and personalized type 1 diabetes (T1D) treatment. However, the reward function – a most critical component in RL – is a component that is in most cases hand-designed and often overlooked. In this paper we show that different reward functions can dramatically influence the final result when using RL to treat in-silico T1D patients.

1 Introduction

Reinforcement learning (RL) is a separate direction in machine learning where the aim is to understand and automate goal-directed learning and decision-making [13]. In combination with recent advances in deep learning, deep reinforcement learning has emerged as a very powerful tool for difficult control tasks [11, 6].

The artificial pancreas (AP) is a system involving an insulin pump, a continuous glucose monitor and a control algorithm to release insulin in response to changing blood glucose (BG) levels mimicking a human pancreas. Several works have shown promising results using RL for the AP [2, 7, 8, 12], but the main focus of these algorithms have been on fitting the RL framework to the case of type 1 diabetes (T1D). In this work we focus on the *reward function*, an often overlooked component of empirical reinforcement learning. It is well known that the success of a RL application strongly depends on how well the reward signal frames the goal of the application’s designer and how well the signal assesses progress in reaching that goal [18]. In the diabetes case it is particularly the contrasting problems of hyper- and hypoglycemia – too high or too low BG levels – that is problematic for RL

applications. In fact, hypoglycemia is a commonly reported problem and one of the acutest complications of all types of diabetes. We propose several new reward functions suited for T1D, and perform in-silico experiments testing different reward functions on the trust-region policy optimization (TRPO) algorithm [9] using the Hovorka model [4].

Our experiments demonstrate that focusing on reward functions that contain more domain knowledge, such as stronger penalties for reaching low BG levels, is crucial.

2 Deep reinforcement learning: Policy optimization and TRPO

Policy gradient algorithms consider *parametric policies* which are optimized using gradient ascent on a given *performance measure*. The most common choice for the performance measure is the expected return of the start state s_0 , given as $J(\theta) = v_\pi(s_0) = \mathbb{E}_\pi [R_0 + \gamma R_1 + \gamma^2 R_2 + \dots]$.

Using policy gradient algorithms yield several benefits: the *policy gradient theorem*, application of RL to continuous action spaces and a naive extension to deep learning using neural network to parameterize the policies.

Furthermore, a key point of using policy gradient algorithms is the policy gradient theorem [13]:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla \pi(a|s, \theta).$$

This states that the gradient of the performance measure is proportional to the gradient of the policy itself. This allows the use of any differentiable policy parameterization. Furthermore, the

policy gradient theorem is constructive, so it directly yields a simple sample-based algorithm, REINFORCE [16], omitted here for brevity. This algorithm has been well studied and a number of improvements and suggestions have been proposed, see e.g. [9, 10, 5]. The current state-of-the-art in *model free* policy gradient algorithms is *Trust Region Policy Optimization* (TRPO) by Schulman et al. [9] and a simplified version, *Proximal Policy Optimization* [10]. In this work we restrict our attention to TRPO.

Trust region policy optimization is a policy gradient algorithm where each update of the policy is guaranteed to improve the performance. This guarantee is achieved, by enforcing the Kullback-Leibler divergence between the old and the updated policy to be small:

$$\begin{aligned} & \underset{\theta}{\text{maximize}} && \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} Q_{\theta_{old}}(s,a) \right] \\ & \text{subject to} && \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} [D_{KL}(\pi_{\theta_{old}}, \pi_{\theta})] \leq \delta. \end{aligned} \quad (1)$$

We refer the reader to Schulman et al. [9] for further details. The policy $\pi_{\theta}(a|s)$ is a Gaussian policy:

$$\pi_{\theta}(a|s) = \frac{1}{\sigma(s, \theta) \sqrt{2\pi}} \exp\left(-\frac{(a - \mu(s, \theta))^2}{2\sigma(s, \theta)^2}\right),$$

where $\sigma(s, \theta)$ and $\mu(s, \theta)$ are feature extractors. We use neural network feature extractors in this work.

3 Reward functions

We consider hyperglycemia as values above $bg_{hyper} = 180$ mg/dL, hypoglycemia as values below $bg_{hypo} = 72$ mg/dL and severe hypoglycemia as values below $bg_{hypo-} = 54$ mg/dL. Thus, normoglycemic range are values between $[bg_{hypo}, bg_{hyper}]$ mg/dL, with a target value $bg_{ref} = 108$ mg/dL. The nine different proposed and tested reward functions can be further divided into two categories: (1) Symmetric reward functions – hyper- and hypoglycemia are equally penalized by the rewards.

Absolute reward [17]: $|bg - bg_{ref}|$

Binary:

$$\begin{cases} 1 & : bg \in [bg_{hypo}, bg_{hyper}] \\ 0 & : otherwise \end{cases}$$

Binary tight:

$$\begin{cases} 1 & : bg \in [bg_{ref} - 10, bg_{ref} + 10] \\ 0 & : otherwise \end{cases}$$

Gaussian reward [2]: $\exp\left(-\frac{1}{\sigma^2}(bg - bg_{ref})^2\right)$

Squared reward: $-(bg - bg_{ref})^2$

(2) The second category is asymmetric reward functions – hand-designed reward functions including external knowledge from the diabetes disease to give more penalty to hypoglycemic events.

T1D reward: Linear function with positive reward for normoglycemic range. Exponential function with negative reward for hypoglycemia, while 0 reward for hyperglycemia. Really negative reward for severe hypoglycemia.

$$\begin{cases} -100 & : bg < bg_{hypo-} \\ \exp\left(\frac{\log(140.9)}{bg_{hypo}}bg\right) - 140.9 & : bg \in [bg_{hypo-}, bg_{hypo}] \\ \frac{1}{36}bg - 2 & : bg \in [bg_{hypo}, bg_{ref}] \\ -\frac{1}{72}bg + \frac{5}{2} & : bg \in [bg_{ref}, bg_{hyper}] \\ 0 & : bg > bg_{hyper} \end{cases}$$

Tight T1D reward: Hypoglycemia considered as values below $bg_{hypo_t} = 90$ mg/dL in order to be even more aggressive against hypoglycemic events.

$$\begin{cases} -100 & : bg < bg_{hypo-} \\ \exp\left(\frac{\log(117.5)}{bg_{hypo_t}}bg\right) - 117.5 & : bg \in [bg_{hypo-}, bg_{hypo_t}] \\ \frac{1}{18}bg - 5 & : bg \in [bg_{hypo_t}, bg_{ref}] \\ -\frac{1}{72}bg + \frac{5}{2} & : bg \in [bg_{ref}, bg_{hyper}] \\ 0 & : bg > bg_{hyper} \end{cases}$$

Hovorka reward: Based on the nonlinear model predictive control from [4].

$$-(bg - y(t))^2$$

$y(t)$ is the desired glucose profile. When BG levels are above the desired level $y(t)$ linearly decrease, while for BG values below target value $y(t)$ exponentially increases [4].

Risk reward [1]: $-10(1.509(\log(bg))^{1.084} - 5.381)^2$

4 Experimental setup

Simulation environment We use the Hovorka simulator as described in Wilinska et al. [15] and Hovorka et al. [4]. The simulator is implemented in Python and the TRPO agent is trained using the open source reinforcement learning toolbox *garage*¹ [3].

Experiment protocol and scenarios Each episode of the simulations consists of a single day

¹<https://github.com/rlworkgroup/garage>.

plus 12 hours into the next day. Four meals are given at [08:00, 12:00, 18:00, 22:00] with a uniform chance of moving the meal back or forward 30 minutes. Each meal is fixed at 40, 80, 60 and 30 grams of carbohydrates with a uniform ± 20 gram disturbance. Finally, we have a $\pm 30\%$ carbohydrate counting error, meaning that the carbohydrate amount used to calculate the bolus insulin dose might be 30% higher or lower than the true carbohydrate amount.

Performance measures and testing We test the algorithm on a fixed scenario consisting of 100 random meal-days generated with a fixed random seed. To measure the performance of our simulations, we use time-in-range and time-in-hypoglycemia as the performance measures, where we want to maximize the former and minimize the latter. We also include low blood glucose risk index (LBGI), high blood glucose risk index (HBGI), risk index (RI) and the coefficient of variation (CoV), all described in Clarke and Kovatchev [1].

5 Results and discussion

In this work we test and compare different reward functions using TRPO on the original Hovorka in-silico patient, [4], in order to show the importance of the reward function design.

In the experiments we consider two cases, with different insulin-to-carbohydrate ratio (ICR) used to calculate pre-meal bolus insulin doses. This ratio specifies the number of grams of carbohydrate covered by each unit of insulin, see e.g. [14].

Given the fact that we are in this work considering a single-hormone AP, the only available action for the algorithm when the BG is too low or approaching low levels is to turn off the insulin pump. Due to this the actual ICR used during meals will have a strong influence on the overall result. Especially the severity of carbohydrate counting errors, which we include in our simulations, will be affected by different ICRs.

5.1 Case 1: 30g/U ICR

We start with a 30g/U ICR. This translates to the in-silico Hovorka patient taking 1 unit of insulin for each 30 grams of carbohydrate intake. We run

the TRPO algorithm for 100 iterations using all the reward functions described in Section 3. Figure 1 shows mean BG level values for the different reward functions used within TRPO and the basal-bolus regimen. The mean BG values show good performance for all the different reward functions and basal-bolus regimen when using 30 g/U ICR as shown in figure 1, spending most of the time within range. However, most of the symmetric rewards show lower values than the asymmetric rewards, resulting in a higher hypoglycemia risk. Only the *tight binary reward function* shows comparable results to the asymmetric reward functions, keeping mean BG values closer to the target value. Results from these experiments are summarized in Table 1.

TRPO outperforms the basal-bolus regimen in terms of time-in-range for all the reward functions tested. However, that is not the case in terms of hypoglycemic events, where the symmetric rewards struggle to avoid hypoglycemia. Only the symmetric *binary tight reward* function presents competitive results avoiding hypoglycemic excursions in similar terms to asymmetric rewards. The *risk reward* function actually increases the time spent in hypoglycemia, showing worse results than the rest of the asymmetric rewards. The opposite happens with hyperglycemic excursions, where the symmetric reward functions show better performance avoiding hyperglycemia. This is because the symmetric reward functions deal equally with hypo- and hyperglycemia events, while asymmetric reward functions are designed taking into account external knowledge from the diabetes problem. In this work, this external information consists of higher penalty to hypoglycemia than to hyperglycemia, which is translated into safer behaviour reducing the time spent in hypoglycemic events. This is also reflected in the risk factors, where the asymmetric reward functions are more robust against risk of hypoglycemia than the symmetric reward functions, while both kind of functions show similar performance in terms of hyperglycemic risk. Therefore, the overall risk factor is lower for the asymmetric rewards. Finally, the asymmetric reward functions where hypoglycemia is penalized more than hyperglycemia also present lower CoV, and only the asymmetric risk reward function show similar results to the symmetric functions.

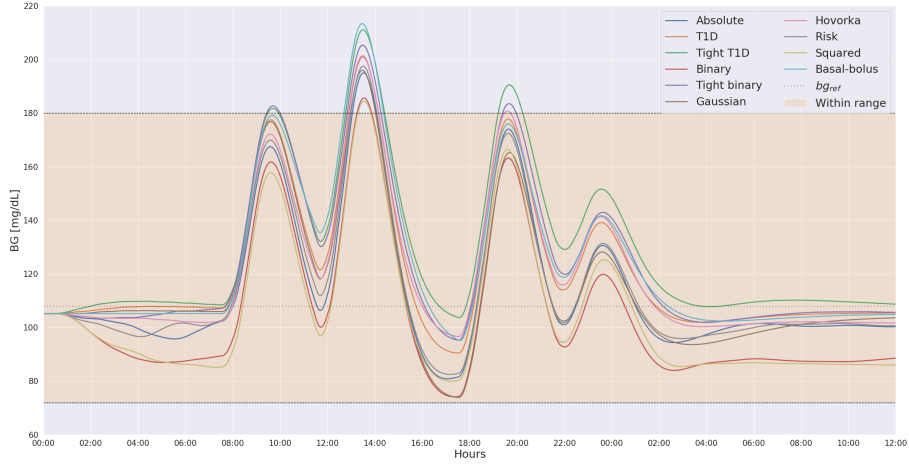


Figure 1: Mean blood glucose levels using TRPO with different reward functions, averaged over 100 episodes. Each test episode runs for one and a half day, a total of 36 hours, to include the effects of the algorithm after the last meal. The Insulin-to-carbohydrate ratio is fixed at 30 g/U.

Treatment	Time-in-range	-hypo	-hyper	LBGI	HBGI	RI	CoV
Basal-bolus	83.45±7.38	2.42±4.9	14.13±7.07	0.87±0.83	4.62±1.45	5.5±1.63	27.61
Absolute	86.45±6.04	4.50±5.74	9.06±4.31	1.30±0.82	4.23±0.96	5.53±1.09	27.31
T1D	88.10±4.78	0.54±1.8	11.36±4.58	0.47±0.31	4.34±0.98	4.81±1.02	25.27
Tight T1D	83.57±5.31	0.0±0.0	16.43±5.31	0.12±0.09	4.70±1.17	4.82±1.17	25.13
Binary	86.92±6.47	6.68±6.39	6.40±3.96	2.38±0.6	3.83±0.94	6.20±1.0	29.56
Tight binary	85.03±5.51	0.55±2.09	14.41±5.31	0.41±0.3	4.79±1.11	5.19±1.16	26.43
Gaussian	84.39±7.16	6.15±6.47	9.46±4.13	1.52±0.98	4.24±1.06	5.76±1.44	27.64
Hovorka	88.95±3.94	0.0±0.0	11.05±3.94	0.41±0.16	4.20±0.81	4.61±0.82	25.34
Risk	86.60±5.79	3.75±4.81	9.65±4.31	1.13±0.63	4.35±1.0	5.48±1.06	27.28
Squared	89.81±5.22	4.13±5.12	6.06±3.79	2.30±0.55	3.62±0.89	5.91±0.89	29.01

Table 1: Summary of results for 30g/U insulin-to-carbohydrates ratio. Mean values ± standard deviation of 100 runs with each episode running for one and a half day, a total of 36 hours.

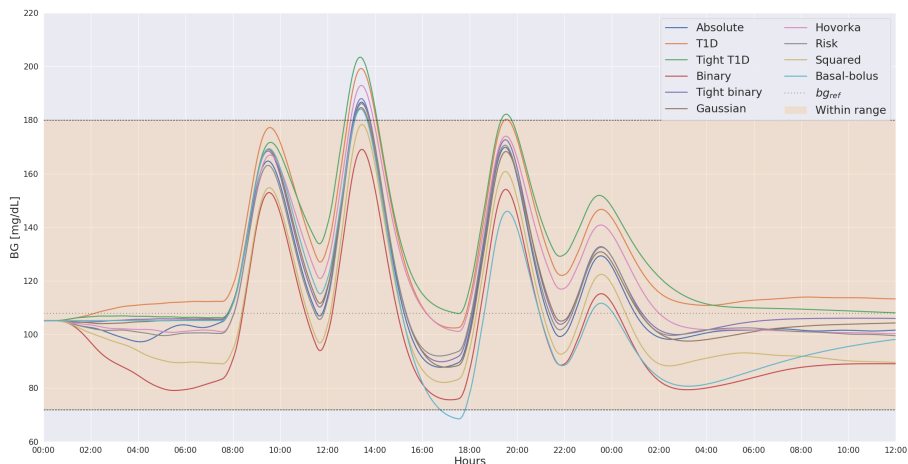


Figure 2: Mean blood glucose levels using TRPO with different reward functions, averaged over 100 runs. Each test episode runs for one and a half day, a total of 36 hours. Insulin-to-carbohydrate ratio fixed to 25g/U.

5.2 Case 2: 25g/U ICR

We select a 25g/U ICR for the second set of experiments. That means the in-silico Hovorka patient uses 1 unit of insulin for each 25 grams of carbohydrates intakes. Therefore, in this set of experiments the patient uses more units of insulin to deal with the same amount of carbohydrates. The mean BG level values for the the basal-bolus regimen and the different reward functions used within TRPO during these experiments are shown in figure 2.

TRPO shows good performance with mean BG values within range most of the time. However, symmetric reward functions lead to lower BG values and then higher risk of hypoglycemia, while asymmetric reward functions stay in safer glucose levels.

Results summarized in table 2 show TRPO clearly improving time spent in target range while reducing hypoglycemic events in comparison with the basal-bolus regimen, which in this case is not able to maintain safe BG values.

Furthermore, the asymmetric reward functions taking into account the importance of avoiding hypoglycemia perform better than symmetric reward functions, reducing hypoglycemic events. This is also reflected in the reduced overall risk index. The symmetric reward functions deals better with high BG values, reducing the time spent in hy-

perglycemia. However, in spite of this reduction in time spent in hyperglycemia, the risk of hyperglycemia is similar for symmetric and asymmetric reward functions, with the *asymmetric T1D reward* function showing the lowest risk. Therefore, asymmetric reward functions results in lower total risk factor. Regarding the coefficient of variation, the asymmetric T1D reward function shows better performance decreasing variance, while symmetric binary reward function presents a CoV value closer to the basal-bolus strategy. The rest of the reward functions present similar results, reducing the CoV with respect to the basal-bolus regimen.

6 Conclusions

In this work we have shown that changing the reward function will have an impact on the overall performance of RL agents for the AP framework. Furthermore, we tested the influence of including domain knowledge in the reward function, and we observed that this both reduces the hypoglycemic events and risk indices in general, ultimately improving the safety of the in-silico T1D patients.

Treatment	Time-in-range	-hypo	-hyper	LBGI	HBGI	RI	CoV
Basal-bolus	79.22±12.14	14.65±12.73	6.14±4.32	2.99±1.98	3.63±1.15	6.62±2.32	29.1
Absolute	91.41±4.69	1.62±3.41	6.98±3.74	0.79±0.42	3.73±0.83	4.52±0.87	24.28
T1D	87.81±4.92	0.08±0.62	12.11±4.85	0.25±0.31	2.87±0.74	3.12±0.79	22.39
Tight T1D	86.90±5.4	0.15±0.89	12.95±5.42	0.19±0.21	3.91±1.0	4.11±1.04	23.57
Binary	91.99±5.74	4.86±5.72	3.15±2.70	2.72±0.48	2.93±0.78	5.66±0.85	27.29
Tight binary	90.48±4.82	1.78±3.71	7.74±3.7	0.59±0.45	3.73±0.80	4.33±0.88	23.47
Gaussian	91.44±4.95	1.44±3.27	7.12±3.75	0.70±0.42	3.59±0.8	4.29±0.88	23.71
Hovorka	90.97±4.23	0.09±0.55	8.94±4.23	0.46±0.22	3.66±0.79	4.11±0.80	24.02
Risk	91.52±4.49	1.57±3.08	6.91±3.69	0.74±0.41	3.53±0.76	4.26±0.80	23.85
Squared	92.82±4.73	2.54±4.4	4.64±3.2	1.67±0.45	3.28±0.79	4.95±0.81	25.79

Table 2: Summary of results for 25g/U insulin-to-carbohydrates ratio. Mean values \pm standard deviation of 100 runs with each episode running for one and a half day, a total of 36 hours.


References

- [1] W. Clarke and B. Kovatchev. Statistical tools to analyze continuous glucose monitor data. *Diabetes technology & therapeutics*, 11(S1):S-45, 2009.
- [2] M. De Paula, L. O. Avila, and E. C. Martinez. Controlling blood glucose variability under uncertainty using reinforcement learning and gaussian processes. *Applied Soft Computing*, 35:310–332, 2015.
- [3] Y. Duan, X. Chen, R. Houthoofd, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1329–1338, 2016.
- [4] R. Hovorka, V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. O. Federici, T. R. Pieber, H. C. Schaller, L. Schaupp, T. Vering, et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological measurement*, 25(4):905, 2004.
- [5] S. M. Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.
- [6] S. S. Mousavi, M. Schukat, and E. Howley. Deep reinforcement learning: an overview. In *Proceedings of SAI Intelligent Systems Conference*, pages 426–440. Springer, 2016.
- [7] J. N. Myhre, I. K. Launonen, S. Wei, and F. Godtliebsen. Controlling blood glucose levels in patients with type 1 diabetes using fitted q-iterations and functional features. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.
- [8] P. D. Ngo, S. Wei, A. Holubová, J. Muzik, and F. Godtliebsen. Reinforcement-learning optimal control for type-1 diabetes. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 333–336. IEEE, 2018.
- [9] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.
- [10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [11] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.
- [12] Q. Sun, M. V. Jankovic, and S. G. Mougiakakou. Reinforcement learning-based adaptive insulin advisor for individuals with type 1 diabetes patients under multiple daily injections therapy. *arXiv preprint arXiv:1906.08586*, 2019.
- [13] R. S. Sutton and A. G. Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.
- [14] J. Walsh and R. Roberts. *Pumping insulin: everything you need for success on a smart insulin pump*. Torrey Pines Press, 2006.
- [15] M. E. Wilinska, L. J. Chassin, C. L. Acerini, J. M. Allen, D. B. Dunger, and R. Hovorka. Simulation environment to evaluate closed-loop insulin delivery systems in type 1 diabetes. *Journal of diabetes science and technology*, 4(1):132–144, 2010.
- [16] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [17] S. Yasini, M. Naghibi-Sistani, and A. Karimpour. Agent-based simulation for blood glucose control in diabetic patients. *International Journal of Applied Science, Engineering and Technology*, 5(1):40–49, 2009.
- [18] H. Zou, T. Ren, D. Yan, H. Su, and J. Zhu. Reward shaping via meta-learning. *arXiv preprint arXiv:1901.09330*, 2019.

Paper III

Article

In-Silico Evaluation of Glucose Regulation Using Policy Gradient Reinforcement Learning for Patients with Type 1 Diabetes Mellitus

Jonas Nordhaug Myhre ^{1,*†}, Miguel Tejedor ^{2†}, Ilkka Kalervo Launonen ³, Anas El Fathi ⁴  and Fred Godtliebsen ⁵

¹ Department of Physics and Technology, UiT-The Arctic University of Norway, 9019 Tromsø, Norway

² Department of Computer Science, UiT-The Arctic University of Norway, 9019 Tromsø, Norway; miguel.tejedor@uit.no

³ Department of Clinical Research, The University Hospital of North-Norway, 9019 Tromsø, Norway; ilkka.launonen@unn.no

⁴ The McGill Artificial Pancreas Lab, McGill University, Montreal, QC H3A 2B4, Canada; anas.elfathi@mail.mcgill.ca

⁵ Department of Mathematics and Statistics, UiT-The Arctic University of Norway, 9019 Tromsø, Norway; fred.godtliebsen@uit.no

* Correspondence: jonas.n.myhre@uit.no

† These authors contributed equally to this work.

Received: 11 August 2020; Accepted: 10 September 2020; Published: 11 September 2020



Abstract: In this paper, we test and evaluate policy gradient reinforcement learning for automated blood glucose control in patients with Type 1 Diabetes Mellitus. Recent research has shown that reinforcement learning is a promising approach to accommodate the need for individualized blood glucose level control algorithms. The motivation for using policy gradient algorithms comes from the fact that adaptively administering insulin is an inherently continuous task. Policy gradient algorithms are known to be superior in continuous high-dimensional control tasks. Previously, most of the approaches for automated blood glucose control using reinforcement learning has used a finite set of actions. We use the Trust-Region Policy Optimization algorithm in this work. It represents the state of the art for deep policy gradient algorithms. The experiments are carried out in-silico using the Hovorka model, and stochastic behavior is modeled through simulated carbohydrate counting errors to illustrate the full potential of the framework. Furthermore, we use a model-free approach where no prior information about the patient is given to the algorithm. Our experiments show that the reinforcement learning agent is able to compete with and sometimes outperform state-of-the-art model predictive control in blood glucose regulation.

Keywords: reinforcement learning; Type 1 Diabetes Mellitus; policy gradient; deep learning; artificial pancreas

1. Introduction

Type 1 Diabetes Mellitus (T1DM) is a metabolic disease caused by the autoimmune destruction of insulin-producing beta cells in the pancreas [1]. The role of insulin is to utilize and transport glucose [2]. T1DM patients need life-long external insulin therapy to regulate their blood glucose concentrations. Without insulin, T1DM patients suffer from chronic high blood glucose levels (hyperglycemia) and, conversely, too much insulin causes hazardous low blood glucose levels (hypoglycemia). In fact, fear of hypoglycemia is a major limiting factor of glucose regulation in T1DM [3].

Treatment of T1DM mainly consists of either multiple daily insulin injections (MDI), or through a pump providing a continuous insulin infusion (CSII) [4]. MDI therapy consists of a basal-bolus insulin regimen, where patients take a basal long-acting insulin dose approximately once a day to regulate fasting blood glucose levels, and short-acting insulin boluses around mealtimes to quickly reduce the impact of carbohydrate intake. Bolus insulin is also used for minor adjustments during the day when the blood glucose level is too high. CSII treatment is a different strategy where the patient instead has an insulin pump that continuously infuses insulin. The pump delivers both basal and bolus doses, where the basal rate consists of regularly infused short-acting insulin doses, while the boluses are activated by the user together with meal intakes and to account for hyperglycemia. In both cases, the insulin is administered subcutaneously, i.e., in the fatty tissue just below the skin. In combination with this, the blood sugar levels have to be monitored. This is either done several times per day via manual finger-prick measurements, or via a continuous glucose monitor (CGM) embedded in the subcutaneous tissue [5]. Finally, in collaboration with a physician, the T1DM patient will design a treatment plan based on their individual needs and self-administer insulin according to the plan and self-measured blood glucose concentrations. The goal of the insulin treatment strategy is to keep the blood glucose levels within the normoglycemic range between 70 and 180 mg/dL [6,7].

Due to the demands of everyday life and the fact that patients to a large degree are responsible for treating themselves, the decisions related to the insulin treatment are thus based partly on hard calculations, personal and medical experience, rules of thumb, and, in some cases, just pure guesswork. Although this results in effective treatment when done correctly, it is extremely time-consuming and a constant burden for the patients.

With the improvement of modern treatment equipment, the combination of an insulin pump and CGM invites the addition of a third element, namely a control algorithm to substitute the operation of beta cells in the healthy pancreas. These three elements constitute the artificial pancreas [8,9]. A pump delivers the insulin subcutaneously, which causes delay in the insulin's action compared to normal insulin secretion where the pancreas releases it to the liver via the portal vein. A simple reactive controller based on momentary blood glucose change cannot thus keep up with the delay to avoid high glucose levels after meals. There exists also a delay associated to the subcutaneous blood glucose measurements from the CGM. Besides the insulin action and CGM delays, there are also dynamic factors that cause variation in the patient-specific parameters and complicate the automation of the control process. The effect of exercise on the blood glucose and insulin dynamics is particularly difficult to model and it is a major source of hypoglycemia [10]. .

The only commercial available artificial pancreas system, the Medtronic 670G [11], as well as several do-it-yourself systems, see e.g., [12] and academic systems, e.g., [13] are all hybrid closed loop systems. A hybrid system means that the patient has to provide information to the system about the number of carbohydrates ingested during a meal. A bolus can then be provided, either automatically by the system or by the patient itself based on the estimated carbohydrate amount. This setup is highly prone to errors due to the difficulties of carbohydrate counting in everyday situations [14]. This difficulty is well established in the scientific literature, where the true effect of these errors is still a topic of debate. Among others, Deeb et al. [15] report that carbohydrate-counting errors are not correlated with meal size, while Vasiloglou et al. [16] found that larger meals led to larger estimation errors. On the other hand, Kawamura et al. [17] found that meals with small amounts of carbohydrate tended to be overestimated. Finally, Reiterer et al. [14] note that random errors, such as faulty carb-counting, as opposed to systematic bias errors, are more detrimental to glycemic control. Under- and over-bolusing due to these difficulties presents a significant risk of postprandial hyperglycemia and hypoglycemia. The current strategy to compensate for the counting errors is to let the artificial pancreas temporarily change the basal insulin rate. Despite these issues, the artificial pancreas is currently the most promising option for persons struggling with T1DM with multiple studies showing promising results, both clinical and in-silico [12,18–20].

There are currently two dominant artificial pancreas controller algorithm paradigms, proportional-integral-derivative (PID) control, [11,21], and MPC [22,23]. Model predictive control, in particular, uses a dynamic model with patient-specific parameters to predict the blood glucose curve into the future, where the prediction window is typically four hours, after which the fast-acting insulin's effect has mostly subsided [24]. Afterwards, if the predicted blood glucose curve and its final value is off the glucose target, MPC calculates an optimized sequence of basal rate actions on the model to correct the prediction towards the target while avoiding hypoglycemia. The first action of this sequence is then picked to change the basal rate momentarily, and the whole process is repeated after a while, usually every five or 10 min. The MPC approaches require a good model of the dynamics. In the artificial pancreas system, MPC algorithms are based on glucose-insulin regulatory models that are not able to capture external perturbations, so these algorithms are limited to compensate for the incomplete model used in the artificial pancreas application [25].

In addition to PID control and MPC, there have been investigations into fuzzy logic [26], and more recently techniques from machine learning and statistics, such as Aiello et al. [27], who proposed a blood glucose forecasting approach based on recurrent neural networks. Similarly, Li et al. [28] created a deep learning based forecasting framework based on convolutional neural networks. The control algorithm used in the artificial pancreas system has to learn models that are rich enough and adapt to the system as a whole [25]. Particularly, reinforcement learning (RL), a branch of machine learning that is based on interactive learning from an unknown environment [29] has, in recent years, gained increased attention in artificial pancreas research [30–39]. A complete systematic review of reinforcement learning application in diabetes blood glucose control can be found in [40]. Outside of diabetes-related research, it has been particularly successful in achieving performance that exceeds the level of top human players in strategy games. The examples range from Backgammon in the early 1990's and more recently in the game of Go in 2015, where RL was combined with deep neural networks and Monte Carlo tree search [41,42]. RL allows us to introduce model-free and data driven algorithms that can enable another level of patient individualization [25]. Finally, previous works from the authors have shown promise for the use of RL in the artificial pancreas [32]. In that work, the amount of infused insulin was selected from a fixed and finite list of values, while the blood sugar level was treated as a continuous variable. In addition, there are several recent works using similar methodology [30,33,34,36–39].

In this work we extend the evaluation of RL algorithms for the artificial pancreas and study the performance of Policy Gradient RL algorithms. It is well known in RL literature that policy gradient algorithms are the most suitable for problems where the action space is continuous. This is an important step in the intersection between the RL and diabetes research. Furthermore, we focus on deep Policy Gradient methods due to the flexibility, power and availability of modern neural network approaches [43–46].

We perform in-silico experiments while using the Hovorka model [22] and the trust-region policy optimization of Schulman et al. [45]. Our experiments demonstrate that RL can adapt to carbohydrate counting errors and that RL is flexible enough to treat a population of 100 patients using a single set of training hyperparameters. We consider MPC to be the current state-of-the-art approach and, thus, we compare the performance of the RL agents to that of MPC. Performance is measured through time-in-range (time spent on healthy blood glucose levels), time in hypo-/hyperglycemia, as well as blood glucose level plots for visual inspection.

1.1. Related Work

We include a quick overview over the most recent developments in deep reinforcement learning and the artificial pancreas. Particularly, Sun et al. [35] used reinforcement learning to learn the parameters of the insulin pump, specifically the insulin to carb-ratio, and not the insulin action itself. They do not use neural networks in the process. Zhu et al. [38] is quite similar to our work; however, they use PPO, a simpler version of TRPO, and they use the blood glucose level, bg rate, and an

estimate of insulin-on board in the state space. The main difference between their work and this work is that they design a reward function that mimics the natural behaviour of the missing beta-cells, whereas our work focuses on a reward that encodes a more direct approach towards well-established performance measures for T1D therapy (time-in-range, etc.). Finally, Lee et al. [39] proposed a Q-learning approach, where a discrete number of actions modify the given basal rate. They also operate in a dual-hormone approach, where the infusion of glucagon is one of the actions. Their reward function however is quite similar to ours. Finally, they provide an alternative approach to training, where a population level policy is first introduced, followed by individual adaptation to each in-silico patient.

1.2. Structure of Paper

We begin with a short introduction to RL in Section 2 followed by a section about in-silico simulation for T1DM in Section 3. In Section 4 we present results and discussions. Section 5 provides concluding remarks and directions of possible future work.

2. Theoretical Background

In this section, we present the relevant theoretical background. We start with an introduction to RL, followed by a short section on MPC.

2.1. Reinforcement Learning

Informally, RL concerns the behavior of a decision-making agent interacting with its unknown environment. In this framework, the goal is to train an agent to take actions that result in preferable states. Figure 1 shows the agent-environment interaction, where at each time step the agent observes the current state of the environment and performs an action based on that state. As a consequence of this action, the environment transitions to a new state. In the next time step, the agent will receive a positive or negative reward from the environment due to the previous action taken [29].

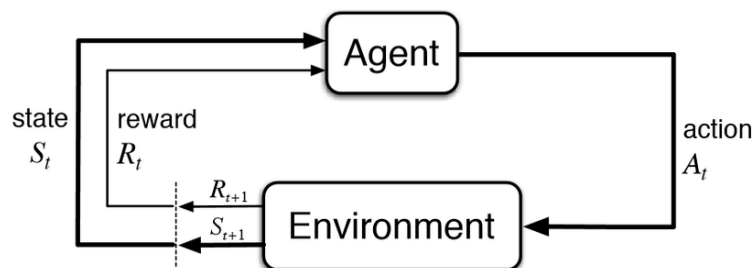


Figure 1. The reinforcement learning framework.

The mathematical basis of reinforcement learning is the Markov decision process, which is represented by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. \mathcal{S} and \mathcal{A} are the state and the action spaces, respectively, \mathcal{P} contains the state transition probabilities $p(s'|s, a)$ and represents the transition to state s' from s using action a . \mathcal{R} contains the rewards, represented by the reward function $r(s, a, s')$, which defines the goal of the problem, and $0 < \gamma \leq 1$ is a discount factor. The mapping from state to action is called the policy, which can be either a deterministic function $\pi : \mathcal{S} \rightarrow \mathcal{A}$ or a set of conditional distributions $\pi(a|s)$, depending on the environment the agent is interacting with. The goal of any RL algorithm is to learn an optimal policy π^* that maximizes the expected return it receives over time, which is the accumulated reward over time $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$, where $R_t = r(s_t, a_t, s_{t+1})$. The expected return assuming that the agent starts from the state s and thereafter follows the policy π

is called the value function $v_\pi(s)$. Concretely, the value function specifies the long-term desirability of states, indicating the total amount of reward that is expected by the agent:

$$v_\pi(s) = E_\pi[G_t | S_t = s] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right].$$

Similarly, the expected return assuming that the agent starts from the state s , takes action a , and thereafter follows the policy π is called the *action-value function* $q_\pi(s, a)$:

$$q_\pi(s, a) = E_\pi[G_t | S_t = s, A_t = a] \quad (1)$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\right]. \quad (2)$$

The ultimate goal of RL is to find an optimal policy, a policy that is better than or equal to all other policies based on the values of the states. Realizing this goal in practice has led to two different main branches of RL algorithms, value based algorithms and policy based algorithms, see e.g., [47].

Value based algorithms aim to estimate the value of each state the agent observes. Decisions are then made such that the agent spends as much time as possible in valuable states. A policy in value based RL is often simply a greedy search over each action in the given state, where the action that gives the highest value is chosen. In the case of a RL agent controlling e.g., an insulin pump in the T1DM case, such states could be safe blood glucose levels, while states with lower value would be either high or low blood glucose values.

Policy based algorithms change the viewpoint from looking at how valuable each separate state is, to evaluating how good the policy itself is. Given some parametric policy, a performance measure for the policy is defined—most commonly how much reward the agent can get over a certain amount of time. This measure is then optimized using gradient-based methods. For the T1DM case, this performance measure could for example be time-in-range.

2.2. Policy Gradient Methods

Policy gradient algorithms consider a parametric policy, $\pi(a|s, \theta) = P(a|s, \theta)$, and the goal is to optimize this policy using gradient ascent with a given performance measure $J(\theta)$ with parameter updates $\theta_{t+1} = \theta_t + \alpha \nabla J(\theta_t)$ [29]. The most common choice for the performance measure is the expected return of the initial state s_0 , given as

$$J(\theta) = v_\pi(s_0) = \mathbb{E}_\pi \left[R_0 + \gamma R_1 + \gamma^2 R_2 + \dots \right].$$

This is equivalent to optimize the value of the initial state—a policy is thus considered to be good if it can generate a lot of reward during the course of an episode.

There are multiple benefits of using policy gradient algorithms; they can be applied directly on continuous action spaces, the policy gradient theorem, introduced below, shows that any differentiable parametric policy can be used and, in the limit deterministic policies, can be modeled by policy gradients, which is useful if we do not want stochastic actions in an online setting—such as in the diabetes case.

One of the key points of policy gradient algorithms is the policy gradient theorem [48]:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla \pi(a|s, \theta). \quad (3)$$

where the distribution μ is the stationary distribution of the states succeeding s_0 when following π . This theorem states that the gradient of the performance measure is proportional to the gradient of the policy itself. This is of great benefit, as it allows the use of any differentiable policy parameterization. The policy gradient theorem allows, with some simple modifications to Equation (3), the formulation

of a simple sample-based algorithm, called REINFORCE. Instead of updating based on summing over all actions, the policy gradient is rewritten using a single sample S_t, A_t , and the gradient update rule becomes

$$\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla \pi(a|s, \theta)}{\pi(a|s, \theta)}. \tag{4}$$

The complete derivation can be found in Sutton and Barto [29] and the entire algorithm is shown in Algorithm 1.

Algorithm 1 REINFORCE

- 1: **Input:** differentiable policy $\pi(a|s, \theta)$.
 - 2: Generate episode from environment (See Section 3)
 - 3: **while** True **do** ▷ Loop until some convergence criteria is met.
 - 4: Generate a sample, $S_0, A_0, R_0, \dots, S_{T-1}, A_{T-1}, R_{T-1}, S_T$ from $\pi(a|s, \theta)$
 - 5: **for** $t = 0, 1, \dots, T$ **do**
 - 6: $G_t \leftarrow \sum_{k=t+1}^T R_k$
 - 7: $\theta_{t+1} \leftarrow \theta_t + \alpha G_t \nabla \ln \pi(A_t|S_t, \theta)$.
 - 8: **Return:** optimized policy $\pi(a|s, \theta)$.
-

The REINFORCE algorithm has been well studied and a number of improvements and suggestions have been proposed [45,46,49]. The current state-of-the-art in model free policy gradient algorithms is Trust-Region Policy Optimization by Schulman et al. [45] and a simplified version of the same algorithm called Proximal Policy Optimization [46]. In this work, we restrict our attention to the former.

Trust-region policy optimization (TRPO) is an algorithm that is based on the fact that if the policy gradient update is constrained by the total variation divergence, $D_{TV}(\pi_1, \pi_2) = \max_{s \in \mathcal{S}} |\pi_1(\cdot|s) - \pi_2(\cdot|s)|$, between the old policy and the new policy, the performance of the policy is guaranteed to increase monotonically [45]. Rewriting the total variation divergence using the Kullback-Leibler divergence and introducing approximations using importance sampling, the trust-region policy optimization reduces to solving the following optimization problem:

$$\begin{aligned} & \underset{\theta}{\text{maximize}} && \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} q_{\theta_{old}}(s,a) \right] \\ & \text{subject to} && \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} [D_{KL}(\pi_{\theta_{old}}, \pi_{\theta})] \leq \delta. \end{aligned} \tag{5}$$

where $q_{\theta_{old}}(s, a)$ is the action-value function, i.e., the value of taking action a in state s when following the policy $\pi_{\theta_{old}}(s, a)$, D_{KL} is the Kullback–Leibler divergence, and δ is the bound on Kullback–Leibler divergence. See Schulman et al. [45] for a complete description of the algorithm.

2.3. Parameterized Policies

The most common way to generate a parametric policy in a continuous action space is to use the Gaussian density function:

$$\pi(a|s, \theta) = \frac{1}{\sigma(s, \theta) \sqrt{2\pi}} \exp \left(-\frac{(a - \mu(s, \theta))^2}{2\sigma(s, \theta)^2} \right). \tag{6}$$

where $\mu(s, \theta)$ and $\sigma(s, \theta)$ are both state dependent parametric feature extractors. We use neural network feature extractors for both μ and σ in this work. In this way, $\mu = nn_{\mu}(s, \theta)$ is a multilayer perceptron with three hidden layers with 100, 50, and 25 hidden neurons, respectively, where θ are the weights of the neural network, mapping the state space into the mean of the Gaussian function. We decided

to use this neural network architecture following [43], where a feedforward neural network policy with the same number of layers and hidden neurons is used to test and evaluate several tasks with continuous action spaces. σ can either be a fixed vector, $\sigma = r \in \mathbb{R}^d$, where d is the dimension of the state space, or the output of a different neural network, $\sigma = nn_\sigma(s, \theta)$. In this case, the multilayer perception used for σ consists of two hidden layers, each with 32 hidden neurons. It is common to take the exponential of σ to ensure a positive standard deviation [29,45]. In the multivariate case, a diagonal covariance matrix is used. For both neural networks, μ and σ , we used a non-linear \tanh intermediate-layer activation functions, while linear activation functions are used in the output layers. Thus, the action is a sample from $\mathcal{N}(\mu, \sigma^2)$. An illustration is shown in Figure 2.

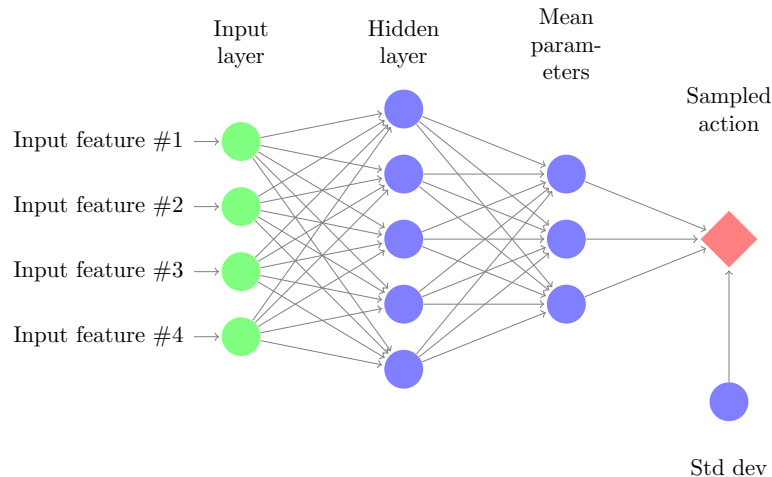


Figure 2. Neural network policy parameterization. The neural net maps the state, in this case a 4 dimensional space, to the mean, μ , of the Gaussian policy. The output is then a sample from the Gaussian policy $\mathcal{N}(\mu, \sigma^2)$. The σ parameter is in this work the output from a neural network.

2.4. Model Predictive Control

Model predictive control (MPC) is currently the state-of-the-art for artificial pancreas systems [50–52], and is used in commercial systems including the recently FDA approved Control-IQ™ advanced hybrid closed loop technology [53]. In general, MPC is a collection of algorithms where a model of the process is used to predict the system’s future behavior. Optimal actions are then computed, while using an objective function, to ensure that the predicted behavior matches the optimal desired behaviour [54]. Algorithms typically differ in the type of model and the objective function used [54]. MPC has the advantage of incorporating constraints in the objective function. This is particularly beneficial for the artificial pancreas system case that is characterized by long delay times [54]. Because the quality of the results is completely driven by the ability of the model to describe the true process state, most MPC algorithms are used in conjunction with state estimation techniques, such as the Kalman filter [24]. The main drawback of MPC is that adapting the model to each patient individually and accounting for intra-day variability is completely dependent on the structure of the predictive model [25]. If the model is not expressive enough to capture the situation, the algorithm will fail. Several recent works have tried to lessen this burden by using multiple predictive models [55–57].

3. In-Silico Simulation

Most in-silico T1DM research is centered around three physiological models: the Bergman (minimal) model [58], the Hovorka model [22] and the UVA/Padova model [59], see also [60]. The minimal model is a simplified model consisting of two equations describing the internal dynamics of glucose and insulin and does not account for the significant delay involved in subcutaneous

insulin infusion. The Hovorka model and the UVA/Padova both include this delay as well as the delay in the subcutaneous glucose measurement. In this work, we use the Hovorka model.

3.1. Simulator

The Hovorka model consists of five compartments that describe the dynamics of glucose kinetics and insulin action [61], two external, and three internal compartments. The three internal compartments describe insulin action, glucose kinetics and glucose absorption from the gastrointestinal tract. The two external compartments describe subcutaneous insulin absorption and interstitial glucose kinetics. The original model includes one virtual patient, which we use in our experiments. In addition, we follow Boiroux et al. [24] and use model equations, parameters and distributions as given in Hovorka et al. [22] and Wilinska et al. [62] to simulate further virtual patients. Unconstrained sampling from these distributions can lead to unrealistic virtual patients, as was also pointed out in Boiroux et al. [24]. To cope with this, the samples were constrained to the following set of rules [63].

- Patient weight is sampled from a uniform distribution between 55–95 kg.
- When the basal rate is delivered and the patient is in fasting conditions, glucose levels are constant and are between 110–180 mg/dL.
- The patient's basal rates were sampled from a uniform distribution between 0.2–2.5 U.
- The patient's carbohydrate ratios were sampled from a uniform distribution between 3–30 g/U.
- Each patient is characterized with a unique insulin sensitivity factor (*ISF*) S_i mg/dL/U, i.e., if an insulin bolus of size 1 U is delivered, glucose levels will drop by *ISF* mg/dL.
- The patient's insulin sensitivities were sampled from a uniform distribution between 0.5–6.5 mmol/L.
- A theoretical total daily dose (*TDD*) of insulin is computed assuming a daily diet of carbohydrates between 70–350 g. This value is then compared to sampled insulin sensitivity to ensure that the 1800 rule holds: $ISF = \frac{1800}{TDD}$.
- A theoretical total fraction of basal insulin is computed and is compared to *TDD* to ensure that the proportion of basal insulin is between 25–75% of *TDD*.
- All Hovorka's parameters, [62], are sampled using a log-normal distribution (to avoid negative values) around published parameters.

3.2. Reinforcement Learning, T1DM and the Artificial Pancreas

Because of the fact that applying reinforcement learning to any problem assumes an underlying Markov decision process, we need to take this into account when designing the state and action spaces for the T1DM case. There are several factors influencing whether or not we can interpret the glucose insulin dynamics as a Markov decision process, most notably the delayed action caused by the use of subcutaneous insulin infusion. Depending on the type of insulin used, the maximum effect of insulin is delayed and can last up to four hours [64]. On top of this comes the delay between the subcutaneous CGM measurements and the true blood glucose values, which is typically between 5 and 15 min. [65]. One of the fundamental properties of reinforcement learning algorithms, is the fact that they can control systems with delayed reward [66]. This implies that an action in a state can still be considered to be good even if the immediate reward from taking that action is not considered good. Furthermore, we note that, since the insulin infusion is the action taken by the RL agent, the environment will not change its state immediately because of the delayed insulin effect. Therefore, in this work we consider 30 min. time intervals as the time between each updated state from the environment. The insulin basal rate is kept constant during these 30 min. This will allow for the environment enough time to change significantly between each time step.

The final component involved is the reward function. In this work, we used two different reward functions, a symmetric Gaussian reward function and an asymmetric reward function, previously introduced in [67]. The Gaussian reward is given as:

$$r(g) = \exp \left\{ -\frac{1}{2h^2} (g - g_{ref})^2 \right\},$$

where g is the current blood glucose value, h is a smoothing parameter, and g_{ref} is the reference blood glucose value fixed at 108 mg/dL. The asymmetric reward function was, in [67], designed to reduce hypoglycemia and, at the same time, encouraging time-in-range. It is built as a piecewise smooth function and gives a strong negative reward for severe hypoglycemia, followed by an exponentially decreasing negative reward for hypoglycemic events starting at severe hypoglycemia, and zero reward when hyperglycemia occurs. Positive rewards from a symmetric linear function are given for glucose values in normoglycemic range. Concretely, the function is given as:

$$r(g) = \begin{cases} -100 & : g < g_{hypo-} \\ \exp\left(\frac{\log(140.9)}{g_{hypo}}g\right) - 140.9 & : g \in [g_{hypo-}, g_{hypo}] \\ \frac{1}{36}g - 2 & : g \in [g_{hypo}, g_{ref}] \\ -\frac{1}{72}g + \frac{5}{2} & : g \in [g_{ref}, g_{hyper}] \\ 0 & : g > g_{hyper}, \end{cases}$$

where hyperglycemia is defined as values above $g_{hyper} = 180$ mg/dL, hypoglycemia as values below $g_{hypo} = 72$ mg/dL and severe hypoglycemia as values below $g_{hypo-} = 54$ mg/dL. Thus, the normoglycemic range are values between $[g_{hypo}, g_{hyper}]$ mg/dL. The parameters of the reward were found experimentally. Figure 3 shows a graphical representation of the reward function.

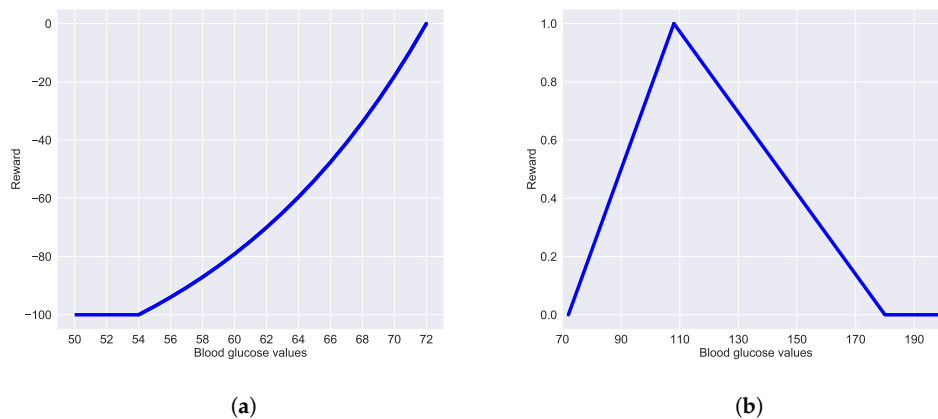


Figure 3. The asymmetric reward function. Low blood glucose levels (a) are more penalized than (b) high blood glucose levels.

3.3. Experiment Setup

The reinforcement learning agent controls the basal insulin rate of the pump which is updated every 30 min. In this work, we use two different action spaces during our experiments. Following Boiroux et al. [24], we define the action space of the agent ranging from zero, where the controller stops the insulin pump, to twice the optimal basal insulin rate (designated as TRPO in the results section). In addition, we use an extended version of the action space, which ranges from zero to three times the optimal basal insulin rate (designated as TRPOe in the results section). It is further assumed that the patient estimates and manually announces the amount of carbohydrate taken at each meal, and a bolus is given according to the patient’s individual carbohydrate to insulin ratio (The number of carbohydrates that one unit of insulin will counteract). The state space of the RL agent is defined

as the concatenation of the last 30 min. of blood glucose values, the last 2 h of insulin values (last 4 actions/basal rates given in 30 min. intervals), the insulin bolus on board, which is the calculation of how much insulin is still active in the patient's body from previous bolus doses, and the size of the last given bolus if a bolus was given during the last 30 min. The insulin on board was calculated while using a decay exponential model as described in Loop (<https://github.com/LoopKit/Loop>). We note that using the previous insulin taken is violating the MDP assumption. We chose to keep this compromise for two reasons: (1) the insulin and carbohydrate dynamics operate on fundamentally different time scales, see e.g., [62] and (2) information about previous insulin and insulin on board is essential knowledge that the agent cannot do without.

We use time-in-range (TIR) and time-in-hypoglycemia (TIH) as the performance measures, where we want to maximize the former and minimize the latter, in order to measure the performance of our simulations. We consider the normoglycemic range as values between 72–180 mg/dL and hypoglycemia as values below 72 mg/dL, see Danne et al. [68] for further details (We ended up using 72 mg/dL as the threshold instead of 70 due to converting from the local standard of using 4 mmol/L as the hypoglycemia threshold). In addition we use the Coefficient of Variation (CoV), σ/μ , to measure glycemic variability [69], defined as the ratio of the standard deviation to the mean of the blood glucose, and the risk index (RI), including high and low blood glucose risk indices, as described in Clarke and Kovatchev [70]. The RI measures the overall glucose variability and risks of hyper- and hypoglycemic events, while the high and low blood glucose indices (HBGI and LBGI) measure the frequency and extent of high and low blood glucose readings, respectively.

To train the algorithms, we use a standard reinforcement learning setup: (1) the agent collects episodes from the environment, followed by (2) the agent updates its parameters based on the rewards ((4) and (5)) and the process repeats until training is done (e.g., when the policy stops improving or stops changing) or the maximum number of iterations is reached. Inspired by the experiments in the original TRPO work [45], where between 50 to 200 iterations was used, we fix the number of policy update iterations to 100. This was also empirically found to provide convergence for the policies that are involved in most experiments. Furthermore, each episode is defined as starting at 00:00 and ending the next day at 12:00, at a total of 36 h. For each episode during training, the virtual patient is given meals from a fixed-seed random meal generator to ensure that each agent is trained on the same data set. This meal generator creates four virtual meals at ± 30 min. of 08:00, 12:00, 18:00, and 22:00 h with 40, 80, 60, and 30 g of carbohydrates. $\mathcal{U}[-20, 20]$ uniform noise is added to simulate meal variation. For simplicity, the meal times are kept concurrent to the start times of each state—every 30 min. Because of the delayed meal response and the generally high variation in the bg curve, we assume that this will generalize well to meals that are taken within a state-space time interval.

To test the agents, we use a fixed set of 100 episodes with 100 daily meals scenarios, sampled from the meal generator with a different seed than the training meals. Finally, to simulate carbohydrate counting errors, all meals—both training and testing—have a counting error of $\pm 30\%$ of the exact carbohydrate count. The reinforcement learning agent was implemented using the open source reinforcement learning toolbox *garage* <https://github.com/rlworkgroup/garage>. [43]. The in-silico simulator was wrapped in the OpenAI Gym framework for simplified testing [71].

4. Results

We now present the results and discuss the performance of a simulated artificial pancreas running the TRPO algorithm described in Section 2.2 in-silico. We show the results on the original Hovorka simulated patient, [22,62], as well as a cohort of 100 simulated patients according to the parameter distributions, as given in Wilinska et al. [62]. To illustrate its potential, we compare its performance to standard basal-bolus strategy and model predictive control algorithm, as described in [6]. We begin by comparing the TRPO agent to a simple basal-bolus treatment strategy on the original Hovorka patient.

4.1. TRPO versus Open Loop Basal-Bolus Treatment—Hovorka Patient and Carbohydrate Counting Errors

In this simulation, we consider open loop basal-bolus therapy, i.e., a fixed optimal basal insulin rate with manually administered meal-time bolus insulin, where the optimal basal rate is calculated as the minimum amount of insulin that is required to manage normal daily blood glucose fluctuations for this particular patient, while keeping the patient at target blood glucose value during steady state. We compared the basal-bolus therapy with a hybrid closed loop system in which the TRPO agent is controlling the basal insulin rate while meal-time bolus insulin are manually administered, both strategies running the same 100 test meal scenarios. Figure 4 shows the two previously mentioned treatments superimposed over each other, where we can see the blood glucose levels for the 100 test meal scenarios. The average blood glucose values for TRPO and basal-bolus strategies are highlighted in dashed and continuous curves, respectively. The dark gray shaded area shows the maximum and minimum values for each individual step of the simulation for the TRPO agent, while the light gray shaded area does likewise with the basal-bolus regimen. We are using the maximum and minimum blood glucose values instead of a confidence interval to include all possible curves in the envelope. This is due to the severe clinical implications of even a single blood glucose curve going too low.

We see in Figure 4 that the baseline performance of the basal-bolus controller is quite good, with a high portion of time being spent within range. Still, there are several hypoglycemic events, especially after the second meal, and the variation is high, as seen in the point-wise maximum and minimum band.

In the case of the TRPO controller, we see that the hypoglycemic events after meals and the overall variance have been reduced.

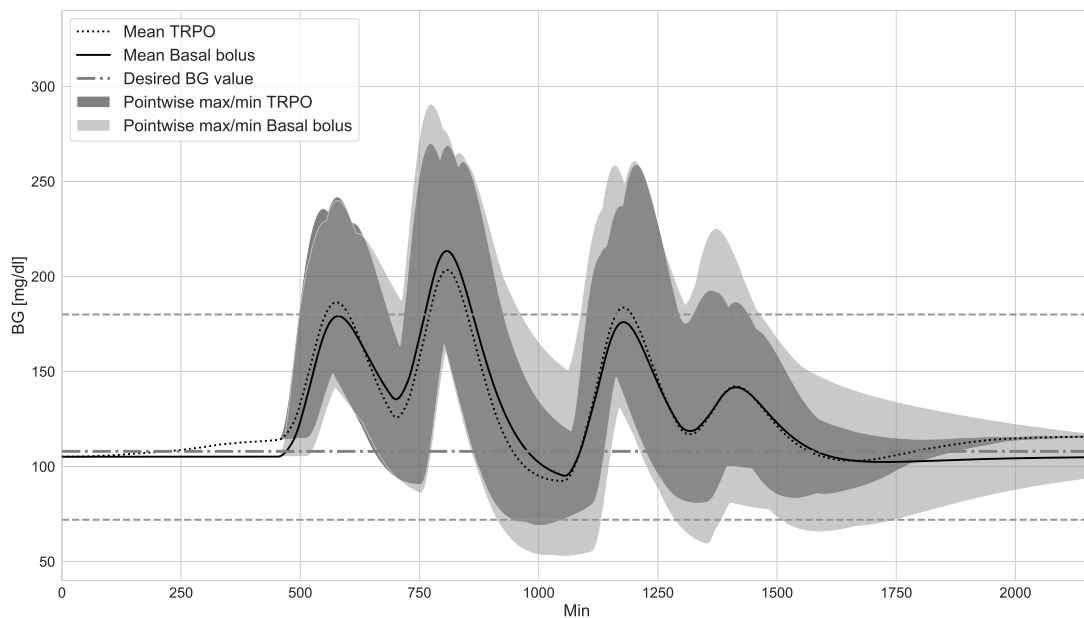


Figure 4. Blood glucose levels of Trust-region policy optimization (TRPO) reinforcement learning (RL) agent and standard basal-bolus therapy for the Hovorka patient. The dashed and continuous curves represent the average blood glucose values over 100 test meal scenarios for the TRPO agent and the basal-bolus regimen, respectively. The shaded dark and light gray envelopes represent the minute-wise maximum and minimum blood glucose level of the simulation for the TRPO agent and the basal-bolus regimen respectively. Each test episode runs for one and a half day, a total of 2160 min.

When comparing the two results, we see that the TRPO agent has improved the results; reducing variance in general and showing better overall within range performance. Especially with respect to hypoglycemia and the glucose levels after the second meal. The TRPO agent is able to get

back to the optimal blood glucose level much quicker and with less variation than the basal-bolus strategy. An interesting observation from Figure 4 is that we see how the TRPO agent chooses to keep the steady state blood glucose value slightly higher than the desired value of 108 mg/dL (this can be observed from approximately minute 250 to 500 and from min. 1700 and onward). This helps to avoid the hypoglycemic incident that often happens after the second meal during the basal-bolus regimen.

The max-min envelope of Figure 4 is not showing the full picture with respect to the standard deviation of the two treatment options. To further illustrate this, we include kernel density plots in Figure 5, showing the distribution of the blood glucose shortly after meals, between meals and during the steady state long after any meals (equivalent to nighttime). The kernel density estimate was calculated using the *seaborn* python package (<https://seaborn.pydata.org/>).

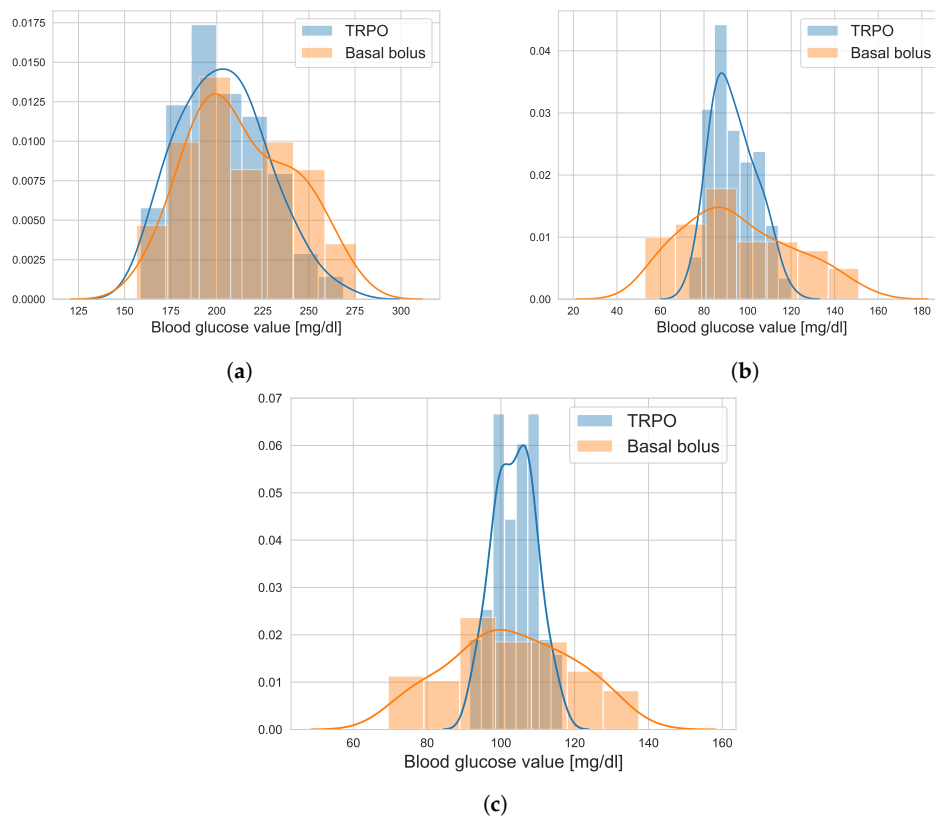


Figure 5. Kernel density estimation plot comparing the distribution of the results when comparing basal-bolus control to the TRPO agent. (a) shortly after a large carbohydrate intake, (b) between meals, (c) during nighttime (close to steady state).

It is clear, especially between meals and during night-time, that the TRPO agent treatment is superior to the basal-bolus strategy in this case.

We test the performance of the treatments in the case where the patient forgets to take the bolus insulin during a meal in order to conclude the comparison of the TRPO agent and the basal-bolus controller. To simulate this, we keep the 100 test meal scenarios, but let each meal during testing have a 0.1 probability of containing a skipped bolus. Table 1 shows a summary of all the performance measures for the experiments with the random skipped boluses (RSB) for both TRPO and basal-bolus treatments. We also include additional experiments, as shown in the Table 1, where the agent, denoted now as TRPOe, was trained with an extended action space (from zero to three times the optimal basal rate) and tested on the RSB scenarios as well as the ordinary 100 test scenarios.

Observing the Table 1, we again see that the TRPO agent is superior to the basal-bolus treatment, increasing time-in-range while decreasing time spent in hypoglycemia. It has lower variation and risk indices, and it is overall more robust towards skipped boluses. We note that the low LBG1 for the skipped bolus experiment is most likely an artifact due to the blood glucose level being higher in general when there are skipped boluses involved. The same goes for the overall percentage of time spent in hypoglycemia.

Table 1. Summary of basal-bolus, TRPO and TRPOe results for the Hovorka patient. low blood glucose indices (LBGI) and high blood glucose indices (HBGI) is low and high blood glucose index respectively, RI is risk index, Std is the overall standard deviation and CoV is the coefficient of variation. All 100 test meal scenarios are included in the performance measures. A lower score is better for all measures, except time-in-range.

Treatment	Time-in-Range	-Hypo	-Hyper	LBGI	HBGI	RI	Std	CoV
Basal-bolus	83.45	2.42	14.13	0.87	4.62	5.5	40.35	0.3
TRPO	86.12	0.1	13.78	0.46	3.17	3.62	36.55	0.27
TRPOe w/ 300 itr	86.33	0.49	13.18	0.42	4.14	4.56	36.71	0.28
Random skipped boluses:								
Basal-bolus	79.59	2.27	18.13	0.85	5.8	6.65	50.35	0.36
TRPO	82.91	0.0	17.09	0.2	5.55	5.75	41.06	0.29
TRPOe w/ 300 itr	84.68	0.49	14.84	0.43	4.68	5.11	40.36	0.3

When it comes to the results using the extended action space TRPOe, we found that the results using 100 policy gradient iterations are inferior to the other results. Therefore, we extended the number of training iterations to 300, which lead to an improvement over the original action space. The extended action space also leads to a treatment that is more robust to skipped boluses. However, the effect of increasing the number of policy gradient iterations from 100 to 300 represents a significant increase in data used for training the policy. There is a trade-off between the size of the action space and the number of training data/simulations needed.

4.2. Virtual Population Experiment: Undertreated Patients

The virtual population, as described in Section 3, have basal and bolus rates that are sub-optimal, keeping the patients within 110–180 mg/dL at steady state. We consider the virtual patients with high steady state glucose values as patients that are undertreated, i.e., their current treatment regimen does not give the desired blood glucose levels. We show a random sample of four patients in Figure 6 to illustrate the improvements made by letting a TRPO agent train and control each virtual patient. Each figure contains the original sub-optimal basal-bolus treatment as well as the results using TRPO agent superimposed over each other.

In all four cases, the TRPO agent improves the sub-optimal basal-bolus treatment. For virtual population patient #4, the performance of the basal-bolus is already close to optimal, but we still see a reduction in variance, especially later in the episode, during nighttime.

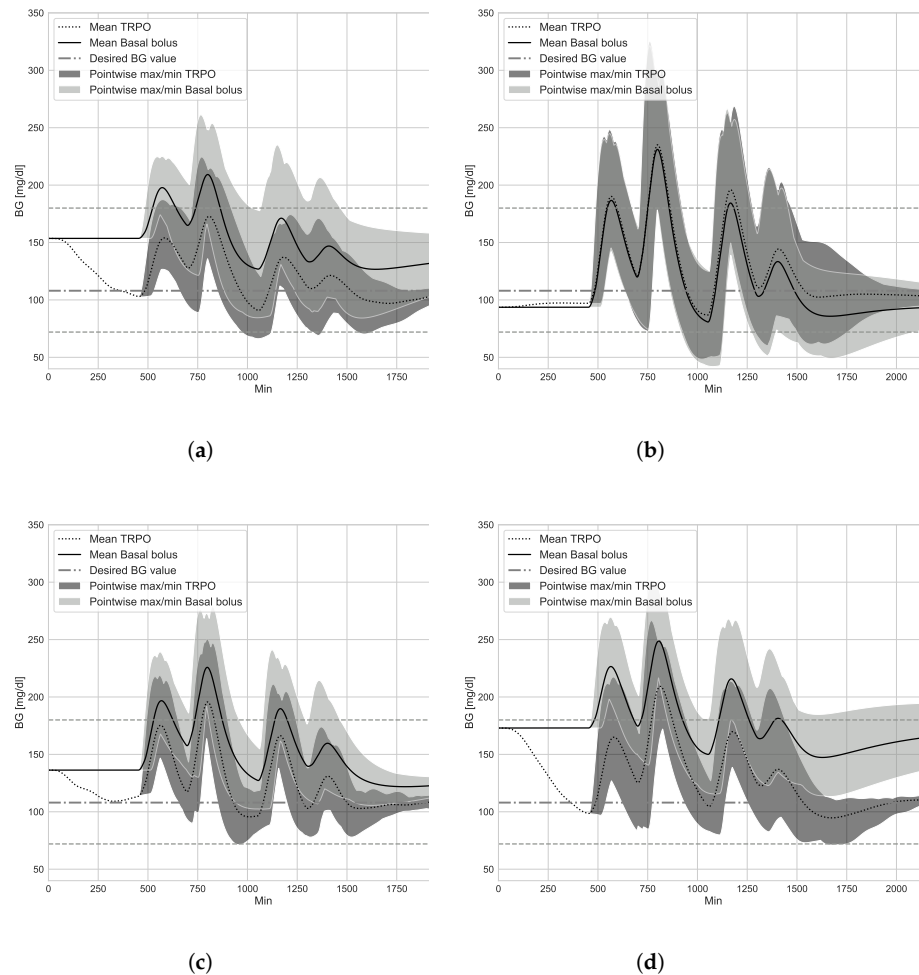


Figure 6. A random sample from the 100 virtual patients (a) patient #0, (b) patient #4, (c) patient #17, and (d) patient #38. All figures show results from the sub-optimal basal-bolus treatment and the TRPO agent trained on each patient individually.

4.3. Virtual Population Experiment: TRPO versus Model Predictive Control

We compare the TRPO agent to the open source MPC implementation (<https://github.com/McGillDiabetesLab/artificial-pancreas-simulator>) provided by the McGill Diabetes Lab (<https://www.mcgill.ca/haidar/>). The TRPO agent is individually trained for each virtual patient. The MPC controller is adapted to each patient using the total daily insulin, basal rate, and carb-ratio. As many of the patients are undertreated, some of these parameters might represent poor choices. We note that this leaves MPC at a disadvantage from the outset, since it is not able to tune the parameters during training.

In Figure 7, we see a scatterplot of the mean of the minimum and the mean of the maximum blood glucose of the 100 virtual patients controlled by MPC, the TRPO agent, and a basal-bolus strategy. This is similar to control-variability grid analysis plot [72], which is often used for measuring the quality of closed loop glucose control on a group of subjects, see e.g., [24]. The undertreated patients are left out of bounds for standard CVGA, thus requiring a different kind of analysis, as shown here.

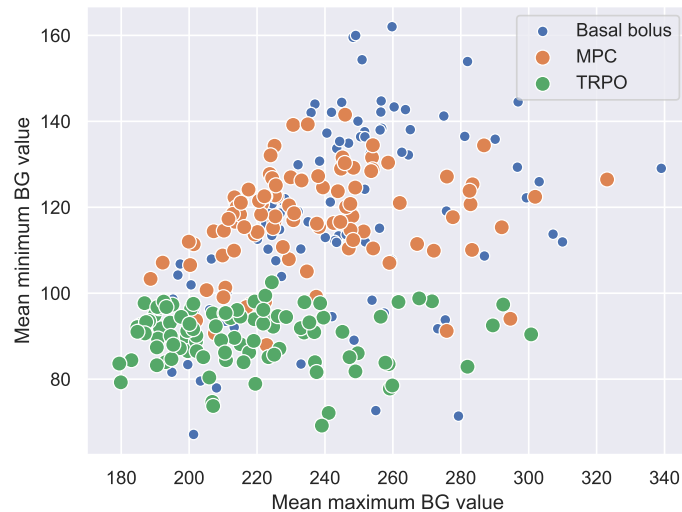


Figure 7. Scatterplot showing the mean of the maximum and minimum values over the 100 meal scenarios for each virtual patient on the x and y axes, respectively. MPC, basal-bolus treatment, and TRPO control is included in the plot. Tight glycemic control is in the mid to lower left area of the plot (a low maximum value and minimum value above 72 is desired).

The virtual population moves from both high mean maximum and minimum values in the basal-bolus case to lower mean maximum in MPC and even lower for the TRPO agent. We see that, in general, MPC stays at higher blood glucose levels as compared to TRPO, but conversely the TRPO agent is in some cases on the borderline low side.

To obtain a more complete picture, kernel density estimates of the same maxima and minima is shown for the entire population in Figure 8.

It is obvious that the TRPO again is outperforming the basal-bolus strategy. It shows tighter overall control and lower maximum values, while most minima are above the hypoglycemia threshold. The MPC is also tighter and improves over basal-bolus, but still the mean maximum values are, in general, higher. In addition, some of the mean minimum values are quite high, which indicates a mean blood glucose value that is generally high.

Finally, Table 2 shows the mean performance measures for the entire virtual population for basal-bolus, MPC and the TRPO agent. It also shows best and worst cases for all three treatments in terms of time-in-range (TIR) and time-in-hypo (TIH). TRPO improves the time spent in normoglycemia, while reducing the overall risk of hypo- and hyperglycemic events. However, MPC is more robust towards hypoglycemic events. Note that, in this case, in-silico patients spend less time in hypoglycemia following basal-bolus strategy than under TRPO control algorithm. This is because these patients are using sub-optimal basal-bolus treatment and therefore have higher steady state glucose values, spending most of the time close to hyperglycemia with almost no risk of hypoglycemic excursions. In this situation, the TRPO agent learns new basal rates to compensate the undertreated in-silico patients, improving the time spent in target range, but also at the same time slightly increasing the risk of hypoglycemia. Although the latter is, in general, considered to be negative, this comes down to how to design the control goals. There will always be a trade-off between better time-in-range and risk of hypo. A future study, with e.g., a parametric reward function, could help determine the exact trade-off for each patient, and take advantage of that. However, this is beyond the scope of this work.

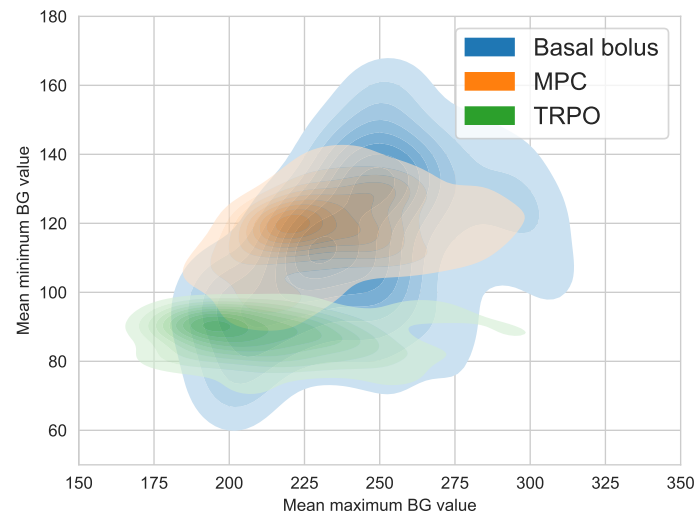


Figure 8. Kernel density estimate showing the approximate distribution of the same mean maxima and minima shown in Figure 7. The axes have been increased with respect to Figure 7 to fully cover the tails of the distributions.

Table 2. Mean performance measures for all 100 patients with all 100 test meal scenarios. Best results for each category is marked bold. For time-in-range, higher is better, for all other measures, lower is better.

Treatment	Time-in-Range	-Hypo	LBGI	HBGI	RI	Std	CoV
Basal-bolus	73.67	0.30	0.51	6.12	6.63	32.73	0.21
TRPO	88.72	0.50	0.78	3.80	4.57	32.75	0.24
MPC	79.25	0.003	0.13	5.14	5.27	30.11	0.19
Best and worst cases:	Best TIR	Worst TIR	Worst TIH				
Basal-bolus	95.59	43.80	7.11				
TRPO	97.18	63.63	5.01				
MPC	96.02	55.27	0.15				

5. Conclusions and Future Work

In this work, we have shown that policy gradient reinforcement learning using TRPO outperforms standard basal-bolus treatment and compares favourably to MPC in our experiments. We consider this work to be a strong proof of concept for the use of policy gradient algorithms in the artificial pancreas framework; the TRPO agent is able to cope with both carbohydrate counting errors and to a certain degree skipped boluses. Furthermore, the control is tighter than using a fixed optimal basal rate and risk indices are, in general, lower than both MPC and basal-bolus insulin therapy.

The main disadvantage of using RL, which has not been fully explored in this work, is the computational complexity of training. In this work, we fixed the number of policy gradient iterations to 100 for all experiments, but we empirically observed that, in many cases, far fewer iterations were required for convergence. Finally, we observed that a larger action space can lead to better control, but the increase in training data needed for convergence is significant.

All of the TRPO agents were trained model free, so from the agent's perspective the diabetes simulator is simply a black box that returns a reward when an input is given. Due to the fact that

T1DM is a well studied disease and multiple treatment strategies already exist, there is a lot of domain knowledge that gets lost in a model free setting. An obvious direction of research is including domain knowledge into the RL framework for T1DM, as in e.g., [73,74].

Finally, state-of-the-art RL contains a plethora of directions that can be explored, the perhaps most important ones for the artificial pancreas framework are inverse reinforcement learning [75], safe reinforcement learning (safe exploration) [76] and hierarchical reinforcement learning [77].

Author Contributions: Conceptualization, J.N.M., M.T. and I.K.L.; Funding acquisition, F.G.; Investigation, J.N.M. and M.T.; Methodology, J.N.M., M.T., I.K.L. and A.E.F.; Project administration, F.G.; Software, J.N.M., M.T. and A.E.F.; Supervision, F.G.; Validation, M.T. and A.E.F.; Writing—original draft, J.N.M., M.T. and I.K.L.; Writing—review & editing, J.N.M., M.T., I.K.L., A.E.F. and F.G. All authors have read and agreed to the published version of the manuscript.

Funding: J.N.M. and I.L. were funded by the Tromso Research Foundation during the course of this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. Diabetes. 2017. Available online: <http://www.webcitation.org/719KGYXpa> (accessed on 8 August 2018).
2. What is Insulin? Available online: <https://www.endocrineweb.com/conditions/type-1-diabetes/what-insulin> (accessed on 23 January 2020).
3. Diabetes Control and Complications Trial Research Group; The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial. *Diabetes* **1995**, *44*, 968–983. [CrossRef]
4. Misso, M.L.; Egberts, K.J.; Page, M.; O'Connor, D.; Shaw, J. Continuous subcutaneous insulin infusion (CSII) versus multiple insulin injections for type 1 diabetes mellitus. *Cochrane Database Syst. Rev.* **2010**, *20*, CD005103.
5. Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. Continuous glucose monitoring and intensive treatment of type 1 diabetes. *N. Engl. J. Med.* **2008**, *359*, 1464–1476. [CrossRef] [PubMed]
6. El Fathi, A.; Smaoui, M.R.; Gingras, V.; Boulet, B.; Haidar, A. The artificial pancreas and meal control: An overview of postprandial glucose regulation in type 1 diabetes. *IEEE Control. Syst. Mag.* **2018**, *38*, 67–85.
7. ADA. Diabetes. Available online: <https://www.diabetes.org/newsroom/press-releases/2019/new-recommendations-for> (accessed on 16 September 2019).
8. Hovorka, R. Closed-loop insulin delivery: From bench to clinical practice. *Nat. Rev. Endocrinol.* **2011**, *7*, 385–395. [CrossRef] [PubMed]
9. Cinar, A. Artificial pancreas systems: An introduction to the special issue. *IEEE Control. Syst. Mag.* **2018**, *38*, 26–29.
10. Basu, R.; Johnson, M.L.; Kudva, Y.C.; Basu, A. Exercise, Hypoglycemia, and Type 1 Diabetes. *Diabetes Technol. Ther.* **2014**, *16*, 331–337. [CrossRef]
11. Messer, L.H.; Forlenza, G.P.; Sherr, J.L.; Wadwa, R.P.; Buckingham, B.A.; Weinzimer, S.A.; Maahs, D.M.; Slover, R.H. Optimizing hybrid closed-loop therapy in adolescents and emerging adults using the MiniMed 670G system. *Diabetes Care* **2018**, *41*, 789–796. [CrossRef]
12. Petruzelkova, L.; Soupal, J.; Plasova, V.; Jiranova, P.; Neuman, V.; Plachy, L.; Pruhova, S.; Sumnik, Z.; Obermannova, B. Excellent glycemic control maintained by open-source hybrid closed-loop AndroidAPS during and after sustained physical activity. *Diabetes Technol. Ther.* **2018**, *20*, 744–750. [CrossRef]
13. Chase, H.P.; Doyle, F.J., III; Zisser, H.; Renard, E.; Nimri, R.; Cobelli, C.; Buckingham, B.A.; Maahs, D.M.; Anderson, S.; Magni, L.; et al. Multicenter closed-loop/hybrid meal bolus insulin delivery with type 1 diabetes. *Diabetes Technol. Ther.* **2014**, *16*, 623–632. [CrossRef]
14. Reiterer, F.; Freckmann, G.; del Re, L. Impact of Carbohydrate Counting Errors on Glycemic Control in Type 1 Diabetes. *IFAC-PapersOnLine* **2018**, *51*, 186–191. [CrossRef]
15. Deeb, A.; Al Hajeri, A.; Alh mouidi, I.; Nagelkerke, N. Accurate carbohydrate counting is an important determinant of postprandial glycemia in children and adolescents with type 1 diabetes on insulin pump therapy. *J. Diabetes Sci. Technol.* **2017**, *11*, 753–758. [CrossRef] [PubMed]

16. Vasiloglou, M.; Mouggiakakou, S.; Aubry, E.; Bokelmann, A.; Fricker, R.; Gomes, F.; Guntermann, C.; Meyer, A.; Studerus, D.; Stanga, Z. A comparative study on carbohydrate estimation: GoCARB vs. Dietitians. *Nutrients* **2018**, *10*, 741. [[CrossRef](#)] [[PubMed](#)]
17. Kawamura, T.; Takamura, C.; Hirose, M.; Hashimoto, T.; Higashide, T.; Kashiwara, Y.; Hashimura, K.; Shintaku, H. The factors affecting on estimation of carbohydrate content of meals in carbohydrate counting. *Clin. Pediatr. Endocrinol.* **2015**, *24*, 153–165. [[CrossRef](#)] [[PubMed](#)]
18. Kovatchev, B.; Cheng, P.; Anderson, S.M.; Pinsky, J.E.; Boscari, F.; Buckingham, B.A.; Doyle, F.J., III; Hood, K.K.; Brown, S.A.; Breton, M.D.; et al. Feasibility of long-term closed-loop control: A multicenter 6-month trial of 24/7 automated insulin delivery. *Diabetes Technol. Ther.* **2017**, *19*, 18–24. [[CrossRef](#)]
19. Boughton, C.K.; Hovorka, R. Advances in artificial pancreas systems. *Sci. Transl. Med.* **2019**, *11*, 4949. [[CrossRef](#)]
20. Turksoy, K.; Hajizadeh, I.; Samadi, S.; Feng, J.; Sevil, M.; Park, M.; Quinn, L.; Littlejohn, E.; Cinar, A. Real-time insulin bolusing for unannounced meals with artificial pancreas. *Control. Eng. Pract.* **2017**, *59*, 159–164. [[CrossRef](#)]
21. Steil, G.M.; Rebrin, K.; Darwin, C.; Hariri, F.; Saad, M.F. Feasibility of automating insulin delivery for the treatment of type 1 diabetes. *Diabetes* **2006**, *55*, 3344–3350. [[CrossRef](#)]
22. Hovorka, R.; Canonico, V.; Chassin, L.J.; Haueter, U.; Massi-Benedetti, M.; Federici, M.O.; Pieber, T.R.; Schaller, H.C.; Schaupp, L.; Vering, T.; et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiol. Meas.* **2004**, *25*, 905. [[CrossRef](#)]
23. Harvey, R.A.; Dassau, E.; Bevier, W.C.; Seborg, D.E.; Jovanović, L.; Doyle, F.J., III; Zisser, H.C. Clinical evaluation of an automated artificial pancreas using zone-model predictive control and health monitoring system. *Diabetes Technol. Ther.* **2014**, *16*, 348–357. [[CrossRef](#)]
24. Boiroux, D.; Duun-Henriksen, A.K.; Schmidt, S.; Nørgaard, K.; Poulsen, N.K.; Madsen, H.; Jørgensen, J.B. Assessment of model predictive and adaptive glucose control strategies for people with type 1 diabetes. *IFAC Proc. Vol.* **2014**, *47*, 231–236. [[CrossRef](#)]
25. Bothe, M.K.; Dickens, L.; Reichel, K.; Tellmann, A.; Ellger, B.; Westphal, M.; Faisal, A.A. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Biomed. Signal Process. Control.* **2013**, *10*, 661–673. [[CrossRef](#)] [[PubMed](#)]
26. Atlas, E.; Nimri, R.; Miller, S.; Grunberg, E.A.; Phillip, M. MD-logic artificial pancreas system: A pilot study in adults with type 1 diabetes. *Diabetes Care* **2010**, *33*, 1072–1076. [[CrossRef](#)] [[PubMed](#)]
27. Aiello, E.M.; Lisanti, G.; Magni, L.; Musci, M.; Toffanin, C. Therapy-driven Deep Glucose Forecasting. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103255. [[CrossRef](#)]
28. Li, K.; Liu, C.; Zhu, T.; Herrero, P.; Georgiou, P. GluNet: A deep learning framework for accurate glucose forecasting. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 414–423. [[CrossRef](#)] [[PubMed](#)]
29. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
30. Ngo, P.D.; Wei, S.; Holubová, A.; Muzik, J.; Godtliebsen, F. Reinforcement-learning optimal control for type-1 diabetes. In Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, USA, 4–7 March 2018; pp. 333–336.
31. Bastani, M. Model-Free Intelligent Diabetes Management Using Machine Learning. Master's Thesis, University of Alberta Libraries, Edmonton, AB, Canada, 2014.
32. Myhre, J.N.; Launonen, I.K.; Wei, S.; Godtliebsen, F. Controlling Blood Glucose Levels in Patients with Type 1 Diabetes Using Fitted Q-Iterations and Functional Features. In Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), Aalborg, Denmark, 17–20 September 2018; pp. 1–6.
33. Fox, I.; Wiens, J. Reinforcement Learning for Blood Glucose Control: Challenges and Opportunities. In Proceedings of the Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning, Long Beach, CA, USA, 30 May 2019.
34. Daskalaki, E.; Diem, P.; Mouggiakakou, S.G. An Actor–Critic based controller for glucose regulation in type 1 diabetes. *Comput. Methods Programs Biomed.* **2013**, *109*, 116–125. [[CrossRef](#)]
35. Sun, Q.; Jankovic, M.V.; Mouggiakakou, S.G. Reinforcement learning-based adaptive insulin advisor for individuals with type 1 diabetes patients under multiple daily injections therapy. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 3609–3612.

36. Yasini, S.; Naghibi-Sistani, M.; Karimpour, A. Agent-based simulation for blood glucose control in diabetic patients. *Int. J. Appl. Sci. Eng. Technol.* **2009**, *5*, 40–49.
37. Sun, Q.; Jankovic, M.V.; Budzinski, J.; Moore, B.; Diem, P.; Stettler, C.; Mougiakakou, S.G. A dual mode adaptive basal-bolus advisor based on reinforcement learning. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 2633–2641. [[CrossRef](#)]
38. Zhu, T.; Li, K.; Herrero, P.; Georgiou, P. Basal Glucose Control in Type 1 Diabetes using Deep Reinforcement Learning: An In Silico Validation. *arXiv* **2020**, arXiv:2005.09059.
39. Lee, S.; Kim, J.; Park, S.W.; Jin, S.M.; Park, S.M. Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation. *IEEE J. Biomed. Health Inform.* **2020**. [[CrossRef](#)]
40. Tejedor, M.; Woldaregay, A.Z.; Godtliebsen, F. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artif. Intell. Med.* **2020**, *104*, 101836. [[CrossRef](#)] [[PubMed](#)]
41. Tesauro, G. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Comput.* **1994**, *6*, 215–219. [[CrossRef](#)]
42. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484. [[CrossRef](#)] [[PubMed](#)]
43. Duan, Y.; Chen, X.; Houthoofd, R.; Schulman, J.; Abbeel, P. Benchmarking Deep Reinforcement Learning for Continuous Control. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1329–1338.
44. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
45. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust region policy optimization. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1889–1897.
46. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
47. Nachum, O.; Norouzi, M.; Xu, K.; Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 2775–2785.
48. Sutton, R.S.; Barto, A.G. *Introduction to Reinforcement Learning*; MIT Press: Cambridge, MA, USA, 1998; Volume 135.
49. Kakade, S.M. A natural policy gradient. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2002; pp. 1531–1538.
50. Shi, D.; Dassau, E.; Doyle, F.J. Adaptive Zone Model Predictive Control of Artificial Pancreas Based on Glucose-and Velocity-Dependent Control Penalties. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 1045–1054. [[CrossRef](#)]
51. Del Favero, S.; Place, J.; Kropff, J.; Messori, M.; Keith-Hynes, P.; Visentin, R.; Monaro, M.; Galasso, S.; Boscaro, F.; Toffanin, C.; et al. Multicenter outpatient dinner/overnight reduction of hypoglycemia and increased time of glucose in target with a wearable artificial pancreas using modular model predictive control in adults with type 1 diabetes. *Diabetes Obes. Metab.* **2015**, *17*, 468–476. [[CrossRef](#)] [[PubMed](#)]
52. Incremona, G.P.; Messori, M.; Toffanin, C.; Cobelli, C.; Magni, L. Model predictive control with integral action for artificial pancreas. *Control. Eng. Pract.* **2018**, *77*, 86–94. [[CrossRef](#)]
53. Brown, S.A.; Kovatchev, B.P.; Raghinaru, D.; Lum, J.W.; Buckingham, B.A.; Kudva, Y.C.; Laffel, L.M.; Levy, C.J.; Pinsky, J.E.; Wadwa, R.P.; et al. Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. *N. Engl. J. Med.* **2019**, *381*, 1707–1717. [[CrossRef](#)]
54. Camacho, E.F.; Bordons, C.; Johnson, M. *Model Predictive Control. Advanced Textbooks in Control and Signal Processing*; Springer: Berlin/Heidelberg, Germany, 1999.
55. Toffanin, C.; Aiello, E.; Del Favero, S.; Cobelli, C.; Magni, L. Multiple models for artificial pancreas predictions identified from free-living condition data: A proof of concept study. *J. Process. Control* **2019**, *77*, 29–37. [[CrossRef](#)]
56. Cameron, F.; Niemeyer, G.; Wilson, D.M.; Bequette, B.W.; Benassi, K.S.; Clinton, P.; Buckingham, B.A. Inpatient trial of an artificial pancreas based on multiple model probabilistic predictive control with repeated large unannounced meals. *Diabetes Technol. Ther.* **2014**, *16*, 728–734. [[CrossRef](#)] [[PubMed](#)]

57. Turksoy, K.; Quinn, L.; Littlejohn, E.; Cinar, A. Multivariable adaptive identification and control for artificial pancreas systems. *IEEE Trans. Biomed. Eng.* **2013**, *61*, 883–891. [[CrossRef](#)] [[PubMed](#)]
58. Bergman, R.N. Toward physiological understanding of glucose tolerance: Minimal-model approach. *Diabetes* **1989**, *38*, 1512–1527. [[CrossRef](#)] [[PubMed](#)]
59. Dalla Man, C.; Rizza, R.A.; Cobelli, C. Meal simulation model of the glucose-insulin system. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 1740–1749. [[CrossRef](#)]
60. Kanderian, S.S.; Weinzimer, S.A.; Steil, G.M. The identifiable virtual patient model: Comparison of simulation and clinical closed-loop study results. *J. Diabetes Sci. Technol.* **2012**, *6*, 371–379. [[CrossRef](#)]
61. Wilinska, M.E.; Hovorka, R. Simulation models for in silico testing of closed-loop glucose controllers in type 1 diabetes. *Drug Discov. Today Dis. Model.* **2008**, *5*, 289–298. [[CrossRef](#)]
62. Wilinska, M.E.; Chassin, L.J.; Acerini, C.L.; Allen, J.M.; Dunger, D.B.; Hovorka, R. Simulation environment to evaluate closed-loop insulin delivery systems in type 1 diabetes. *J. Diabetes Sci. Technol.* **2010**, *4*, 132–144. [[CrossRef](#)]
63. Walsh, J.; Roberts, R. *Pumping Insulin: Everything You Need for Success on a Smart Insulin Pump*; Torrey Pines Press: San Diego, CA, USA, 2006.
64. Gingras, V.; Taleb, N.; Roy-Fleming, A.; Legault, L.; Rabasa-Lhoret, R. The challenges of achieving postprandial glucose control using closed-loop systems in patients with type 1 diabetes. *Diabetes Obes. Metab.* **2018**, *20*, 245–256. [[CrossRef](#)]
65. Schmelzeisen-Redeker, G.; Schoemaker, M.; Kirchsteiger, H.; Freckmann, G.; Heinemann, L.; del Re, L. Time delay of CGM sensors: Relevance, causes, and countermeasures. *J. Diabetes Sci. Technol.* **2015**, *9*, 1006–1015. [[CrossRef](#)]
66. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [[CrossRef](#)]
67. Tejedor, M.; Myhre, J.N. Controlling Blood Glucose For Patients With Type 1 Diabetes Using Deep Reinforcement Learning—The Influence Of Changing The Reward Function. *Proc. North. Light. Deep. Learn. Workshop* **2020**, *1*, 1–6.
68. Danne, T.; Nimri, R.; Battelino, T.; Bergenstal, R.M.; Close, K.L.; DeVries, J.H.; Garg, S.; Heinemann, L.; Hirsch, I.; Amiel, S.A.; et al. International consensus on use of continuous glucose monitoring. *Diabetes Care* **2017**, *40*, 1631–1640. [[CrossRef](#)] [[PubMed](#)]
69. Suh, S.; Kim, J.H. Glycemic variability: How do we measure it and why is it important? *Diabetes Metab. J.* **2015**, *39*, 273–282. [[CrossRef](#)] [[PubMed](#)]
70. Clarke, W.; Kovatchev, B. Statistical tools to analyze continuous glucose monitor data. *Diabetes Technol. Ther.* **2009**, *11*, S-45–S-54. [[CrossRef](#)] [[PubMed](#)]
71. Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zaremba, W. Openai gym. *arXiv* **2016**, arXiv:1606.01540.
72. Magni, L.; Raimondo, D.M.; Dalla Man, C.; Breton, M.; Patek, S.; De Nicolao, G.; Cobelli, C.; Kovatchev, B.P. Evaluating the Efficacy of Closed-Loop Glucose Regulation via Control-Variability Grid Analysis. *J. Diabetes Sci. Technol.* **2008**, *2*, 630–635. [[CrossRef](#)]
73. Gu, S.; Lillicrap, T.; Sutskever, I.; Levine, S. Continuous deep q-learning with model-based acceleration. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2829–2838.
74. Berkenkamp, F.; Turchetta, M.; Schoellig, A.; Krause, A. Safe model-based reinforcement learning with stability guarantees. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 908–918.
75. Ho, J.; Ermon, S. Generative adversarial imitation learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4565–4573.

76. Garcia, J.; Fernandez, F. A Comprehensive Survey on Safe Reinforcement Learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
77. Bacon, P.L.; Harb, J.; Precup, D. The option-critic architecture. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.




© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Paper IV

Article

Risk-Averse Food Recommendation Using Bayesian Feedforward Neural Networks for Patients with Type 1 Diabetes Doing Physical Activities

Phuong Ngo ^{1,*}, Miguel Tejedor ^{2,*}, Maryam Tayefi ¹, Taridzo Chomutare ¹ 
and Fred Godtliebsen ^{1,3}

¹ Norwegian Centre for E-health Research, University Hospital of Northern Norway, 9019 Tromsø, Norway; Maryam.Tayefi@ehealthresearch.no (M.T.); Taridzo.Chomutare@ehealthresearch.no (T.C.); fred.godtliebsen@uit.no (F.G.)

² Department of Computer Science, Faculty of Science and Technology, UiT The Arctic University of Norway, 9019 Tromsø, Norway

³ Department of Mathematics and Statistics, Faculty of Science and Technology, UiT The Arctic University of Norway, 9019 Tromsø, Norway

* Correspondence: phuong.dinh.ngo@ehealthresearch.no (P.N.); miguel.tejedor@uit.no (M.T.)

Received: 7 October 2020; Accepted: 9 November 2020; Published: 12 November 2020



Abstract: *Background.* Since physical activity has a high impact on patients with type 1 diabetes and the risk of hypoglycemia (low blood glucose levels) is significantly higher during and after physical activities, an automatic method to provide a personalized recommendation is needed to improve the blood glucose management and harness the benefits of physical activities. This paper aims to reduce the risk of hypoglycemia and hyperglycemia (high blood glucose levels), and empowers type 1 diabetes patients to make decisions regarding food choices connected with physical activities. *Methods.* Traditional and Bayesian feedforward neural network models are developed to provide accurate predictions of the blood glucose outcome and the risks of hyperglycemia and hypoglycemia with uncertainty information. Using the proposed models, safe actions that minimize the risk of both hypoglycemia and hyperglycemia are provided as food recommendations to the patient. *Results.* The predicted blood glucose responses to the optimal and safe food recommendations are significantly better and safer than by taking random food. *Conclusions.* Simulations conducted on the state-of-the-art UVA/Padova simulator combined with Brenton's physical activity model show that the proposed methodology is safe and effective in managing blood glucose during and after physical activities.

Keywords: type-1 diabetes; machine learning; feedforward neural networks; Bayesian neural networks; physical activities

1. Introduction

Type 1 diabetes (T1D) is a chronic metabolic disorder characterized by elevated blood glucose levels over a prolonged period, leading to long-term damage to the heart, kidneys, eyes, nerves, and blood vessels. High blood glucose (hyperglycemia) in T1D is a consequence of the lack of insulin produced by the pancreas so that T1D patients require external insulin administration to regulate their blood glucose concentrations [1], while too much insulin dangerously reduces blood glucose level (hypoglycemia).

Treatment of T1D mainly consists of either a basal-bolus insulin regimen, where patients take basal insulin dose to regulate fasting blood glucose levels and insulin boluses around mealtimes to quickly reduce the impact of carbohydrate intake, or through an insulin pump providing a continuous insulin

infusion. The insulin pump infuses both regular basal insulin rate and meal boluses activated by the user during food intake to avoid hyperglycemia [2]. In both therapies, the insulin is administered subcutaneously in the fatty tissue just below the skin. In addition to the external insulin administration, monitoring blood glucose levels is required and can be done via a continuous glucose monitor (CGM) embedded in the subcutaneous tissue, or several times per day via manual finger-prick measurements [3]. Finally, T1D patients in collaboration with a physician will design a treatment based on individual patient needs. Insulin doses will be self-administered by the patients according to the treatment plan and self-measured blood sugar concentrations. External insulin therapy aims to keep the blood glucose concentrations within the normoglycemic range between 70 and 180 mg/dL [4,5].

Regular physical exercise has many proven health benefits and is therefore widely recommended as part of a healthy lifestyle. However, exercise significantly alters glucose homeostasis in patients with T1D. In addition, physical activities increase glucose uptake by muscles leading to a drop in blood glucose concentration, which can reach the hazardous hypoglycemic values. Increased insulin sensitivity, during and several hours or even days after the exercise [6], creates long-lasting effects on daily activities of patients.

The common method for preventing hypoglycemia is to reduce insulin doses. However, the slow absorption of insulin from the subcutaneous tissue and the physical limit of insulin reduction make this method insufficient to prevent hypoglycemia. Additional to the insulin treatment, many T1D related conditions can be mitigated by a nutrition therapy, which arises as an important solution to prevent, manage and control diabetes, as well as relieve complications associated with T1D by adjusting the quantity, quality and methods of nutrient intake [7]. Along with a healthy diet, physical activity plays a vital role in diabetes treatment, producing multiple general and diabetes-specific health benefits [8]. Despite the evidenced benefits, many people are physically inactive [9], since exercise is a major source of hypoglycemia in diabetic patients [10], and risk of hypoglycemia is a significant limiting factor of blood glucose regulation in T1D patients [11]. For most diabetic people, exercise has far less adverse health consequences than sedentary lifestyle [12].

Healthcare approaches change from the traditional relationship between providers and patients to a paradigm that gives patients a crucial role in guiding their care [13]. The change emphasizes the importance of self-management, which is considered a necessary part of chronic disease management and secondary prevention [14], especially for diabetes patients. Evidence shows that supporting patients to manage their health will improve clinical outcomes, reduce the economic burden, and improve quality of life [14]. Food recommendation systems emerge as a new self-management solution that can suggest the best diets according to patients' health situation and preferences, solving the physical activity paradigm while following a nutrition therapy for diabetic patients [15]. Among others, Phanich et al. [16] used a food-clustering analysis to propose a food recommendation system for patients with diabetes, while Norouzi et al. [17] developed a smartphone application for managing diabetic patient nutrition using artificial intelligence techniques. In [18], Lee et al. develop a diabetic food recommendation agent that, according to a personal lifestyle and particular health needs, can create a meal plan. Mohammed and Hagraas [19] present a type-2 fuzzy logic-based diet recommendation system to help achieve a healthy lifestyle to control diabetes. A complete systematic review of nutrition recommendation systems with a focus on technical aspects can be found in [20]. Previous work from the authors has shown promising results for using machine learning techniques in a food recommendation system, maintaining healthy blood glucose levels on a T1D simulator during exercise [21].

The term machine learning is considered a large family of mathematical and statistical methods that have historically focused on prediction [22]. With the development of new technology, a vast amount of health-related data is continuously generated. However, data availability is varied among various dimensions and quality. Machine learning and statistical techniques such as feedforward neural networks and the Bayesian inferencing mechanism become powerful tools to understand and

quantify data quality into uncertainties, which is a crucial step to make use of the increasingly available data safely and effectively.

Due to the importance and clinical benefits of the diet in diabetes, different studies have been conducted to develop diet recommendation systems to diabetes patients. Unlike [16,23], our work focused on preventing the complications related to physical activity in type-1 diabetes patients, alleviating the risks associated with doing exercise while having diabetes. Xie and Wang [24] proposed a food recommendation system with a similar purpose using a Nonlinear Auto Regressive Moving Average. However, the method did not provide a measurement of uncertainties in the data and how to compensate for these uncertainties. In this work, we introduced a new technique for food recommendation using Bayesian feedforward neural networks that can minimize the risk of hypoglycemia and hyperglycemia during and after physical activities while improving blood glucose regulation. We performed in-silico experiments, including the exercise model described in [25] on the UVA/Padova simulator [26,27]. Our experiments demonstrate that the proposed food recommendation system is able to reduce the risks of hypoglycemia and hyperglycemia while maintaining the blood glucose levels in the healthy range during and after the exercise.

Structure of Paper

We describe the methods in Section 2, where we introduce the experimental setup, defined the outcome and risk functions, and describe the models used to predict those functions. Section 3 presents the results and discussion. Section 4 provides the clinical significance and limitations of the method. In Section 5, we present concluding remarks and directions of possible future work.

2. Methods

The risk-averse food recommendation system presented in this paper used three criteria: the blood glucose outcome, the risk of hypoglycemia, and the risk of hyperglycemia. Recommendations were derived such that the risk of taking an action must be lower than a specified level and within a measure of probabilistic confidence level, while maximizing the outcome. This section starts with a description of the in-silico simulations used in the paper, followed by a formulation of the outcome and risks. Finally, implementation of deterministic and Bayesian feedforward neural networks prediction models are described.

2.1. In-Silico Simulation

In the simulation scenario used in the paper, the food recommendation is given before the beginning of each exercise session for a virtual patient with no meal boluses associated with the recommended amount of carbohydrates. During the experiments, the basal insulin dose was constant and equal to the optimal value, which means the virtual patient stays at the healthy reference blood glucose concentration $BG_{ref} = 108$ mg/dL in steady-state. Training data was obtained by repeated simulations from the blood glucose simulator under scenarios where a patient with T1D performs physical exercises with the same intensity but consuming a different amount of food.

The Physical Activity Guidelines for Americans recommends adults do vigorous enough exercise to raise their heart rate to 50–85% of their maximum heart rate, defined as 220 beats per minute minus their age, during 75 to 150 min a week—values might vary for younger people [28]. The virtual patient used in our simulations is 24 years old, with a maximum heart rate of 196 beats per minute and recommended heart rate during exercise between 98 and 167 beats per minute, based on the guidelines from the American Heart Association. The duration and intensity of physical activities are set to be constant at 50 min and 157 beats per minute (80% of its maximum heart rate), respectively. In absence of carbohydrate intake, a hypoglycemic excursion is induced as a consequence of the exercise session setup. The virtual patient always eats at 15 min before the exercise starts to avoid the hypoglycemic event. The outcome of each exercise is evaluated by measuring the average scores of the blood glucose over the course of four hours starting at 15 min before the exercise (mealtime). The blood glucose

is sampled every 5 min during the simulations, which is similar to the sampling time of common CGM devices.

There exist mainly three physiological models in the T1D research field, namely the Bergman (minimal) model [29], the Hovorka model [30] and the UVA/Padova model [26], see also [31]. The minimal model is a simplified model consisting of two equations describing the internal dynamics of glucose and insulin and does not account for the significant delay involved in subcutaneous insulin infusion. The Hovorka and the UVA/Padova models both include this delay as well as the delay in the subcutaneous glucose measurement. In this work, we used the UVA/Padova model, since this is the only computer simulator of the dynamics of the human metabolic glucose-insulin system which is FDA approved as a substitute for the pre-clinical testing of certain control strategies in T1D [27]. An extension of the UVA/Padova model has been used in this paper where the effect of physical activity is included [25].

The UVA/Padova model consists of seven internal compartments describing the dynamics of glucose kinetics, insulin kinetics, glucagon kinetics and secretion, glucose rate of appearance, endogenous glucose production, glucose utilization and renal excretion, while three external compartments describe subcutaneous glucose, insulin and glucagon kinetics [27]. In addition, a physical activity model was included in the glucose-utilization subsystem, modifying the insulin-dependent utilization component to simulate exercise sessions describing changes in glucose-insulin dynamics [25]. The original UVA/Padova simulator includes ten children patients, ten adolescents patients, and ten adults patients, as well as one average child patient, one average adolescent patient, and one average adult patient. In this work, we used the adult patient number four during our experiments, since this patient presents acute hypoglycemia as a consequence of our physical activity experimental setup. Finally, we use a CGM for glucose measurements during our simulations, where the CGM sensor noise is generated based on the model and the parameters determined by [32]. The non-Gaussian sensor noise is given by:

$$e_n = 0.8(e_{n-1} + v_n), \quad n > 0 \quad (1)$$

$$v_n \sim N_{iid}(0, 1) \quad (2)$$

$$\varepsilon_n = \xi + \lambda \sinh\left(\frac{e_n - \gamma}{\delta}\right), \quad (3)$$

with the initial condition $e_0 \sim N_{iid}(0, 1)$. Note that in the original model from [32], the error e_n introduced in the sensor noise is multiplied by a factor of 0.7, while in this work we increased the CGM noise multiplied the error e_n by a factor of 0.8, adding more uncertainty information to the Bayesian feedforward neural network. The CGM sensor noise ε_n was added to the blood glucose values obtained from the UVA/Padova simulator with physical activity. The numerical values used in this paper for ξ , λ , γ and δ are shown in Table 1.

Table 1. Parameters for continuous glucose monitor (CGM) sensor noise extracted from [32].

Parameter	Value
ξ	−5.471
λ	15.96
γ	−0.5444
δ	1.6898

2.2. Outcome Function and Risk Definition

This subsection defines the outcome and risk functions for the designed food recommendation system before physical activities for T1D patients. The blood glucose outcome is represented by the average score assigned for the blood glucose levels during and after physical activities. The score was calculated based on the asymmetric reward function previously introduced in [33], designed to reduce

hypoglycemia while rewarding time spent in normoglycemia. The reward function is designed as a piecewise smooth function and gives a strong negative reward for severe hypoglycemia, followed by an exponentially decreasing negative reward for hypoglycemic events starting at severe hypoglycemia, and negative reward when hyperglycemia occurs. Positive rewards from a symmetric linear function are given for glucose values in normoglycemic range. To be specific, the function is given as:

$$r(BG) = \begin{cases} -10 & : BG < BG_{hypo-} \\ \exp\left(\frac{\log(19.157)}{BG_{hypo}} BG\right) - 19.157 & : BG \in [BG_{hypo-}, BG_{hypo}] \\ \frac{1}{36}BG - 2 & : BG \in [BG_{hypo}, BG_{ref}] \\ -\frac{1}{72}BG + \frac{5}{2} & : BG \in [BG_{ref}, BG_{hyper}] \\ -5 & : BG > BG_{hyper} \end{cases}, \quad (4)$$

where BG is the current blood glucose value. The parameters of the reward function were selected based on the guidelines as described in [33] and can be found in Table 2. A graphical representation of the reward function is shown in Figure 1.

Table 2. Parameters of the reward function [33].

Parameter	Description	Value
BG_{ref}	Reference blood glucose	108 mg/dL
BG_{hyper}	Hyperglycemia blood glucose	180 mg/dL
BG_{hypo}	Hypoglycemia blood glucose	72 mg/dL
BG_{hypo-}	Severe hypoglycemia blood glucose	54 mg/dL

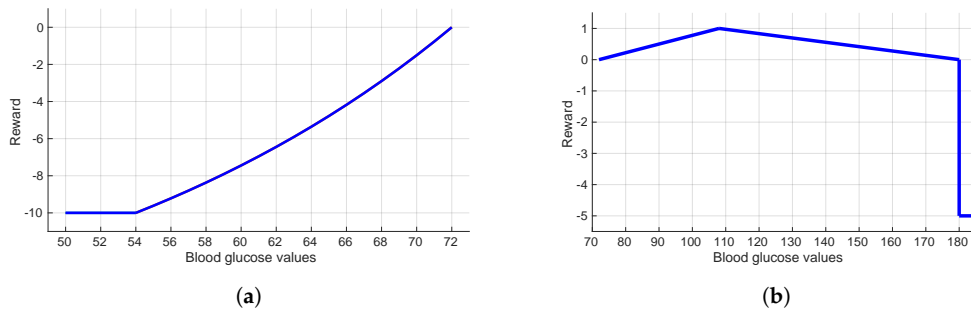


Figure 1. The asymmetric reward function. Low blood glucose levels (a) are more penalized than (b) high blood glucose levels.

In order to quantify the risks of hypo- and hyperglycemia, we used clinically defined low and high blood glucose indices (LBGI and HBGI, respectively) as described in Clarke and Kovatchev [34], which measure the frequency and extent of low and high blood glucose readings based on a symmetrization of the blood glucose measurements:

$$LBGI = \frac{1}{n} \sum_{i=1}^n rl(BG_i) \quad \text{and} \quad HBGI = \frac{1}{n} \sum_{i=1}^n rh(BG_i), \quad (5)$$

where BG_i is the measurement i in mg/dL and n is the number of measurements during and after the physical activity. The risk functions $rl(BG)$ and $rh(BG)$ are defined as follows:

$$rl(BG) = 10 \times f(BG)^2 \quad \text{if} \quad f(BG) < 0 \quad \text{and} \quad 0 \quad \text{otherwise} \quad (6)$$

$$rh(BG) = 10 \times f(BG)^2 \quad \text{if} \quad f(BG) > 0 \quad \text{and} \quad 0 \quad \text{otherwise} \quad (7)$$

where

$$f(BG) = 1.509 \times [(\ln(BG))^{1.084} - 5.381], \quad (8)$$

and BG is the current blood glucose value.

2.3. Modelling of the Outcome, Risk and Uncertainties

This subsection describes the methods to predict the blood glucose outcome, the hypoglycemia and the hyperglycemia risks of actions based on a predefined set of variables. Uncertainty estimation using Bayesian techniques will be described with these methods.

The blood glucose outcome, hyperglycemia and hypoglycemia risks were predicted in this paper by using a deterministic feedforward neural networks (FFNN) and two Bayesian feedforward neural networks (BFNN) with the same structure. FFNN and BFNN are types of artificial neural networks which are constructed by neurons organized into layers. The networks estimate the outcome and risks based on inputs such as the intensity of physical activity, historical blood glucose and dietary information (carbohydrate intake) before the physical activity. The structure of the deterministic FFNN and the two BFNN implemented in this paper are similar and can be found in Figure 2. For demonstration, we include only the food amount (carbohydrate intake) as the input for the networks. It is assumed that other conditions are the same every time physical activity is conducted.

Information from the input also flows through hidden nodes, which are nonlinear functions of the carbohydrate intake amount D :

$$h_j = \sigma \left(w_j^{(1)} D + b_j^{(1)} \right), \quad (9)$$

where h_j is the output value of the j^{th} , $j = 1, \dots, n$ hidden node, w_j and b_j are the weights and biases. The function σ is a rectified linear unit activation function (ReLU):

$$\sigma(x) = \max(0, x). \quad (10)$$

The model output y , (either predicted risks or blood glucose outcome), is calculated from the outputs of the hidden layers by a linear function:

$$\hat{y} = w^{(2)} D + b^{(2)}, \quad (11)$$

where $w^{(2)}$ and $b^{(2)}$ are the weight and bias for the output layer. The task of training the outcome model is to find the optimal values of w and b such that the following mean squared error (MSE) cost function is minimized:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2, \quad (12)$$

where N is the number of training samples, \hat{y}_i is the predicted value and y_i is the actual blood glucose score for each training sample i . Training of the outcome model was done using least square estimation and implemented through the Pytorch library [35].

Two models were built to predict the hypoglycemia and hyperglycemia risks based on a Bayesian feedforward neural network in order to capture the uncertainties in the risk prediction. The networks have similar structure like the outcome prediction model with six hidden nodes. However, the weight and bias of the hidden and output layers are represented by Gaussian distributions instead of discrete numbers. The prior distribution of all the parameters were chosen to be normal distributions with mean of zero and standard deviation of one. The task of training the hyperglycemia and hypoglycemia models is to determine the posterior distribution for the weights and biases using Bayesian inference. The posterior distribution was estimated by using the stochastic variational inference method [36] and implemented with Pyro [37]. For each weight and bias z , a family of distributions is generated with each distribution characterized by parameters θ . Then, an approximation of the posterior distribution

$p_\theta(z)$ is conducted by selecting from this family a distribution that is closest to the true posterior $p(z|y, x)$, where x and y are the training data set. The selection is conducted by finding the optimal set of parameters θ that minimize the Kullback–Leibler divergence, which is a measurement of how two probability distributions are different from each other [36,38]:

$$\begin{aligned}
 KL(p_\theta(z), p(z|y, x)) &= \int_z p_\theta(z) \log \left(\frac{p_\theta(z)}{p(z|y, x)} \right) \\
 &= \int_z p_\theta(z) \log \left(\frac{p(z)}{p(y|z, x)p(z)} \right) + \log(p(y|x)) = -L(\theta) + \log(p(y|x)),
 \end{aligned}
 \tag{13}$$

where $L(\theta)$ is defined as the evidence lower bound (ELBO) function [36]:

$$L(\theta) = \int_z p_\theta(z) \log \left(\frac{p(y|z, x)p(z)}{p(z)} \right).
 \tag{14}$$

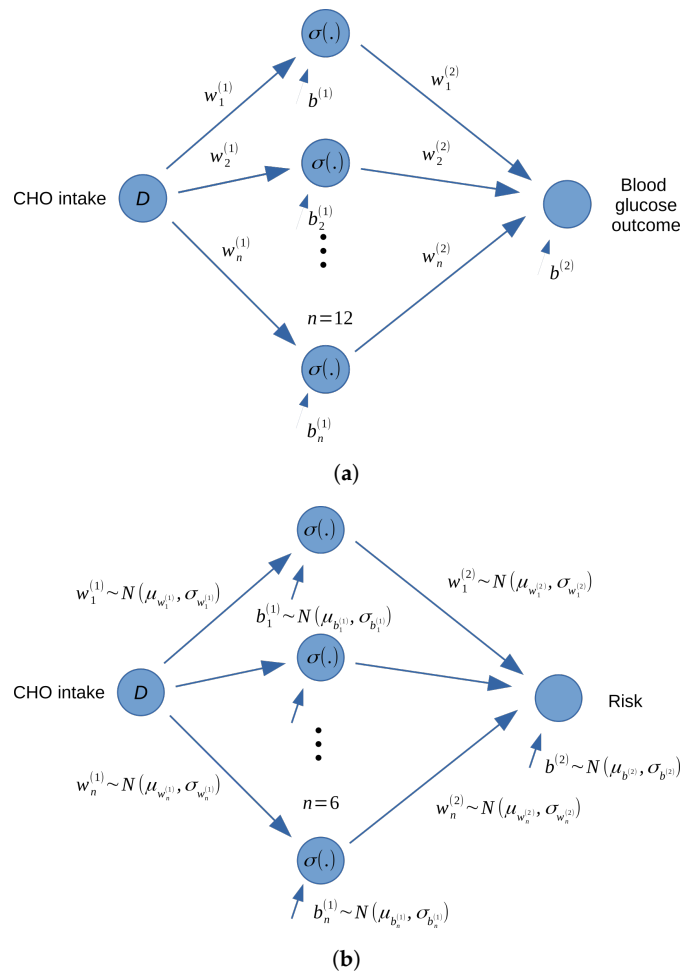


Figure 2. Structure of the feedforward neural networks for predicting the blood glucose outcome and risks (a) blood glucose outcome (b) risks.

Since the Kullback–Leibler divergence is negative, the task of determining the posterior distributions of all the weights and biases can be obtained by maximizing the ELBO by using the stochastic gradient descend method, through variational inferencing [36].

2.4. Validation

In order to validate the performances of the neural networks, parts of the data (20% for the FFNN and 50% for the BFNN) were randomly selected for testing the model's accuracy and performances. By propagating the carbohydrate information in each sample set through the neural networks, the prediction values are calculated and compared with the testing data. For the FFNN that is used to blood glucose outcome, the distributions of the estimated risks and the actual risks are compared by using the MSE (Equation (12)), which is also used as the validation criteria. For the BFNNs that are used to estimate the hyperglycemia and hypoglycemia risks, the distributions of the estimated risks and the actual risks are compared by using the ELBO criteria (Equation (14)). Since more data is need for the ELBO than for the MSE calculation, a higher validation portion is assigned for the BFNN than the FFNN.

3. Results and Discussion

By using data generated from the UVA/Padova simulator, the outcome and risk prediction models were built based on gradient descend and the stochastic variational inference. Figure 3 shows the MSE of the training and testing data during the training process of the feedforward neural networks. Table 3 shows the final ELBO (defined by Equation (14)) of the Bayesian neural networks under without and with noise in the CGM measurements. Figure 4 shows the training data and predictions made by the models for different carbohydrate values without and with CGM noise, respectively. The deterministic reward prediction is represented by the solid red line. The predicted risks with 90%-confidence intervals are represented by a green and a blue-shaded uncertainty bands in the figure, correspondingly to the hypoglycemia and hyperglycemia risks, respectively.

It can be seen from Figure 3 and Table 3 that the prediction errors are higher when measurement noise is introduced. Consequently, Figure 4 shows that the predicted risk uncertainty bands are wider when there is measurement noise, indicating that the predicted risk cannot be accurately predicted compared to nominal condition.

Based on the prediction models of the reward and risks, safe and optimal actions are derived as shown in Table 4. Safe actions are defined as actions that have a 90% probability of the hypoglycemia risk under 5 and the hyperglycemia risk under 7. The risk thresholds reflect that higher emphasis was put on hypoglycemia due to the higher adverse effects of getting hypoglycemia. It can be seen that the lower and upper bound for safe actions are tighter under noise conditions, since the risk models capture the uncertainties and provide more conservative recommendations under more uncertain condition. The optimal recommendation is chosen among the safe actions that maximize the blood glucose outcome predicted by the deterministic feedforward neural network. Hence the value is the same for both conditions when there is CGM noise and when there is no CGM noise.

Table 3. Evidence lower bound (ELBO) of the risk models for training and validation data.

Risk	Training		Validation	
	without Noise	with Noise	without Noise	with Noise
Hyperglycemia	0.4198	0.0632	0.5340	0.0307
Hypoglycemia	−0.4650	−0.7172	−0.3112	−1.0458

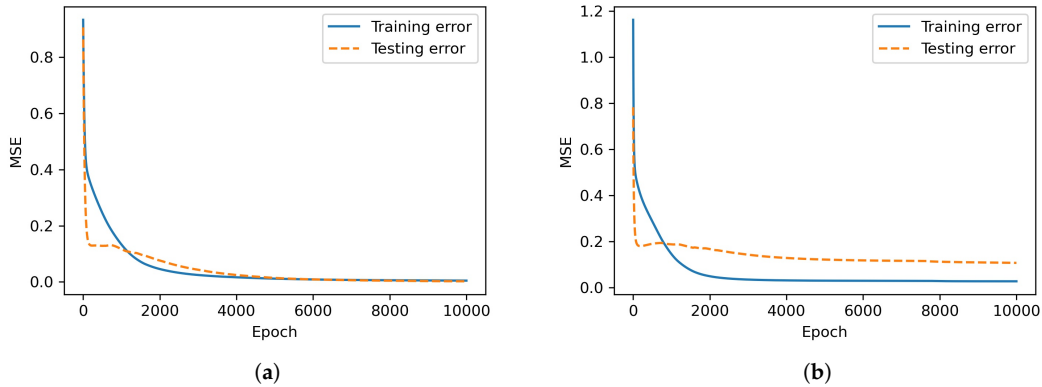


Figure 3. Training (solid blue line) and testing (orange dashed line) error without CGM measurement noise (a) and with CGM measurement noise (b).

Figures 5 and 6 show different ranges of blood glucose when the patient follows the optimal recommendation, consumes food within the safe recommendation actions, or takes food with carbohydrate amounts between 0 and 200 g. The distribution of blood glucose measurement according to food intakes at certain time points can also be seen in the figures. The results show that the range of blood glucose responses to the optimal and safe actions are significantly better than the range of blood glucose when any action can be selected. The blood glucose range is slightly narrower in the case with CGM noise since the recommendation is more conservative in order to compensate for uncertainties in the responses.

Table 4. Recommended actions (grams of carbohydrates).

Action Type	No Noise	with Noise
Safe lower	26.3	34.3
Safe upper	74.7	60.6
Optimal	50.2	50.2

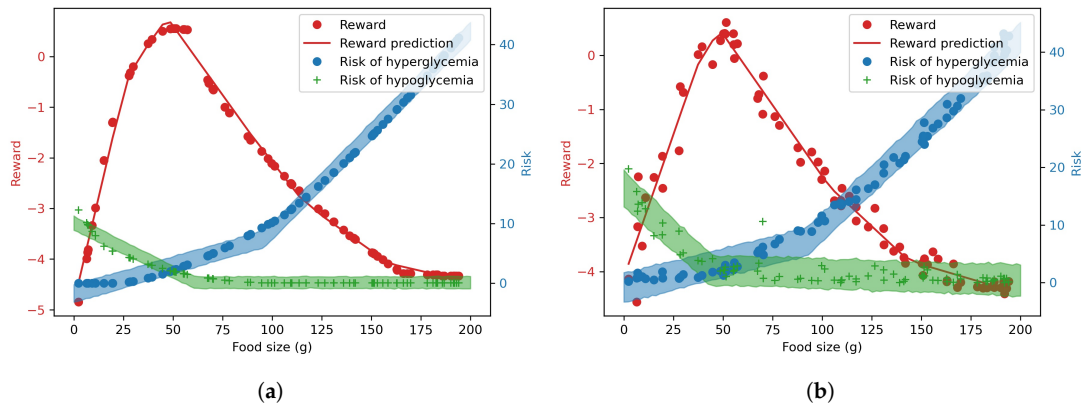


Figure 4. Training data (dots and cross), predicted outcome and risks under no CGM noise (a), and CGM noise (b). Green and blue shaded uncertainty bands represent the 90% confident intervals for hypoglycemia and hyperglycemia risks respectively.

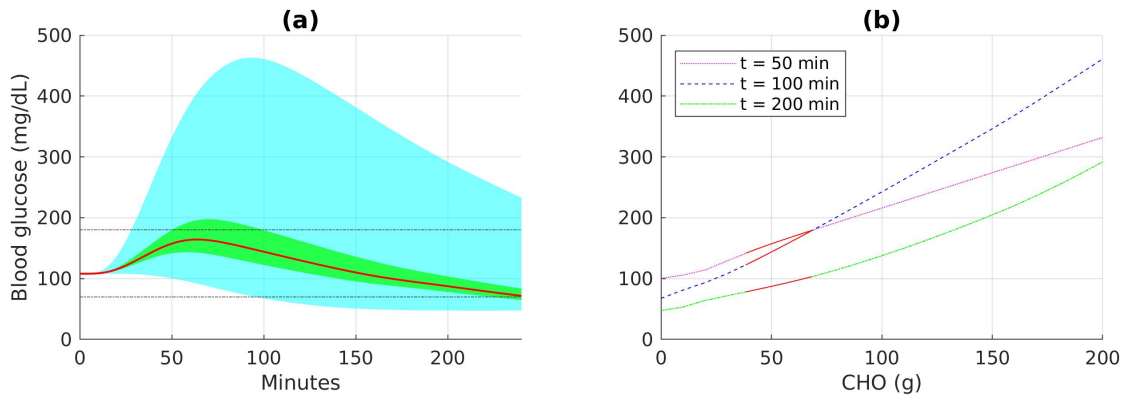


Figure 5. Blood glucose responses with no CGM noise in training data: (a) overtime as the effects of all possible actions (blue shade), safe actions (green shade), and optimal action (red line) (b) over the amount of carbohydrates at certain time points (red solid lines indicate safe actions).

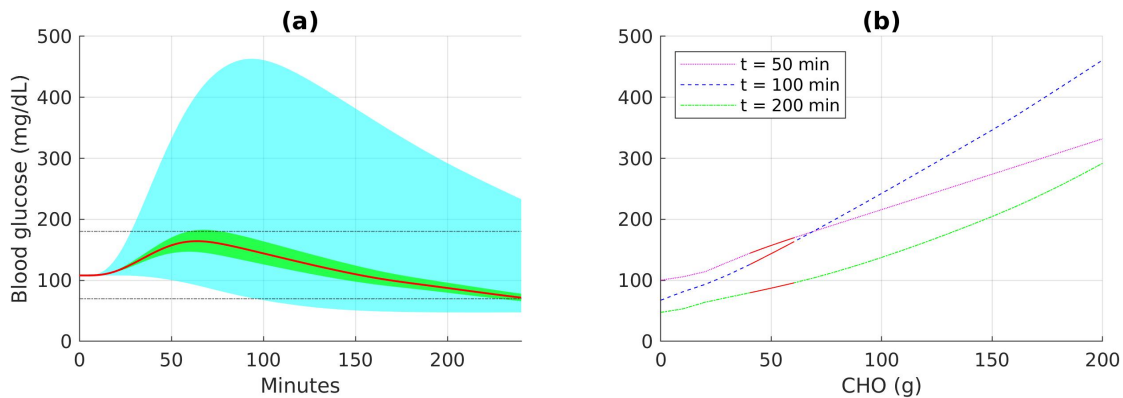


Figure 6. Blood glucose responses with CGM noise in training data: (a) overtime as the effects of all possible actions (blue shade), safe actions (green shade), and optimal action (red line) (b) over the amount of carbohydrates at certain time points (red solid lines indicate safe actions).

4. Clinical Significance and Limitations

The methods presented in the paper demonstrate that a clinical decision support tool can be built using machine learning. Uncertainties in the data can be estimated and recommendations are provided with a statistically confident level. The tools have the potentials to be applicable in many safety-critical applications beyond type-1 diabetes and contribute towards the development of safe artificial intelligence. However, future efforts are needed to ensure the algorithms can be applied beyond in silico simulations.

As a demonstration of how risk-averse machine learning methods can be developed and how the relationship between food intake and reward and risk can be visualized, only one variable (carbohydrate) was chosen as the input of the neural networks. Future studies should incorporate more variables that can affect blood glucose during and after physical activities such as previous food intakes, insulin doses, blood glucose levels before exercises and heart rate. The uncertainties can also be expanded to include other sources such as incorrect carbohydrate measurements, changes in intensity of physical activities and blood glucose kinetics. It is also noted that more complex neural networks require more data to obtain a satisfactory accuracy or performance, therefore, new methods that can effectively reduce the dependence on a large dataset is needed. Propagation of uncertainty analysis also contributes and make the methods more reliable. The current technique assumes that the outputs of the Bayesian neural networks are Gaussian-distributed. New methods that can accurately predict the output distribution will increase the accuracy of the prediction interval, and also the recommendations provided to the patients.

5. Conclusions

This paper provides a method to predict blood glucose outcome and risks associated with physical activities. The prediction models were used to select safe and optimal food amounts for patients to consume before physical activities. Simulation results show that the feedforward neural networks accurately predicted the blood glucose outcome while the Bayesian neural networks effectively capture the uncertainties due to measurement noise in the risk predictions. The blood glucose responses to the safe and optimal actions are significantly better than random actions within the input range. The results also present a potential direction for the future development of safe AI methods that are not only effective but also minimizing potential risks to the patients.

Author Contributions: P.N. developed the food recommendation algorithms, performed training and validating the deterministic and Bayesian feedforward neural networks for predicting blood glucose outcome, hyperglycemia and hypoglycemia risks, conducted numerical simulations and lead the writing process. M.T. (Miguel Tejedor) implemented the UVA/Padova simulator with physical activities, developing risk and reward functions, CGM noise model, performed numerical simulations, and was involved in the writing process. F.G. acquired funding and resource, managed and supervised the project leading to this publication. M.T. (Maryam Tayefi) and T.C. provided critical feedback, analysed results, read and approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tromso Research Foundation under project “A smart controller for T1D using RL and SS representation” (Grant/award number: A3327). The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway.

Conflicts of Interest: The authors declare no conflict of interest

References

1. World Health Organization. Diabetes. Available online: <https://www.who.int/health-topics/diabetes> (accessed on 2 September 2020).
2. Misso, M.L.; Egberts, K.J.; Page, M.; O'Connor, D.; Shaw, J. Continuous subcutaneous insulin infusion (CSII) versus multiple insulin injections for type 1 diabetes mellitus. *Cochrane Database Syst. Rev.* **2010**. [[CrossRef](#)] [[PubMed](#)]
3. Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. Continuous glucose monitoring and intensive treatment of type 1 diabetes. *N. Engl. J. Med.* **2008**, *359*, 1464–1476. [[CrossRef](#)] [[PubMed](#)]
4. El Fathi, A.; Raef Smaoui, M.; Gingras, V.; Boulet, B.; Haidar, A. The Artificial Pancreas and Meal Control: An Overview of Postprandial Glucose Regulation in Type 1 Diabetes. *IEEE Control Syst. Mag.* **2018**, *38*, 67–85.
5. Breton, M.; Farret, A.; Bruttomesso, D.; Anderson, S.; Magni, L.; Patek, S.; Man, C.D.; Place, J.; Demartini, S.; Toffanin, C.; et al. Fully Integrated Artificial Pancreas in Type 1 Diabetes. *Diabetes* **2012**, *61*, 2230–2237. [[CrossRef](#)] [[PubMed](#)]
6. Yardley, J.E.; Brockman, N.K.; Bracken, R.M. Could Age, Sex and Physical Fitness Affect Blood Glucose Responses to Exercise in Type 1 Diabetes? *Front. Endocrinol.* **2018**, *9*, 1–11. [[CrossRef](#)] [[PubMed](#)]
7. Bantle, J.P.; Wylie-Rosette, J.; Albright, A.L.; Apovian, C.M.; Clark, N.G.; Franz, M.J.; Hoogwerf, B.J.; Lichtenstein, A.H.; Mayer-Davis, E.; Mooradian, A.D.; et al. Nutrition recommendations and interventions for diabetes: A position statement of the american diabetes association. *Am. Diabetes Assoc.* **2008**, *31*, 61–78.
8. Hayes, C.; Kriska, A. Role of Physical Activity in Diabetes Management and Prevention. *J. Am. Diet. Assoc.* **2008**, *108*, 19–23. [[CrossRef](#)]
9. LaMonte, M.J.; Blair, S.N.; Church, T.S. Physical activity and diabetes prevention. *J. Appl. Physiol.* **2005**, *99*, 1205–1213. [[CrossRef](#)]
10. Basu, R.; Johnson, M.L.; Kudva, Y.C.; Basu, A. Exercise, Hypoglycemia, and Type 1 Diabetes. *J. Am. Diet. Assoc.* **2014**, *16*, 331–337. [[CrossRef](#)]
11. Diabetes Control; Complications Trial Research Group. The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial. *Diabetes* **1995**, *44*, 968–983. [[CrossRef](#)]
12. Sigal, R.J.; Armstrong, M.J.; Colby, P.; Kenny, G.P.; Plotnikoff, R.C.; Reichert, S.M.; Riddell, M.C. Physical Activity and Diabetes. *Can. J. Diabetes* **2018**, *37*, 54–63. [[CrossRef](#)] [[PubMed](#)]

13. Gobeil-Lavoie, A.P.; Chouinard, M.C.; Danish, A.; Hudon, C. Characteristics of self-management among patients with complex health needs: A thematic analysis review. *BMJ Open* **2019**, *9*, e028344. [[CrossRef](#)] [[PubMed](#)]
14. Grady, P.A.; Gough, L.L. Self-Management: A Comprehensive Approach to Management of Chronic Conditions. *Am. J. Public Health* **2014**, *104*, e25–e31. [[CrossRef](#)] [[PubMed](#)]
15. Norouzi, S.; Nematy, M.; Zabolinezhad, H.; Sistani, S.; Etminani, K. Food recommender systems for diabetic patients: A narrative review. *Rev. Clin. Med.* **2017**, *4*, 128–130.
16. Phanich, M.; Pholkul, P.; Phimoltares, S. Food Recommendation System Using Clustering Analysis for Diabetic Patients. In Proceedings of the 2010 International Conference on Information Science and Applications, Seoul, Korea, 21–23 April 2010; pp. 1–8.
17. Norouzi, S.; Ghalibaf, A.K.; Sistani, S.; Banazadeh, V.; Keykhaei, F.; Zareishargh, P.; Amiri, F.; Nematy, M.; Etminani, K. A Mobile Application for Managing Diabetic Patients' Nutrition: A Food Recommender System. *Arch. Iran. Med.* **2018**, *21*, 466–472. [[PubMed](#)]
18. Lee, C.S.; Wang, M.H.; Li, H.C.; Chen, W.H. Intelligent ontological agent for diabetic food recommendation. In Proceedings of the 2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–6 June 2008; pp. 1803–1810.
19. Mohammed, H.A.; Hagra, H. Towards Developing Type 2 Fuzzy Logic Diet Recommendation System for Diabetes. In Proceedings of the 2018 10th Computer Science and Electronic Engineering (CEECE), Colchester, UK, 19–21 September 2018; pp. 56–59.
20. Abhari, S.; Safdari, R.; Azadbakht, L.; Lankarani, K.B.; Niakan Kalhori, S.R.; Honarvar, B.; Abhari, K.; Ayyoubzadeh, S.M.; Karbasi, Z.; Zakerabasali, S.; et al. A Systematic Review of Nutrition Recommendation Systems: With Focus on Technical Aspects. *J. Biomed. Phys. Eng.* **2019**, *9*, 591–602. [[CrossRef](#)]
21. Ngo, P.D.; Tayefi, M.; Nordsletta, A.T.; Godtliebsen, F. Food Recommendation Using Machine Learning for Physical Activities in Patients with Type 1 Diabetes. In Proceedings of the 2019 Scandinavian Conference on Health Informatics (EHIN), Oslo, Norway, 12–13 November 2019; pp. 45–49.
22. Crown, W.H. Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health* **2015**, *18*, 137–140. [[CrossRef](#)]
23. Othman, M.; Zain, N.M.; Muhamad, U.K. e-Diet Meal Recommender System for Diabetic Patients. In Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) 2017; Saian, R., Abbas, M.A., Kangar, Malaysia, 15 - 16 August 2017; Volume 2, pp. 155–164.
24. Xie, J.; Wang, Q. A personalized diet and exercise recommender system for type 1 diabetes self-management: An in silico study. *Smart Health* **2019**, *13*, 100069. [[CrossRef](#)]
25. Dalla Man, C.; Breton, M.; Cobelli, C. Physical Activity into the Meal Glucose—Insulin Model of Type 1 Diabetes: In Silico Studies. *J. Diabetes Sci. Technol.* **2009**, *3*, 56–67. [[CrossRef](#)]
26. Dalla Man, C.; Rizza, R.A.; Cobelli, C. Meal Simulation Model of the Glucose-Insulin System. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 1740–1749. [[CrossRef](#)]
27. Dalla Man, C.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; Cobelli, C. The UVA/PADOVA Type 1 Diabetes Simulator: New Features. *J. Diabetes Sci. Technol.* **2014**, *8*, 26–34.
28. Piercy, K.L.; Troiano, R.P.; Ballard, R.M. The Physical Activity Guidelines for Americans. *J. Am. Med. Assoc. (JAMA)* **2018**, *320*, 2020–2028. [[CrossRef](#)] [[PubMed](#)]
29. Bergman, R.N. Toward physiological understanding of glucose tolerance: minimal-model approach. *Diabetes* **1989**, *38*, 1512–1527. [[CrossRef](#)] [[PubMed](#)]
30. Hovorka, R.; Canonico, V.; Chassin, L.J.; Haueter, U.; Massi-Benedetti, M.; Federici, M.O.; Pieber, T.R.; Schaller, H.C.; Schaupp, L.; Vering, T.; et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiol. Meas.* **2004**, *25*, 905. [[CrossRef](#)]
31. Kanderian, S.S.; Weinzimer, S.A.; Steil, G.M. The identifiable virtual patient model: Comparison of simulation and clinical closed-loop study results. *J. Diabetes Sci. Technol.* **2012**, *6*, 371–379. [[CrossRef](#)]
32. Breton, M.; Kovatchev, B. Analysis, Modeling, and Simulation of the Accuracy of Continuous Glucose Sensors. *J. Diabetes Sci. Technol.* **2008**, *2*, 853–862. [[CrossRef](#)]
33. Tejedor, M.; Myhre, J.N. Controlling Blood Glucose for Patients with Type 1 Diabetes Using Deep Reinforcement Learning—The Influence Of Changing The Reward Function. In Proceedings of the Northern Lights Deep Learning Workshop 2020, Tromsø, Norway, 10 January 2020; Volume 1, pp. 1–6.

34. Clarke, W.; Kovatchev, B. Statistical tools to analyze continuous glucose monitor data. *Diabetes Technol. Ther.* **2009**, *11*, S45–S54. [[CrossRef](#)]
35. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2019; pp. 8024–8035.
36. Wingate, D.; Weber, T. Automated Variational Inference in Probabilistic Programming. *arXiv* **2013**, arXiv:1301.1299.
37. Bingham, E.; Chen, J.P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; Goodman, N.D. Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.* **2019**, *20*, 973–978.
38. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86, [[CrossRef](#)]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Bibliography

- [1] Steven Truong. When going to the bathroom becomes scary. In *Ilona Karmel Writing Prizes, MIT*, 2017.
- [2] James Norman. Normal regulation of blood glucose. the important roles of insulin and glucagon: Diabetes and hypoglycemia. In *[cited 18 November 2020]; Available from: <https://www.endocrineweb.com/conditions/diabetes/normal-regulation-blood-glucose>*, 2016.
- [3] Thomas F Lotz. High Resolution Clinical Model-Based Assessment of Insulin Sensitivity. 2007.
- [4] Spyros G Tzafestas. *Feedback Control in Life and Society*, pages 575–625. Springer International Publishing, Cham, 2018.
- [5] Faculty Of Medicine For Doctors and Medical Students. All you need to know about the glucose tolerance test. In *[cited 11 February 2021]; Available from: <https://forum.facmedicine.com/threads/all-you-need-to-know-about-the-glucose-tolerance-test.25348/>*, 2016.
- [6] Sohaib Mehmood, Imran Ahmad, Hadeeqa Arif, Umm E Ammara, and Abdul Majeed. Artificial Pancreas Control Strategies Used for Type 1 Diabetes Control and Treatment: A Comprehensive Analysis. *Applied System Innovation*, 3(3):31, jul 2020.
- [7] María F Villa-Tamayo and Pablo S Rivadeneira. Adaptive Impulsive Offset-Free MPC to Handle Parameter Variations for Type 1 Diabetes Treatment. *Industrial & Engineering Chemistry Research*, 59(13):5865–5876, apr 2020.

-
- [8] Robles M, Gautier C, Mendoza L, Peugnet P, Dubois C, Dahirel M, et al. Maternal Nutrition during Pregnancy Affects Testicular and Bone Development, Glucose Metabolism and Response to Overnutrition in Weaned Horses Up to Two Years. *PLoS ONE*, 12(1), 2017.
- [9] Malgorzata E Wilinska, Ludovic J Chassin, Carlo L Acerini, Janet M Allen, David B Dunger, and Roman Hovorka. Simulation Environment to Evaluate Closed-Loop Insulin Delivery Systems in Type 1 Diabetes. *Journal of Diabetes Science and Technology*, 4(1):132–144, 2010.
- [10] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA Type 1 Diabetes Simulator: New Features. *Journal of Diabetes Science and Technology*, 8(1):26–34, 2014.
- [11] Sutton, R S; Barto, A G. *Reinforcement Learning: An introduction*. MIT press, second edition, 2018.
- [12] Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource Management with Deep Reinforcement Learning. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks, HotNets '16*, pages 50–56. Association for Computing Machinery, 2016.
- [13] Syeed Adnan Raheel Shah, Hunain Arshad, Muhammad Farhan, Syed Safdar Raza, Mudasser Muneer Khan, Sunera Imtiaz, Gullnaz Shahzadi, Muhammad Ahmed Qurashi, and Muhammad Waseem. Sustainable Brick Masonry Bond Design and Analysis: An Application of a Decision-Making Technique. *Applied Sciences*, 9(20), 2019.
- [14] G S Yadav and S K Dubey. Analytical Approach towards Electrocardiogram Signal using Machine Learning and Data Mining Techniques. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 1210–1214, 2020.
- [15] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. A Batched Scalable Multi-Objective Bayesian Optimization Algorithm. *CoRR*, abs/1811.01323, 2018.
- [16] Forouzanfar M, Afshin A, Alexander L, Anderson H, Bhutta Z, Biryukov S, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and

- metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053):1659–1724, 2016.
- [17] He Chen, Gong Chen, Xiaoying Zheng, and Yan Guo. Contribution of specific diseases and injuries to changes in health adjusted life expectancy in 187 countries from 1990 to 2013: retrospective observational study. *BMJ*, 364, 2019.
- [18] Yang J, Yu D, Wen W, Saito E, Rahman S, Shu X, et al. Association of Diabetes With All-Cause and Cause-Specific Mortality in Asia: A Pooled Analysis of More Than 1 Million Participants. *JAMA Network Open*, 2(4):e192696–e192696, apr 2019.
- [19] Xiling Lin, Yufeng Xu, Xiaowen Pan, Jingya Xu, Yue Ding, Xue Sun, Xiaoxiao Song, Yuezhong Ren, and Peng-Fei Shan. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Scientific Reports*, 10(1):14790, 2020.
- [20] Jawad A Al-Lawati. Diabetes Mellitus: A Local and Global Public Health Emergency! *Oman medical journal*, 32(3):177–179, may 2017.
- [21] Armando Arredondo, Alejandra Azar, and Ana Lucía Recamán. Diabetes, a global public health challenge with a high epidemiological and economic burden on health systems in Latin America. *Global Public Health*, 13(7):780–787, 2018.
- [22] Georg Serfling, Hannes Kalscheuer, Sebastian M Schmid, and Hendrik Lehnert. New technologies in diabetes treatment. *Der Internist*, 60(9):912–916, sep 2019.
- [23] Elisa Giani, Andrea Enzo Scaramuzza, and Gian Vincenzo Zuccotti. Impact of new technologies on diabetes care. *World journal of diabetes*, 6(8):999–1004, jul 2015.
- [24] Timothy S Bailey, John Walsh, and Jenine Y Stone. Emerging Technologies for Diabetes Care. *Diabetes Technology & Therapeutics*, 20(S2):S2–78–S2–84, 2018.

-
- [25] Neesha Ramchandani and Rubina A Heptulla. New technologies for diabetes: a review of the present and the future. *International Journal of Pediatric Endocrinology*, 2012(1):28, 2012.
- [26] Melanie K Bothe, Luke Dickens, Katrin Reichel, Arn Tellmann, Björn Ellger, Martin Westphal, and Ahmed A Faisal. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Review of Medical Devices*, 10(5):661–673, 2013.
- [27] Ian Fox and J Wiens. Reinforcement Learning for Blood Glucose Control: Challenges and Opportunities. In *Proceedings of the Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning, Long Beach, CA, USA, 2019*.
- [28] Ivan Contreras and Josep Vehi. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. *J Med Internet Res*, 20(5):e10775, 2018.
- [29] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15:104–116, 2017.
- [30] Silvia Oviedo, Josep Vehí, Remei Calm, and Joaquim Armengol. A review of personalized blood glucose prediction strategies for T1DM patients. *International Journal for Numerical Methods in Biomedical Engineering*, 33(6):e2833, 2017.
- [31] Katrin Lunze, Tarunraj Singh, Marian Walter, Mathias D Brendel, and Steffen Leonhardt. Blood glucose control algorithms for type 1 diabetic patients: A methodological review. *Biomedical Signal Processing and Control*, 8(2):107–119, 2013.
- [32] P D Ngo, S Wei, A Holubová, J Muzik, and F Godtlielsen. Reinforcement-learning optimal control for type-1 diabetes. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 333–336, 2018.
- [33] Meysam Bastani. Model-Free Intelligent Diabetes Management Using Machine Learning. Master’s Thesis, University of Alberta Libraries, Edmonton, AB, Canada. 2014.

- [34] J N Myhre, I K Launonen, S Wei, and F Godtliebsen. CONTROLLING BLOOD GLUCOSE LEVELS IN PATIENTS WITH TYPE 1 DIABETES USING FITTED Q-ITERATIONS AND FUNCTIONAL FEATURES. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2018.
- [35] Elena Daskalaki, Peter Diem, and Stavroula G Mougiakakou. An Actor–Critic based controller for glucose regulation in type 1 diabetes. *Computer Methods and Programs in Biomedicine*, 109(2):116–125, 2013.
- [36] Qingnan Sun, Marko V Jankovic, and Stavroula G Mougiakakou. Reinforcement Learning-Based Adaptive Insulin Advisor for Individuals with Type 1 Diabetes Patients under Multiple Daily Injections Therapy(). *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2019:3609–3612, jul 2019.
- [37] Sh Yasini, MB Naghibi-Sistani, and A Karimpour. Agent-based simulation for blood glucose control in diabetic patients. *International Journal of Applied Science, Engineering and Technology*, 5(1):40–49, 2009.
- [38] Q Sun, M V Jankovic, J Budzinski, B Moore, P Diem, C Stettler, and S G Mougiakakou. A Dual Mode Adaptive Basal-Bolus Advisor Based on Reinforcement Learning. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2633–2641, nov 2019.
- [39] T Zhu, K Li, P Herrero, and P Georgiou. Basal Glucose Control in Type 1 Diabetes using Deep Reinforcement Learning: An In Silico Validation. *IEEE Journal of Biomedical and Health Informatics*, page 1, 2020.
- [40] S Lee, J Kim, S W Park, S M. Jin, and S M. Park. Toward a Fully Automated Artificial Pancreas System Using a Bioinspired Reinforcement Learning Design: In Silico Validation. *IEEE Journal of Biomedical and Health Informatics*, page 1, 2020.
- [41] WHO. Diabetes. In *[cited 21 October 2020]*; Available from: <https://www.who.int/health-topics/diabetes>, 2020.

- [42] David M Nathan. Long-Term Complications of Diabetes Mellitus. *New England Journal of Medicine*, 328(23):1676–1685, 1993.
- [43] IDF. Diabetes atlas 9th edition. In [cited 21 October 2020]; Available from: <https://www.diabetesatlas.org/en/>, 2019.
- [44] Dieren Susan Van, Joline W J Beulens, Schouw Yvonne T Van der, Diederick E Grobbee, and Bruce Nealb. The global burden of diabetes and its complications: an emerging pandemic. *European Journal of Cardiovascular Prevention & Rehabilitation*, 17(1_suppl):s3–s8, 2010.
- [45] ADA. Changing our future through research. In [cited 22 October 2020]; Available from: <https://www.diabetes.org/research>, 2020.
- [46] Pia V Röder, Bingbing A Wu, Yixian Liu, and Weiping Han. Pancreatic regulation of glucose homeostasis. *Experimental & molecular medicine*, 48(3):e219, 2016.
- [47] Iben Rix, Christina Nexøe-Larsen, Natasha C Bergmann, Asger Lund, and Knop Filip K. Glucagon physiology. In [cited 19 November 2020]; Available from: <https://www.ncbi.nlm.nih.gov/books/NBK279127/>, 2019.
- [48] The Global Diabetes Community. Blood sugar level ranges. In [cited 8 February 2021]; Available from: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html, 2019.
- [49] Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. Continuous glucose monitoring and intensive treatment of type 1 diabetes. *New England Journal of Medicine*, 359(14):1464–1476, 2008.
- [50] Marie L Misso, Kristine J Egberts, Matthew Page, Denise O’Connor, and Jonathan Shaw. Continuous subcutaneous insulin infusion (csii) versus multiple insulin injections for type 1 diabetes mellitus. *Cochrane Database of Systematic Reviews*, (1), 2010.
- [51] Michele Heisler, Sandeep Vijan, Robert M Anderson, Peter A Ubel, Steven J Bernstein, and Timothy P Hofer. When do patients and their physicians agree on diabetes treatment goals and strategies, and

- what difference does it make? *Journal of General Internal Medicine*, 18(11):893–902, 2003.
- [52] Sandeep Vijan, Rodney A Hayward, David L Ronis, and Timothy P Hofer. BRIEF REPORT: The burden of diabetes therapy. *Journal of General Internal Medicine*, 20(5):479–482, 2005.
- [53] Wayne Katon, Elizabeth H B Lin, and Kurt Kroenke. The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *General Hospital Psychiatry*, 29(2):147–155, 2007.
- [54] Fadia T Shaya, Xia Yan, Pei-Jung Lin, Linda Simoni-Wastila, Morgan Bron, Robert Baran, and Thomas W Donner. US Trends in Glycemic Control, Treatment, and Comorbidity Burden in Patients With Diabetes. *The Journal of Clinical Hypertension*, 12(10):826–832, 2010.
- [55] Riccardo Candido, Kathleen Wyne, and Ester Romoli. Diabetes therapy: research, treatment and education of diabetes and related disorders. *Endocrinología y Nutrición*, 9(3):927–949, 2018.
- [56] Jino Y. Howard and Sharon A. Watss. Bolus insulin prescribing recommendations for patients with type 2 diabetes mellitus. *Federal practitioner: for the health care professionals of the VA, DoD, and PHS*, 34(8):26–31, 2017.
- [57] Irene Vinagre, Juan Sánchez-Hernández, José L. Sánchez-Quesada, Miguel Á. María, Alberto de Leiva, and Antonio Pérez. Switching to basal-bolus insulin therapy is effective and safe in long-term type 2 diabetes patients inadequately controlled with other insulin regimens. *Endocrinología y Nutrición*, 60(5):249–253, 2013.
- [58] Nicoleta D Sora, Fnu Shashpal, Elizabeth A Bond, and Alicia J Jenkins. Insulin Pumps: Review of Technological Advancement in Diabetes Management. *The American Journal of the Medical Sciences*, 358(5):326–331, 2019.
- [59] Guillermo E. Umpierrez and David C. Klonoff. Diabetes Technology Update: Use of Insulin Pumps and Continuous Glucose Monitoring in the Hospital. *Diabetes Care*, 41(8):1579–1589, 2018.

-
- [60] Ana Isabel Silva, António Norton de Matos, I Gabrielle Brons, and Marília Mateus. An overview on the development of a bio-artificial pancreas as a treatment of insulin-dependent diabetes mellitus. *Medicinal Research Reviews*, 26(2):181–222, 2006.
- [61] Kaitlin M Bratlie, Roger L York, Michael A Invernale, Robert Langer, and Daniel G Anderson. Materials for diabetes therapeutics. *Advanced healthcare materials*, 1(3):267–284, may 2012.
- [62] R. Millstein, N. M. Becerra, and J. H. Shubrook. Insulin pumps: Beyond basal-bolus. *Cleveland Clinic Journal of Medicine*, 82(12):835–842, 2015.
- [63] John Pickup. Insulin Pumps. *Diabetes Technology & Therapeutics*, 16(S1):S-17–S-22, 2014.
- [64] Lutz Heinemann, G Alexander Fleming, John R Petrie, Reinhard W Holl, Richard M Bergenstal, and Anne L Peters. Insulin Pump Risks and Benefits: A Clinical Appraisal of Pump Safety Standards, Adverse Event Reporting, and Research Needs. *Diabetes Care*, 38(4):716–722, 2015.
- [65] J.-P. Riveline, S Franc, M Biedzinski, F.-X. Jollois, N Messaoudi, F Lagarde, B Lormeau, S Pichard, M Varroud-Vial, A Deburge, E Dresco, and G Charpentier. Sexual activity in diabetic patients treated by continuous subcutaneous insulin infusion therapy. *Diabetes & Metabolism*, 36(3):229–233, 2010.
- [66] Kathryn Graff Low, Lori Massa, Dana Lehman, and Jerrold S Olshan. Insulin pump use in young adolescents with type 1 diabetes: a descriptive study. *Pediatric Diabetes*, 6(1):22–31, 2005.
- [67] Fatemah M Alsaleh, Felicity J Smith, Rebecca Thompson, Mohammad A Al-Saleh, and Kevin M G Taylor. Insulin pump therapy: impact on the lives of children/young people with diabetes mellitus and their parents. *International journal of clinical pharmacy*, 36(5):1023–1030, oct 2014.
- [68] P.-Y. Benhamou, V Melki, R Boizel, F Perreal, J.-L. Quesada, S Bessieres-Lacombe, J.-L. Bosson, S Halimi, and H Hanaire. One-year efficacy and safety of Web-based follow-up using cellular phone

- in type 1 diabetic patients under insulin pump therapy: the PumpNet study. *Diabetes & Metabolism*, 33(3):220–226, 2007.
- [69] Katharine Barnard and T Chas Skinner. Qualitative study into quality of life issues surrounding insulin pump use in type 1 diabetes. *Practical Diabetes International*, 24(3):143–148, 2007.
- [70] L Ratheau, N Jeandidier, F Moreau, S Sigrist, and M Pinget. How technology has changed diabetes management and what it has failed to achieve. *Diabetes & metabolism*, 37 Suppl 4:S57–64, dec 2011.
- [71] Jannik Kruse Nielsen, Christian Born Djurhuus, Claus Højbjerg Gravholt, Andreas Christiansen Carus, Jacob Granild-Jensen, Hans Ørskov, and Jens Sandahl Christiansen. Continuous Glucose Monitoring in Interstitial Subcutaneous Adipose Tissue and Skeletal Muscle Reflects Excursions in Cerebral Cortex. *Diabetes*, 54(6):1635–1639, 2005.
- [72] Günther Schmelzeisen-Redeker, Michael Schoemaker, Harald Kirchsteiger, Guido Freckmann, Lutz Heinemann, and Luigi del Re. Time Delay of CGM Sensors: Relevance, Causes, and Countermeasures. *Journal of Diabetes Science and Technology*, 9(5):1006–1015, 2015.
- [73] William H Polonsky and Danielle Hessler. What Are the Quality of Life-Related Benefits and Losses Associated with Real-Time Continuous Glucose Monitoring? A Survey of Current Users. *Diabetes Technology & Therapeutics*, 15(4):295–301, 2013.
- [74] David Rodbard. Continuous Glucose Monitoring: A Review of Recent Studies Demonstrating Improved Glycemic Outcomes. *Diabetes Technology & Therapeutics*, 19(S3):S–25–S–37, 2017.
- [75] Laurel H Messer, Paul F Cook, Molly L Tanenbaum, Sarah Hanes, Kimberly A Driscoll, and Korey K Hood. CGM Benefits and Burdens: Two Brief Measures of Continuous Glucose Monitoring. *Journal of Diabetes Science and Technology*, 13(6):1135–1141, 2019.
- [76] Nurul A Mohd Asarani, Andrew N Reynolds, Sara E Boucher, Martin de Bock, and Benjamin J Wheeler. Cutaneous Complications With Continuous or Flash Glucose Monitoring Use: Systematic Review of Trials and Observational Studies. *Journal of Diabetes Science and Technology*, 14(2):328–337, 2020.

- [77] Vidita Divan, Margaret Greenfield, Christopher P Morley, and Ruth S Weinstock. Perceived Burdens and Benefits Associated with Continuous Glucose Monitor Use in Type 1 Diabetes Across the Lifespan. *Journal of Diabetes Science and Technology*, 1(69):768, 2020.
- [78] Eleni Bekiari, Konstantinos Kitsios, Hood Thabit, Martin Tauschmann, Eleni Athanasiadou, Thomas Karagiannis, Anna-Bettina Haidich, Roman Hovorka, and Apostolos Tsapas. Artificial pancreas treatment for outpatients with type 1 diabetes: systematic review and meta-analysis. *BMJ*, 361, 2018.
- [79] Roman Hovorka. Closed-loop insulin delivery: from bench to clinical practice. *Nature Reviews Endocrinology*, 7(7):385–395, 2011.
- [80] Kavita Kumareswaran, Mark L Evans, and Roman Hovorka. Artificial pancreas: an emerging approach to treat Type 1 diabetes. *Expert review of medical devices*, 6(4):401–410, jul 2009.
- [81] Ali Cinar. Artificial pancreas systems: An introduction to the special issue. *IEEE Control Systems Magazine*, 38(1):26–29, 2018.
- [82] Sasan Adibi. *Mobile Health: A Technology Road Map*. Springer, 2015.
- [83] Ashenafi Zebene Woldaregay, Ilkka Kalervo Launonen, Eirik Årsand, David Albers, Anna Holubová, and Gunnar Hartvigsen. Toward Detecting Infection Incidence in People With Type 1 Diabetes Using Self-Recorded Data (Part 1): A Novel Framework for a Personalized Digital Infectious Disease Detection System. *J Med Internet Res*, 22(8):e18911, aug 2020.
- [84] Gary L. Wnek, Gary E.; Bowlin. *Encyclopedia of Biomaterials and Biomedical Engineering*. Informa Healthcare, Boca Raton, 2nd edition, 2008.
- [85] Terry G Jr Farmer, Thomas F Edgar, and Nicholas A Peppas. The future of open- and closed-loop insulin delivery systems. *The Journal of pharmacy and pharmacology*, 60(1):1–13, jan 2008.
- [86] Richard M Bergenstal, Satish Garg, Stuart A Weinzimer, Bruce A Buckingham, Bruce W Bode, William V Tamborlane, and Francine R Kaufman. Safety of a Hybrid Closed-Loop Insulin Delivery System in Patients With Type 1 Diabetes. *JAMA*, 316(13):1407–1408, oct 2016.

-
- [87] Florian Reiterer, Guido Freckmann, and Luigi del Re. Impact of Carbohydrate Counting Errors on Glycemic Control in Type 1 Diabetes. *IFAC-PapersOnLine*, 51(27):186–191, 2018.
- [88] Tomoyuki Kawamura, Chihiro Takamura, Masakazu Hirose, Tomomi Hashimoto, Takashi Higashide, Yoneo Kashihara, Kayako Hashimura, and Haruo Shintaku. The factors affecting on estimation of carbohydrate content of meals in carbohydrate counting. *Clinical Pediatric Endocrinology*, 24(4):153–165, 2015.
- [89] Maria F Vasiloglou, Stavroula Mougiakakou, Emilie Aubry, Anika Bokelmann, Rita Fricker, Filomena Gomes, Cathrin Guntermann, Alexa Meyer, Diana Studerus, and Zeno Stanga. A Comparative Study on Carbohydrate Estimation: GoCARB vs. Dietitians. *Nutrients*, 10(6):741, jun 2018.
- [90] Asma Deeb, Ahlam Al Hajeri, Iman Alhmoudi, and Nico Nagelkerke. Accurate Carbohydrate Counting Is an Important Determinant of Postprandial Glycemia in Children and Adolescents With Type 1 Diabetes on Insulin Pump Therapy. *Journal of Diabetes Science and Technology*, 11(4):753–758, 2017.
- [91] Kiyoo Mibu, Tomoaki Yatabe, and Kazuhiro Hanazaki. Blood glucose control using an artificial pancreas reduces the workload of ICU nurses. *Journal of Artificial Organs*, 15(1):71–76, 2012.
- [92] Takehiro Okabayashi, Yasuo Shima, Tatsuaki Sumiyoshi, Akihito Kozuki, Teppei Tokumaru, Tasuo Iiyama, Takeki Sugimoto, Michiya Kobayashi, Masataka Yokoyama, and Kazuhiro Hanazaki. Intensive Versus Intermediate Glucose Control in Surgical Intensive Care Unit Patients. *Diabetes Care*, 37(6):1516–1524, 2014.
- [93] Tsutomu Namikawa, Masaya Munekage, Hiroyuki Kitagawa, Tomoaki Yatabe, Hiromichi Maeda, Yuuki Tsukamoto, Kenichi Hirano, Takuji Asano, Yoshihiko Kinoshita, and Kazuhiro Hanazaki. Comparison between a novel and conventional artificial pancreas for perioperative glycemic control using a closed-loop system. *Journal of artificial organs : the official journal of the Japanese Society for Artificial Organs*, 20(1):84–90, mar 2017.

- [94] H Thabit and R Hovorka. Bridging technology and clinical practice: innovating inpatient hyperglycaemia management in non-critical care settings. *Diabetic Medicine*, 35(4):460–471, 2018.
- [95] Lenka Petruzalkova, Jan Soupal, Veronika Plasova, Pavlina Jiranova, Vit Neuman, Lukas Plachy, Stepanka Pruhova, Zdenek Sumnik, and Barbora Obermannova. Excellent Glycemic Control Maintained by Open-Source Hybrid Closed-Loop AndroidAPS During and After Sustained Physical Activity. *Diabetes Technology & Therapeutics*, 20(11):744–750, 2018.
- [96] Boris Kovatchev, Peiyao Cheng, Stacey M Anderson, Jordan E Pinsker, Federico Boscari, Bruce A Buckingham, Francis J Doyle, Korey K Hood, Sue A Brown, Marc D Breton, Daniel Chernavvsky, Wendy C Bevier, Paige K Bradley, Daniela Bruttomesso, Simone Del Favero, Roberta Calore, Claudio Cobelli, Angelo Avogaro, Trang T Ly, Satya Shanmugham, Eyal Dassau, Craig Kollman, John W Lum, and Roy W Beck. Feasibility of Long-Term Closed-Loop Control: A Multicenter 6-Month Trial of 24/7 Automated Insulin Delivery. *Diabetes Technology & Therapeutics*, 19(1):18–24, 2017.
- [97] Charlotte K Boughton and Roman Hovorka. Advances in artificial pancreas systems. *Science Translational Medicine*, 11(484), 2019.
- [98] Kamuran Turksoy, Iman Hajizadeh, Sediqeh Samadi, Jianyuan Feng, Mert Sevil, Minsun Park, Laurie Quinn, Elizabeth Littlejohn, and Ali Cinar. Real-time insulin bolusing for unannounced meals with artificial pancreas. *Control Engineering Practice*, 59:159–164, 2017.
- [99] Alyson Weiner, Elizabeth Robinson, and Rachelle Gandica. 1017-P: Effects of the T:slim X2 Insulin Pump with Basal-IQ Technology on Glycemic Control in a Pediatric Urban Academic Diabetes Practice. *Diabetes*, 69(Supplement 1), 2020.
- [100] Tauschmann M, Thabit H, Bally L, Allen J, Hartnell S, Wilinska M, et al. Closed-loop insulin delivery in suboptimally controlled type 1 diabetes: a multicentre, 12-week randomised trial. *The Lancet*, 392(10155):1321–1329, 2018.
- [101] Laurel H Messer, Gregory P Forlenza, Jennifer L Sherr, R Paul Wadwa, Bruce A Buckingham, Stuart A Weinzimer, David M Maahs, and

- Robert H Slover. Optimizing Hybrid Closed-Loop Therapy in Adolescents and Emerging Adults Using the MiniMed 670G System. *Diabetes Care*, 41(4):789–796, 2018.
- [102] Lalantha Leelarathna, Pratik Choudhary, Emma G Wilmot, Alistair Lumb, Tim Street, Partha Kar, and Sze M Ng. Hybrid closed-loop therapy: Where are we in 2021? *Diabetes, Obesity and Metabolism*, 2020.
- [103] H Peter Chase, Francis J Doyle, Howard Zisser, Eric Renard, Revital Nimri, Claudio Cobelli, Bruce A Buckingham, David M Maahs, Stacey Anderson, Lalo Magni, John Lum, Peter Calhoun, Craig Kollman, and Roy W Beck. Multicenter Closed-Loop/Hybrid Meal Bolus Insulin Delivery with Type 1 Diabetes. *Diabetes Technology & Therapeutics*, 16(10):623–632, 2014.
- [104] K Kumareswaran, M L Evans, and R Hovorka. Closed-loop insulin delivery: towards improved diabetes care. *Discovery medicine*, 13(69):159–170, 2012.
- [105] Rita Basu, Matthew L Johnson, Yogish C Kudva, and Ananda Basu. Exercise, Hypoglycemia, and Type 1 Diabetes. *Diabetes Technology & Therapeutics*, 16(6):331–337, may 2014.
- [106] Eran Atlas, Revital Nimri, Shahar Miller, Eli A Grunberg, and Moshe Phillip. MD-Logic Artificial Pancreas System. *Diabetes Care*, 33(5):1072–1076, 2010.
- [107] Eleonora Maria Aiello, Giuseppe Lisanti, Lalo Magni, Mirto Musci, and Chiara Toffanin. Therapy-driven Deep Glucose Forecasting. *Engineering Applications of Artificial Intelligence*, 87:103255, 2020.
- [108] K Li, C Liu, T Zhu, P Herrero, and P Georgiou. GluNet: A Deep Learning Framework for Accurate Glucose Forecasting. *IEEE Journal of Biomedical and Health Informatics*, 24(2):414–423, feb 2020.
- [109] Revital Nimri, Eran Atlas, Michal Ajzensztejn, Shahar Miller, Tal Oron, and Moshe Phillip. Feasibility study of automated overnight closed-loop glucose control under MD-logic artificial pancreas in patients with type 1 diabetes: the DREAM Project. *Diabetes technology & therapeutics*, 14(8):728–735, aug 2012.

- [110] Garry M Steil, Kerstin Rebrin, Christine Darwin, Farzam Hariri, and Mohammed F Saad. Feasibility of Automating Insulin Delivery for the Treatment of Type 1 Diabetes. *Diabetes*, 55(12):3344–3350, 2006.
- [111] Roman Hovorka, Valentina Canonico, Ludovic J Chassin, Ulrich Haueter, Massimo Massi-Benedetti, Marco Orsini Federici, Thomas R Pieber, Helga C Schaller, Lukas Schaupp, Thomas Vering, and Malgorzata E Wilinska. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement*, 25(4):905–920, jul 2004.
- [112] Rebecca A Harvey, Eyal Dassau, Wendy C Bevier, Dale E Seborg, Lois Jovanovič, Francis J Doyle, and Howard C Zisser. Clinical Evaluation of an Automated Artificial Pancreas Using Zone-Model Predictive Control and Health Monitoring System. *Diabetes Technology & Therapeutics*, 16(6):348–357, 2014.
- [113] Garry M Steil. Algorithms for a Closed-Loop Artificial Pancreas: The Case for Proportional-Integral-Derivative Control. *Journal of Diabetes Science and Technology*, 7(6):1621–1631, 2013.
- [114] MJ Willis. Proportional-integral-derivative control. *Dept. of Chemical and Process Engineering University of Newcastle*, 1999.
- [115] Garry M Steil, Cesar C Palerm, Natalie Kurtz, Gayane Voskanyan, Anirban Roy, Sachiko Paz, and Fouad R Kandeel. The effect of insulin feedback on closed loop glucose control. *The Journal of clinical endocrinology and metabolism*, 96(5):1402–1408, may 2011.
- [116] Trang T Ly, Stuart A Weinzimer, David M Maahs, Jennifer L Sherr, Anirban Roy, Benyamin Grosman, Martin Cantwell, Natalie Kurtz, Lori Carria, Laurel Messer, Rie von Eyben, and Bruce A Buckingham. Automated hybrid closed-loop control with a proportional-integral-derivative based system in adolescents and adults with type 1 diabetes: individualizing settings for optimal performance. *Pediatric Diabetes*, 18(5):348–355, 2017.
- [117] Jessica L Ruiz, Jennifer L Sherr, Eda Cengiz, Lori Carria, Anirban Roy, Gayane Voskanyan, William V Tamborlane, and Stuart A Weinzimer. Effect of insulin feedback on closed-loop glucose control: a crossover

- study. *Journal of diabetes science and technology*, 6(5):1123–1130, sep 2012.
- [118] Stuart A Weinzimer, Garry M Steil, Karena L Swan, Jim Dziura, Natalie Kurtz, and William V Tamborlane. Fully Automated Closed-Loop Insulin Delivery Versus Semiautomated Hybrid Control in Pediatric Patients With Type 1 Diabetes Using an Artificial Pancreas. *Diabetes Care*, 31(5):934–939, 2008.
- [119] Eduardo F Camacho and Carlos Bordons Alba. *Model predictive control*. Springer Science & Business Media, 2013.
- [120] Hyunjin Lee, Bruce A Buckingham, Darrell M Wilson, and B Wayne Bequette. A closed-loop artificial pancreas using model predictive control and a sliding meal size estimator. *Journal of diabetes science and technology*, 3(5):1082–1090, sep 2009.
- [121] Dimitri Boiroux, Anne Katrine Duun-Henriksen, Signe Schmidt, Kirsten Nørgaard, Niels Kjølstad Poulsen, Henrik Madsen, and John Bagterp Jørgensen. Assessment of Model Predictive and Adaptive Glucose Control Strategies for People with Type 1 Diabetes. *IFAC Proceedings Volumes*, 47(3):231–236, 2014.
- [122] Ricardo S. Colmegna, Patricio; Sánchez-Peña. Simulators of Diabetes Mellitus Dynamics. In *23^o Congreso Argentino de Control Automático*, page 6, 2012.
- [123] C Dalla Man, R A Rizza, and C Cobelli. Meal Simulation Model of the Glucose-Insulin System. *IEEE Transactions on Biomedical Engineering*, 54(10):1740–1749, 2007.
- [124] Richard N Bergman. Toward Physiological Understanding of Glucose Tolerance: Minimal-Model Approach. *Diabetes*, 38(12):1512–1527, 1989.
- [125] R N Bergman, Y Z Ider, C R Bowden, and C Cobelli. Quantitative estimation of insulin sensitivity. *American Journal of Physiology-Endocrinology and Metabolism*, 236(6):E667, 1979.
- [126] Pasquale Palumbo, Susanne Ditlevsen, Alessandro Bertuzzi, and Andrea De Gaetano. Mathematical modeling of the glucose–insulin system: A review. *Mathematical Biosciences*, 244(2):69–81, 2013.

-
- [127] R N Bergman. Minimal Model: Perspective from 2005. *Hormone Research in Paediatrics*, 64(suppl 3)(Suppl. 3):8–15, 2005.
- [128] Roman Hovorka, Fariba Shojaee-Moradie, Paul V Carroll, Ludovic J Chassin, Ian J Gowrie, Nicola C Jackson, Romulus S Tudor, A Margot Umpleby, and Richard H Jones. Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT. *American journal of physiology. Endocrinology and metabolism*, 282(5):E992–1007, may 2002.
- [129] Malgorzata E Wilinska and Roman Hovorka. Simulation models for in silico testing of closed-loop glucose controllers in type 1 diabetes. *Drug Discovery Today: Disease Models*, 5(4):289–298, 2008.
- [130] Boris P Kovatchev, Marc Breton, Chiara Dalla Man, and Claudio Cobelli. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. *Journal of diabetes science and technology*, 3(1):44–55, jan 2009.
- [131] Sami S Kanderian, Stuart A Weinzimer, and Garry M Steil. The Identifiable Virtual Patient Model: Comparison of Simulation and Clinical Closed-Loop Study Results. *Journal of Diabetes Science and Technology*, 6(2):371–379, 2012.
- [132] Marc D Breton. Physical Activity—The Major Unaccounted Impediment to Closed Loop Control. *Journal of Diabetes Science and Technology*, 2(1):169–174, 2008.
- [133] Chiara Dalla Man, Marc D Breton, and Claudio Cobelli. Physical Activity into the Meal Glucose—Insulin Model of Type 1 Diabetes: In Silico Studies. *Journal of Diabetes Science and Technology*, 3(1):56–67, 2009.
- [134] The Diabetes Control and Complications Trial Research Group. The Relationship of Glycemic Exposure (HbA1c) to the Risk of Development and Progression of Retinopathy in the Diabetes Control and Complications Trial. *Diabetes*, 44(8):968–983, 1995.
- [135] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

-
- [136] Véronique Gingras, Nadine Taleb, Amélie Roy-Fleming, Laurent Legault, and Rémi Rabasa-Lhoret. The challenges of achieving post-prandial glucose control using closed-loop systems in patients with type 1 diabetes. *Diabetes, Obesity and Metabolism*, 20(2):245–256, 2018.
- [137] RICHARD BELLMAN. A Markovian Decision Process. *Journal of Mathematics and Mechanics*, 6(5):679–684, dec 1957.
- [138] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.
- [139] Morten Frydenberg. The chain graph markov property. *Scandinavian Journal of Statistics*, pages 333–353, 1990.
- [140] Xiaoxiao Guo. *Deep Learning and Reward Design for Reinforcement Learning*. PhD thesis, 2017.
- [141] Michael L Littman. Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2(1):55–66, 2001.
- [142] Sudharsan Ravichandiran. *Hands-On Reinforcement Learning with Python*. Packt Publishing, 2018.
- [143] Oded Berger-Tal, Jonathan Nathan, Ehud Meron, and David Saltz. The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE*, 9(4):e95693, apr 2014.
- [144] Volodymyr O’Donoghue, Brendan; Osband, Ian; Munos, Remi; Mnih. The Uncertainty Bellman Equation and Exploration. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018.
- [145] K Zhang and W Pan. The Two Facets of the Exploration-Exploitation Dilemma. In *2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 371–380, 2006.
- [146] Gerald Tesauro. Practical Issues in Temporal Difference Learning. *Machine Learning*, 8(3):257–277, 1992.
- [147] G. A. Rummery and M. Niranjan. On-line q-learning using connectionist systems. Technical report, 1994.
- [148] Christopher J C H Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

-
- [149] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS'99*, pages 1057–1063, Cambridge, MA, USA, 1999. MIT Press.
- [150] Jan Peters and J Andrew Bagnell. Policy gradient methods. *Scholarpedia*, 5(11):3698, 2010.
- [151] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.
- [152] John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In Francis Bach and David Blei, editors, *32nd International Conference on Machine Learning, ICML 2015*, volume 3 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 2015. PMLR.
- [153] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.0, 2017.
- [154] James M Joyce. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
- [155] Christian Robert Shelton. *Importance sampling for reinforcement learning with multiple objectives*. PhD thesis, MIT - Massachusetts Institute of Technology, 2001.
- [156] Yan Duan, Xi Chen, Rein Houthoofd, John Schulman, and Pieter Abbeel. Benchmarking Deep Reinforcement Learning for Continuous Control. In Maria Florina Balcan and Kilian Q Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1329–1338. PMLR, 2016.
- [157] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *arXiv*, 11:219–354, 2018.

-
- [158] James A Anderson. *An introduction to neural networks*. MIT press, 1995.
- [159] Tomasz Szandała. Bio-inspired Neurocomputing. Studies in Computational Intelligence. chapter : Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks, pages 203–224. Springer Singapore, 2021.
- [160] W Thomas Miller, Paul J Werbos, and Richard S Sutton. *Neural networks for control*. MIT press, 1995.
- [161] Michael Kampffmeyer. *Advancing Segmentation and Unsupervised Learning Within the Field of Deep Learning*. PhD thesis, UiT The Arctic University of Norway, Tromsø, Norway, 2018.
- [162] Vikram Mullachery, Aniruddh Khera, and Amir Husain. Bayesian Neural Networks. *CoRR*, abs/1801.0, 2018.
- [163] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv*, 2015.
- [164] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on bayesian neural networks – a tutorial for deep learning users. *arXiv*, 2020.
- [165] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.
- [166] Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-End Safe Reinforcement Learning through Barrier Functions for Safety-Critical Continuous Control Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3387–3395, 2019.
- [167] S Arora and P Doshi. A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress. *ArXiv*, abs/1806.0, 2018.
- [168] Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 4565–4573. Curran Associates, Inc., 2016.

-
- [169] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The Option-Critic Architecture. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 1726–1734. AAAI Press, 2017.
- [170] Andrew G Barto and Sridhar Mahadevan. Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems*, 13(1):41–77, 2003.
- [171] T M Peters and A Haidar. Dual-hormone artificial pancreas: benefits and limitations compared with single-hormone systems. *Diabetic medicine : a journal of the British Diabetic Association*, 35(4):450–459, apr 2018.
- [172] Ahmad Haidar, Michael A Tsoukas, Sarah Bernier-Twardy, Jean-Francois Yale, Joanna Rutkowski, Anne Bossy, Evelyne Pytka, Anas El Fathi, Natalia Strauss, and Laurent Legault. A Novel Dual-Hormone Insulin-and-Pramlintide Artificial Pancreas for Type 1 Diabetes: A Randomized Controlled Crossover Trial. *Diabetes Care*, 2020.
- [173] Ashenafi Zebene Woldaregay, Ilkka Kalervo Launonen, David Albers, Jorge Igual, Eirik Årsand, and Gunnar Hartvigsen. A Novel Approach for Continuous Health Status Monitoring and Automatic Detection of Infection Incidences in People With Type 1 Diabetes Using Machine Learning Algorithms (Part 2): A Personalized Digital Infectious Disease Detection Mechanism. *J Med Internet Res*, 22(8):e18912, aug 2020.

